Brain-Inspired fMRI-to-Text Decoding via Incremental and Wrap-Up Language Modeling

Wentao Lu[1], **Dong Nie**[2], **Pengcheng Xue**[1], **Zheng Cui**[1], **Piji Li**[1], **Daoqiang Zhang**[1], **Xuyun Wen**[1,*]
[1]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]ChatAlpha AI, California, USA
{luwentao,charles1231,cuizheng,pjli,dqzhang,wenxuyun}@nuaa.edu.cn
dongnie@cs.unc.edu

Abstract

Decoding natural language text from non-invasive brain signals, such as functional magnetic resonance imaging (fMRI), remains a central challenge in brain-computer interface research. While recent advances in large language models (LLMs) have enabled open-vocabulary fMRI-to-text decoding, existing frameworks typically process the entire fMRI sequence in a single step, leading to performance degradation when handling long input sequences due to memory overload and semantic drift. To address this limitation, we propose a brain-inspired sequential fMRI-to-text decoding framework that mimics the human cognitive strategy of segmented and inductive language processing. Specifically, we divide long fMRI time series into consecutive segments aligned with optimal language comprehension length. Each segment is decoded incrementally, followed by a wrap-up mechanism that summarizes the semantic content and incorporates it as prior knowledge into subsequent decoding steps. This sequence-wise approach alleviates memory burden and ensures semantic continuity across segments. In addition, we introduce a text-guided masking strategy integrated with a masked autoencoder (MAE) framework for fMRI representation learning. This method leverages attention distributions over key semantic tokens to selectively mask the corresponding fMRI time points, and employs MAE to guide the model toward focusing on neural activity at semantically salient moments, thereby enhancing the capability of fMRI embeddings to represent textual information. Experimental results on the two datasets demonstrate that our method significantly outperforms state-of-the-art approaches, with performance gains increasing as decoding length grows. The code is available at https://github.com/WENXUYUN/CogReader.

1 Introduction

Language serves as a window into cognitive processes, conveying vast amounts of information through its syntactic and semantic structures [25]. Advances in non-invasive neuroimaging, such as functional magnetic resonance imaging (fMRI), have enabled researchers to measure brain activity patterns associated with language processing. Translating cognitive signals into natural language not only deepens our understanding of the neural basis of the language system, but also facilitates the development of practical brain-computer interfaces (BCIs) by leveraging insights into the decoding process [34, 22].

In recent years, advances in deep learning have led to significant progress in short text generation from brain signals, such as mapping fMRI activity to semantic representations of individual words or short

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

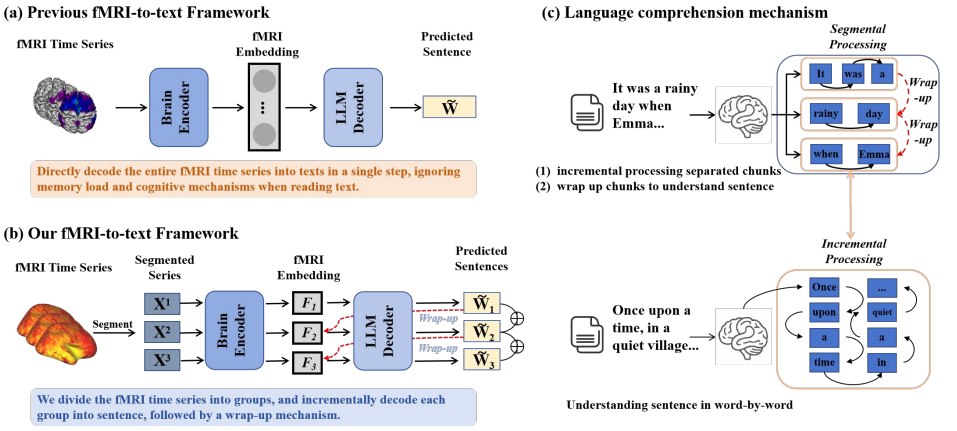

Figure 1: Comparison of fMRI-to-text decoding frameworks. (a) Existing frameworks directly decode the entire fMRI sequence corresponding to the target text in a single step. (b) Our proposed segment-based sequential decoding framework. (c) Cognitive mechanisms of human language comprehension, where incremental processing and segmental wrap-up operating in parallel.

phrases [1, 4]. However, these approaches are limited in scope, often restricted to closed vocabularies or single-word outputs, lacking the capacity to decode full natural language sentences. With the rapid development of large language models (LLM), researchers have begun to explore their application in various brain decoding tasks [7, 19, 21, 28, 8]. Recent studies have shown encouraging progress in open-vocabulary fMRI-to-text generation by incorporating large language models (LLMs)[35, 3]. However, these methods still face significant challenges in decoding long sequences. As illustrated in Figure1(a), most current approaches process the entire fMRI sequence corresponding to a given text in a single step, overlooking the segmented and inductive processing strategy that the human brain adopts to manage memory load during language comprehension [13, 10]. As the length of the input increases, these approaches lead to excessive memory burden and semantic drift [24], ultimately impairing decoding performance. Unlike the traditional machine translation paradigm employed in existing methods, which enables one-to-one mapping between two independent modalities, fMRI-to-text decoding reconstructs textual content from neural activity patterns elicited during human language comprehension. Consequently, models tailored for cross-modal translation are not directly transferable to this task. Guided by this distinction, we hypothesize that a decoding framework better aligned with human language processing mechanisms would more effectively address this challenge. Therefore, it is essential to incorporate insights from human language comprehension mechanisms into the design of fMRI-to-text decoding models.

Human language understanding is neither a passive nor a strictly linear process. Instead, it emerges from a complex interplay between incremental processing and segmental integration [29], as shown in Figure 1(c). Incremental processing enables the brain to construct semantic representations in real time, interpreting linguistic input on a word-by-word basis. While this allows for immediate comprehension, it imposes a heavy load on working memory and becomes less effective for complex or long-form text. In contrast, segmental integration provides a complementary mechanism, wherein the brain periodically aggregates information across semantically coherent segments. This wrap-up process facilitates semantic consolidation and disambiguation at key structural boundaries, thereby reducing cognitive load and enhancing comprehension accuracy. Despite growing evidence supporting the importance of these dual mechanisms, they remain largely underexplored in existing fMRI-to-text decoding frameworks.

Inspired by human cognitive mechanisms for language processing, we propose a novel fMRI-to-text decoding framework that combines incremental processing with a wrap-up-based semantic integration strategy, named as **CogReader**. As shown in Figure 1 (b), we first divide the continuous fMRI time series into multiple sequential segments. For each segment, the model performs incremental decoding, generating the corresponding text word by word in real time. We also design a wrap-up integration module that summarizes the decoding results of the current segment into a semantic representation. This representation is then passed as prior knowledge to guide the decoding of the next segment,

enabling effective cross-segment information flow. Furthermore, to learn fMRI features containing more textual information, we introduce a text-guided masking strategy, integrated into a Masked Autoencoder (MAE)-based framework for fMRI representation learning. Our main contributions are summarized as follows:

1. Motivated by human language comprehension mechanisms, we design a new fMRI-to-text decoding framework that integrates incremental processing and wrap-up semantic integration. Our model enables real-time decoding for each segment and progressively incorporates cross-segment knowledge, offering an effective solution for decoding long-form text from neural activity.
2. We propose a text-guided masking strategy. By leveraging attention distributions over key semantic tokens, our method selectively masks corresponding fMRI time points and incorporates MAE to encourage the model to focus on neural activity at key time points to learn brain representations with more key textual information.
3. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches on standard fMRI-to-text decoding benchmarks. Moreover, the performance advantage becomes increasingly pronounced as sentence length grows, underscoring the feasibility and effectiveness of our cognitively inspired decoding framework.

2 Related Works

2.1 fMRI Representation Learning

Due to the complex spatiotemporal structure of fMRI data and the variability across subjects, learning robust and high-quality fMRI representations remains a significant challenge. In recent years, a variety of deep learning paradigms have been proposed to improve fMRI representation quality. For example, Kim et al. [14] utilized a variational autoencoder (VAE) [16] to model the distribution of fMRI signals while disentangling spatial and temporal components. Asadi et al. [2] proposed a hybrid model that combines spatial attention with temporal Transformers, to better model long-range spatiotemporal dependencies. While model architecture innovation has advanced fMRI representation learning, complementary learning paradigm improvements have emerged as another research focus. In fMRI-to-text decoding tasks, fully supervised learning has become standard for enhancing semantic richness of neural representations [1, 30]. Building on this foundation, contrastive learning frameworks [1, 3] further optimize cross-modal alignment by treating paired fMRI-text data as positive samples. However, the reliance on scarce paired datasets limits these supervised approaches. To mitigate this bottleneck, recent work integrates self-supervised pre-training paradigms that leverage abundant unlabeled data. For example, masked autoencoding (MAE)-inspired methods [11] have demonstrated effectiveness in capturing spatiotemporal features from raw fMRI signals through reconstruction-based learning [35], establishing a synergistic pipeline with task-specific supervision.

2.2 fMRI-to-text Decoding

Decoding natural language from non-invasive brain imaging modalities such as functional magnetic resonance imaging (fMRI) has long posed a core challenge in brain-computer interface (BCI) research. Early efforts predominantly focused on closed-vocabulary decoding, wherein brain signals were mapped to a fixed set of candidate words. For instance, Brain2Word [1] employed a classification-based approach to decode individual words from fMRI activity, while Défossez et al. [4] utilized contrastive learning to decode words and short phrases from auditory-evoked brain signals. With the advent of large-scale pretrained language models (LLMs), recent studies have pivoted toward open-vocabulary decoding, aiming to reconstruct fluent and unconstrained natural language from brain activity. For example, UniCoRN[35] treated fMRI time series as a foreign language and leveraged a BART-style translation architecture to generate continuous text. In addition, Tang et al. [30] employed a hybrid model that combines linear regression with a generative pretrained transformer (GPT) to perform similarity-based decoding. Most recently, BP-GPT [3] introduced a prompt-based decoding paradigm, where embeddings derived from fMRI sequences serve as prompts to condition large language models (e.g., GPT-2) for coherent text generation. To better align the modalities of brain signals and natural language, BP-GPT further integrates contrastive learning to align fMRI-derived and text-derived prompts, significantly improving decoding accuracy.

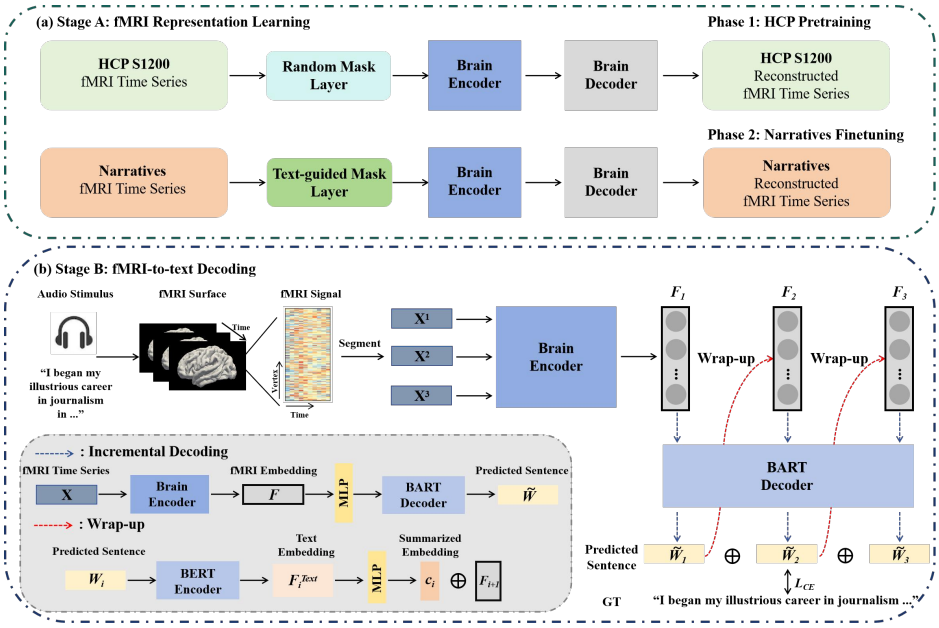

Figure 2: Framework of **CogReader**, comprising two main components: (A) fMRI representation learning and (B) fMRI-to-text decoding.

3 Method

In this section, we introduce the overall structure of the proposed fMRI-to-text decoding method, i.e., **CogReader**. As illustrated in Figure 2, CogReader consists of two main components: (A) fMRI representation learning and (B) fMRI-to-text decoding. In stage A, we employ a text-guided masking strategy within the MAE framework to train the brain encoder in a self-supervised manner. In stage B, a brian-inspired decoding framework is applied to generate natural language descriptions from the learned fMRI representations. Specifically, given an fMRI signal X with T time frames (each time frame with TR seconds), denoted as $X = \{x_1, x_2, ...x_T\}$, our goal is to decode the corresponding natural language sequence presented during scanning, represented as a sequence of words $W = \{w_1, w_2, ...w_n\}$, where each token $w_i \in V$, and V denotes an open vocabulary. The implementation details of each component are described in the following subsections.

3.1 fMRI Representation Learning

Given the limited size of currently available fMRI-text paired datasets, it is challenging to jointly optimize both the fMRI representation learning module and the decoding module. We thus introduce a self-supervised MAE-based pretraining task (i.e., fMRI reconstruction task) prior to the decoding stage to obtain a robust brain encoder for fMRI representation learning. However, existing MAE-based approaches for fMRI representation learning suffer from a critical limitation: they fail to account for the varying semantic importance of different grouped text corresponding to different time frames. To address the above issues, we adopt a two-stage training framework for fMRI representation learning and introduce a text-guided masking strategy within the MAE to enhance the stability of the learned fMRI embeddings and improve their ability to capture the semantic information of the corresponding text.

Two-Stage Training Strategy: To address the data scarcity in fMRI-text paired datasets, we adopt a two-stage training process consisting of pretraining and fine-tuning, as illustrated in Figure 2(a). In the first phase, we pretrain the model on a large-scale public fMRI dataset from the Human Connectome Project (HCP), using a random masking strategy to guide the reconstruction of missing signals (mask ratio = 75%). This encourages the model to learn spatial coherence and temporal dynamics across brain regions, allowing the encoder and decoder to capture generalizable neural patterns and providing well-initialized parameters for downstream task. In the second phase, we

finetuned the HCP pretrained model on the target fMRI-text paired dataset (Narratives). Unlike conventional MAE frameworks, we introduce a text-guided masking strategy in this stage to replace the random masking approach. Through the two-phase training paradigm, we incorporate large-scale fMRI data, thereby enhancing the general representation capability of the brain encoder.

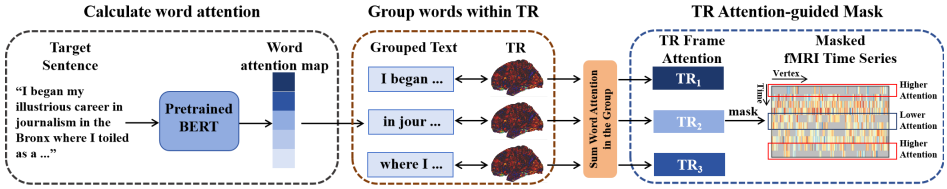

Figure 3: Illustration of text-guided masking strategy.

Text-Guided Masking: To enhance the semantic relevance of learned fMRI representations, we propose a text-guided masking strategy that directs the model's attention toward fMRI signals associated with highly informative semantic content, thereby improving the semantic quality of the learned representations. Given a paired input consisting of an fMRI time series $X = \{x_1, x_2, ..., x_T\}$, its corresponding stimulus text $W = \{w_1, w_2, ..., w_n\}$, we first compute the semantic importance of each word in the text using a pre-trained BERT model. Specifically, for each word w_i, we calculate its attention-based relevance score:

$$A_{w_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{exp(\text{score}(w_j, w_i))}{\sum_{k=1}^{n} exp(\text{score}(w_j, w_k))} \tag{1}$$

where $\text{score}(w_i, w_j) = Q_i \cdot K_j^T$ represents the attention score between word w_i and word w_j, and Q_i and K_j are the query and key vectors from the BERT self-attention layer. Since BERT uses multi-head attention, we average the attention scores across all heads and apply normalization to obtain the final importance score $\overline{A}_{w_i}$ for each word w_i:

$$\overline{A}_{w_i} = \frac{\frac{1}{H} \sum_{h=1}^{H} A_{w_i}^h}{\sum_{i=1}^{n} \frac{1}{H} \sum_{h=1}^{H} A_{w_i}^h} \tag{2}$$

where $A_{w_i}^h$ donates the attention score of word w_i from head h, and H is the total number of attention heads. To align word-level attention scores with fMRI frames (TRs), we associate each fMRI time point x_t with a group of words according to fMRI-text pairs. The attention score for each TR is then computed as the sum of its associated word importance: $A_{x_t} = \sum_{w_i \in x_t} \overline{A}_{w_i}$. Using the attention vector $A_X = \{A_{x_1}, A_{x_2}, \ldots, A_{x_T}\}$, we apply a differentiable masking strategy across fMRI frames to encourage the model to focus more on neural activity during key moments. Specifically, for the top 40% of fMRI frames with the highest attention scores, we randomly mask 75% of the vertex signals. For the remaining 60% of fMRI frames with lower importance, we randomly mask only 25% of the vertex signals. This strategy encourages the model to focus more heavily on reconstructing fMRI signals from semantically rich time points, improving the quality of the learned representations and ultimately boosting downstream decoding performance.

Autoencoder: We employ a transformer-based autoencoder to capture the overall fMRI representation [5]. The brain encoder consists of a spatial module and a temporal module. The spatial module is designed to capture the spatial structural relationships among cortical vertices, while the temporal module models the dynamic changes across time. Both modules are composed of multiple stacked Transformer blocks, each receiving positional embeddings corresponding to either spatial or temporal locations. Specifically, given the input fMRI time series $X = \{x_1, x_2, ..., x_T\}$, the spatial module first projects it through a linear layer, followed by a stack of 8 Transformer blocks, producing a feature matrix $F_t^s \in \mathbb{R}^{N \times d_{\text{spat}}}$, where N is the number of vertices and d_{spat} is the dimension of the spatial module. A global average pooling layer then aggregates these vertex-level features into a single vector $s_t \in \mathbb{R}^{d_{\text{spat}}}$ representing the spatial embedding at time t. The temporal module then takes the sequence of spatial embeddings $\{s_t\}_{t=1}^T$ and encodes temporal dependencies via 8 Transformer layers, producing a latent sequence $\{h_t\}_{t=1}^T$, $h_t \in \mathbb{R}^{d_{\text{temp}}}$. The brain decoder takes $\{h_t\}_{t=1}^T$ as input and adopts 4 transformer blocks and a linear layer to reconstruct the fMRI time series $\tilde{X} = \{\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_T}\}$.

Reconstruction Loss: Considering that fMRI signals are time series with strong temporal dependencies, our reconstruction loss is designed to account for both the masked and unmasked portions of the sequence. Specifically, given the original fMRI signal X and the reconstructed fMRI signal $\tilde{X}$, the reconstruction loss is formatted as $L_{\text{recon}} = \text{MSE}(X, \tilde{X})$.

3.2 fMRI-to-Text Decoding

Current fMRI-to-text decoding frameworks typically generate complete text segments directly from the entire fMRI sequence, neglecting the segmented and inductive processing strategy adopted by the human brain to manage memory load during language comprehension. This often results in performance degradation when decoding long text sequences. To address this issue, we propose a brain-inspired sequence-by-sequence fMRI-to-text decoding framework that integrates incremental decoding with a semantic wrap-up mechanism, as illustrated in Figure 2(b). Specifically, we divide the long fMRI sequence into consecutive segments, each aligned with the optimal length for human language processing (discussed in section 4.4). Within each segment, we perform incremental decoding to generate partial textual outputs. After decoding each segment, a wrap-up mechanism summarizes its semantic content, which is then incorporated as prior knowledge into the subsequent segment's decoding process. This sequence-wise decoding strategy not only mitigates the memory burden and performance drop associated with long-text decoding but also ensures semantic continuity across successive segments.

Incremental Decoding within fMRI Segments: Given an input fMRI time series $X = \{x_1, x_2, \ldots, x_T\}$, we first divide it into consecutive segments of equal length, where each segment contains N_s fMRI frames. The segmented sequence is denoted as $\overline{X} = \{X^1, X^2, \ldots, X^K\}$, where $K = \frac{T}{N_s}$. The optimal N_s is discussed in Section 4.2. For each fMRI segment, we first employ a brain encoder to extract its corresponding representation, which is then fed into a well-established brain-to-text decoder to directly generate the associated text. This process is referred to as incremental decoding, resembling the human brain's real-time comprehension of incoming language input. In this study, we adopt the BART model as the fMRI-to-text decoder, due to its well performance on language understanding tasks and its suitability for sequence-to-sequence reconstruction [32, 18]. This aligns well with our task structure, where the fMRI representations are translated into corresponding textual sequences [35]. Specifically, given the fMRI segment X^i, we first use brain encoder to extract its fMRI representation $F_i \in \mathbb{R}^{N_s \times d_{fMRI}}$, where d_{fMRI} is the number of feature dimensions. Next, a linear projection layer is used to map F_i into the embedding space of the BART decoder, resulting in $F_i^{\text{BART}} \in \mathbb{R}^{N_s \times d_{\text{BART}}}$, where d_{BART} represents the dimensionality of the decoder's embedding space. Finally, the projected embedding F_i^{BART} is fed into the BART model to generate the predicted text $\tilde{W}_i$ corresponding to the current fMRI segment.

Semantic Wrap-Up across fMRI Segments: To address the potential issue of semantic discontinuity across segments in the final decoded text, we incorporate a wrap-up mechanism into the sequential decoding framework, inspired by the human brain's cognitive strategy for integrating information during language comprehension. Specifically, after obtaining the decoding text $\tilde{W}_i$ from the fMRI segment X^i via incremental decoding, we employ a pretrained BERT [6] model to extract its contextualized embedding representation, denoted as F_i^{text}. This embedding is then passed through an MLP layer P_θ to derive a semantic summary vector:

$$c_i = P_\theta(F_i^{text}) \tag{3}$$

This process is designed to simulate the inductive summarization mechanism observed in human reading. During the decoding of the next fMRI segment X^{i+1}, we incorporate the summarized embedding c_i into its corresponding representation vector F_{i+1}, guiding the decoding process for X^{i+1}. By incorporating the summarized semantic knowledge from the previous text segment into the decoding of the subsequent segment, the model enhances semantic continuity across successive segments. The dimensionality of the MLP P_θ is discussed in detail in Section 4.2.

Decoding Loss: Through incremental decoding and semantic wrap-up, we decode the complete text $\tilde{W} = \left\{\tilde{W}_1, \tilde{W}_2, ...\tilde{W}_K\right\}$ from fMRI segments $\overline{X} = \{X^1, X^2, \ldots, X^K\}$ in a sequence-by-sequence manner. The objective function for the decoding stage is defined as the cross-entropy loss between

the generated text $\tilde{W}$ and the corresponding ground-truth text W, which is formatted as

$$L_{decoding} = \sum_{i=1}^{K} CE\left(W_i, \tilde{W}_i\right) \tag{4}$$

Here, $CE(\overline{w}_i, w_i) = -\sum_{i=1}^{{N_t}^i} w_i \log(\tilde{w}_i)$, where ${N_t}^i$ is the number of words in text squence W_i.

4 Experiments

4.1 Experimental Setup

Datasets This study employs three neuroimaging datasets: HCP S1200 [31], Narratives [23] and Huth dataset [17]. The Human Connectome Project's HCP S1200 dataset provides extensive fMRI data from 1,206 healthy young adults across seven cognitive domains. We primarily use this dataset to pretrain the brain encoder for fMRI representation learning, addressing the scarcity of fMRI-text paired data and enhancing the encoder's generalization. The Narratives dataset, a paired fMRI-text benchmark, contains fMRI recordings from 345 participants during naturalistic auditory comprehension of 27 real-world narrative stories, totaling approximately 6.4 days of functional imaging data. The Huth dataset comprises fMRI data from 8 subjects recorded while they passively listened to naturally spoken English stories, and the stories were sourced from The Month and New York Times Modern Love podcasts. Narratives dataset and Huth Dataset are used for the decoding task. All fMRI data from these datasets were preprocessed [9] and projected onto the cortical surface using the standardized preprocessing pipelines provided by each source.

Implementaion Details Our model is built using the PyTorch framework [27] and the Huggingface Transformers package [33]. All models utilize the Adam optimizer [15], with a warmup strategy. All experiments are conducted on CUDA 12.2 and the computer with NVIDIA GeForce RTX 3090 GPU. Additional implementation details can be found in the Appendix.

Evaluation Metrics To comprehensively evaluate decoding performance, we adopt three text generation metrics: BLEU-N [26], ROUGE-1 [20], and BERTScore [37]. Among them, BLEU-N and ROUGE-1 assess word-level overlaps between the generated and reference texts, while BERTScore evaluates semantic similarity based on contextual embeddings. Specifically, we report BLEU-1 to BLEU-4 scores for BLEU-N; ROUGE-F, ROUGE-P, and ROUGE-R scores for ROUGE-1; and BERTScore-F, BERTScore-P, and BERTScore-R for BERTScore.

4.2 Parameter Settings

This section discusses the optimal configuration of two key parameters for the fMRI-to-image decoding task. The evaluation is conducted on the Narrative dataset using BLEU-1, ROUGE-R, and BERTScore-R as performance metrics.

Segment Length $\mathbf{N_s}$: To determine the optimal segment length N_S, we vary it from 10 to 70 in steps of 10. For each value of N_s, we train and test the CogReader framework accordingly. As shown in Figure 4, all metrics exhibit a trend of first increasing and then decreasing with longer segment lengths, reaching peak performance at $N_S = 20$. Therefore, we set $N_S = 20$ for all subsequent experiments.

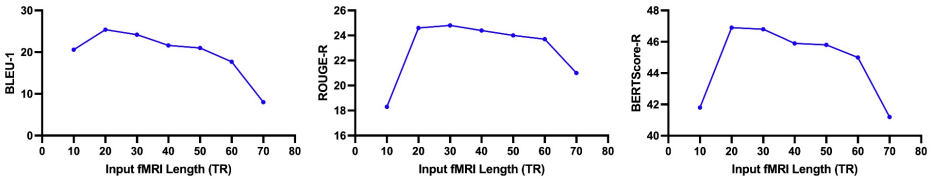

Figure 4: Parameter settings for segment length N_s in incremental decoding.

Dimensionality of the MLP: For the MLP dimensionality, we evaluate five configurations: 0, 32, 64, 128, and 256. The corresponding performance of CogReader under each setting is shown in Figure 5. Taking into account both word-level and semantic-level evaluation metrics, we ultimately set the

MLP dimension to 128, as this configuration demonstrates consistently good and stable performance across all metrics. In comparison, other configurations perform well on at most a single metric.

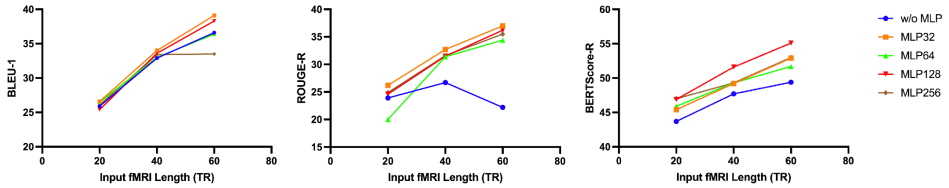

Figure 5: Parameter settings for dimensionality of the MLP in semantic wrap-up.

4.3 Comparison with State-of-the-art methods

We compared our proposed fMRI-to-text method with four state-of-the-art (SOTA) approaches: UniCoRN [35], EEG-Text [32], BP-GPT [3] and PREDFT [36]. Similar to our method, UniCoRN adopts a two-stage decoding framework consisting of fMRI representation learning followed by fMRI-to-text decoding. EEG-Text was originally designed for EEG signal generation tasks. We retrained its encoder using fMRI data to adapt it for fMRI decoding. BP-GPT is a prompt-based decoding method that guides text generation by aligning fMRI representations with text embeddings via contrastive learning. PREDFT is an end-to-end fMRI-to-text decoding model that jointly models neural decoding and brain predictive coding. In our experiments, we conducted comparisons under three different text decoding lengths, corresponding to fMRI time series of 20 TRs, 40 TRs, and 60 TRs. The comparative results in Narratives are given in Tables 1 and 2. The results in Huth dataset can be found in the Appendix A.3. Further qualitative analyses of decoded text cases are included in Appendix A.4.

Table 1: Comparison of our method and SOTA methods under different text decoding lengths on the Narrative dataset.

Length	Method	BLEU-N(%)				ROUGE-1(%)			BERTScore(%)		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-F	ROUGE-P	ROUGE-R	BERTScore-F	BERTScore-P	BERTScore-R
20TR	UniCoRN	22.9	2.5	0.3	0	20.3	19.6	21	43.9	44.2	42.8
	EEG-Text	24.6	9.3	4.4	1.9	21.9	21.1	23.4	44.6	43.9	45.4
	BP-GPT	21.6	3.8	2.5	1.7	21.6	20.9	23.4	44.1	42.1	46.3
	PREDFT	24.3	4.2	0.7	0.1	20.1	22.3	18.3	45.9	45.5	46.7
	CogReader(ours)	**25.4**	**10.5**	**4.7**	**2.6**	**23.4**	**22.6**	**24.6**	**46.3**	**45.7**	**46.9**
40TR	UniCoRN	19.1	2.3	0.5	0.1	17.8	18.2	17.6	43.8	44.8	42
	EEG-Text	20.1	7.3	3	1.3	24.4	25.1	24.7	45.4	45.8	45.5
	BP-GPT	19.9	3.6	2.3	1.5	21.1	19.3	22.9	42.6	39.4	46.1
	PREDFT	25.9	4.8	1.4	0.4	21.1	24.8	18.6	46.3	46.2	46.8
	CogReader(ours)	**31.2**	**15.3**	**10.3**	**8.2**	**29.6**	**28.7**	**30.4**	**50**	**49.3**	**51.1**
60TR	UniCoRN	18	1.7	0.2	0.4	16.5	15.9	17	43.2	43.7	42.7
	EEG-Text	22.1	8.2	3.4	1.6	28.1	29.4	28.1	47.7	47.8	47.7
	BP-GPT	19.3	3.4	1.3	0.6	19.4	19.6	19.3	41.6	38.2	45.3
	PREDFT	26.4	6.1	1.9	0.6	28.1	25.5	20.5	48.1	47.7	48.5
	CogReader(ours)	**36.2**	**20.4**	**14.7**	**12.1**	**36.2**	**35.6**	**37.2**	**53.5**	**52.6**	**54.5**

Quantitative Comparison As shown in Table 1, our method consistently outperforms all SOTA methods across all decoding lengths and evaluation metrics, demonstrating the overall effectiveness of our proposed brain-inspired framework. From the perspective of word-level metrics (BLEU-N and ROUGE-1), the performance of SOTA methods degrades as the length of the decoded text increases, whereas our method exhibits an upward trend. In terms of semantic-level evaluation (BERTScore), the SOTA methods show relatively stable performance, while our method continues to improve with longer decoding sequences. Similar results are also observed on the Huth dataset (Appendix A.3). Under the 60TR time window, our method achieves significantly better text decoding performance than other approaches.

Decoding Text Comparison As shown in Table 2, our method consistently outperforms others in both semantic information extraction and narrative consistency. For instance, our model accurately identifies and generates key phrases such as “Pie Man”, “Ego sum non an bestia”, “scene”, “said”, and “fled”, while producing more coherent and contextually appropriate sentences. In contrast, other methods exhibit greater deviations from the target content. These results suggest that the advantage of our approach becomes increasingly evident as the decoding length grows, highlighting the effectiveness and feasibility of the brain-inspired sequential decoding strategy we adopt.

Table 2: Cases analysis for fMRI-text decoding on Narrative dataset. Exact matches between the target and predicted sentences are indicated in **bold**, while semantic similarity is shown in *italic* font.

Results
Target: *was* **I** gave **the** figure a name **I** called **him Pie Man** capital P capital M **And I** described **him as** a cape wearing masked avenger Though **in** fact **he**'d **been** capeless **And said that as he fled the scene he** clicked **his** heels **in** rakish glee **And I** gave **him** a catchphrase **in** Latin **I said that he cried out Ego sum non an bestia which** means **I am** *not* an animal **Which makes** no sense
UniCoRN: **and** me gray to **the** little **Man and** no taller than my knee, with a coat too big for **his** frame lingered by **the** gate. It's **been** nigh of years **he** to **the** then **said** a case here. **The** leather sat weathered on **the** step, its lock rusted shut, but **I** knew what **was** inside: **the** girl of **the** story, **the** one who scribbled tales of dragons **in the** margins of her notebooks, had tucked her most precious drawings there before she
EEG-Text: **and** this is where it gets really interesting **I***'m* trying to think of a title for **the** book **and I** finally come up with **an** idea **that** sounds like something **out** of one of those old pulp stories by John Wilmot who lived **in** Waukegan Illinois but **he was** actually born **in** Westchester County New York **which** is about twenty five miles outside Chicago So **I am** going to describe my novel **as The Man** Who Walks into Times Square **and** stops before **he** can get too lost
BP-GPT: **I***'m* a teacher **and I** teach people how to make things better by using my own talents **And** so one day **I was** walking down **the** street **and** this old lady came up to me **and** she **said** You know you sound like **an** intelligent black woman So **I** told her **that**'s exactly what **I***'m* here to do
PREDFT: **the** a **said the** girl a **man I said the Man and** use of best **and** she me **and I man** fl of Litgo be **and the** It's flirting **and** she guy of **the** eyes thelee raised **non** crate **and** she littleiving it then to of crate **non** a best owan is best me best to
Ours: **I** think **that** you realize what happened next **Pie Man** emerged from **the** late night library drop **made his** delivery **and fled** away **crying Ego sum non an bestia** Or that*'s* what it **said in** my story **in the** newspaper next day **which** ran with photos of **him** leaving **the scene** cape flowing behind **him** doing this **And I'm** just like praying my life does*n't* flash before my eyes **and** ruins

4.4 Ablation Study

In this subsection, we conduct an ablation study to evaluate the effectiveness of three key components: the HCP pretraining phase, the text-guided masking strategy, and the sequential fMRI-to-text decoding framework. For testing the masking strategy, we replace the proposed text-guided masking with a conventional random masking scheme. The experiments are conducted on decoding tasks with fMRI time series of 60 TRs on the Narrative dataset. The results are reported in Table 3. The results show a consistent improvement in performance as each module is incrementally added, validating the effectiveness of each individual component. Notably, the brain-inspired sequential decoding framework yields the most significant performance gain, further demonstrating the feasibility and impact of our proposed decoding approach.

Table 3: Ablation Study of our method

Method			BLEU-N(%)				ROUGE-1(%)			BERTScore(%)		
Sequential Decoding	Pretraining	Text-guided Masking	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-F	ROUGE-P	ROUGE-R	BERTScore-F	BERTScore-P	BERTScore-R
✗	✗	✗	17.7	6.5	2.4	1.1	29.2	32.9	24.7	46.7	47.6	45
✓	✗	✗	32.5	16.5	11.1	8.9	28.2	25.8	30.6	51.1	50.3	51.8
✓	✓	✗	34.0	18.1	12.7	10.2	34.1	33.7	35.7	52.3	51.0	53.7
✗	✓	✓	21.6	7.9	3.2	1.5	26.6	29.4	25	47.4	47.7	47.2
✓	✓	✓	**36.2**	**20.4**	**14.7**	**12.1**	**36.2**	**35.6**	**37.2**	**53.5**	**52.6**	**54.5**

4.5 Evaluation on fMRI Representation Learning

In this section, we evaluate the effectiveness of fMRI representation learning on the Narrative dataset. To this end, we design two experiments, including a comparison with other representation learning methods and an analysis against noise data.

Comparison with other fMRI Representation Learning Method Since the overall framework of UniCoRN is similar to ours in the current fMRI-to-text decoding paradigm, we take UniCoRN as a baseline and replace its fMRI representation learning module with our proposed method, while keeping the fMRI-to-text decoding strategy unchanged. This design allows us to evaluate the effectiveness of our representation learning approach. We conduct comparison experiments on fMRI time series ranging from 10 TRs to 50 TRs. The results, shown in Table 4, demonstrate that

under the same decoding strategy, replacing the representation learning component with our method consistently improves decoding performance across all sequence lengths, validating the effectiveness of the proposed representation learning framework.

Table 4: Comparison results of our fMRI representation learning method with other methods

Length	Method	BLEU-N(%)				ROUGE-1(%)			BERTScore(%)		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-F	ROUGE-P	ROUGE-R	BERTScore-F	BERTScore-P	BERTScore-R
10TR	UniCoRN	18.1	2.9	0.4	0	10.5	10.2	16.6	40.2	40.1	40.4
	Ours	20.6	7	2.8	1.3	17.1	16.2	18.3	41.1	40.5	41.8
20TR	UniCoRN	22.9	2.5	0.3	0	20.3	19.6	21	43.9	44.2	42.8
	Ours	25.4	10.5	4.7	2.6	23.4	22.6	24.6	46.3	45.7	46.9
30TR	UniCoRN	20.3	2.8	0.5	0.1	18.3	18.3	18.4	41.4	41.5	41.4
	Ours	24.2	9.1	3.9	1.8	25.1	26.2	24.8	47	47.1	46.8
40TR	UniCoRN	19.1	2.3	0.5	0.1	17.8	18.2	17.6	43.8	44.8	42
	Ours	21.6	7.9	3.2	1.5	25.2	27	24.4	46.1	46.2	45.9
50TR	UniCoRN	18.9	1.9	1.8	1.1	17.3	16.8	17.4	44.8	43.9	45.7
	Ours	21	7.7	3.2	1.5	26.1	29.4	24	46.5	47.7	45.8

Comparison with Noise Data Previous work [12] has shown that existing open-vocabulary brain decoding methods often yield similar performance on both real and noise data, suggesting that these approaches fail to effectively capture the semantic information encoded in brain signals and instead rely heavily on the memory capacity of the large language model (LLM) decoder. To address this concern, we evaluate the effectiveness of our proposed CogReader model using both real fMRI inputs and noise data as input. The experiment is conducted on fMRI time series with a length of 60 TRs. The experimental results are presented in Table 5. The results show that decoding performance is significantly higher when using real fMRI data compared to noise input, providing strong evidence that our brain-inspired method is capable of extracting meaningful semantic information from fMRI time series, rather than depending solely on the memorization ability of the LLM.

Table 5: Comparison results between real fMRI data and noise data

Data		BLEU-N(%)				ROUGE-1(%)			BERTScore(%)		
Train	Test	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-F	ROUGE-P	ROUGE-R	BERTScore-F	BERTScore-P	BERTScore-R
Noise	Noise	27.5	9.4	4.6	1.8	25.6	26.1	25.5	48.2	48.0	48.4
Noise	fMRI	25.3	7.2	2.4	1.4	23.9	23.5	24.8	47.7	47.2	48.3
fMRI	Noise	26.8	7.7	2.7	1.2	23.9	23.1	24.9	47.9	47.4	48.5
fMRI	fMRI	**36.2**	**20.4**	**14.7**	**12.1**	**36.2**	**35.6**	**37.2**	**53.5**	**52.6**	**54.5**

5 Discussion

Limitations and Future Work In the current method, the segment length is determined based on optimal decoding performance in a fixed setting and cannot dynamically adjust to the complexity of different stimulus texts. This static segmentation strategy may constrain further improvements in decoding accuracy. Future work could explore content-adaptive segmentation approaches that dynamically predict segment boundaries based on narrative complexity, enabling more flexible adaptation to diverse textual inputs. In light of the relatively low temporal resolution of fMRI, where each TR corresponds to multiple words, making it harder to generate coherent and complete sentences. Future work could explore the strengths of integrating fMRI with EEG, as fMRI offers semantic representation capabilities, while EEG provides high temporal resolution, which may help improve decoding accuracy. Moreover, due to the limited availability of paired fMRI-to-text datasets, our model was evaluated on a single public dataset. We plan to validate the robustness and generalizability of our approach on multiple datasets in future studies.

Conclusion This work proposes a brain-inspired sequential fMRI-to-text decoding framework that mimics the human cognitive strategy of segmented and incremental language processing. This method divides long fMRI sequences into optimal-length segments, each of which is decoded incrementally. A wrap-up mechanism is employed between segments to integrate and propagate semantic information, thereby alleviating memory burden and preserving semantic coherence across the entire sequence.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62476129) and the STI 2030-Major Projects (2022ZD0209000).

References

[1] Nicolas Affolter, Beni Egressy, Damian Pascual, and Roger Wattenhofer. Brain2word: decoding brain activity for language generation. *arXiv preprint arXiv:2009.04765*, 2020.

[2] Nima Asadi, Ingrid R Olson, and Zoran Obradovic. A transformer model for learning spatiotemporal contextual representation in fmri data. *Network Neuroscience*, 7(1):22–47, 2023.

[3] Xiaoyu Chen, Changde Du, Che Liu, Yizhe Wang, and Huiguang He. Open-vocabulary auditory neural decoding using fmri-prompted llm. *arXiv preprint arXiv:2405.07840*, 2024.

[4] 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.

[5] Xin Deng, Jiahao Zhang, Rui Liu, and Ke Liu. Classifying asd based on time-series fmri using spatial–temporal transformer. *Computers in biology and medicine*, 151:106320, 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[8] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*, 2023.

[9] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[10] Edward Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[12] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024.

[13] Marcel Adam Just and Patricia A Carpenter. A capacity theory of comprehension: Individual differences in working memory. *Psychological review*, 99(1):122, 1992.

[14] Jung-Hoon Kim, Yizhen Zhang, Kuan Han, Zheyu Wen, Minkyu Choi, and Zhongming Liu. Representation learning of resting state fmri with variational autoencoder. *NeuroImage*, 241:118423, 2021.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023.

[18] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[19] Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6–14, 2022.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[21] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*, 35:29624–29636, 2022.

[22] Shiv Kumar Mudgal, Suresh K Sharma, Jitender Chaturvedi, and Anil Sharma. Brain computer interface advancement in neurosciences: Applications and issues. *Interdisciplinary Neurosurgery*, 20:100694, 2020.

[23] Samuel A Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J Honey, Yaara Yeshurun, Mor Regev, et al. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250, 2021.

[24] Alexander J Oh, Roger Levy, and Richard Futrell. Modeling memory effects in neural language models. In *NAACL*, 2022.

[25] Mark Pagel. Q&a: What is human language, when did it evolve and why should we care? *BMC biology*, 15:1–6, 2017.

[26] 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*, pages 311–318, 2002.

[27] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[28] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

[29] Elizabeth AL Stine-Morrow and Giavanna S McCall. Reading comprehension is both incremental and segmental—and the balance may shift with aging. In *Psychology of Learning and Motivation*, volume 77, pages 263–290. Elsevier, 2022.

[30] 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.

[31] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80: 62–79, 2013.

[32] 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, pages 5350–5358, 2022.

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[34] Jonathan R Wolpaw. Brain-computer interfaces (bcis) for communication and control. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 1–2, 2007.

[35] Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. *arXiv preprint arXiv:2307.05355*, 2023.

[36] Congchi Yin, Ziyi Ye, and Piji Li. Language reconstruction with brain predictive coding from fmri data. *arXiv preprint arXiv:2405.11597*, 2024.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Method.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is available at https://anonymous.4open.science/r/CogReader-A42C/.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: See Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhered to the relevant guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Implementation Details

BLEU-N Metric Implementation We use the BLEU-N scores as the main evaluation metrics for word-level similarity. This evaluation method is summarized in the algorithm below:

Algorithm 1 BLEU-N Score calculation process

Require: Predicted sentence $\tilde{W}$, Reference sentence(s) W, n-gram order n
Ensure: BLEU-N score list P
Initialize: $P \leftarrow$ list of n-gram precisions
for $i = 1$ to n **do**
    $num_{pred} \leftarrow$ Count all i-grams in $\tilde{W}$
    $num_{ref} \leftarrow$ Count all i-grams in W
    $num_{overlap} \leftarrow \sum_{w \in \tilde{W} \cap W} \min(num_{ref}, num_{pred})$,
    $p_i \leftarrow \frac{num_{overlap}}{num_{pred}}$
    Add p_i to P
end for

ROUGE-1 Metric Implementation We compute the ROUGE-1 precision, recall, and F1 score to evaluate unigram overlap between the predicted and reference sentences. The computation process is as follows:

Algorithm 2 ROUGE-1 Score (Precision, Recall, F1) calculation process

Require: Predicted sentence $\tilde{W}$, Reference sentence W
Ensure: ROUGE-1 Recall R, Precision P, and F1 Score F
Extract all unigrams w from $\tilde{W}$ and W
Count overlapping unigrams between $\tilde{W}$ and W
$num_{overlap} \leftarrow \sum_{w \in \tilde{W} \cap W} \min(\text{count}_{\tilde{W}}, \text{count}_{W})$
$num_{ref} \leftarrow$ total number of unigrams in W
$num_{pred} \leftarrow$ total number of unigrams in $\tilde{W}$
$R \leftarrow \frac{num_{overlap}}{num_{ref}}$
$P \leftarrow \frac{num_{overlap}}{num_{pred}}$
if $R + P = 0$ **then**
    $F \leftarrow 0$
else
    $F \leftarrow \frac{2 \cdot R \cdot P}{R + P}$
end if

BERTScore Metric Implementation We choose BERTScore to evaluate the similarity between the predicted sentences and the reference sentences by computing the alignment-based similarity of contextualized token embeddings. The computation steps are as follows:

A.2 Experimental Settings

In the fMRI Repersentation Learning Stage, during HCP pretraining phase, we split the HCP dataset into training and testing sets in a 4:1 ratio. While for the Narratives dataset, to avoid text leakage, we adopt a stimulus split approach to ensure that the train, validation and test sets use different story content, with a ratio of 60%, 20% and 20%, during Narratives pretraining stage and fMRI-to-text decoding stage. We finalized the model parameters as detailed Table 6.

During the phase 2 in fMRI Representation Stage, we adopt the proposed Text-guided masking strategy to finetuned the HCP pretrained model on the target fMRI-text paired dataset. The parameters are shown in Table 7.

Algorithm 3 BERTScore (Precision, Recall, F1) Calculation

Require: Pretrained language model $M(\cdot)$, Predicted sentence $\tilde{W}$, Reference sentence W
Ensure: BERTScore Precision P, Recall R, F1 Score F
Tokenize $\tilde{W}$ and W into subword tokens
Compute contextual embeddings: $E_{\tilde{W}} \leftarrow M(\tilde{W})$, $E_W \leftarrow M(W)$
Compute cosine similarity matrix $S[i,j] = \cos(E_{\tilde{W}}[i], E_W[j])$
$P \leftarrow \frac{1}{|\tilde{W}|} \sum_{i=1}^{|\tilde{W}|} \max_{1 \le j \le |W|} S[i,j]$
$R \leftarrow \frac{1}{|W|} \sum_{j=1}^{|W|} \max_{1 \le i \le |\tilde{W}|} S[i,j]$
if $P + R = 0$ **then**
    $F \leftarrow 0$
else
    $F \leftarrow \frac{2PR}{P+R}$
end if

Table 6: **Parameter setting in HCP Pretraining.**

parameter	value	parameter	value	parameter	value
mask ratio	0.75	epochs	20	Temporal encoder embed dim	64
batch size	32	warm-up epochs	5	Temporal encoder depth	8
optimizer	Adam	initial LR	1e-4	Temporal encoder heads	2
LR scheduler	StepLR	Spatial encoder embed dim	128	decoder embed dim	64
step size	5	Spatial encoder depth	8	decoder depth	4
gamma	0.5	Spatial encoder heads	2	decoder heads	2

Table 7: **Parameter setting in Narratives Finetuning.**

parameter	value	parameter	value	parameter	value
high attention ratio	0.4	batch size	16	LR scheduler	StepLR
mask ratio for high attention	0.75	epochs	40	step size	5
low attetnion ratio	0.6	optimizer	Adam	gamma	0.5
mask ratio for low attention	0.25	initial LR	1e-4	warm-up epochs	5

For the fMRI-to-text decoding, we use the pretrained spatial-temporal encoder from Stage A and BART as the decoder for text generation from fMRI embedding, with a pretrained BERT as the encoder for summarized embedding. The detailed parameters in this stage are shown in Table 8.

Table 8: **Parameter setting in fMRI-to-text decoding.**

parameter	value	parameter	value	parameter	value
batch size	16	LR scheduler	StepLR	fMRI embed dim	256
epochs	20	step size	5	BART embed dim	1024
optimizer	Adam	gamma	0.5	BERT embed dim	768
initial LR	1e-5	warm-up epochs	5		

The equipment used in the experiment is configured as follows: AMAX Tower Workstation TS40-X3, equipped with dual Intel Xeon 4316 CPUs (2.3 GHz, 20 cores), 256 GB of DDR4 memory (32 GB modules at 3200 MHz), a 480 GB SSD for the system disk, a 3.84 TB SSD for hot data, and a 16 TB 7200 RPM SATA enterprise HDD for data storage. The system is powered by dual power supplies rated at 2000W and 1650W.

A.3 Other Results

We evaluate our method on Huth dataset obtained during apassive natural language listening task [17]. The comparative results with four SOTA methods in 60 TR are given in Table 9 below.

Table 9: Comparison between our method and SOTA methods

Length	Method	BLEU-N(%)				ROUGE-1(%)			BERTScore(%)		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-F	ROUGE-P	ROUGE-R	BERTScore-F	BERTScore-P	BERTScore-R
60TR	UniCoRN	21.5	5.4	2.3	0.6	17.1	14.1	19.7	40.1	37.7	42.7
	EEG-Text	23.6	7.5	2.2	1.0	27.6	28.1	26.1	47.2	47.3	47.2
	BP-GPT	20.7	8.1	2.6	1.2	27.1	30.4	23.5	50.7	51.3	49.2
	PREDFT	22.6	6.7	2.5	1.1	24.6	23.5	26.6	45.5	42.8	48.9
	CogReader(ours)	**35.2**	**18.0**	**10.2**	**7.1**	**32.2**	**34.0**	**30.9**	**53.9**	**54.5**	**53.3**

A.4 Text cases

Table 10 and 11 present the comparision results of representative examples in Narratives, comparing the predicted sentence and the reference sentence in different fMRI time series lengths, including 20 TRs, 40 TRs, and 60 TRs between our propoesed method CogReader and SOTA methods.

Table 10: Cases analysis for fMRI2text. Exact matches between the target and predicted sentences are indicated in **bold**, while semantic similarity is shown in *italic* font.

fMRI Length	Results
20TR	**Target**: ... on campus People started dressing ... and quoting him in class The Ram ran ... stories about Pie Man all ... me And toward ... and I saw I came and I saw in the corner Angela from my Brit Lit class drinking with some friends And now Angela and I had been...
	UniCoRN: ... **and** the McG first **and** *wearing* best **and** the **Man** ... the **story Man** ... **drinking** to the thing ... **and** I know ...
	EEG-Text: ... the boy climbs ... **with two** long arms ... **and two** ... pounces on **him and** tries **to** ...
	BP-GPT: ... **and** he's **like** look**ing** at **me** ... give **him** ... he *sees* this girl *coming* after ... that looks **like** a ... on the ground ...
	Ours: ... **Pie Man** So **I** was ... from **the Ram And** sure ... *Sheila Beale* student ... And now *Sheila* was different **from me and all** the other Fordham *students* ... *Sheila* ... was the kind of *student* ...
40TR	**Target** : ... I called him Pie Man as he fled ... And I gave ... said that he cried out Ego sum non an bestia ... and he says Pie Man I love it ...
	UniCoRN: ... and **the scene** a platform **and I** and the rumor ... **stories** a *Sherila* ... non my place ... **my story**
	EEG-Text: ... **I'm** looking at her **and I** ... kind of **figure** as ... than *his* **and** then **it** ... got **good** educations ... went **to** work ...
	BP-GPT: ... know how you do **it** ... end of a long **day** ... *Clara* is hold**ing** on**to** my **hand** ... **The** gray haired **man** ...
	Ours: ... you **Pie Man And I** ... **And** wasn't **I** really **Pie Man** ... made *his* delivery and **fled** away *crying* **Ego sum non an bestia** ... what it **said** in ...
60TR	**Target** : ... moved to New York ... And I went and looked ... what he did in Brazil And he said ... he wanted Bob to go hire ... because Alan was going to do this stupid thing ... And so I'm J Jhon Moscow ...
	UniCoRN: ... **he's** the girl You't **want** ... *Sherlock* so *Watson* ... two **things** with not .. Well *Arthur* ... almost days when before ...
	EEG-Text: ... **with** the press and **came to** *Boston* ... **And he said** ...he asked *Ryan* to find the same ... So *Ryan* contacted four officials and ...
	BP-GPT: ... **went** downtown **and** started ... **he knew** people **in the** group ... he helped **move** items but ... It **was** unclear who **was** ... from *Clara* ...
	Ours: ... **I did** not ... **I'm going to** *start* ... who **I am And so** ... **he** got plenty of ... But the fact that *Sheila* had collaborated with the *Dean* to ... radio station **in** *St Louis* ...

As shown in Table 10 and 11, our proposed method, **CogReader**, outperforms SOTA methods in terms of capturing semantics and syntax in tokens, with more accuracy of individual words, which indicated in bold. Specifically, our decoding results capture more key content words ranging from verbs (such as "fled" and "cried") to nouns ("scene" and "night"), including more accurate named entities such as person and place names ("Pie Man" and "St Louis"), and produce sentences that are semantically more aligned with the intended meaning. Despite other SOTA methods are still able to decode some accurate information such as "Man" and "came", they fail to decode more meaningful words such as "The Ram" and "Ego sum non an bestia", which is paired in our method. Therefore, our method produces decoded outputs that exhibit higher word-level overlap and better semantic alignment with the reference texts, demonstrating superior decoding performance.

Table 11: Cases analysis for fMRI2text. Exact matches between the target and predicted sentences are indicated in **bold**, while semantic similarity is shown in *italic* font.

fMRI Length	Results
20TR	**Target**: ...I gave the figure ... I called him Pie Man ... as he fled ... And I gave ... said that he cried out Ego sum non an bestia ...
	UniCoRN: ... **and** *me* gray to little **the Man** and ..of years ... to the then **said** a case ... girl of the story ...
	EEG-Text: ... **I'm** trying to ... a *title* for the *book* ... stories by ... in *Waukegan* but he **was** ... **which** is about ... **So I am** going to ...
	BP-GPT: ... **and I** teach people *how* to **make** ... one day **I was** walk**ing** ... and she **said** ... you sound like ...
	Ours: ... **I** think ... happened next **Pie Man** ... **and fled** away *crying* **Ego sum non an bestia** Or ... **said** in my story...
40TR	**Target** : ... she and her friends ... all the way ... at night .. She drives around ... I ever saw my mother cry ... My mom and dad ... was a government worker ...
	UniCoRN: ... **she** remembered how they used to ... the *evenings* sometimes ... when **she drove** back ... **and** that **was** when **she** felt ...
	EEG-Text: ... walking **down the** street **and** ... high**way** And then ... home from **work she** says to me ... There's something wrong with ...
	BP-GPT: ... about to **come off** ... **coming down** from the top **of** structure ... for no *good* reason ... playing with this guy ...
	Ours: ... **My mom** as **she cried** ... **she** stopped **and she drove** us home ... later that **night** .. An *organization* one of many ... onto **the** high**way** ...
60TR	**Target** : ... he uh like moves ... Um the scene then ... the lady and she uh she looks ... come in Merlin like ... and leaves Um so ...
	UniCoRN: ... I says at takes and **then** of ... and um says ... him guy **like** ... **the scene** of ... she is the man of the friending to ...
	EEG-Text: ... **he** slowly walks ... **and moves** ... under **the** books she ... where something powerful had ... **he** tries to *walk away* ...
	BP-GPT: ... **uh he** stands rather ... when **she**'s about to open ... **and then so** ... when *Arthur* yells ... **and he** finally responds ...
	Ours: ... **like he** hears ... this is **uh** a **scene** takes ... Probably **um she looks like the** *woman* who ... **So** when *Mutarelli* **and** I ...