SEMANTIC-GUIDED CONTRASTIVE LEARNING FOR EEG MULTIMODAL DECODING OF LISTENING AND WATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic decoding poses a fundamental challenge in brain-computer interfaces (BCIs) aiming for naturalistic brain-machine communication. Current approaches remain modality-specific (e.g., tailored for visual, auditory, or linguistic stimuli), lacking a universal framework that generalizes across diverse modalities. To address this, we propose Semantic-Guided Contrastive Learning (Semantic-CL), a unified framework for EEG-based semantic decoding that adapts to multiple modalities—including auditory and visual—without architectural modifications. The framework leverages semantic-guided soft contrastive learning to align EEG representations using stimulus semantic similarity metrics, augmented by inter-subject contrastive alignment to harmonize neural patterns across subjects. Evaluated on two modality-distinct benchmarks—SEED-DV (dynamic video) and Broderick (natural speech)—Semantic-CL achieves the state-of-the-art performance in semantic decoding, especially in the more challenging cross-subject settings. This study establishes a modality-agnostic EEG semantic decoding framework, enabling deployable BCIs in naturalistic contexts. Our code is available at `https://anonymous.4open.science/r/Semantic-CL-Cross_subject-anonymous-DBD7`.

1 INTRODUCTION

Semantic decoding aims to recover semantic meaning from brain activity. The millisecond temporal resolution of the electroencephalogram (EEG) makes it well-suited for decoding rapidly unfolding semantic processes. Neuroscience studies show robust EEG correlates of semantic processing, such as the N400 event-related response to semantic incongruity (Kutas & Hillyard, 1980; Kutas & Federmeier, 2011), which lays the foundations for semantic decoding from EEG. More recent work starts to decode the concrete concepts or words from listening (Défossez et al., 2023), reading (Duan et al., 2023) and image (Song et al., 2024) or video (Liu et al., 2024) viewing tasks. Advancing semantic decoding can clarify how meaning is represented and transformed in the brain, and enable more seamless brain-computer interactions (Rybář & Daly, 2022).

Current EEG semantic-decoding approaches face three persistent issues. First, most methods are modality-specific: they are tailored to visual, auditory, or linguistic stimuli, so the learned features may largely reflect modality-dependent signals rather than higher-level semantic representations, while also hindering reuse across different modalities. Second, although recent works adopt contrastive or representation-based objectives, supervision in EEG semantic decoding remains largely discrete (e.g., class labels or one-to-one pairing), with limited use of graded semantic similarity—despite evidence that neural responses may vary continuously with meaning (Huth et al., 2016; Pereira et al., 2018). Third, inter-subject variability remains a major obstacle: differences in anatomy, noise, and individual cognitive processing often distort EEG embeddings. Learning subject-invariant embeddings is critical for isolating shared semantic representation, suppressing idiosyncratic noise, and enabling robust transfer to unseen participants with greater practical utility (Nastase et al., 2019; Shen et al., 2024).

We address these gaps with Semantic-guided Contrastive Learning (Semantic-CL), a unified framework for EEG-based semantic decoding that applies without architectural modification to heteroge-

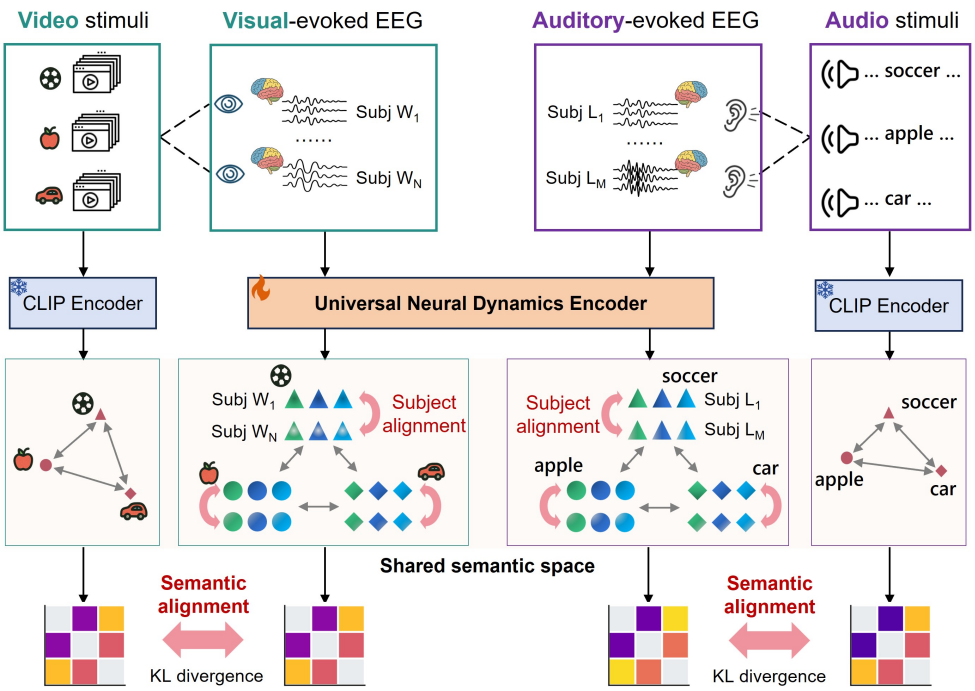

Figure 1: The semantic-guided contrastive learning framework for EEG semantic decoding of watching and listening.

neous paradigms (here, dynamic video in SEED-DV and natural speech in the Broderick corpus). The key component is a semantic-guided soft contrastive objective whose pairwise weights reflect stimulus-level semantic similarity, aligning the EEG embedding space with the geometry of meaning rather than categorical labels. Further, an inter-subject contrastive term ties this semantic geometry to cross-participant alignment, yielding collective subject-invariant semantic representations (Figure 1). Evaluated on SEED-DV (video) and Broderick (speech) datasets, the same architecture and objectives achieve state-of-the-art within-subject and cross-subject semantic decoding. The contributions of this paper are threefold:

- We establish a modality-agnostic EEG semantic-decoding framework that can be instantiated for watching and listening without redesign, enabling consistent training and evaluation across paradigms.
- We propose a semantic-guided soft contrastive loss that encodes graded semantic relatedness to learn EEG embeddings with faithful global semantic geometry.
- We introduce an inter-subject semantic alignment that binds subject invariance to the learned semantic geometry, producing collective representations that improve cross-subject robustness.

2 Related Work

EEG Semantic Decoding. Semantic decoding from EEG has seen recent advances in multiple modalities. For continuous speech stimuli, Défossez et al. (2023) trained a model with a contrastive objective that aligns EEG/MEG signals with paired speech representations, using data from participants listening to stories or isolated sentences. Jayalath et al. (2025) pointed out that many prior EEG semantic decoding methods fail to significantly exceed chance baselines on standard metrics without teacher forcing. They combined word classification with LLM rescoring and predictive in-filling of out-of-vocabulary words, achieving BLEU improvements of about 1.4-2.6 times. For semantic decoding from visual stimuli, contrastive frameworks such as NICE-EEG align EEG responses with visual embeddings to achieve open-set object recognition, later extended with large language model

guidance for richer semantic alignment (Song et al., 2024; 2025). Li et al. (2024) further enhanced the image generation performance by diffusion models. Recently, dynamic visual scenes were also introduced for semantic decoding from videos (Liu et al., 2024). These works represent meaningful progress but still lack modality generalizability, fine-grained semantic-aware supervision, and robust cross-subject generalization.

EEG Pre-training. Pre-training for EEG representation learning has followed the trajectory of vision and language, adopting self-supervised objectives such as masked autoencoding or contrastive learning. Representative approaches include BENDR (Kostas et al., 2021), which combines convolutional encoders with contrastive prediction; BIOT (Yang et al., 2023), which tokenizes biosignals for Transformer modeling; LaBraM (Jiang et al., 2024b), which uses vector-quantized autoencoding; and EEGPT (Wang et al., 2024b), which integrates masked reconstruction with latent spatiotemporal alignment. These models are typically trained on massive resting-state or task-agnostic EEG collections and demonstrate improvements on generic benchmarks such as sleep staging or anomaly detection. However, pre-trained EEG encoders are rarely evaluated on semantically demanding tasks. Their task-agnostic objectives risk diluting meaningful semantic structure. Our work differs by introducing a pre-training strategy that is explicitly semantic-aware and subject-invariant, tailored to decode the semantic meaning rather than generic signal features.

3 METHODS

We design an end-to-end framework that jointly optimizes three objectives. A semantic-guided soft contrastive loss captures fine-grained semantic relationships in different stimulus modalities, while an inter-subject alignment loss encourages a shared semantic space across participants. In addition, a cross-entropy loss is applied to enable direct supervision for semantic category prediction.

Let $\mathcal{B} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be a mini–batch of N EEG segments $\mathbf{x}_i \in \mathbb{R}^{C\times T}$ with semantic annotations $\mathbf{y}_i \in \mathbb{R}^{d_s}$. The Neural Dynamics Encoder (NDE) f_θ maps raw EEG to a hidden representation $\mathbf{h}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^{d_h}$. The outputs are ℓ_2–normalized as

$$\mathbf{z}_i = \frac{\mathbf{h}_i}{||\mathbf{h}_i||_2} \in \mathbb{S}^d,$$

where $\mathbb{S}^d$ denotes the d-dimensional unit sphere. The embeddings $\mathbf{z}_i$ are used to compute both the semantic-guided soft contrastive loss and the inter-subject alignment loss, while simultaneously serving as inputs to the MLP classifier for semantic classification.

3.1 SEMANTIC-GUIDED SOFT CONTRASTIVE LEARNING

We align EEG embeddings with semantic representations by matching their similarity distributions, so that the learned space reflects graded semantic relationships rather than discrete categories.

Semantic affinity matrix. A pre-trained semantic encoder g_s based on CLIP produces normalized semantic codes $\mathbf{v}_i = \frac{g_s(\mathbf{y}_i)}{||g_s(\mathbf{y}_i)||_2} \in \mathbb{S}^{d_s}$. Pair-wise semantic affinity is computed as

$$w_{ij} \;=\; \frac{1+\cos(\mathbf{v}_i, \mathbf{v}_j)}{2} \;\in\; [0,1]. \tag{1}$$

We only kept the top-k similar entries for each row (the diagonal entries are excluded), and set other entries to zero in w_{ij}. w_{ij} is then normalized by each row:

$$\tilde{w}_{ij} = w_{ij} / \sum_{k\neq i} w_{ik}. \tag{2}$$

KL divergence loss. Given a mini-batch of N EEG embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$, we first compute the pairwise feature similarity matrix

$$s_{ij} \;=\; \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\tau}, \tag{3}$$

where $\tau > 0$ is a temperature hyperparameter. To avoid trivial self-matching, the diagonal entries are masked. Applying the log-softmax row-wise yields a probability distribution over candidates:

$$\log p_{ij} \;=\; \frac{s_{ij}}{\tau} - \log \sum_{k \neq i} \exp\!\left(\tfrac{s_{ik}}{\tau}\right), \quad j \neq i. \tag{4}$$

The semantic-guided contrastive loss is defined as the Kullback–Leibler (KL) divergence between the EEG latent distribution p_{ij} and the semantic target distribution $\tilde{w}_{ij}$:

$$\mathcal{L}_{\text{Sem-CL}} \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}\big(\tilde{w}_{i,:} \,\big\|\, p_{i,:}\big) \;=\; \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \tilde{w}_{ij} \, \log \frac{\tilde{w}_{ij}}{p_{ij}}. \tag{5}$$

This loss drives the EEG similarity distribution to match the semantic similarity distribution.

In our implementation, we further integrate the subject information into the semantic affinity matrix. $\tilde{w}_{ij}$ is adjusted by subject-aware weights: similarities between samples from the same participant are down-weighted (factor 0.8), while cross-subject similarities retain full weight (factor 1). This emphasizes aligning semantic structure across participants rather than overfitting to subject-specific patterns.

3.2 Inter-subject alignment strategy

Sampling strategy of a mini-batch. To facilitate inter-subject alignment, each mini-batch $\mathcal{B}$ is constructed by paired sampling across participants. Specifically, for a given set of K stimuli, we draw K EEG segments from one participant and the corresponding K segments from another participant who experienced the same stimuli, yielding $N = 2K$ samples in total. Thus, a pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$ may correspond to different participants but share the same semantic annotation $\mathbf{y}$ if they are elicited by the same stimulus.

Inter-subject alignment loss. To explicitly align embeddings across participants, we treat cross-subject pairs elicited by the *same* stimulus as positives and all other cross-sample pairs in the batch as negatives following Shen et al. (2022). Recall that each mini–batch is formed by paired sampling from two subjects over K shared stimuli, yielding $N = 2K$ embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$ on the unit sphere. For each anchor i, let $\pi(i)$ denote its counterpart recorded from the other subject under the same stimulus. We define the temperature-scaled similarity as $a_{ij} = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}$, and the induced probability of selecting the cross-subject positive $\pi(i)$ among all candidates in the row of i:

$$p^{\text{align}}_{i,\pi(i)} \;=\; \frac{\exp\!\big(a_{i,\pi(i)}/\tau\big)}{\sum_{j \neq i} \exp\!\big(a_{ij}/\tau\big)}. \tag{6}$$

The inter-subject alignment loss is the InfoNCE objective that maximizes the likelihood of all cross-subject positives:

$$\mathcal{L}_{\text{Sub-CL}} \;=\; -\frac{1}{N} \sum_{i=1}^{N} \log \, p^{\text{align}}_{i,\pi(i)}. \tag{7}$$

This pulls embeddings of the same stimulus across subjects together while pushing apart embeddings of different stimuli (within- or cross-subject), thereby promoting subject-invariant alignment of the shared semantic structure.

3.3 The classification head

On top of the Neural Dynamics Encoder, we employ a lightweight multilayer perceptron (MLP) as the classification head. The MLP comprises two linear layers with batch normalization and GELU activation in between. The EEG embeddings are projected to predict semantic classes.

Overall objective. Our training jointly optimizes three complementary objectives: the semantic-guided soft contrastive loss (Eq. 5), the inter-subject alignment loss (Eq. 7), and the cross-entropy loss for semantic classification. The unified objective is

$$\mathcal{L} \;=\; \lambda \, \mathcal{L}_{\text{Sem-CL}} \;+\; (1-\lambda) \, \mathcal{L}_{\text{Sub-CL}} \;+\; \mathcal{L}_{\text{CE}}, \tag{8}$$

where the weight of the classification loss is fixed to 1 and the trade-off between the two contrastive terms is controlled by λ. In this work, we set $\lambda = 0.5$. This joint formulation enforces that embeddings capture graded semantic geometry, remain aligned across participants, and are directly optimized for semantic category prediction.

3.4 THE ARCHITECTURE OF NEURAL DYNAMICS ENCODER

The Neural Dynamics Encoder (NDE) is designed to be physiologically plausible and computationally efficient. It consists of three main modules: an MLLA-based channel encoder, a spatiotemporal dynamics module, and a nonlinear projector. These three components are designed to extract single-channel long-term temporal features, model global cross-channel dynamics, and produce representations suitable for semantic and subject alignment, respectively.

MLLA channel encoder. We denote a preprocessed EEG sample as $\mathbf{x} \in \mathbb{R}^{C \times L}$. Each channel sequence $\mathbf{x}^{(c)}$ is divided into overlapping patches of length P with stride S, producing $\mathbf{x}_p^{(c)} \in \mathbb{R}^{N_1 \times P}$. Patches from each channel are processed independently by a Mamba-like Linear Attention (MLLA) encoder (Han et al., 2024), which combines (i) a linear+convolutional input gate for local temporal features, (ii) a linear attention operator for long-range dependencies, and (iii) a forget gate to filter salient activity. Parameters of the MLLA are shared across channels. The output $\hat{\mathbf{x}} \in \mathbb{R}^{C \times N_1 \times K_1}$ preserves rich temporal dynamics per channel.

Spatiotemporal dynamics module. To capture coordinated dynamics across electrodes, $\hat{\mathbf{x}}$ is first mapped by a trainable spatial projector $\mathbf{W}_s \in \mathbb{R}^{C \times C}$. A spatial-transition convolution with multiple dilations then extracts cross-channel patterns over different temporal scales, yielding latent representations $\tilde{\mathbf{h}}$. A local attention module—comprising temporal convolutions, pooling, linear mixing, and sigmoid activations—assigns time-varying importance weights to latent dimensions. The weighted representation $\hat{\mathbf{h}}$ highlights salient spatiotemporal patterns of brain activity. The spatiotemporal dynamics module architecture follows the dynamic attention design proposed by Shen et al. (2025).

Projector. For semantic-guided loss and inter-subject alignment loss, a projector transforms $\hat{\mathbf{h}}$ into a compact representation. It consists of average pooling followed by two depthwise temporal convolution layers with ReLU activations. The design also follows Shen et al. (2025).

4 EXPERIMENTS

4.1 DATASETS AND PREPROCESSING

SEED-DV video EEG. SEED-DV dataset (Liu et al., 2024) contains EEG recordings from 20 subjects watching 1400 natural video clips covering 40 everyday concepts across 9 coarse classes (transportation, musical, food, natural scene, human, exercise, plant, water animal, and land animal). Each concept is represented by 35 clips, organized into 7 blocks per subject with resting periods in between. Each video clip has a time duration of 2 seconds. 62-channel EEG recordings (ESI NeuroScan System) were obtained at a sampling rate of 1000 Hz during video watching.

Broderick natural-speech EEG. The Broderick dataset (Broderick et al., 2018) contains 128-channel EEG data from 19 participants listening to an audiobook (a popular American fiction novel from the mid-20th century). Each subject completed 20 trials of approximately 180 s, following the audiobook's original timeline without repetitions or discontinuities. The public released data is downsampled from 512 Hz to 128 Hz and segmented into trials.

Data preprocessing. We implemented a fully automatic pipeline using Python MNE package for SEED-DV and Broderick datasets. EEG data were first resampled to 125 Hz and band-passed to 1–47 Hz. For quality control, we detected noisy channels with a robust median-based rule: for each channel, if the proportion of samples whose absolute amplitude exceeded m times the channel median was greater than n of the trial duration, the channel was flagged as bad. We used two thresholds to capture long/short artifacts, $(m{=}3, n{=}0.4)$ and $(m{=}30, n{=}0.01)$. Flagged channels were then interpolated on a custom 10–20 montage, while frontal electrodes were temporarily *excluded* from

this first interpolation (Fp1, Fp2, F7, F8, AF3, AF4) to avoid suppressing ocular components prior to ICA. Next, we performed ICA (*infomax* method) and used mne_icalabel to automatically label components. Non-brain classes (e.g., ocular, muscle, line noise) were rejected. After ICA, noisy channels were re-detected with the same thresholds.

Data slicing. For SEED-DV, continuous EEG recordings were segmented into 1 s windows with a 0.5 s stride following Liu et al. (2024). For the Broderick natural-speech dataset, segments were locked to word onsets: each epoch extended from 0.3 s before to 0.7 s after the onset of each word, thereby capturing anticipatory and early lexical processing activity.

4.2 Experimental setup

Data spliting. We consider two training settings: *within-subject* and *cross-subject*. For SEED-DV, the *within-subject* setting uses 5 blocks for training, 1 block for validation, and 1 block for testing per subject, following Liu et al. (2024). The *cross-subject* setting randomly assigns 16 subjects for training, 2 for validation, and 2 for testing. For the Broderick dataset, the *within-subject* setting follows Jayalath et al. (2025), using 14 sessions for training, 3 sessions for validation, and 3 sessions for testing. In the *cross-subject* setting, subjects 1–15 are used for training, subjects 16–17 for validation, and subjects 18–19 for testing. All comparison methods adopt the same split protocol to ensure fairness.

Baselines. For SEED-DV, we compared our approach against GLMNet (Liu et al., 2024), the current state-of-the-art model for video concept decoding. To ensure fairness, we re-implemented GLMNet on our own preprocessed data. Additionally, we evaluated representative large pre-trained models, including MMM (Yi et al., 2023), LaBraM (Jiang et al., 2024a), and EEGPT (Wang et al., 2024a) under the same settings. These baselines cover both task-specific SOTA and general large pre-trained models, ensuring a comprehensive comparison. For the Broderick dataset, we introduced a random baseline, which predicts each word by sampling uniformly from the vocabulary at each position. Prior speech decoding models either failed to outperform the random baseline when trained on this dataset (Jayalath et al., 2025), or adopted data splits that risk data leakage (Défossez et al., 2023), so they were not included in the comparison table.

Evaluation metrics. For SEED-DV, we evaluated semantic decoding performance using both fine-grained and coarse labels, reporting top-1 and top-5 accuracy on the 40-category concept classification, and top-1 and top-3 accuracy on the 9-category coarse semantic classification, following Liu et al. (2024). For the Broderick dataset, we evaluated top-1, top-5, and top-10 accuracy on the word classification task. To ensure sufficient samples per class, we restricted the vocabulary to the 250 most frequent words in the dataset.

4.3 Comparison with state-of-the-art

Table 1 and Table 2 report within-subject and cross-subject performance on the SEED-DV dataset. In the within-subject setting (Table 1), our method achieves the best performance on 40-category classification, reaching 6.30±2.18% top-1 and 17.09±4.36% top-5 accuracy, which clearly surpasses all baselines. On the 9-category coarse classification, our model attains comparable results, with 18.83±2.13% top-1 and 43.73±3.71% top-3 accuracy, competitive with the state-of-the-art LaBraM and EEGPT models.

In the cross-subject setting (Table 2), the advantage of our approach is more pronounced. Our method substantially outperforms all baselines on both 40-category and 9-category tasks, with 6.22% top-1 and 24.49% top-5 accuracy on the 40-category classification, and 20.96% top-1 and 50.92% top-3 accuracy on the 9-category classification. These results highlight the effectiveness of our semantic-guided contrastive learning framework in improving both fine-grained and coarse-grained semantic decoding, especially under the more challenging cross-subject scenario.

Table 3 shows the performance on the Broderick dataset. This dataset is very challenging for speech decoding, even the best existing models failed to achieve above-chance performance, where the top-10 accuracy was approximately 4% (see Figure 3B in Jayalath et al. (2025)). In the within-subject setting, our method also failed to outperform the random baseline. In contrast, in the cross-subject

Table 1: Within-subject performance on the SEED-DV dataset.

Method	40-c top1	40-c top5	9-c top1	9-c top3
Random	2.50	12.50	11.11	33.33
MMM	2.61±0.70	12.69±0.98	17.47±0.23	44.85±0.61
LaBraM	3.70±1.75	15.94±2.89	18.10±3.24	**46.14±4.07**
EEGPT	3.03±1.24	13.43±2.23	17.21±1.56	44.15±2.05
GLMNet	3.95±1.85	16.72±4.07	15.67±2.70	40.99±4.22
Ours	**6.30±2.18**	**17.09±4.36**	**18.83±2.13**	43.73±3.71

Table 2: Cross-subject performance on the SEED-DV dataset.

Method	40-c top1	40-c top5	9-c top1	9-c top3
Random	2.50	12.50	11.11	33.33
MMM	2.56	12.51	17.50	45.10
LaBraM	3.16	14.50	18.62	46.71
EEGPT	2.70	13.70	16.98	44.88
GLMNet	3.03	14.80	16.69	43.29
Ours	**6.22**	**24.49**	**20.96**	**50.92**

setting, our framework demonstrates clear advantages. It achieves 1.09% top-1 accuracy, more than twice the random baseline (0.47%), and similarly outperforms the baseline on top-5 and top-10 accuracy. These results highlight that by leveraging data across participants, our model is able to extract shared semantic structure and yield substantial improvements on this highly challenging task.

Table 3: Within-subject and cross-subject semantic decoding performance on Broderick dataset.

Method	Within-subject			Cross-subject		
	top1	top5	top10	top1	top5	top10
Random	0.0052	0.0197	**0.0415**	0.0047	0.0225	0.0366
Ours	**0.0060**	**0.0201**	0.0363	**0.0109**	**0.0546**	**0.1001**

4.4 Ablation experiments

To validate the effectiveness of the loss designs, we conducted ablation experiments by removing either the semantic-guided soft contrastive loss ($\mathcal{L}_{\text{Sub-CL}}$) or the inter-subject alignment loss ($\mathcal{L}_{\text{Sub-CL}}$) (Table 4). The experiment was conducted under the cross-subject setting of the SEED-DV dataset. Without semantic alignment loss, top-1 accuracy drops from 6.22% to 5.49% on the 40-category task and from 20.96% to 18.97% on the 9-category task, indicating that semantic alignment is critical for capturing fine-grained semantic geometry.

Similarly, removing subject alignment reduces performance to 4.83% (40-c top-1) and 17.61% (9-c top-1), showing that explicit cross-subject alignment is essential for learning subject-invariant representations. The complete model consistently outperforms both ablated variants, demonstrating that the two alignment objectives are complementary and jointly contribute to robust semantic decoding.

4.5 Interpretable analysis

Latent structures of EEG embeddings. To examine whether our model learns semantic structure consistent with external language embeddings, we analyzed the pairwise correlation matrices of concept representations. Figure 2 compares CLIP-derived concept embeddings (left) with the similarity matrix of model-extracted EEG embeddings (right). EEG embeddings can recover several apparent

Table 4: Ablations of the semantic alignment or subject alignment loss.

Method	40-c top1	40-c top5	9-c top1	9-c top5
w/o $\mathcal{L}_{\text{Sem-CL}}$	5.49	22.54	18.97	47.90
w/o $\mathcal{L}_{\text{Sub-CL}}$	4.83	20.08	17.61	45.56
Complete	**6.22**	**24.49**	**20.96**	**50.92**

clusters of the concept embeddings, and reveal richer correlation patterns across concepts. Significant correlation was observed between the two matrices (Pearson $r = 0.192$, $p = 6.3 \times 10^{-8}$), suggesting that the EEG encoder aligns neural representations with the underlying semantic structure of concepts.

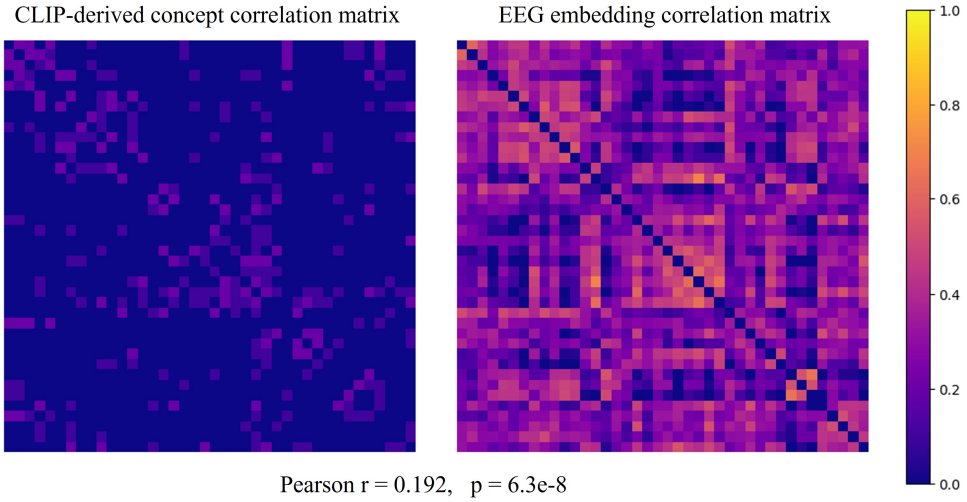

Figure 2: Comparison of correlation matrices between CLIP-derived concept embeddings and EEG embeddings.

The importance of brain regions. To evaluate the importance of different brain regions in semantic decoding, we quantified the drop in accuracy when the corresponding EEG electrodes were masked. As shown in the topographic map (Figure 3a) and the bar plot (Figure 3b), the **frontal** and **parietal** regions contributed most strongly to decoding accuracy, with accuracy drops of 0.0074 and 0.0069, respectively. Central and parieto-occipital regions also showed moderate contributions, while temporal and occipital regions had comparatively smaller effects. According to previous studies (Binder et al., 2009), the parietal region might support higher-level sensory integration, and the frontal region engages in abstract, modality-agnostic semantic processing. The relatively low contribution of the occipital primary visual areas indicates that the model does not rely solely on low-level visual signals but instead leverages distributed, higher-order neural representations.

5 Discussion and Conclusion

This work introduces Semantic-CL, a modality-agnostic framework for EEG-based semantic decoding that integrates semantic-guided soft contrastive learning with inter-subject alignment. Across both SEED-DV and Broderick datasets, the framework achieves consistent improvements in within- and cross-subject evaluation, with prominent benefits in the more challenging cross-subject task. By combining graded semantic alignment with subject-invariant representation learning, Semantic-CL offers a principled step toward unifying EEG semantic decoding across heterogeneous paradigms.

Beyond raw decoding accuracy, our analyses provide insights into the learned representations. The semantic-guided objective was validated by the significant correlation between EEG embeddings and CLIP-derived concept embeddings, indicating that the model captures graded semantic geome-

try. Region-importance analysis further highlights the roles of frontal and parietal regions, consistent with prior evidence that these areas contribute to semantic control and multimodal integration (Binder et al., 2009; Jackson, 2021). These findings suggest that the model leverages distributed higher-order representations, rather than relying solely on low-level sensory activity.

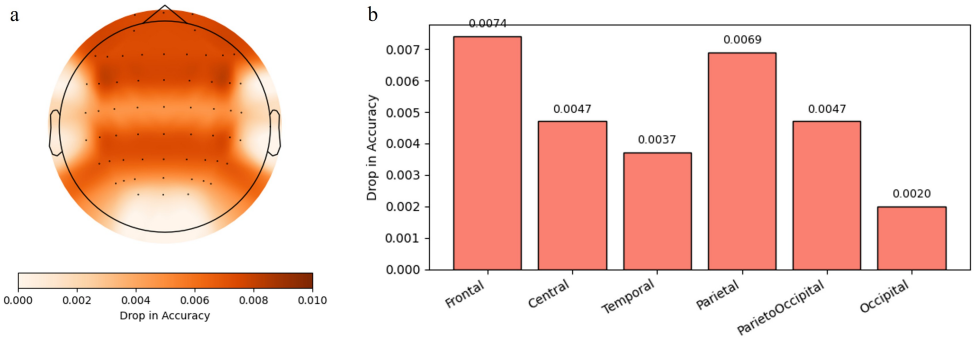

Figure 3: Importance of brain regions in semantic decoding. **a**, the topography map of accuracy drops when masking a specific region in semantic decoding. **b**, the bar plot of accuracy drops when masking different regions.

Limitations. Despite these advances, the absolute performance on Broderick remains low, reflecting the intrinsic difficulty of natural-speech decoding and the limited size of current datasets. In addition, scaling the framework to larger, more diverse corpora and incorporating temporal priors from language or vision models may further improve generalization.

Conclusion. Overall, this study provides a step toward a general framework for EEG semantic decoding across modalities and participants. By explicitly binding semantic geometry with inter-subject alignment, Semantic-CL enables more robust and interpretable decoding, advancing the development of brain–computer interfaces that capture shared semantic representations in naturalistic contexts.

6 Reproducibility Statement

We described the details of data preprocessing and data slicing in Section 4.1, the detailed training-validation-testing split methods and baseline implementations in Section 4.2, and the hyperparameter settings in Appendix A. The code is available at `https://anonymous.4open.science/r/Semantic-CL-Cross_subject-anonymous-DBD7`.

References

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796, 2009.

Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. 36:9907–9918, 2023.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097–1107, 2023.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *arXiv preprint arXiv:2405.16605*, 2024.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Rebecca L Jackson. The neural correlates of semantic control revisited. *NeuroImage*, 224:117444, 2021.

Dulhan Jayalath, Gilad Landau, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text (b2t) models across datasets. *arXiv preprint arXiv:2505.13446*, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *Advances in Neural Information Processing Systems*, 2024a.

Weibo Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for generic eeg representation learning. *arXiv preprint arXiv:2405.18765*, 2024b.

Dimitrios Kostas, Sebastian Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and contrastive self-supervised learning to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Marta Kutas and Kara D. Federmeier. Thirty years and counting: Finding meaning in the n400 component. *Annual Review of Psychology*, 62:621–647, 2011.

Marta Kutas and Steven A. Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102822–102864, 2024.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from EEG signals. In *Advances in Neural Information Processing Systems*, 2024.

Samuel A Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. Measuring shared responses across subjects using intersubject correlation, 2019.

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

Milan Rybář and Ian Daly. Neural decoding of semantic concepts: A systematic literature review. *Journal of Neural Engineering*, 19(2):021002, 2022.

Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, 2022.

Xinke Shen, Lingyi Tao, Xuyang Chen, Sen Song, Quanying Liu, and Dan Zhang. Contrastive learning of shared spatiotemporal eeg representations across individuals for naturalistic neuroscience. *NeuroImage*, 301:120890, 2024.

Xinke Shen, Runmin Gan, Kaixuan Wang, Shuyi Yang, Qingzhu Zhang, Quanying Liu, Dan Zhang, and Sen Song. Dynamic-attention-based eeg state transition modeling for emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*, 2024.

Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from eeg with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024a.

Guanhua Wang, Wei Liu, Yifan He, Chengcheng Xu, Lin Ma, and Hui Li. Eegpt: Pretrained transformer for universal and reliable eeg representation. In *Advances in Neural Information Processing Systems*, 2024b.

Chen Yang, M. Westover, and J. Sun. Biot: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

A IMPLEMENTATION DETAILS

The hyperparameter settings are summarized in Table S1 to ensure reproducibility.

Table S1: Hyperparameters of pre-training and fine-tuning.

Hyperparameters		Values
Training settings	training epochs	200
	NDE learning rate	0.0002
	MLP learning rate	0.0005
	weight decay	0.0002
	loss temperature	0.1
MLLA channel encoder	patch size	32
	patch stride	6
	hidden dim	128
	out dim	32
	depth	2
	number of attention heads	8
Spatiotemporal dynamics model	number of spatial-transition convolution	4
	time length of spatial-transition convolution	3
	dilations	[1,3,6,12]
	length of temporal filters	15
	length of average pooling	15
Projector	length of temporal filters	3
MLP	hidden dim	128

B THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this paper, we use LLMs for paper polishing, translation of the draft from Chinese to English, and proof reading.