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

Gilad D. Landau [1] **Dulhan Jayalath** [1] **Oiwi Parker Jones** [1]

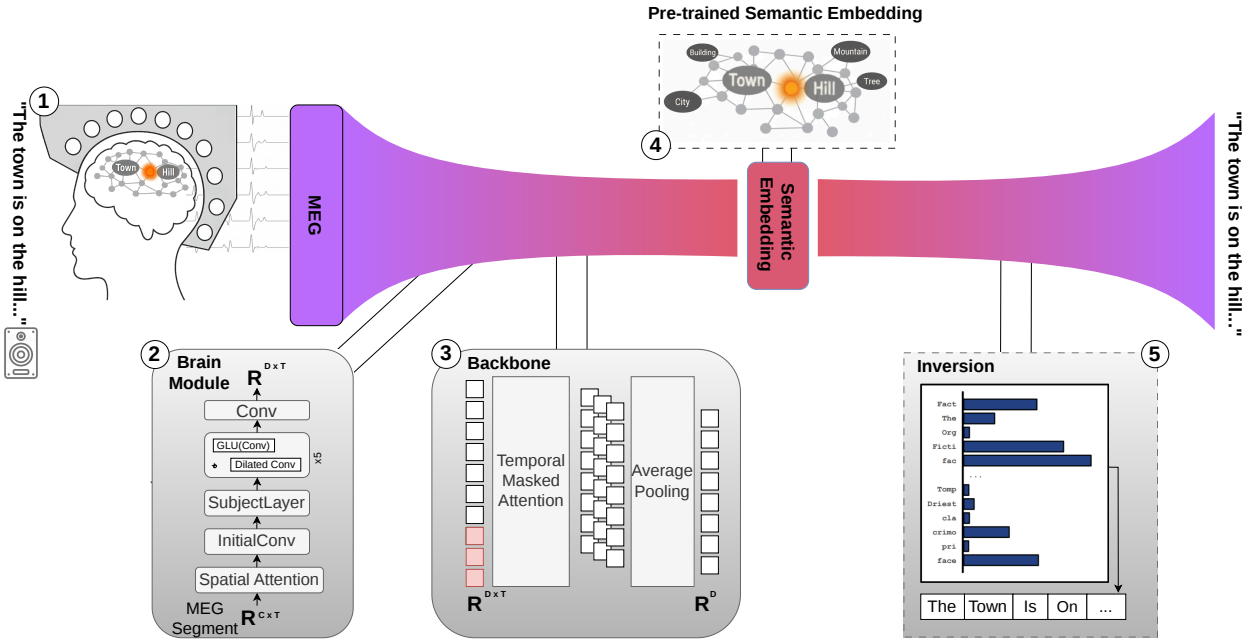

Figure 1. The *Brain2Semantics2Text* method.

1. MEG responses are collected while the participant listens to spoken sentence stimuli. **2.** The Brain Module 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

Non-invasive speech decoding remains constrained by the low signal-to-noise ratio of neural recordings, which makes fine-grained reconstruction of phonemes or individual words difficult. Motivated by neuroscientific evidence that high-level semantic representations are distributed across cortical regions and evolve over slower temporal scales, we hypothesize that semantic content may provide a more suitable target for non-invasive decoding than low-level acoustic or lexical features. We introduce **Brain2Semantics2Text**, a method that reconstructs text through an intermediate semantic embedding space. Our model maps sentence-level MEG responses into a semantic manifold and then inverts the predicted embeddings into natural language. This semantic bottleneck enables recovery of high-level meaning without word-level alignment. We describe the core principles of the approach, its implementation, and the strategies used to mitigate the challenges of learning a reliable neural-to-semantic mapping. Finally, we compare against prior non-invasive *Brain2Text* methods and show improved sentence-level decoding results.

[1] PNPL, University of Oxford. Correspondence to: Gilad D. Landau <gilad.landau@jesus.ox.ac.uk>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

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), are redundant, and evolve over slower temporal scales (Gwilliams et al., 2025). These properties suggest that semantic content may be more amenable to decoding from non-invasive neural signals.

This work directly targets semantic representations in brain activity, with an explicit emphasis on compositional sentence-level meaning rather than acoustic or lexical features. The *Brain2Semantics2Text* method maps sentence-length MEG responses 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., 2024), which enable text reconstruction from semantic embeddings either as an intrinsic property of newer embedding models (Duquenne et al., 2023) or through 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 representations near the top of the representational hierarchy (Gwilliams et al., 2025). This shifts the decoding objective away from exact word or phoneme recovery and toward reconstruction of sentence-level meaning.
- **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. (2024) later adapted the approach to MEG, demonstrating partial but promising reconstruction performance.

Défossez et al. (2023) developed a system that decodes speech from MEG by learning a mapping from brain signals to Wav2Vec representations (Baevski et al., 2020), specifically targeting 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 a standard backbone for MEG-based speech decoding.

Recent work has substantially advanced MEG-based *Brain2Text* 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 Brain2Text 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., 2024) 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 the target manifold of semantic embeddings.

Previous Brain2Text 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, 2024) and introduce several auxiliary losses in addition to the SigLIP loss that encourage the model to preserve global properties of the target manifold and prevent collapse. The resulting training objective consists of the following components (invariance, covariance and variance losses are adopted from VICReg (Bardes et al., 2021)):

Let $\hat{\mathbf{y}}_i$ denote the predicted semantic embedding for example i, $\mathbf{y}_i$ the corresponding target embedding, and $\hat{\mathbf{Y}}$ the batch of predicted embeddings.

SigLIP Loss: A contrastive alignment term that formulates predicted–target matching as independent pairwise classification rather than a batch-wise softmax objective. This is useful for sentence-level semantic decoding, where different non-matching sentences may still be semantically related and should not necessarily be treated as mutually exclusive classes. We also found this objective more stable in the low-data MEG setting, particularly with small batches.

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

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

$$\mathcal{L}_{\text{inv}} = \frac{1}{n} \sum_{i=1}^{n} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 \,. \quad (2)$$

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

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

$$C(\hat{\mathbf{Y}}) = \frac{1}{n-1} \sum_{k=1}^{n} (\hat{\mathbf{y}}_k - \bar{\hat{\mathbf{y}}})(\hat{\mathbf{y}}_k - \bar{\hat{\mathbf{y}}})^\top . \quad (4)$$

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

$$\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max\big(0, \, s_0 - S(\hat{y}^j, \epsilon)\big) \,, \quad (5)$$

$$S(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon} . \quad (6)$$

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

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

The final loss is:

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

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 3.

3.3. Inverting Semantic Embeddings Back to Text

Embedding inversion (Morris et al., 2024) 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 and evaluation, we use LibriBrain (Özdogan et al., 2025), the largest single-subject speech-decoding MEG dataset available at the time of writing.

Table 1. LibriBrain data splits used in our experiments. Validation and test sets are held-out recording sessions.

Stimuli	Words	Unique	Sent.	Hours
Sherlock (Train)	600,107	20,837	40,659	61.80
Sherlock (Validation)	3,427	1,155	198	0.36
Sherlock (Test)	3,577	1,210	172	0.38
Sherlock (Total)	607,111	20,971	41,029	62.54

Text and audio were manually corrected, normalized, and force-aligned, with sentence boundaries defined by corpus punctuation. 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; Li et al., 2025) using standard text-generation metrics: WER, BLEU, ROUGE, and BERTScore. While WER, BLEU, and ROUGE primarily measure lexical overlap and word-level reconstruction accuracy, BERTScore provides a complementary estimate of sentence-level semantic similarity. This is particularly important for our setting, where successful decoding may preserve the meaning of a sentence even when its exact wording is not recovered.

At the same time, text-generation metrics alone cannot determine whether a decoded sentence was driven by neural information or by linguistic and dataset-level priors in the generation model. Following Jo et al. (2024), we therefore include a noise-control analysis to estimate how much of the decoded output is attributable to the neural input rather than to textual priors alone.

4.1. Results

Comparison with prior methods. Table 2 compares the proposed *Brain2Semantics2Text* approach with prior *Brain2Text* decoding methods. The word-level metrics reveal a clear performance gap between word-level decoding, represented by d'Ascoli (d'Ascoli et al., 2024), and sentence-level decoding. This gap is expected, since word-level decoding operates in a much more constrained setting, with a closed vocabulary and access to exact word-level alignment. By contrast, when compared with acoustic-based sentence-level decoding, represented by BrainECHO (Li et al., 2025), our method performs favorably, achieving better results on BLEU-1, ROUGE-1, and BERTScore. These improvements hold both in absolute terms and when measured as neural-signal uplift relative to the noise control (Figure 2). Our method tends to generate more words than appear in the ground-truth sentence, making WER less informative; recall-based metrics such as ROUGE-1, however, show a distinct signal-based improvement.

Signal uplift. Figure 2 evaluates the contribution of the neural signal across methods, measured using both BERTScore and ADA cosine similarity. BERTScore serves as the standard metric for comparing sentence-level reconstructions, while ADA cosine similarity provides a complementary measure of semantic similarity that more directly reflects the explicit training target of *Brain2Semantics2Text*. Because all methods may exploit text priors and corpus-level regularities that are not driven by the brain signal, we measure performance relative to a noise-control baseline. This signal-dependent uplift estimates the contribution of the neural signal itself, rather than reconstruction quality attributable to the inversion model, language prior, or corpus-level bias.

On BERTScore, our method yields a 1.2-point uplift over its noise-control baseline, exceeding the uplift of the previous sentence-level method, BrainECHO, but remaining below the 3.2-point uplift of d'Ascoli et al. However, d'Ascoli et al. operate at the word level and require exact word-aligned supervision, whereas our method performs sentence-level

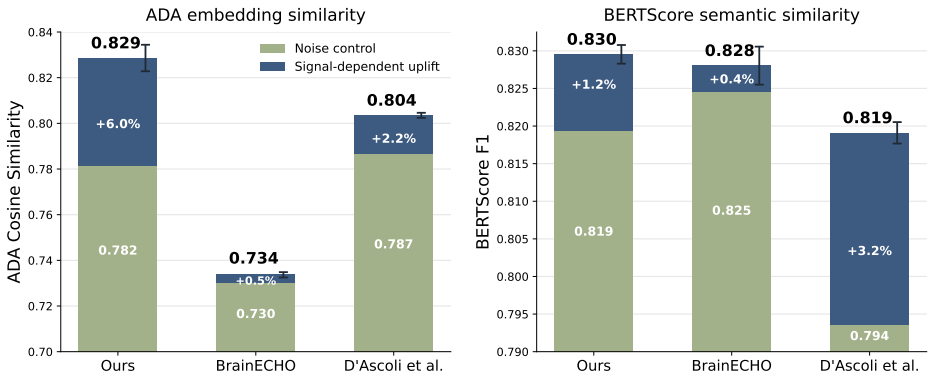

Figure 2. **Signal uplift.** Comparison of signal-dependent improvement over noise-control baselines across decoding methods.

Table 2. Comparison of decoding methods (mean ± s.d.). Brain2Semantics2Text and BrainECHO operate at the sentence level and do not rely on word-level alignment, whereas d'Ascoli et al. (2024) use word-aligned supervision. Lexical-overlap metrics such as WER, BLEU-1, and ROUGE-1 therefore favor methods with access to word-level timing or lexical supervision, while BERTScore better reflects the sentence-level semantic reconstruction objective of our method.

Method	**WER** ↓	**BLEU-1** ↑	**ROUGE-1** ↑	**BERTScore** ↑
Sentence-level decoding, no word-level alignment				
Ours	1.925 ± 0.100	**0.100 ± 0.008**	**0.132 ± 0.003**	**0.830 ± 0.001**
Ours (noise control)	2.648 ± 0.415	0.086 ± 0.009	0.113 ± 0.004	0.819 ± 0.006
BrainECHO	1.018 ± 0.030	0.061 ± 0.008	0.091 ± 0.007	0.828 ± 0.002
BrainECHO (noise control)	1.012 ± 0.005	0.056 ± 0.006	0.085 ± 0.008	0.825 ± 0.002
Word-level decoding with word-aligned supervision				
d'Ascoli et al.	**0.871 ± 0.002**	**0.190 ± 0.004**	**0.172 ± 0.004**	0.820 ± 0.001
d'Ascoli et al. (noise control)	0.994 ± 0.003	0.072 ± 0.024	0.057 ± 0.021	0.794 ± 0.004

Table 3. Loss ablation results measured as mean decoded-sentence BERTScore F1 on the test set. Values are mean ± standard deviation across runs. Higher values indicate better decoded-sentence semantic similarity.

Method	BERTScore
Full objective	**0.8297 ± 0.0008**
w/o global cosine loss	0.8103 ± 0.0077
w/o VICReg invariance term	0.8263 ± 0.0025
w/o VICReg variance term	0.8267 ± 0.0015
w/o VICReg covariance term	0.8270 ± 0.0012

decoding without word-level alignment. On ADA cosine similarity, our method shows the largest signal-dependent uplift, improving by 6.0 points and exceeding the corresponding uplift of the word-level method. This discrepancy between semantic metrics suggests that current evaluation measures capture different aspects of sentence-level decoding performance. More broadly, it highlights the need for better standardized metrics for evaluating semantic reconstruction in brain-decoding models.

5. Discussion

5.1. Limitations and Future Work

This work provides an initial demonstration of speech decoding through a direct mapping of MEG signals into semantic representations. Learning this semantic manifold proved to be challenging. Beyond the usual constraints of non-invasive neural recordings, such as low SNR and limited training data, reliable semantic encoding is likely to require greater semantic variability in the training corpus. We observed that the model learned semantic structure that was strongly biased toward the specific corpus used in this study. Therefore, future work should involve data collection protocols that emphasize variability in topics and concepts, perhaps guided by the 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, offering a way to mitigate the fundamental noise limitations of non-invasive neural recordings.

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-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.

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.

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

Gwilliams, L., Marantz, A., Poeppel, D., and King, J.-R. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122, 2025. doi: 10.1073/pnas.2422097122. URL https://www.pnas.org/doi/10.1073/pnas.2422097122.

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.

Jo, H., Yang, Y., Han, J., Duan, Y., Xiong, H., and Lee, W. H. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024. doi: 10.48550/arXiv.2405.06459. URL https://arxiv.org/abs/2405.06459.

Li, J., Song, Z., Wang, J., Zhang, M., Liu, H., Zhang, M., and Zhang, Z. BrainECHO: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2762–2778, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.142. URL https://aclanthology.org/2025.findings-acl.142/.

Meilă, M. and Zhang, H. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11:393–417, 2024. doi: 10.1146/annurev-statistics-040522-115238.

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. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=t9dWHpGkPj.

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.

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. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. doi: 10.1109/ICASSP48485.2024.10448281. 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.

A. Impact Statement

This paper presents work toward 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 is still below what communication aids require, and substantial further work is needed. We also note that neural decoding technologies raise clear privacy concerns, since 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

Table 4. Hyperparameters used for MEG-to-semantic embedding training in **Brain2Semantics2Text**.

Parameter	Value
Dropout	0.6
Learning rate	3×10^{-5}
Optimizer	AdamW
Loss type	SigLIP
VICReg weight	5.0
Cosine loss weight	6.0
Contrastive loss weight	2.0
Aggregation	Attention (4 heads)

C. Compute Resources

All MEG-to-semantic embedding models were trained on a single NVIDIA GPU using 4 CPU cores and 64 GiB of system memory. A typical full-data run took approximately 16–18 GPU-hours, corresponding to about 6 minutes per epoch for 150–170 epochs. Final training across seven random seeds used approximately 110–130 GPU-hours, excluding exploratory runs, failed jobs, and downstream evaluation or decoding.

D. Sampled Decodings

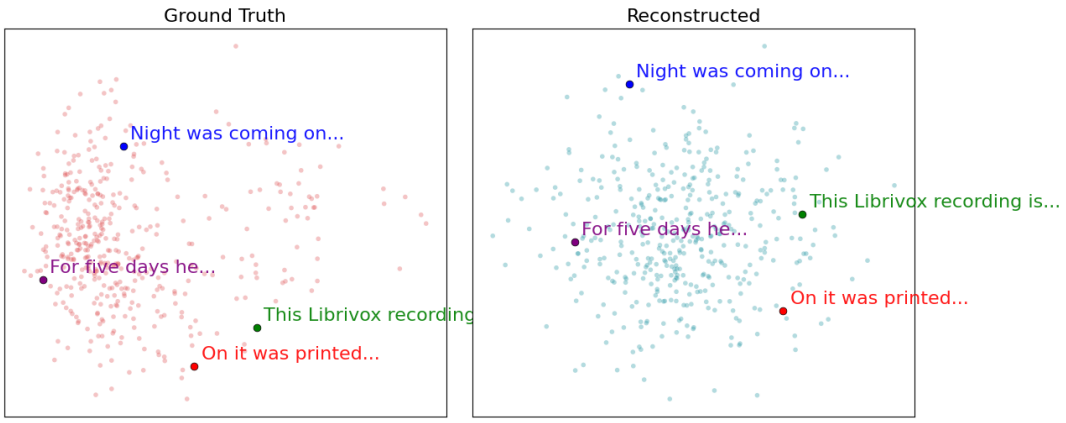

Figure 3. **Semantic geometry of reconstructed sentence embeddings.** PCA projections of sentence-level embeddings for ground-truth sentences and reconstructed predictions. Highlighted examples illustrate that relative positions and global structure are largely preserved under reconstruction. Examples are shown for geometric comparison only, and textual reconstructions may differ from the ground truth.

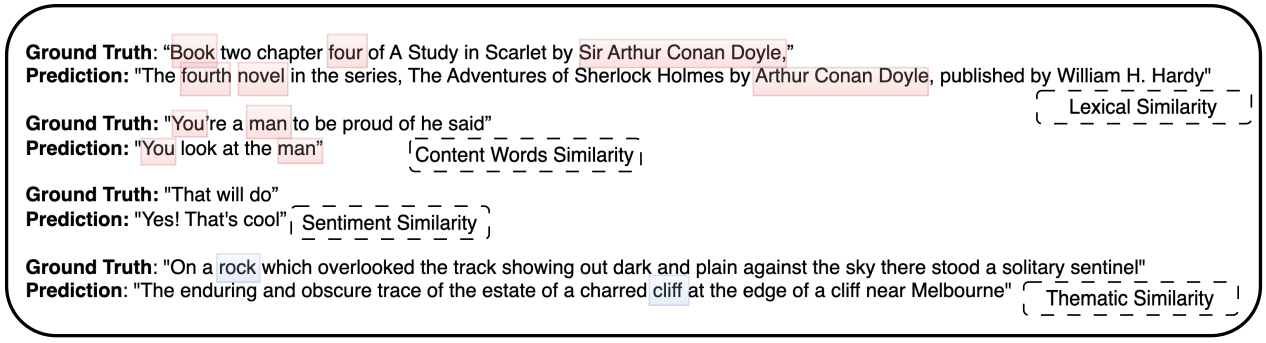

Figure 4. **Qualitative decoding examples.** Example decoded sentences shown for qualitative illustration of semantic similarity between the target sentence and the reconstructed output.

E. Neural Signal Analysis

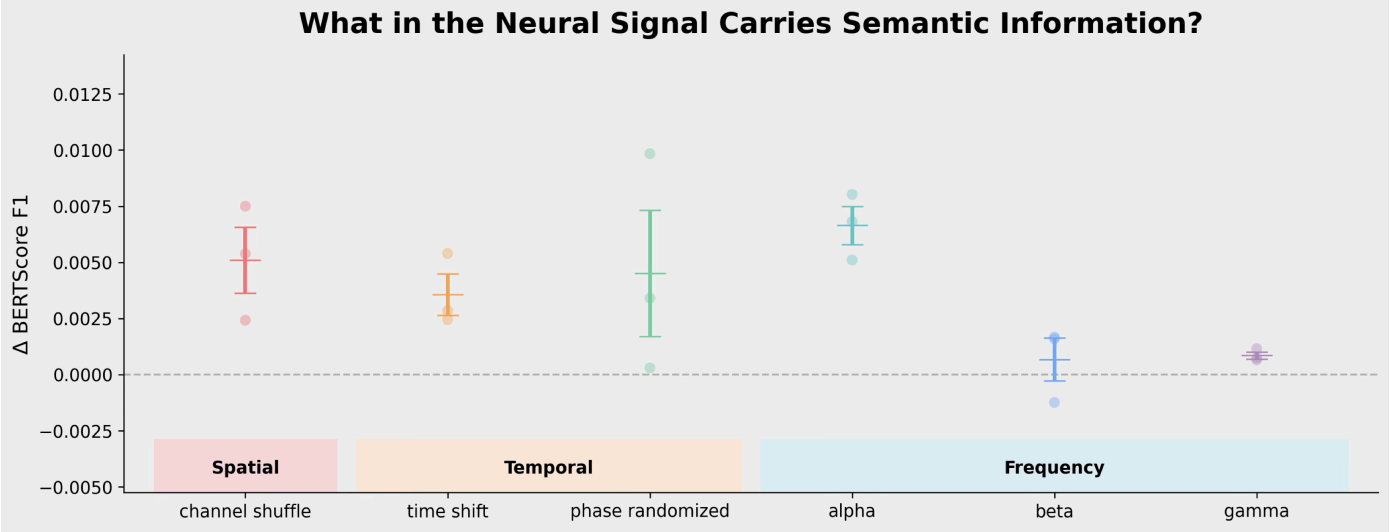

Figure 5. **Neural signal properties supporting semantic decoding.** We examine how different properties of the MEG signal affect semantic mapping performance, providing an interpretable view of which aspects of the neural response contribute most to decoding.