The Semantic Bottleneck: Leveraging Semantic Representations for Non-Invasive Speech Decoding

Anonymous Authors[1]

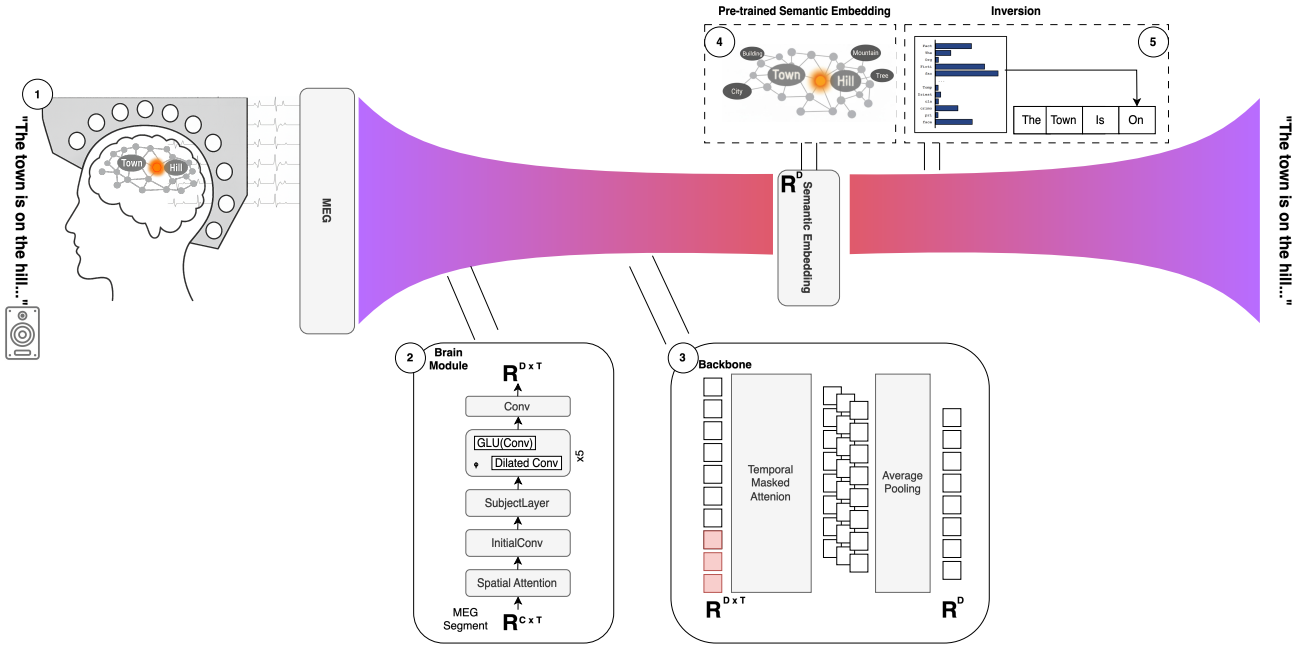

Figure 1. The Brain2Semantics2Text Method

1. MEG responses are collected while the participant listens to spoken sentence stimuli. **2.** The BrainModule extracts time-resolved neural features using spatial attention and dilated temporal convolutions. **3.** A backbone with temporal masked attention pools the neural sequence into a fixed-dimensional vector. **4.** The vector is trained to align with a pre-trained sentence-embedding space representing semantic content. **5.** The predicted semantic embedding is inverted back into natural language text.

Abstract

Making non-invasive brain–computer interfaces (BCIs) for speech decoding a practical reality has the potential to substantially improve quality of life for many individuals. However, the low signal-to-noise ratio that characterizes all non-invasive recording modalities remains a fundamental constraint, significantly hindering progress in this domain.

We introduce **"Brain2Semantics2Text"**, a new approach to non-invasive speech decoding. Rather than reconstructing speech directly at the word or phoneme level, the proposed method extends existing non-invasive Brain2Text efforts by explicitly targeting high-level semantic representations. The approach is motivated by neuroscientific evidence that semantic information is represented across distributed cortical systems, and is enabled by recent progress in semantic embedding models that can be approximately inverted back into text. We describe the central principle of the method: a *semantic bottleneck* that constrains the decoding pipeline to map neural activity into a sentence-level semantic space before text reconstruction. Finally, we report improved BERTScore performance and compare our results against prior baselines.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.**AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Speech decoding BCIs have long been a sought-after goal in both neuroscience and healthcare. Recent progress in invasive *Brain2Text* systems has demonstrated remarkable success in translating neural activity into language(Willett et al. (2023),Card et al. (2024)). However, these approaches require surgical implantation of intracranial electrodes, creating a strong incentive to develop non-invasive speech decoding solutions that are safer, more accessible, and easier to deploy.

Extending this paradigm to non-invasive modalities remains a major challenge, largely due to their lower signal-to-noise ratios. These constraints make it difficult to reliably decode fine-grained linguistic units such as phonemes or individual words. In contrast, higher-order, contextual semantic representations are spatially distributed across the cortex(Huth et al., 2016), maintain high redundancy and processed slowly(Gwilliams et al., 2025), therefore they may be more amenable to decoding from non-invasive neural signals.

This work directly targets the decoding of semantic representations from brain activity, with an explicit emphasis on the higher-level semantic branches of the proposed representational hierarchy. The *"Brain2Semantics2Text"* method maps sentence-length segments of MEG neural activity during heard speech, directly into a semantic embedding space, which is subsequently decoded into text.

Several aspects distinguish our method from prior work on *Brain2Text*:

- **Semantic Embedding Inversion.** We build on recent advances in semantic embedding inversion (Morris et al., 2023), which enable the reconstruction of text from semantic embeddings either as an intrinsic property of newer embedding models (Duquenne et al., 2023) or via general inversion techniques applicable to arbitrary pre-trained semantic spaces (Jha et al., 2025).
- **Sentence-level semantic decoding.** Our method operates at the sentence level, targeting compositional semantic from the top of the representational hierarchy (Gwilliams et al., 2023)..
- **Semantic decoding from MEG.** Unlike most prior semantic decoding work, which relies on the high spatial resolution of fMRI (Tang et al., 2023), we leverage the largest single-subject heard-speech MEG dataset of its kind to date.

2. Related Work

Tang et al. (2023) introduced a non-invasive language decoding framework that reconstructs continuous language from fMRI recordings by mapping distributed cortical responses into semantic representations, which are then decoded into word sequences capturing both perceived and imagined speech. They also proposed BERTScore (Zhang et al., 2020) as an evaluation metric to assess the semantic fidelity of reconstructed language. Building on this framework, Wang et al. (2023) later adapted the approach to MEG, demonstrating partial but promising reconstruction performance.

Défossez et al. (2023) developed a system that can decode speech from MEG by learning a mapping function from brain signal to Wav2Vec representation (Baevski et al., 2020). Specifically targeting the auditory aspects of speech processing. This method established a new workflow for speech decoding based on contrastive loss (Radford et al., 2021), as well as an architectural module that became the standard backbone for MEG.

Recent work has substantially advanced MEG-based Brain-to-Text decoding through a variety of modeling strategies. Yang et al. (2024c) align MEG signals with auditory representations to improve speech reconstruction, Yang et al. (2024a) frame Brain-to-Text decoding within a multi-modal foundation-model architecture, and Yang et al. (2024b) adapt a pre-trained Whisper model for MEG-based speech decoding. Complementarily, d'Ascoli et al. (2024) demonstrate closed-vocabulary word classification by leveraging lexical semantic embeddings augmented with sentence-level contextual information. While these approaches are capable of producing sentence-level outputs, they do not explicitly learn a direct mapping from neural signals to sentence-level semantic embeddings.

3. Method

The *Brain2Semantics2Text* method operates in two stages. In the first stage, MEG responses to continuously presented spoken sentences are mapped to vector representations in a pre-trained semantic embedding space, using `text-embedding-ada-002`. Training is guided by objectives that encourage alignment with the target embeddings while preserving their global statistical structure.

In the second stage, the predicted semantic embedding is inverted into natural language using a pre-trained inversion model(Morris et al., 2023) that reconstructs text from semantic vectors.

3.1. Backbone

MEG-to-semantic mapping. The MEG input signal $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C denotes the number of sensors and T the number of temporal samples, is first processed by a spatial attention module, followed by an initial 1×1 convolution that projects the sensor dimension into a latent feature space.

The resulting representation is then passed through a stack of dilated temporal convolutional blocks. Each block consists of dilated convolutions equipped with residual connections and gated linear units (GLUs), enabling the model to capture long-range temporal dependencies.

Temporal aggregation. To obtain a fixed-dimensional semantic representation from the time-resolved features, we use a 4-head self-attention Transformer over the temporal axis. We then pool over time with a masked mean,

$$\hat{\mathbf{y}} = \frac{\sum_{t=1}^{T} m_t \, H_t}{\sum_{t=1}^{T} m_t} \in \mathbb{R}^d,$$

where m_t is a temporal mask based on the true segment lengths, preventing length-related surface confounders from influencing the pooled embedding.

3.2. Objectives for Manifold Learning

At the core of our training setup is a SigLIP-style contrastive loss (Zhai et al., 2023), which has been shown to be effective for aligning representations across modalities with different dimensionalities and statistical characteristics (Radford et al., 2021). However, in the low-data regime typical of non-invasive speech decoding, contrastive objectives alone are insufficient to learn a the target semantic embeddings manifold.

Previous brain-to-text and word-decoding approaches have largely relied on contrastive objectives to align neural signals with semantic embeddings (Défossez et al., 2023; d'Ascoli et al., 2024). However, we observe that in the low-data regimes typical of non-invasive speech decoding, contrastive loss primarily optimizes a retrieval objective. In this setting, the model learns a mapping that enables nearest-neighbor matching under cosine similarity (Minnema & Herbelot, 2019), but does not necessarily preserve the inter-vector distances or the scale of embedding magnitudes.

This limitation is problematic for our setting, where the goal is not merely to retrieve a correct target embedding, but to learn a mapping that faithfully captures the global geometry of the target semantic manifold. To address this, we draw inspiration from the manifold learning literature (Meilă & Zhang, 2023) and introduce several auxiliary losses in addition to the SigLIP that force the model to learn the global properties of the target manifold and prevent collapse. The resulting training objective consists of the following components (Invariance, Covariance and Varaince losses are dopted from VICReg (Bardes et al., 2021)):

SigLip Loss:

$$\mathcal{L}_{\text{ctr}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log\Big(1 + \exp\Big(- t_{ij} \big(\tau \, \hat{\mathbf{y}}_i^{\top} \mathbf{y}_j + b\big)\Big)\Big), \tag{1}$$

Following SigLIP, similarity scores are scaled by a learnable temperature τ and shifted by a learnable bias b, which stabilizes optimization under the large imbalance between matching and non-matching pairs.

Invariance Loss : mean squared distance between predicted and target embeddings.

$$\mathcal{L}_{\text{inv}} = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 . \tag{2}$$

Covariance Loss: a decorrelation term that penalizes off-diagonal covariances between embedding dimensions, reducing redundancy and preventing informational collapse.

$$\mathcal{L}_{\text{cov}} = \frac{1}{d} \sum_{i \neq j} C(\mathbf{x})_{i,j}^2 . \tag{3}$$

$$C(\mathbf{x}) = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^{\top}, \tag{4}$$

Variance Loss: a hinge loss that enforces a minimum standard deviation across the batch for each embedding dimension, preventing collapse:

$$\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max(0, \, \gamma - \sigma_j(\mathbf{x})) , \tag{5}$$

$$\sigma_j(\mathbf{x}) = \sqrt{\text{Var}(x^j) + \epsilon}. \tag{6}$$

Global Cosine Alignment Loss - maximizes the average cosine similarity between predicted and target embedding vectors.

$$\mathcal{L}_{\text{gcs}} = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\mathbf{y}}_i^{\top} \mathbf{y}_i}{\|\hat{\mathbf{y}}_i\|_2 \, \|\mathbf{y}_i\|_2}. \tag{7}$$

The final loss is:

$$\mathcal{L} = \alpha \, \mathcal{L}_{\text{ctr}} + \beta \, \mathcal{L}_{\text{gcs}} + \gamma \, \mathcal{L}_{\text{inv}} + \delta \, \mathcal{L}_{\text{var}} + \eta \, \mathcal{L}_{\text{cov}}. \tag{8}$$

To test whether each component contributes to the final decoding performance, we ablate individual terms from the training objective while keeping the rest of the pipeline fixed, as shown in Table 2.

3.3. Inverting Semantic Embeddings Back To Text

Embedding Inversion (Morris et al., 2023) is formulated as an iterative conditional generation problem, where the objective is to recover a text sequence x^* given only its embedding $e^* = f(x^*)$. The procedure initializes by sampling an initial hypothesis from a base generator,

$$x_0 \sim p_\theta(x \mid e^*).$$

At each iteration t, the current hypothesis x_t is re-embedded to obtain $e_t = f(x_t)$, and a learned correction model generates an improved hypothesis conditioned on the current text and the embedding discrepancy:

$$x_{t+1} \sim p_\phi(x \mid x_t, e_t, e^*).$$

This iterative refinement progressively reduces the embedding distance $\|e_t - e^*\|$, yielding increasingly faithful reconstructions without direct optimization in discrete token space.

3.4. Data

For training our model we chose to use the new dataset: LibriBrain (Özdogan et al., 2025), the largest single-subject speech decoding MEG dataset to date:

Stimuli	Words	Unique	Sent.	Hours
Sherlock Holmes Books	466,230	16,892	32,348	52.32

Text and audio were manually corrected, normalized, and force-aligned, with sentence boundaries defined by corpus punctuation.

3.4.1. Preprocessing

The recordings were originally sampled at 1 kHz and downsampled to 250 Hz to preserve oscillations into the high-gamma range (70–125 Hz).

4. Experiments

We compare our results against prior *Brain2Text* approaches (d'Ascoli et al., 2024). For completeness, we report standard metrics, including BLEU, ROUGE, and Word Error Rate (WER). However, consistent with prior work in semantic decoding (Tang et al., 2023), our primary evaluation focuses on BERTScore (Zhang et al., 2020), which is designed to assess the semantic fidelity of the reconstructed text and is largely decoupled from exact lexical or verbatim matching.

4.1. Results

Neural decoding gain. Figure 2 evaluates the contribution of the neural signal at two stages of the decoding pipeline. On the left, we report ADA embedding cosine similarity between the predicted and ground-truth sentence embeddings. The baseline reflects performance before incorporating neural information, capturing the extent to which the semantic model can exploit corpus-level regularities without access to brain data. The improvement over this baseline therefore reflects the model's actual neural decoding performance.

On the right, we report the corresponding BERTScore after embedding inversion into text. The improvement is smaller than in embedding space, suggesting that gains in ADA similarity are only partially preserved during text reconstruction. Together, these results show a measurable semantic uplift from the neural signal, while highlighting embedding-to-text inversion as a source of variability in the full decoding pipeline.

Comparison with prior methods. Table 1 compares the proposed *Brain2Semantics2Text* approach with prior *Brain2Text* decoding methods. Our method achieves strong BERTScore performance, indicating semantic similarity between reconstructed and reference sentences, while underperforming on word-level metrics such as WER and BLEU. This pattern is expected, since the proposed approach explicitly targets sentence-level semantic representations and does not rely on exact word-level alignment information, unlike the D'Ascolli method. In particular, WER is disproportionately affected by the absence of explicit alignment, since decoded outputs may contain a different number of words than the reference sentences even when semantic content is preserved. By contrast, ROUGE-1, which is recall-based, still indicates a meaningful word-level signal.

5. Discussion

5.1. Limitations and Future Work.

The current work is the first attempt of Speech Decoding based on direct mapping from MEG into semantic representation. Learning the semantic manifold proved to be challenging. Other the normal culprits such as low Signal-To-noise and small data regime, a meaningful capture of Semantic Embeddings will require more semantic variability. We observed that the model learned semantics that is highly biased towards the specific corpus. Future work will require data collection with experiment design that emphasizes the variability of topics and concepts, perhaps even guided by the unique properties of the target semantic manifold (Pereira et al., 2018).

5.2. Conclusion.

While fully non-invasive speech decoding systems remain a long-term goal, decoding contextual semantic content from brain activity represents an important step toward their realization.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2006.11477. arXiv:2006.11477.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-

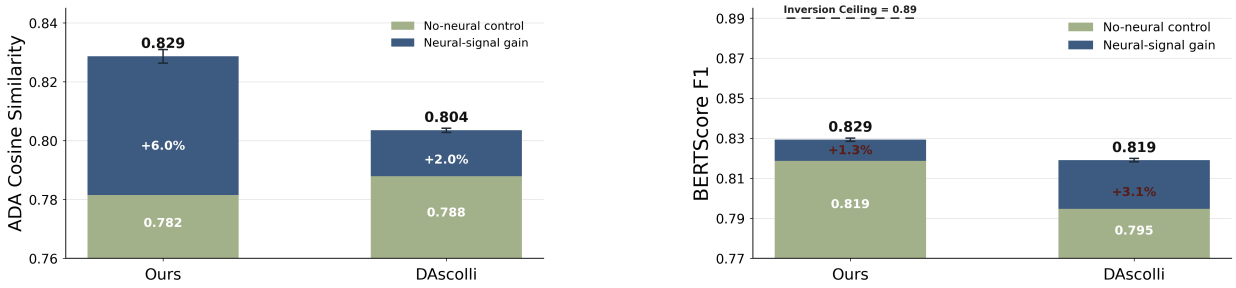

Figure 2. **Embedding-space and text-space results.** Left: ADA cosine similarity between predicted and ground-truth sentence embeddings. Right: BERTScore F1 of the decoded text. The blue segment shows the gain over the no-neural control.

Method	**WER** ↓	**BLEU-1** ↑	**ROUGE-1** ↑	**BERTScore** ↑
Ours	1.925 ± 0.100	0.100 ± 0.008	0.132 ± 0.003	$\mathbf{0.829 \pm 0.001}$
Ours (noise control)	2.648 ± 0.415	0.086 ± 0.009	0.114 ± 0.004	0.819 ± 0.006
D'Ascoli	$\mathbf{0.871 \pm 0.002}$	$\mathbf{0.190 \pm 0.004}$	$\mathbf{0.172 \pm 0.004}$	0.820 ± 0.001

Table 1. Comparison of decoding methods (mean ± s.d.). D'Ascoli performs best on lexical overlap metrics (WER, BLEU-1, and ROUGE-1), whereas Brain2Semantics2Text achieves the highest BERTScore.

Method	BERTScore
Full objective (SigLIP + VICReg + global cosine)	**0.8288**
w/o global cosine loss	0.7964
w/o VICReg invariance term	0.8268
w/o VICReg variance term	0.8264
w/o VICReg covariance term	0.8203

Table 2. Loss ablation results measured by their effect on decoded-sentence BERTScore F1.

invariance-covariance regularization for self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2105.04906. URL https://arxiv.org/abs/2105.04906.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., Fan, C., Vahdati Nia, M., Deo, D. R., Srinivasan, A., Choi, E. Y., Glasser, M. F., Hochberg, L. R., Shahlaie, K., Stavisky, S. D., and Brandman, D. M. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132. URL https://www.nejm.org/doi/full/10.1056/NEJMoa2314132.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C., and King, J.-R. Decoding individual words from non-invasive brain recordings across 723 participants. *arXiv preprint arXiv:2412.17829*, 2024. URL https://arxiv.org/abs/2412.17829.

Duquenne, P.-A., Schwenk, H., and Sagot, B. Sonar: Sentence-level multimodal and language-agnostic representations, 2023. URL https://arxiv.org/abs/2308.11466.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., and King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5:1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

Gwilliams, L., Marantz, A., Poeppel, D., and King, J.-R. Top-down information shapes lexical processing when listening to continuous speech. *Language, Cognition and Neuroscience*, 39(1):1045–1058, 2023. doi: 10.1080/23273798.2023.2171072.

Gwilliams, L. et al. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *bioRxiv Preprint*, 2025. doi: 10.1101/2024.04.19.590280v2. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11042271/. Preprint. PMCID: PMC11042271; PMID: 38659750.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016. doi: 10.1038/nature17637. URL https://doi.org/10.1038/nature17637.

Jha, R., Zhang, C., Shmatikov, V., and Morris, J. X. Harnessing the universal geometry of embeddings, 2025. URL https://arxiv.org/abs/2505.12540.

Meilă, M. and Zhang, H. Manifold learning: what, how, and why. *arXiv preprint arXiv:2311.03757*, 2023. doi: 10.48550/arXiv.2311.03757. URL https://arxiv.org/abs/2311.03757.

Minnema, G. and Herbelot, A. From brain space to distributional space: The perilous journeys of fmri decoding. In Alva-Manchego, F., Choi, E., and Khashabi, D. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 155–161, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2021. URL https://aclanthology.org/P19-2021/.

Morris, J. X., Zhao, W., Chiu, J. T., Shmatikov, V., and Rush, A. M. Language model inversion. *arXiv*, abs/2311.13647, 2023. URL https://arxiv.org/abs/2311.13647. Preprint.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, 2018. doi: 10.1038/s41467-018-03068-4.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Tang, J., LeBel, A., Jain, S., and Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023. doi: 10.1038/s41593-023-01304-9.

Wang, B., Xu, X., Zhang, L., Xiao, B., Wu, X., and Chen, J. Semantic reconstruction of continuous language from meg signals. *arXiv*, abs/2309.07701, 2023. URL https://arxiv.org/abs/2309.07701.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023. doi: 10.1038/s41586-023-06377-x. URL https://www.nature.com/articles/s41586-023-06377-x.

Yang, Y., Duan, Y., Jo, H., Zhang, Q., Xu, R., Parker Jones, O., Hu, X., Lin, C.-t., and Xiong, H. Neugpt: Unified multi-modal neural gpt. *arXiv preprint arXiv:2410.20916*, 2024a. URL https://arxiv.org/abs/2410.20916.

Yang, Y., Duan, Y., Zhang, Q., Jo, H., Zhou, J., Lee, W. H., Xu, R., and Xiong, H. Neuspeech: Decode neural signal as speech. *arXiv preprint arXiv:2403.01748*, 2024b. URL https://arxiv.org/abs/2403.01748.

Yang, Y., Jo, H., Duan, Y., Zhang, Q., Zhou, J., Lee, W. H., Xu, R., and Xiong, H. Mad: Multi-alignment meg-to-text decoding. *arXiv*, abs/2406.01512, 2024c. URL https://arxiv.org/abs/2406.01512.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. Preprint: https://arxiv.org/abs/1904.09675.

Özdogan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., Woolrich, M., and Parker Jones, O. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. *arXiv preprint arXiv:2506.02098*, 2025. doi: 10.48550/arXiv.2506.02098. URL https://arxiv.org/abs/2506.02098.

A. Appendix

Impact Statement

This paper presents work towards non-invasive speech decoding, with potential applications in brain-computer interfaces for individuals who have lost the ability to speak. Clinical deployment remains distant, as current performance remains below what communication aids require, and substantial work remains. We note that neural decoding technologies raise obvious privacy concerns, as they involve inferring mental content from brain activity. Our work uses only publicly available research datasets with their own ethics approvals and decodes perceived speech rather than covert thought. As decoding capabilities improve, the field will need norms around consent, data ownership, and the boundary between assistive and surveillant applications. These are questions we do not resolve here but consider essential.

B. Hyperparameters

Parameter	Value
Dropout	0.6
Learning rate	3e-5
Optimizer	AdamW
Loss type	SigLIP
VICReg weight	5.0
Cosine loss weight	6.0
Contrastive loss weight	2.0
Aggregation	Attention (4 heads)

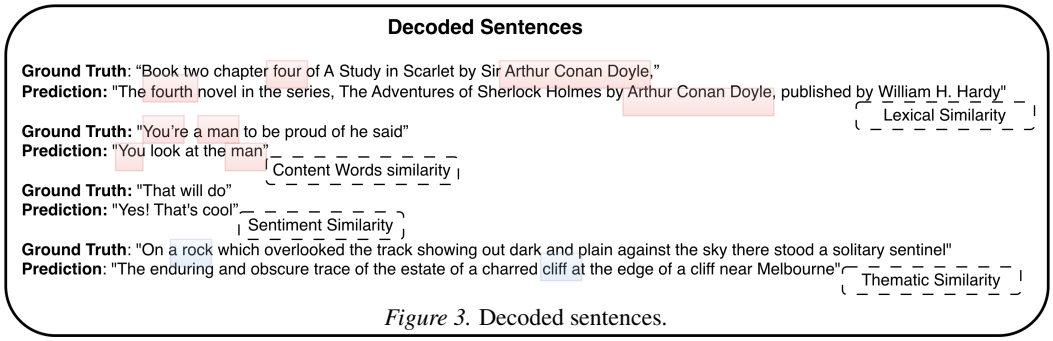

Figure 3. Decoded sentences.

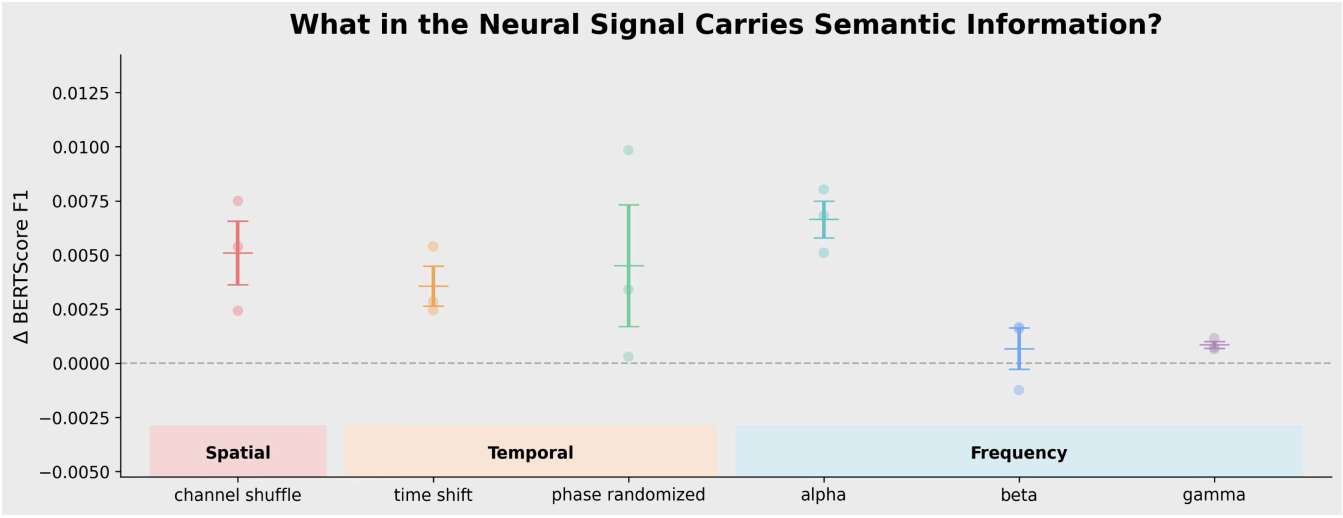

Figure 4. **Neural signal properties supporting semantic decoding.** We examine how different properties of the MEG signal affect semantic mapping performance, providing an interpretable view of which aspects of the neural response contribute most to decoding.