CONTEXT-AWARE OPEN-VOCABULARY MEG-TO-TEXT DECODING WITH LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding magnetoencephalography (MEG) signals into text holds immense potential for restoring communication through non-invasive brain-computer interfaces, but remains highly challenging. The scarcity of paired MEG and text data limits models from learning robust signal to text mappings, causing prior approaches to fall short in the open-vocabulary setting. We address this challenge by reframing open-vocabulary decoding as a context-guided inference task, where MEG signals carry the linguistic content and preceding context guides the language model toward decoding it. From this point of view, we present a novel **Co**ntext-aware **ME**G-to-**T**ext decoding framework (**CoMET**) that leverages preceding context as a semantic anchor, narrowing the textual search space. CoMET first aligns MEG signals with corresponding text at the word-level, and subsequently maps these representations into the latent space of a large language model under contextual conditioning. Extensive evaluations on two MEG-text datasets demonstrate that while prior open-vocabulary decoders collapse to near random performance, CoMET achieves state-of-the-art open-vocabulary decoding performance, highlighting its potential for practical brain-to-text decoding. Code will be made available upon publication.

1 INTRODUCTION

Brain-to-text decoding (Défossez et al., 2023; Feghhi et al., 2025; Willett et al., 2023) has emerged as a promising solution for restoring communication for those who have lost it by translating brain activity into natural language. In particular, non-invasive brain signals such as magnetoencephalography (MEG) are widely used due to their high accessibility (Boyko et al., 2024). MEG-to-text decoding therefore serves as a fundamental testbed in this field for determining whether semantic information can be reliably decoded from non-invasive brain signals. While recent studies (Li et al., 2025; Wang & Ji, 2022; Yang et al., 2024a;b) have demonstrated remarkable success by utilizing pretrained models to map brain signals to text, their success is largely confined to restricted experimental setups. Specifically, dataset splits are often constructed such that the training and test sets share a common pool of sentences. When the text pool is reused across splits, the model can memorize this pool, which reduces the decoding task to a closed vocabulary retrieval problem rather than learning a signal to text mapping as shown in Fig. 1(a).

For practical brain-to-text decoding, open-vocabulary decoding that the model generalizes to unseen text is essential. Early studies (Duan et al., 2023; Feng et al., 2023a; Wang & Ji, 2022; Xi et al., 2023) in this setting heavily rely on teacher-forcing (Williams & Zipser, 1989) during inference, which feeds the ground-truth previous token at each decoding step. However, such ground-truth tokens are unavailable at inference, limiting their real-world viability (Fig. 1(b)). A recent study (Jo et al., 2025) reveals that once teacher-forcing is removed, existing open-vocabulary decoders produce outputs indistinguishable from those generated by replacing brain signals with random noise, indicating that current methods fail to extract meaningful information from non-invasive brain signals. This limitation arises from the difficulty in aligning non-invasive brain signals to text. While open-vocabulary decoding requires massive paired datasets to cover a broad text space, the difficulty of collecting large-scale data limits the learning of correlations between brain signals and text.

To address the challenges of open-vocabulary brain-to-text decoding, we propose a novel **Co**ntext-aware **ME**G-to-**T**ext decoding framework **(CoMET)** that leverages textual context as a semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

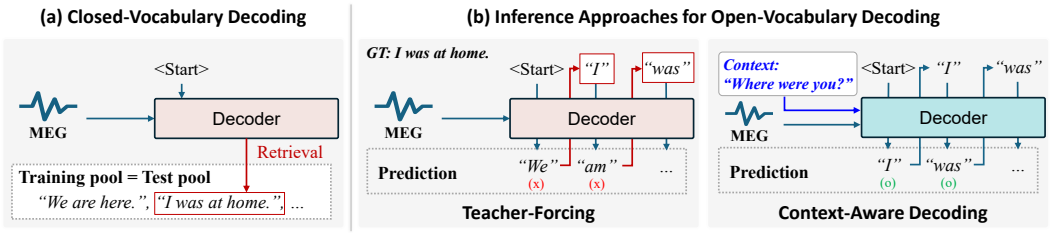

Figure 1: Conceptual comparison of brain-to-text decoding paradigms. The ground-truth is "*I was at home.*". (a) Closed-vocabulary decoding, where target sentences appear in both the training and test sets. (b) Inference approaches for open-vocabulary decoding. Teacher-forcing based decoding (left), where the previous ground-truth word is provided for next-word prediction. The proposed context-aware decoding (right), which utilizes textual context during inference.

anchor (Fig. 1(b)). Our key insight draws from the fundamental nature of human communication. In natural language, the semantic content of an upcoming sentence is strongly guided by the preceding context. Such context is naturally available in real-world scenarios, including ongoing dialogue and continuous reading. Unlike teacher-forcing, which requires access to the target itself, conditioning on preceding context is practically realizable at inference time. From this point of view, the uncertainty of open-vocabulary decoding can be reduced by conditioning the decoding process on such text context (Radford et al., 2019). This aligns with the principle of large language models (LLMs), which predict the next word given previous words. In this study, we define text context as the sentences immediately preceding the target. Providing this context enables the LLM to maximize the conditional probabilities of context-associated tokens, thereby effectively narrowing the search space in open-vocabulary decoding (Brown et al., 2020). However, since multiple plausible continuations may follow from the same context, context alone cannot determine the specific content of the target sentence. MEG signals therefore provide this missing linguistic information, complementing context within the constrained decoding space.

To facilitate this, CoMET first aligns MEG signals with the text embedding space through word-level MEG-to-text alignment, allowing for precise alignment of local semantic representations. The MEG representations are then projected into the LLM latent space via context-aware MEG-LLM alignment. The LLM decodes these representations conditioned on a structured template that combines MEG signals with preceding textual context. Extensive evaluations on two MEG datasets demonstrate that this context-aware strategy significantly enhances performance in open-vocabulary decoding. The main contributions of this work are as follows:

(1) We formulate open-vocabulary brain-to-text decoding as a context-guided inference task. Unlike teacher-forcing, the preceding context is available at inference, which makes the proposed setting practically realizable. To the best of our knowledge, this is the first study to leverage the context for open-vocabulary brain-to-text decoding.

(2) We propose a multi-stage alignment strategy that aligns MEG signal with local semantics and maps the MEG representations into LLM latent space conditioned on the preceding context, ensuring precise cross-modal decoding.

(3) We achieve state-of-the-art performance on two MEG-text pair datasets. While prior decoders collapse to near random performance, CoMET generates semantically meaningful text, demonstrating the potential for practical brain-to-text decoding.

2 METHOD

2.1 PROBLEM SETTINGS.

The model can access both current MEG signals and the preceding sentences for open-vocabulary context-aware MEG-to-text decoding. Each sample consists of four components as follows: (1) MEG signal $x \in \mathbb{R}^{C \times T}$, where C and T respectively denote the number of channels and the number of time points, (2) sentence s, which is a ground-truth sentence corresponding to x, (3) time

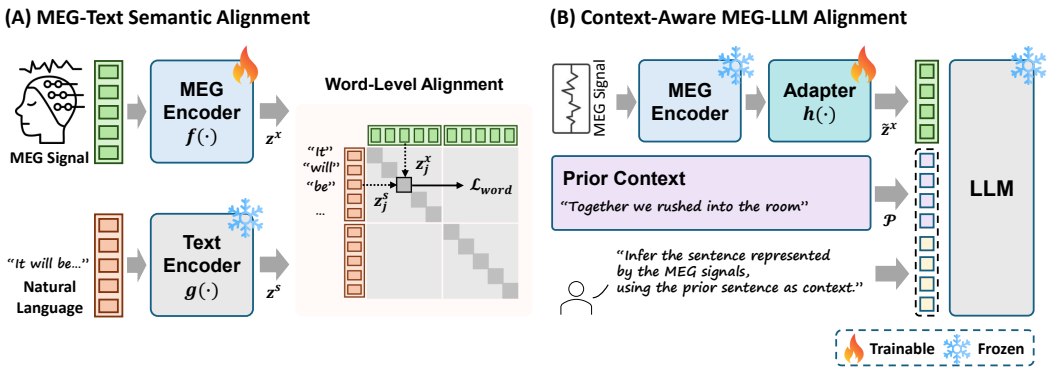

Figure 2: Overview of the proposed framework, consisting of a two-stage process. (A) MEG-Text Semantic Alignment. The MEG signal and its corresponding text are fed into the MEG encoder and the text encoder, respectively. These representations are used to train the MEG encoder via word-level contrastive learning, while the text encoder remains frozen. (B) Context-Aware MEG-LLM Alignment. To align the MEG embeddings with the latent space of the LLM, we train the adapter while freezing the others. LLM takes the MEG representations and the preceding sentences formatted with a structured template.

annotation $\mathcal{Y} = \{y_j\}_{j=1}^{K}$, which provides word-level temporal markers defining the start and end indices for each of the K words in s, and (4) preceding context $\mathcal{P} = \{p_m\}_{m=1}^{M}$, which consists of M sentences that appear immediately before s.

2.2 MEG-Text Local Semantic Alignment.

As illustrated in Fig. 2(A), we first train an MEG encoder f using a contrastive learning strategy (Radford et al., 2021) to align MEG signals to the corresponding text. Prior methods such as MAD (Yang et al., 2024b) and BrainECHO (Li et al., 2025) align global sentence-level representations derived from continuous acoustic features. These enforce consistency only between aggregated global embeddings, capturing coarse semantic correspondence. In contrast, we introduce word-level alignment that explicitly supervises local representations to match individual lexical embeddings, enabling the encoder to capture fine-grained semantics.

Given an input sentence-level MEG signal x, the encoder $f : \mathbb{R}^{C\times T} \to \mathbb{R}^{d\times T}$ produces a sequence of token-level representations z^x. We then extract K word-level brain embeddings $\{z_j^x \in \mathbb{R}^{d\times W}\}_{j=1}^{K}$ from z^x by centering a fixed-size window W at each word interval, identified via time index $\mathcal{Y}$. The fixed-size window captures surrounding signals rather than restricting each embedding to the exact word interval, since neural responses are not sharply aligned with word boundaries. In parallel, the corresponding word s_j is encoded by a frozen text encoder g, yielding $z_j^s = g(s_j)$. We align the two representations using a symmetric InfoNCE loss $\mathcal{L}_{NCE}$ (Oord et al., 2018), where each corresponding MEG-text pair serves as a positive and all other pairs within the batch as negatives:

$$\mathcal{L}_{word} = \frac{1}{2K}\sum_{j=1}^{K}(\mathcal{L}_{NCE}(z_j^x, z_j^s) + \mathcal{L}_{NCE}(z_j^s, z_j^x)). \tag{1}$$

Word-level supervision is used only during this stage. Once trained, the MEG encoder f operates on sentence-level MEG signals and produces a sequence representation z^x that retains local semantics.

2.3 Context-Aware MEG-LLM Alignment

To align the MEG and LLM representation space, we employ an adapter h consisting of a multi-layer projection following (Liu et al., 2023), as shown in Fig. 2(B). Specifically, the adapter is trained under contextual conditioning on $\mathcal{P}$ to project MEG representations into a context-constrained subset of the LLM latent space. For each sentence-level MEG signal x, the frozen MEG encoder produces $z^x = f(x)$, preserving local semantic structure without requiring word-level annotations at this

stage. The adapter then projects it into $\tilde{z}^x = h(z^x)$, matching the hidden dimensionality of the LLM. The LLM input is constructed using a structured template consisting of (1) the projected MEG embedding $\tilde{z}^x$ prepended as a soft prefix, (2) a task instruction, and (3) the preceding sentences $\mathcal{P}$ as prior context. The adapter is optimized to minimize the negative log-likelihood of generating the target sentence s conditioned on $\mathcal{P}$ and $\tilde{z}^x$ as follows:

$$\mathcal{L}_{\text{adapter}} = -\mathbb{E}\left[\sum_{t=1}^{|s|} \log p_{llm}\left(s_t \mid s_{<t}, \mathcal{P}, \tilde{z}^x\right)\right], \tag{2}$$

where s_t denotes the t-th token of the target sentence s, $|s|$ is the total number of tokens in s, and $p_{llm}(\cdot)$ is the conditional probability distribution modeled by the frozen LLM. This ensures that MEG representations are aligned not merely to project into the LLM latent space, but to be compatible with the preceding context, thereby enabling context-aware decoding.

3 Experimental Settings

3.1 Datasets

We use two MEG-text pair datasets such as LibriBrain (Özdogan et al., 2025) and GWilliams (Gwilliams et al., 2023) acquired when the participant listened to the audio book.

(1) LibriBrain dataset (Özdogan et al., 2025) contains 52.32 hour recording of MEG signals, accompanied by paired time-locked annotations for linguistic events including speech, words, and phonemes. The data were acquired over 95 sessions from a single participant during an English audio book listening task. The recordings were obtained from 306-channel system at 1000Hz, but the provided dataset is downsampled to 250Hz.

(2) GWilliams dataset (Gwilliams et al., 2023) comprises data from 27 English speakers with time-locked annotations for words and phonemes. The data were acquired over two identical one-hour sessions per participant during a naturalistic story listening task. The recordings were obtained from a 208-channel system sampled at 1,000 Hz. We used data from 21 of the 27 participants, excluding six whose MEG recordings were unavailable.

For data pre-processing, we apply notch filtering and bandpass filtering to both datasets and downsample the MEG signals to 200Hz. We then normalize MEG signals to $[-1, 1]$ with outlier clipping. Each MEG recording is segmented at the sentence level based on word-level time annotations.

We split the data into training, validation, and test sets at an 8:1:1 ratio based on sentence index order. To prevent leakage from dependencies between adjacent sentences, we avoid random shuffling and exclude sentences adjacent to the boundaries from the validation and test sets. This ensures the preceding sentence used as context at inference time does not overlap with the training set.

3.2 Implementation Details

DeepConvNet (Schirrmeister et al., 2017) with a MLP projection head is used as the MEG encoder to process variable-length brain signal inputs, with the word-level window W set to 128. As the text encoder, we employ an encoder-only Transformer (WhereIsAI/UAE-Large-V1) (Li & Li, 2024) following a BERT-style architecture. The MEG encoder is trained for 20 epochs using AdamW (Loshchilov & Hutter, 2019), a learning rate of 5×10^{-3}, a batch size of 256, and a contrastive-loss temperature τ of 0.07. The trained encoder is frozen during adapter training. All experiments are conducted with a single fixed random seed.

We adopt Qwen2.5-14B-Instruct (Yang & et al., 2024) as the LLM. The adapter h is implemented as two linear layers with GELU activation and trained with a single preceding sentence ($M = 1$), a learning rate of 1e-4, and a batch size of 1 for 5 epochs. The LLM receives the projected MEG embeddings as a soft prefix before the textual instruction, which includes the prior sentence as context followed by the task prompt "*Infer the sentence represented by the MEG signals, using the prior sentence as context.*" At inference, text is generated using beam search with a beam size of 5, a repetition penalty of 5.0, and a no-repeat n-gram size of 2, following the inference configuration of BrainECHO (Li et al., 2025). The other details are shown in Appendix B.

Table 1: Comparison of MEG-to-Text decoding performance on LibriBrain and GWilliams datasets.

	Method	BLEU-N (%) ↑			ROUGE-1 (%) ↑			WER(%) ↓
		N=1	N=2	N=3	P	R	F	
LibriBrain	EEG-to-Text	9.06	2.01	0.48	10.12	**22.57**	12.60	213.82
	NeuSpeech	12.29	2.78	0.59	10.57	12.12	11.13	114.80
	MAD	6.94	1.12	0.00	10.92	7.30	8.07	101.05
	BrainECHO	12.65	3.00	0.62	13.06	16.56	13.68	133.23
	CoMET	**16.58**	**5.51**	**2.28**	**18.23**	16.78	**17.37**	**96.35**
GWilliams	EEG-to-Text	7.02	1.29	0.21	7.48	**14.94**	9.28	196.30
	NeuSpeech	6.52	0.78	0.19	7.10	5.88	6.24	**105.87**
	MAD	6.39	0.94	0.00	6.42	6.59	6.10	118.41
	BrainECHO	4.36	0.82	0.00	6.79	5.16	5.54	110.89
	CoMET	**10.33**	**2.65**	**0.91**	**10.41**	12.13	**10.97**	110.83

4 Experimental Results

4.1 Brain-to-Text Decoding Performance

We evaluate the MEG-to-text decoding performance of CoMET against four baselines including EEG-to-Text (Wang & Ji, 2022), NeuSpeech (Yang et al., 2024a), MAD (Yang et al., 2024b) and BrainECHO (Li et al., 2025). Table 1 presents the decoding performance on two datasets. All methods are trained and evaluated on the same dataset splits for fair comparison, and receive the same sentence-level MEG signals as CoMET. Following their original formulations, the baselines operate on each sentence independently without access to preceding context. These baselines decode each sentence in isolation and provide no mechanism for incorporating preceding context. To the best of our knowledge, no existing MEG-to-text method supports context-aware decoding, and this capability is precisely what CoMET introduces. It is worth noting that the scores remain modest across all models, underscoring the challenges of the open-vocabulary text decoding from non-invasive brain signals. However, while baselines fail to decode MEG signals, CoMET extracts meaningful linguistic patterns and demonstrates the feasibility of context-aware decoding in the open-vocabulary setting.

On the LibriBrain dataset, we observe a severe collapse of the baselines in generating higher-order n-grams, indicating a failure to capture semantics from MEG signals. In contrast, CoMET achieves significant improvements across most evaluation metrics. A similar tendency appears on the GWilliams dataset, which is more challenging due to fewer MEG-text pairs and the inclusion of multiple subjects. The scores of all baselines fall below 7% on most BLEU-N and ROUGE-1 metrics. CoMET consistently maintains substantial gains over the baselines under the same conditions, suggesting that context enables the model to infer semantic representations even when paired data is scarce. Across both datasets, these improvements indicate that the multi-stage alignment allows the LLM to capture semantic information in MEG signals, while prior context serves as a semantic anchor that constrains the decoding distribution toward contextually relevant tokens. Although absolute performance remains modest, the results support context-aware decoding as a promising direction for open-vocabulary brain-to-text decoding.

Baselines occasionally outperform CoMET on ROUGE-1 recall and WER, but these gains reflect length bias rather than improved decoding quality (Lin, 2004; Sun et al., 2019; Morris et al., 2004). ROUGE-1 recall measures how much of the target sentence is covered by the decoded sentence. This makes it susceptible to length bias, as longer outputs are more likely to incidentally include reference words. EEG-to-Text tends to decode longer sentences than CoMET, which inflates its ROUGE-1 recall score. In contrast, CoMET achieves higher BLEU-N, ROUGE-1 precision, and F-measure, indicating stronger overall lexical overlap with the target. WER is also sensitive to the length of outputs, as extra words in the decoded sentence are penalized as insertion errors. On the GWilliams dataset, NeuSpeech achieves a slightly lower WER than CoMET by producing outputs of similar length to the target sentence. However, its substantially lower BLEU-N and ROUGE-1 scores indicate limited contextual correctness. Beyond these length-biased metrics, CoMET consistently outperforms all baselines, demonstrating superior contextual correctness and semantic similarity.

Table 2: Ablation study for input variants on LibriBrain. *Context only* uses the preceding context without MEG signals. *Random MEG + Context* replaces the input MEG with another MEG signal randomly sampled from the dataset. *MEG only* uses MEG signals without the preceding context, and *MEG + Context* is the proposed CoMET.

Input	BLEU-N (%) ↑			ROUGE-1 (%) ↑			WER(%) ↓
	N=1	N=2	N=3	P	R	F	
Context only (pretrained)	4.08	0.99	0.22	4.65	**33.01**	7.82	688.18
Context only (LoRA fine-tuned)	3.82	0.90	0.15	4.49	30.14	7.33	686.76
Random MEG + Context	5.47	1.28	0.31	10.00	13.10	9.81	278.30
MEG + Random Context	13.20	3.22	1.21	11.89	11.63	11.67	101.39
MEG only	12.80	3.98	1.34	14.91	14.01	13.41	113.15
MEG + Context (CoMET)	**16.58**	**5.51**	**2.28**	**18.23**	16.78	**17.37**	**96.35**

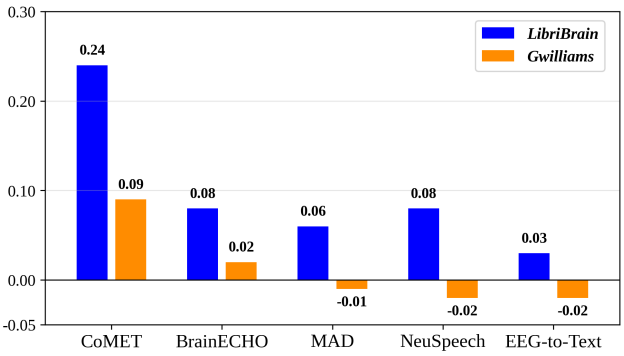

Figure 3: Semantic similarity comparison on LibriBrain and GWilliams datasets.

4.2 Semantic Equivalence Evaluation

While BLEU-N and ROUGE-1 capture exact word and n-gram overlaps, they fail to reflect semantic equivalence when the generated text conveys the same meaning through different surface forms. To complement these surface-level metrics, we additionally report BERT†, the rescaled BERTScore (Zhang et al., 2019) computed via RoBERTa-large (Liu et al., 2019). It measures contextual similarity between the decoded and target sentences in the embedding space of a pretrained language model. The rescaling is defined as BERT† = (BERTScore - b) / (1 - b), where b is the average BERTScore computed over randomly paired sentences. Higher BERT† indicates that the generated text is semantically closer to the target.

Fig. 3 presents BERT† scores on both datasets. On the LibriBrain dataset, all baselines yield scores near zero, with outputs that are semantically indistinguishable from random sentence pairs. In contrast, CoMET achieves a BERT† of 0.24, more than 3$\times$ higher than the best baseline. A similar pattern emerges on the GWilliams dataset, where baselines obtain scores around or below zero. CoMET, on the other hand, achieves a BERT† of 0.09 under the same challenging conditions. These results demonstrate that CoMET produces text that is not only lexically but also semantically aligned with the target, going beyond the surface overlap that lexical metrics may capture by chance.

4.3 Ablation on Input Components

To validate the necessity of the preceding context and to isolate the contribution of the MEG signal, we compare six variants, as shown in Table 2. The two inputs (*i.e.*, MEG and context) play distinct roles in decoding. The preceding context alone (*i.e.*, *Context only*) attains at most 4.08 BLEU-1, and cannot identify the target sentence in the absence of the MEG signal. With only MEG signals (*i.e.*, *MEG only*), the MEG-text local semantic alignment allows the model to outperform the baselines, but the performance remains below that obtained with the preceding context, since the model must generate the target words without any constraint on the candidate tokens. The constraint imposed by the context is effective only when the context is relevant to the target. Pairing the MEG signal with a randomly sampled context (*i.e.*, *MEG + Random Context*) drops the performance to 13.20 BLEU-1,

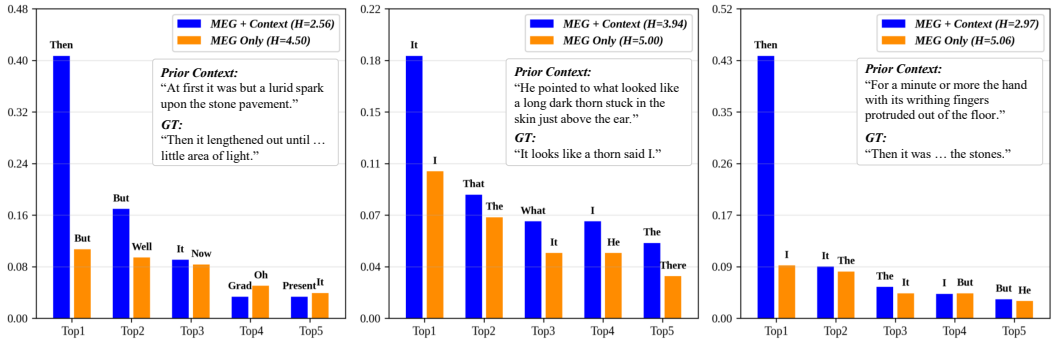

Figure 4: Comparison of the top-5 predicted tokens at the next position with and without context. H indicates the entropy of the predicted probability distribution, where low values reflect concentrated predictions. Across three examples, incorporating context yields a peaked distribution and assigns higher probability to contextually aligned tokens (*e.g.*, "*Then*", "*It*"), whereas MEG Only produces a flat distribution with less distinguishable candidates.

comparable to the *MEG only* setting. When the MEG signal is paired with the relevant preceding context (*i.e.*, *MEG + Context*), the performance reaches 16.58 BLEU-1, as the preceding context reduces the search space of the LLM, and the MEG signal provides the target linguistic content.

The two *Context only* variants further indicate that the gains of CoMET do not stem from the language priors of the LLM or from the memorization of the source texts. *Context only (pretrained)* produces generic continuations driven by language priors, which indicates that the language priors of the LLM alone cannot decode exact words of the target sentence and lead the LLM to hallucinate. Since both datasets derive from publicly available audiobooks, a further concern is that the LLM reproduces memorized targets given the context. The *LoRA fine-tuned* variant addresses this concern, as it fails to recover the targets despite being trained on the target split. It tends to reproduce training sentences due to overfitting as illustrated in Fig. 1 (a).

To verify that the model does not rely on the preceding context alone to decode the target sentence, we replace the input MEG with another MEG signal randomly sampled from the dataset while keeping the preceding context (*i.e.*, *Random MEG + Context*). The input remains a real MEG signal, but it no longer corresponds to the target sentence. The performance drops from 16.58 to 5.47 BLEU-1, close to the *Context only* floor. Corrupting the context instead of the MEG signal is far less damaging, as the *MEG + Random Context* setting retains 13.20 BLEU-1. This asymmetry shows that the decoded sentence relies on the MEG representation obtained from the trained encoder. The preceding context narrows the search space of the LLM, but it cannot compensate for a MEG signal that does not correspond to the target.

4.4 First-Token Probability Analysis under Context

The first-token distribution is examined to isolate how the preceding context narrows the search space toward semantically relevant continuations. This differs from the generation used at inference, which selects the final output from multiple high-scoring sequence candidates, but reveals how probability mass is distributed across plausible next-token candidates and thus provides insight into prediction confidence. Fig. 4 compares the top-5 next-token probabilities and shows that combining the MEG signal with preceding context assigns higher probabilities to tokens that are contextually coherent and aligned with the linguistic content encoded in the MEG signal. For example, when "*At first it was but a lurid spark upon the stone pavement.*" is given as the preceding sentence, "*At first*" implies a temporal progression so that "*Then*" receives the highest probability as the next token. In "*He pointed to what looked like a long dark thorn stuck in the skin just above the ear.*", "*It*" receives a high probability as the next token, as it naturally continues the description of the object being pointed to. Without the preceding context, the MEG signal is decoded over the entire vocabulary rather than over a restricted set of plausible continuations, leading to a less confident prediction distribution. This observation is further supported by the entropy of the next-token distribution. Averaged over the entire test set, the next-token entropy decreases from 5.06 without

Table 3: Ablation studies on LibriBrain Dataset. (a) Comparison of alignment strategies in the MEG-text semantic alignment stage. Sentence aligns sentence-level representations between MEG and text, while Word aligns word-level representations, (b) Performance with varying number of preceding sentences M, and (c) Performance with varying LLM backbone sizes. Qwen2.5-Instruct models at three scales are compared.

Ablation	Variants	BLEU-N (%) ↑			ROUGE-1 (%) ↑			WER(%) ↓
		N=1	N=2	N=3	P	R	F	
(a) Alignment	Sentence	14.33	4.77	2.11	16.13	15.30	15.11	104.83
	Word	16.58	5.51	2.28	18.23	16.78	17.37	96.35
(b) Context M	0	12.80	3.98	1.34	14.91	14.01	13.41	113.15
	1	16.58	5.51	2.28	18.23	16.78	17.37	96.35
	2	17.25	6.24	2.92	16.88	17.20	16.75	103.58
	3	16.95	5.60	2.69	16.16	17.17	16.45	103.82
(c) LLM Size	7B	16.50	5.26	2.26	15.80	16.05	15.82	99.87
	14B	16.58	5.51	2.28	18.23	16.78	17.37	96.35
	32B	17.70	5.83	2.34	18.08	18.10	18.00	99.52

prior context to 4.13 with prior context. These results show that the preceding context reduces the search space and allows the model to select the target token more reliably.

4.5 Analysis of Alignment Strategy

To analyze the effect of alignment at different levels, we compare sentence-level (*i.e.*, Sentence) and word-level (*i.e.*, Word) MEG–text alignment as shown in Table 3 (a). Although sentence-level alignment improves over the baselines, word-level alignment consistently achieves higher performance across all metrics. Since LLMs process text as word tokens (Peng et al., 2025; Wang et al., 2024), learning word-level semantics in MEG representations aligns naturally with the token granularity of the model. Moreover, word-level alignment generalizes better to unseen sentences. Exact sentences rarely recur across splits, whereas individual words appear repeatedly, making alignment more generalizable. While this may bias the model toward generating repetitive sentence patterns composed of frequent words, conditioning generation on prior context helps suppress such generic continuations while preserving the generalization benefits of word-level alignment.

4.6 Varying Number of Preceding Sentences

We vary the number of preceding sentences M to analyze the effect of the amount of preceding context as shown in Table 3 (b). Providing preceding context yields a substantial improvement over the no-context setting (*i.e.*, M=0), confirming that preceding context provides essential semantic guidance. Beyond M=1, CoMET maintains stable performance across different values of M, with all settings substantially outperforming the baselines. This indicates that even a single preceding sentence is sufficient to anchor the decoding process, and that the framework is robust to the amount of preceding context provided. The slight degradation in WER and inflation in ROUGE-1 recall observed at larger M are consistent with the length-related effects, since increasing M tends to produce longer generated outputs. Specifically, the average number of words per generated sentence grows from 12.6 (max 60) for $M = 1$, to 14.6 for $M = 2$, and 15.0 (max 107) for $M = 3$.

4.7 Effect of LLM Scale

Table 3(c) shows that scaling the LLM from 7B to 14B improves performance across metrics, while both models substantially outperform all baselines, demonstrating that CoMET remains effective even with a smaller LLM. Further scaling to 32B improves BLEU-N and ROUGE-1 but slightly worsens WER, accompanied by an increase in ROUGE-1 recall. It aligns with the length-related effects, since larger LLMs tend to generate more elaborate outputs. This pattern is consistent with prior observations that larger LLMs tend to produce more fluent but less faithful outputs, exchanging cautious avoidance for confident generation (Zhou et al., 2024). In our cross-modal setting, stronger language priors in larger backbones may further attenuate the conditioning effect of weaker inputs such as MEG signals (Zhang et al., 2025).

Table 4: MEG-to-Text decoding examples on the LibriBrain dataset. *Context only* denotes outputs generated by the pretrained LLM conditioned solely on the preceding context without MEG signals.

(a) Target: Not one word of truth has he ever told me.	
Preceding Context	He has lied to me the villain in every conceivable way.
EEG-to-Text	It is a matter of great importance to us that we have found our way out of the woods and back into town.
NeuSpeech	Leighton asked me and you told me that I missed it.
MAD	to be who will interpret familiar beings.
BrainECHO	I have no idea what it is in the book that I answered.
Context only	He always [lies/deceives/fibs].
CoMET	**He has never told me the truth in his life.**
(b) Target: Yes sir there is another man upon the moor.	
Preceding Context	You know that there is another man then.
EEG-to-Text	I have to say that I am a little taken aback by what has happened in the last few days. It is all very strange and unexpected indeed.
NeuSpeech	Did you and your friends cry with each other all over the place?
MAD	I believe that it will be the beginning of 10 years.
BrainECHO	in the morning which I had remembered.
Context only	It seems like you're aware that someone else is involved in this situation.
CoMET	**Yes I know that there is another man.**
(c) Target: I have no complaint to make.	
Preceding Context	She treated me as fairly as ever a woman treated a man.
EEG-to-Text	It was a long time ago and I have not seen it again since then. It is as if nothing has changed at all in the intervening years.
NeuSpeech	for a long time.
MAD	and be link to now.
BrainECHO	in the morning, but how he remembered me.
Context only	Her treatment stood out even more given the challenging situation we were in.
CoMET	**I have no complaint to make.**

4.8 Decoding Examples

Table 4 presents qualitative results from the LibriBrain dataset. The baselines produce semantically incoherent outputs that are disconnected from the preceding context and lack well-formed syntactic structure. *Context only* yields plausible continuations, but misses the specific words encoded in MEG signals. This confirms that MEG signals are essential for decoding and that performance gains do not arise solely from the inherent contextual generation ability of the LLM. In contrast, CoMET successfully decodes brain signals into semantically aligned text in (a), recovering specific target words such as "*Yes*" and "*another man*" in (b), and achieving exact word-level reconstruction of the target sentence in (c).

5 Conclusion

In this study, we introduce a novel context-aware MEG-to-text decoding framework designed to facilitate open-vocabulary decoding. The proposed method leverages the preceding context as a semantic anchor to ensure precise cross-modal decoding. By aligning MEG signals with text at the word level and mapping MEG representations into the latent space of an LLM, the proposed method effectively transforms the open-vocabulary decoding problem into the context-guided inference task. Extensive evaluations on the two datasets demonstrate that the proposed method significantly outperforms existing baselines. While baselines completely fail to generate meaningful text in open-vocabulary settings, the proposed method successfully decodes contextually relevant semantic text. Ultimately, our findings validate that the use of context is a promising pathway toward practical brain-to-text decoding for individuals with communication impairments.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting datasets, or interpreting results, and generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. AI-modified code was verified and tested for correctness by the authors, and AI-suggested grammar corrections were reviewed by the authors to ensure the intended meaning was preserved. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our work. The full formulation of CoMET is given in Section 2, including the problem setting and input components (Section 2.1), the word-level MEG-text local semantic alignment with fixed-size windows centered at each word interval and the symmetric InfoNCE objective (Eq. 1), and the context-aware MEG-LLM alignment with the adapter objective (Eq. 2). Section 3.1 describes the datasets and preprocessing, including the participant selection for GWilliams, notch and bandpass filtering, downsampling, normalization with outlier clipping, sentence-level segmentation, and the sentence-index-based 8:1:1 split that excludes boundary-adjacent sentences to prevent context leakage. Both datasets, LibriBrain and GWilliams, are publicly available. Section 3.2 and Appendix B.1 specify the MEG encoder architecture and word-level window size, the pretrained text encoder and LLM backbones, the adapter architecture, the two-stage training schedule, the optimizer and all hyperparameters, and the hardware used. Appendix B.2 provides the decoding configuration, Appendix B.3 gives the exact prompt template fed to the LLM, and Appendix B.4 together with Section 4.2 define the evaluation metrics, including the rescaled BERTScore. All baselines are trained and evaluated on the same data splits as CoMET (Section 4). The anonymized source code is provided in the supplementary material and will be publicly released upon publication.

REFERENCES

Maria Boyko, Polina Druzhinina, Georgii Kormakov, Aleksandra Beliaeva, and Maxim Sharaev. Megformer: enhancing speech decoding from brain activity through extended semantic representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 281–290. Springer, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: discrete eeg waves encoding for brain dynamics to text translation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 9907–9918, 2023.

Ebrahim Feghhi, Shreyas Kaasyap, Nima Ryan Hadidi, and Jonathan Kao. Time-masked transformers with lightweight test-time adaptation for neural speech decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883, 2023a.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883, 2023b. doi: 10.1109/TNSRE.2023.3314642.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating eeg-to-text models through noise-based performance analysis. *Scientific Reports*, 2025.

Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. 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, 2025.

Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1825–1839, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.101.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, number 4-8, pp. 2004, 2004.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. Libribrain: Over 50 hours of within-subject meg to improve speech decoding methods at scale. *arXiv preprint arXiv:2506.02098*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qiwei Peng, Yekun Chai, and Anders Søgaard. Understanding subword compositionality of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22535–22546, 2025.

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

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

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 21–29, 2019.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Dixuan Wang, Yanda Li, Junyuan Jiang, Zepeng Ding, Ziqin Luo, Guochao Jiang, Jiaqing Liang, and Deqing Yang. Tokenization matters! degrading large language models through challenging their tokenization. *arXiv preprint arXiv:2405.17067*, 2024.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. doi: 10.1162/neco.1989.1.2.270.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13277–13291, 2023.

An Yang and et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yiqian Yang, Yiqun Duan, Qiang Zhang, Hyejeong Jo, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Neuspeech: Decode neural signal as speech. *arXiv preprint arXiv:2403.01748*, 2024a.

Yiqian Yang, Hyejeong Jo, Yiqun Duan, Qiang Zhang, Jinni Zhou, Xuming Hu, Won Hee Lee, Renjing Xu, and Hui Xiong. Mad: Multi-alignment meg-to-text decoding. *arXiv preprint arXiv:2406.01512*, 2024b.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1): 346, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4232–4241, 2025.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

APPENDIX

A RELATED WORKS

A.1 BRAIN-TO-TEXT DECODING

Brain-to-text decoding aims to translate neural activity into natural language Willett et al. (2023), and recent studies use non-invasive signals such as EEG and MEG for accessibility. The central difficulty lies in learning a reliable mapping from noisy signals to text with limited paired data. Early studies attach pre-trained sequence-to-sequence language models to a brain encoder. EEG-to-Text Wang & Ji (2022) feeds a sequence of word-level EEG features into a pre-trained BART, where each feature is extracted from an EEG segment delimited by eye-tracking fixation boundaries and therefore presupposes an explicit word-level alignment between the neural signal and the target tokens. Subsequent works inherit this paradigm and extend it through discrete EEG encodings Duan et al. (2023), curriculum contrastive alignment Feng et al. (2023b), and unified cognitive signal reconstruction Xi et al. (2023).

However, these approaches share a common evaluation protocol that applies teacher-forcing Williams & Zipser (1989) at inference, feeding the ground-truth previous token at every decoding step. Since the target text is by definition unavailable at inference, the reported scores do not reflect autonomous generation. Jo et al. (2025) quantify this issue, showing that the output collapses into degenerate templates unrelated to the input once teacher-forcing is removed. This indicates that apparent success of previous studies is largely attributable to the teacher-forcing rather than to information extracted from brain signals.

A.2 OPEN-VOCABULARY BRAIN-TO-TEXT DECODING

In response to these concerns, recent studies have pursued decoding that generalizes to text unseen during training without relying on teacher-forcing. Défossez et al. (2023) match short brain signal segments to speech representations through contrastive learning, although the formulation remains retrieval over a candidate set rather than open-vocabulary text generation. NeuSpeech Yang et al. (2024a) recasts the task as speech recognition by feeding raw signals into a pre-trained Whisper model, generating text autoregressively without teacher-forcing. MAD Yang et al. (2024b) aligns brain features with an auxiliary speech modality at the acoustic and semantic levels. BrainECHO Li et al. (2025) decouples the problem into spectrogram autoencoding, brain-audio latent alignment, and fine-tuning.

Despite these advances, absolute performance in the open-vocabulary setting remains close to the floor, indicating that reliable generation of unseen text from brain signals is still highly challenging. This limitation becomes particularly evident under evaluation splits that exclude any sentence overlap between the training and test sets. In such settings, the performance margins over random baselines largely vanish, suggesting that the models struggle to generalize beyond the linguistic patterns observed during training. Consequently, decoded outputs rarely recover the semantic content conveyed by the target sentence.

A.3 CONTEXT-CONDITIONED BRAIN DECODING

Another line of work leverages the generative capacity of language models to compensate for the limited information carried by brain recordings, conditioning the generation process on text. Tang et al. (2023) condition the generation on the decoder's own previously decoded words, maintained as a truncated window for each candidate in the beam. BrainLLM Ye et al. (2025) instead conditions on ground-truth stimulus text drawn from inside the same utterance as the target, where a sentence is split so that its first part serves as the prompt and the remainder as the target. However, these conditioning schemes face limitations in terms of the information that the context supplies. In Tang et al. (2023), the context carries no information beyond what the decoder has already inferred, so it only keeps the output linguistically coherent. In BrainLLM Ye et al. (2025), the prompt is part of the target sentence itself, so much of the target is already given rather than decoded. In practice, however, the beginning of the target sentence is not available before decoding.

In this work, we use the preceding sentence as context and decode the target sentence in full, so that no part of the target is given in advance. Such context is naturally available in real-world scenarios such as continuous reading and ongoing dialogue, and it narrows down the topic of the target while leaving its actual wording open. Combined with local semantic alignment between brain signals and text, this reframes open-vocabulary decoding as a context-guided inference task, where the brain signal determines which sentence was perceived among those the context allows.

B Details

B.1 Training details

For MEG-text semantic alignment, the fixed window size centered at each word interval is set to $W = 128$. The MEG encoder is trained using the AdamW optimizer Loshchilov & Hutter (2019) with a learning rate of 5e-3 and a batch size of 256 for 20 epochs, and is then frozen during adapter training. The temperature parameter τ in contrastive loss is set to 0.07. The adapter is trained with a preceding sentence $M = 1$, a learning rate of 1e-4, and a batch size of 1 for 5 epochs. The LLM remains frozen throughout training. All experiments are conducted on two NVIDIA RTX A6000 GPUs.

B.2 Inference details

At inference, text is generated using beam search with a beam size of 5, a repetition penalty of 5.0, and no-repeat n-gram size of 2, following the inference configuration of BrainECHO Li et al. (2025).

B.3 Prompt Template

The input prompt fed to the frozen LLM follows the template below. Curly braces denote fields filled at runtime: s_{t-1} is the prior sentence and $\hat{s}_t$ is the predicted target sentence.

`[MEG tokens]` *(continuous prefix embeddings)*
`Prior sentence:` $\{s_{t-1}\}$
`Task: Infer the sentence represented by the MEG signals, using the prior sentence as context.`
`Answer:` $\{\hat{s}_t\}$

The full input sequence is $[\texttt{BOS}] \rightarrow [\text{MEG}] \rightarrow [\text{prompt}] \rightarrow [\hat{s}_t] \rightarrow [\texttt{EOS}]$, where the MEG tokens are produced by the adapter and inserted immediately after the `[BOS]` token as continuous prefix embeddings.

B.4 Evaluation Metrics

For evaluation, we adopt standard metrics that capture different aspects of text generation quality. BLEU-N Papineni et al. (2002) ($N = 1, 2, 3$) measures the n-gram precision between the generated text and the reference at different n-gram orders. It therefore reflects how accurately the model reproduces individual words as well as short local word sequences. ROUGE-1 Lin (2004) reports precision (P), recall (R), and f-measure (F) based on unigram overlap between the generated and reference texts. These three measures capture complementary aspects of lexical matching and coverage. Word error rate (WER) quantifies the word-level edit distance between the prediction and the reference by accounting for the insertions, deletions, and substitutions required for alignment. Lower WER indicates fewer word-level errors, whereas higher values for BLEU-N and ROUGE-1 indicate better generation performance.

Table 5: Ablation study for input variants on GWilliams. *Context only* uses the preceding context without MEG signals. *Random MEG + Context* replaces the input MEG with another MEG signal randomly sampled from the dataset. *MEG only* uses MEG signals without the preceding context, and *MEG + Context* is the proposed CoMET.

Variants	BLEU-N (%) ↑			ROUGE-1 (%) ↑			WER(%) ↓
	N=1	N=2	N=3	P	R	F	
Context only (pretrained)	3.03	0.62	0.18	3.43	**24.50**	5.85	718.04
Context only (LoRA fine-tuned)	3.27	0.71	0.20	3.86	24.09	6.31	708.48
Random MEG + Context	3.91	0.70	0.19	5.87	13.86	6.84	311.60
MEG + Random Context	6.92	2.34	0.71	6.33	9.47	7.18	115.15
MEG only	6.08	1.85	0.42	6.59	10.00	7.72	117.73
MEG + Context (CoMET)	**10.33**	**2.65**	**0.91**	**10.41**	12.13	**10.97**	**110.83**

Table 6: Ablation studies on GWilliams. (a) Comparison of alignment strategies in the MEG-text semantic alignment stage. Sentence aligns sentence-level representations between MEG and text, while Word aligns word-level representations, (b) Performance with varying number of preceding sentences M, and (c) Performance with varying LLM backbone sizes. Qwen2.5-Instruct models at three scales are compared.

Ablation	Variants	BLEU-N (%) ↑			ROUGE-1 (%) ↑			WER(%) ↓
		N=1	N=2	N=3	P	R	F	
(a) Alignment	Sentence	9.16	2.17	0.52	11.19	10.44	10.00	116.0
	Word	10.33	2.65	0.91	10.41	12.13	10.97	110.83
(b) Context M	0	6.08	1.85	0.42	6.59	10.00	7.72	117.73
	1	10.33	2.65	0.91	10.41	12.13	10.97	110.83
	2	10.83	2.73	0.94	11.28	13.07	11.55	112.42
	3	10.75	3.04	0.98	10.99	12.81	11.33	115.47
(c) LLM Size	7B	9.60	2.26	0.68	10.57	12.54	10.93	140.18
	14B	10.33	2.65	0.91	10.41	12.13	10.97	110.83
	32B	9.98	2.40	0.86	10.92	12.37	11.10	112.77

C Ablation in GWilliams Dataset

C.1 Ablation on Input Components

The same ablation on input components is conducted on the GWilliams dataset, as reported in Table 5. Both *MEG only* and *Context only* yield substantially lower performance than CoMET, and *Random MEG + Context* drops to a level comparable to *Context only*, confirming that the LLM cannot exploit unmatched MEG signals. The best performance is obtained only when MEG signals are correctly paired with the preceding context. This consistency across two datasets indicates that the contribution of each input component is not specific to the single-subject setting of LibriBrain, but generalizes to the more challenging multi-subject setting of GWilliams.

C.2 Analysis of Alignment Strategy

Table 6 (a) presents a comparison of the alignment strategies on the GWilliams dataset. Consistent with the trend observed on LibriBrain (Table 3 (a) of the main paper), word-level alignment achieves better performance than sentence-level alignment across most evaluation metrics. These results suggest that word-level alignment allows the encoder to learn local semantic information more effectively within each sentence. As a result, the learned representations remain effective during inference even when the brain signal is processed at the sentence level, without requiring explicit segmentation into individual word-level units.

C.3 Varying Number of Preceding Sentence

Table 6 (b) reports the performance obtained with varying numbers of preceding sentences M on the GWilliams dataset. Consistent with the trend observed on LibriBrain (Table 3 (b) of the main

paper), incorporating preceding context ($M \geq 1$) leads to a substantial improvement over the setting without context ($M = 0$). At the same time, performance remains relatively stable as the number of preceding sentences increases. This result further supports our observation that even a single preceding sentence provides sufficient contextual information to anchor the decoding process.

C.4 Performance with varying LLM backbone sizes

Table 6 (c) reports the performance obtained with varying LLM backbone sizes on the GWilliams dataset. Consistent with the trend observed on LibriBrain (Table 3 (c) of the main paper), CoMET substantially outperforms all baselines across different LLM scales. A similar performance pattern is also observed as the backbone size increases, confirming that the results are consistent across the two datasets. These results indicate that CoMET remains effective even with a smaller LLM and is robust to the choice of LLM scale.