# Rethinking Cross-Subject Data Splitting for Brain-to-Text Decoding

**Anonymous ACL submission**


## Abstract

Recent major milestones have successfully reconstructed natural language from non-invasive brain signals (e.g. functional Magnetic Resonance Imaging (fMRI) and Electroencephalogram (EEG)) across subjects. However, we find current dataset splitting strategies for cross-subject brain-to-text decoding are wrong. Specifically, we first demonstrate that all current splitting methods suffer from data leakage problem, which refers to the leakage of validation and test data into training set, resulting in significant overfitting and overestimation of decoding models. In this study, we develop a right cross-subject data splitting criterion without data leakage for decoding fMRI and EEG signal to text. Some SOTA brain-to-text decoding models are re-evaluated correctly with the proposed criterion for further research.


## 1 Introduction

Brain-to-text decoding aims to recover natural language from brain signals stimulated by corresponding speech. Recent studies (Makin et al., 2020; Wang and Ji, 2022; Xi et al., 2023; Tang et al., 2023; Duan et al., 2024) have successfully decoded non-invasive brain signals (e.g. fMRI, EEG) to text by applying deep neural networks. Most of these works perform within-subject data splitting for training and evaluating decoding models. This subject-specific splitting method causes two main problems. First, it only uses a small part of the whole dataset. For example, Tang et al. (2023) trained and tested model three times on three subjects respectively. Since brain signal collection is costly and time-consuming, such splitting method results in a significant waste of data resources. Second, it leads to poor model generalization. As every brain has unique functional and anatomical structures, subject-specific models may exhibit considerable variability across individuals and fail to generalize to other subjects (Liu et al., 2024). Moreover, decoding models trained from scratch on limited data are prone to facing the overfitting problem.

Human brain responds similarly to the same stimuli, despite the individual discrepancy (Hasson et al., 2004; Pereira et al., 2018). Therefore, some studies (Wang and Ji, 2022; Xi et al., 2023; Duan et al., 2024) begin to shed light on cross-subject brain-to-text decoding, which performs data splitting based on all the subjects, trains and evaluates decoding model once. Cross-subject data splitting effectively compensates for the shortcomings of subject-specific splitting, and has been widely applied in brain-to-image decoding (Wang et al., 2024; Liu et al., 2024). However, unlike datasets for brain-to-image decoding (Allen et al., 2022; Chang et al., 2019) where subjects are guided to see different and unrepeated pictures, different subjects will be stimulated by the same story in common naturalistic language comprehension dataset, which challenges cross-subject data splitting.

Based on our observations, current cross-subject data splitting methods for brain-to-text decoding are wrong because data for validation and test leaks into the training set, rendering the evaluation of the decoding process meaningless. Specifically, we find two types of data leakage: *brain signal leakage* and *text stimuli leakage*. Brain signal leakage refers to test subject's brain signal appears in training set. Text stimuli leakage refers to text in test set appears in the training set. Modern brain-to-text decoding models follow an encoder-decoder manner. We pick two representative models (detailed in Section B): EEG2Text (Wang and Ji, 2022) and UniCoRN (Xi et al., 2023) to reveal data leakage and its damage. Experiments support that data leakage affects model training on both encoder side and decoder side. For the encoder, the encoder will become overfitting and fail to well represent brain signals if brain signal leakage exists. For the decoder, the situation gets worse if text stimuli leakage happens. Any data leakage would cause the auto-regressive decoder to memorize previously seen paragraphs during training stage, resulting in poor generalization to unseen text.

To avoid data leakage and fairly evaluate the performance of cross-subject brain-to-text decoding models, we propose a right data splitting method. We focus on fMRI and EEG signals in this study, although the proposed criterion could be applied to any datasets satisfying the prescribed format. In the proposed method, we follow two basic rules: (1) Brain signals collected from specific subject in validation set and test set will not appear in training set, which means the trained encoder cannot get access to any brain information belonging to subjects in test set. (2) Text stimuli in validation set and test set will not appear in training set. The decoder learns to reconstruct language with brain signals instead of memorizing seen text.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to identify the issue of data leakage in current cross-subject data splitting methods for brain-to-text decoding.
- We define the splitting criterion for cross-subject brain-to-text decoding, and propose a right dataset splitting method.
- Some SOTA brain-to-text decoding models are re-evaluated using the proposed cross-subject data splitting method to ensure a fair assessment of their performance.

## 2 Problem Formulation

### 2.1 Dataset Description

A naturalistic language comprehension dataset $\mathcal{D}$ contains brain signals of $N$ subjects when they passively listen to $K$ spoken stories. Suppose that not all subjects are stimulated by all stories, and different subjects may hear the same story.

Formally, $S_1, S_2, \ldots, S_N$ denotes to the $N$ subjects and $M_1, M_2, \ldots, M_K$ denotes to the $K$ stories in dataset. The $k$-th story $M_k$ consists of $l_k$ text segments $T_{k1}, T_{k2}, \ldots, T_{kl_k}$. If the $i$-th subject $S_i$ hears the $j$-th text segment $T_{kj}$, then his brain signal is denoted as $F_{ijk}$.

### 2.2 Use Graph to Describe Dataset

We use graph to describe the intricate structure of naturalistic language comprehension dataset. We first introduce *multigraph* and *k-partite graph*.

**Definition 2.1.** An directed *multigraph* $\mathcal{G}$ is a type of graph which is permitted to have multiple edges between two vertices. When the edges own identity, $\mathcal{G}$ can be written as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, f)$, where $f : \mathcal{E} \to \mathcal{V} \times \mathcal{V}$ is an incidence function that maps each edge to a pair of vertices.

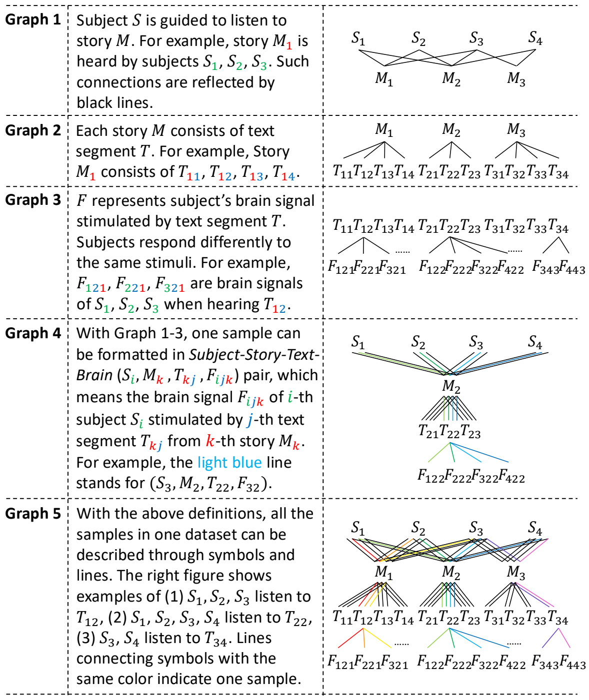


Figure 1: Illustration of how to build graph to describe dataset step by step.

**Definition 2.2.** A *$k$-partite graph* $\mathcal{G}$ is a type of graph that can be divided into $k$ distinct independent sets such that no two vertices in the same set are connected. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \cdots \cup \mathcal{V}_k$ and $\forall i \neq j, \mathcal{V}_i \cap \mathcal{V}_j = \emptyset$.

Following the dataset definition in Section 2.1, we use graph to describe a naturalistic language comprehension dataset with Definition 2.1 and 2.2.

**Definition 2.3.** A naturalistic language comprehension dataset $\mathcal{D}$ can be represented via a directed 4-*partite multigraph* $\mathcal{G}_{\mathcal{D}}$.

How to build the directed 4-*partite multigraph* $\mathcal{G}_{\mathcal{D}}$ step by step is shown in Figure 1. Graph 1 is a 2-*partite* graph indicating subject $S_i$ listening to story $M_k$. Subject $S_i$ and story $M_k$ are viewed as vertices, and edges connecting them indicate certain type of relationship (e.g. $S_i$ "listen to" $M_k$ in this case). Graph 2 illustrates that story $M_k$ consists of text segments $T_{kj}$. Graph 3 shows the brain signals $F_{ijk}$ of subject $S_i$ stimulated by text segment $T_{kj}$. Graph 4 is an example of combining the three 2-*partite* graphs Graph 1-3: $F_{122}, F_{222}, F_{322}, F_{422}$ are brain signals of $S_1, S_2, S_3, S_4$ stimulated by text segment $T_{22}$ from story $M_2$. In this example, four edges between $M_2$ and $T_{22}$ correspond to the different responses of four subjects to the same text segment. There are three edges between $S_2$ and $M_2$ because $M_2$ contains three text segments. Edges of the same color indicate one sample in dataset. Graph 5 shows the complete directed 4-*partite multigraph* $\mathcal{G_D}$ for representing whole dataset. Every sample in dataset can be represented through ordered subject-story-text-brain $(S_i, M_k, T_{kj}, F_{ijk})$ pair. We introduce the formal notation of $\mathcal{G_D}$:

**Notation 2.4.** $\mathcal{G_D} = (\mathcal{V}, \mathcal{E}, f)$, where $\mathcal{V} = \mathcal{S} \cup \mathcal{M} \cup \mathcal{T} \cup \mathcal{F}$, $\mathcal{S} = \{S_i\}_{i=1}^N$, $\mathcal{M} = \{M_k\}_{k=1}^K$, $\mathcal{T} = \{T_{kj}\}_{k,j=1}^{K,l_k}$, $\mathcal{F} = \{F_{ijk}\}_{i,j,k=1}^{N,l_k,K}$ denote subject set, story set, text segment set, and brain signal set. $f : \mathcal{E} \to \mathcal{V} \otimes \mathcal{V}$ is an incidence function that maps each edge to a pair of vertices.

**Notation 2.5.** $\otimes$ is a Cartesian product-like operator. $X \otimes Y = \{(x, y)|x \in X, y \in Y$, there exists relationship between $x$ and $y$ in dataset$\}$. It's designed to describe the connectivity among $\mathcal{S}, \mathcal{M}, \mathcal{T}, \mathcal{F}$. For example, edges in $\mathcal{S} \otimes \mathcal{M}$ indicates certain subjects are stimulated by certain stories as described in dataset.

### 2.3 Brain-to-text Decoding Task

The brain-to-text decoding task seeks to build a decoding model that reconstructs natural language text from brain signals, with the goal of accurately decoding what the subject hears.

Take fMRI and EEG signal for example. fMRI captures brain responses at second level and such interval is known as TR (Repetition Time), whereas EEG samples brain activity at the millisecond level. As a result, the pre-processing for fMRI and EEG input varies. Previous practice in fMRI-to-text decoding (Tang et al., 2023; Xi et al., 2023) concatenated $L$ future fMRI frames and corresponding text segments to form one sample:

$$T^*_{k,j} = concat(T_{k,j}, T_{k,j+1}, \dots, T_{k,j+L}), \quad (1)$$

$$F^*_{i,j,k} = concat(F_{i,j,k}, F_{i,j+1,k}, \dots, F_{i,j+L,k}). \quad (2)$$

In this case, one $(S_i, M_k, T_{kj}, F_{ijk})$ pair in graph $\mathcal{G_D}$ only represents the start point of one sample, while $(S_i, M_k, T^*_{kj}, F^*_{ijk})$ indicates the whole sample. In EEG-to-text decoding, previous methods sampled continuous EEG signal $F_{ijk}$ that corresponds to text $T_{kj}$. So one $(S_i, M_k, T_{kj}, F_{ijk})$ pair is viewed as one sample in our definition.

## 3 Methodology

We first introduce the criterion for cross-subject splitting. Then we discuss current cross-subject data splitting methods, and point out that all existing methods suffer from data leakage problem. Finally, we design a right splitting method that satisfies cross-subject splitting criterion.

### 3.1 Cross-Subject Data Splitting Criterion

Consistent with cross-subject brain-to-image decoding (Wang et al., 2024; Liu et al., 2024), the dataset splitting should obey two basic principles: (1) If brain signal $F_{ijk}$ appears in test set, then any brain signal $F_{i*k}$ belonging to this subject $i$ should not appear in training set. (2) If text segment $T_{kj}$ appears in test set, then it should not appear in training set. Following the definitions in Section 2, graph $\mathcal{G_D}$ is applied to describe data splitting. Training and test set are denoted as $\mathcal{G}_{train}$ and $\mathcal{G}_{test}$. Since the validation samples are split in the same manner as the test samples, we focus solely on the test samples. Therefore, we have $\mathcal{G_D} = \mathcal{G}_{train} \cup \mathcal{G}_{test}$. We formally define cross-subject splitting criterion which is applicable to training set and test set. Test set definition is omitted for simplicity.

**Definition 3.1.** The training set for cross-subject brain-to-text decoding should be formatted in $\mathcal{G}_{train} = \mathcal{S}_{train} \otimes \mathcal{M} \otimes \mathcal{T}_{train} \otimes \mathcal{F}_{train}$, where $\mathcal{S}_{train} = \{S_i|\forall S'_i \in \mathcal{S}_{test}, S_i \neq S'_i\}$; $\mathcal{F}_{train} = \{F_{ijk}|i \in I\}$, $I = \{i|\forall j, \forall k, F_{ijk} \notin \mathcal{F}_{test}\}$; $\mathcal{T}_{train} = \{T_{kj}|\forall T'_{kj} \in \mathcal{T}_{test}, T_{kj} \neq T'_{kj}\}$.

### 3.2 Analysis of Current Splitting Methods

As illustrated in Figure 2, we use different colored edges to represent their classification as either part of the training set or the test set. $(S_i, M_k, T_{kj}, F_{ijk})$ pairs with green edges indicate training samples, and those with orange edges are test samples. Current cross-subject data splitting methods (Wang and Ji, 2022; Xi et al., 2023) can be summarized as five categories:

- Method (a): Split subjects $\mathcal{S}$ randomly with given ratio.

$$\begin{aligned}\mathcal{G}_{train} = \{&(S_i, M_k, T_{kj}, F_{ijk})| \\ &\forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, S_i \neq S'_i\}\end{aligned} \quad (3)$$

- Method (b): Split stories $\mathcal{M}$ randomly with given ratio.

$$\begin{aligned}\mathcal{G}_{train} = \{&(S_i, M_k, T_{kj}, F_{ijk})| \\ &\forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, M_k \neq M'_k\}\end{aligned} \quad (4)$$

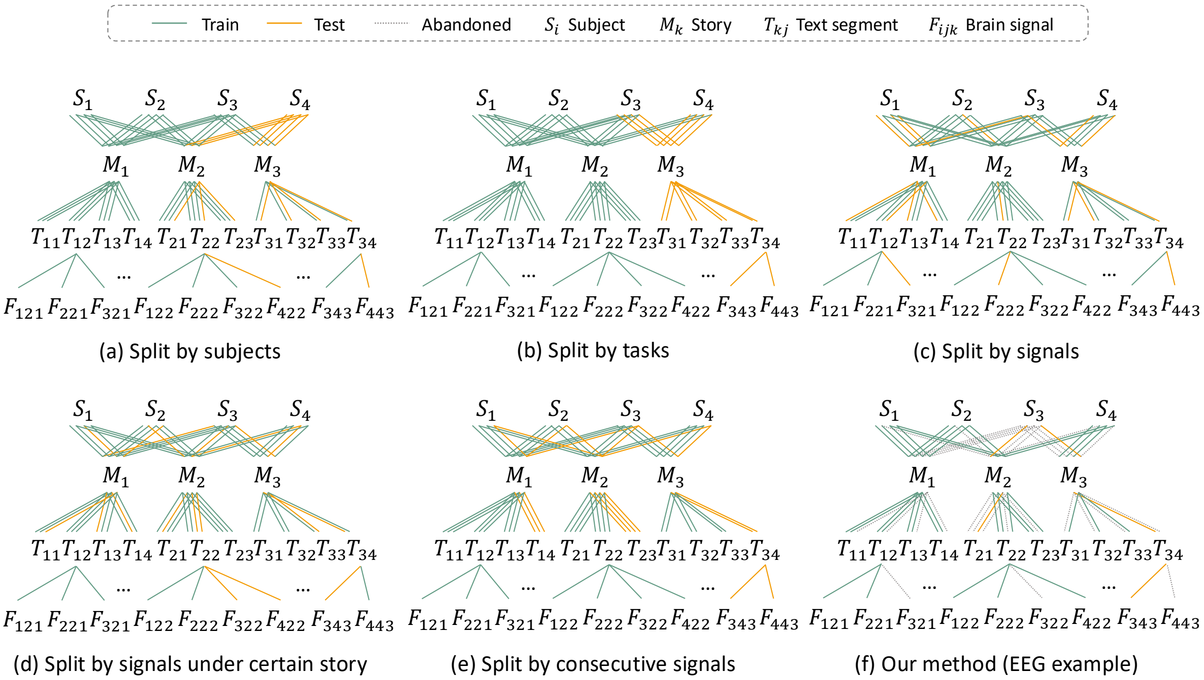


Figure 2: Different splitting methods for cross-subject brain-to-text decoding. (Color printing is preferred.)

- Method (c): Split all the brain signals $\mathcal{F}$ randomly with given ratio.

$$\mathcal{G}_{train} = \{(S_i, M_k, T_{kj}, F_{ijk})| \\ \forall(S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, F_{ijk} \neq F'_{ijk}\} \tag{5}$$

- Method (d): Different from Method (c), it splits brain signals under each story randomly with given ratio, and union them to form the whole training and test set.
- Method (e): Different from Method (d), it splits continuous brain signals under each story with given ratio, and union them to form the whole training and test set.

It's evident that Equation (3), (4), (5) do not meet the criterion outlined in Definition 3.1. To facilitate a thorough analysis, we introduce the concept of *brain signal leakage* and *text stimuli leakage*. Specifically, brain signal leakage refers to test subject's brain signal appears in training set. Text stimuli leakage refers to text segment in test set appears in the training set. Formal definitions of two types of data leakage are given.

**Definition 3.2.** Brain signal leakage happens when

$$\forall(S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_{train}, \\ \exists(S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, S'_i = S_i. \tag{6}$$

**Definition 3.3.** Text stimuli leakage happens when

$$\forall(S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_{train}, \\ \exists(S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, T'_{kj} = T_{kj}. \tag{7}$$

| fMRI | Method(a) | Method(b) | Method(c) | Method(d) | Method(e) |
|---|---|---|---|---|---|
| Brain Signal Leakage | ✗ | ✓ | ✓ | ✓ | ✓ |
| Text Stimuli Leakage | ✓ | ✗ | ✓ | ✓ | ✓ |

Table 1: Data leakage in five different splitting methods applied to fMRI-to-text decoding.

| EEG | Method(a) | Method(b) | Method(c) | Method(d) | Method(e) |
|---|---|---|---|---|---|
| Brain Signal Leakage | ✗ | ✓ | ✓ | ✓ | N/A |
| Text Stimuli Leakage | ✓ | ✗ | ✓ | ✓ | N/A |

Table 2: Data leakage in five different splitting methods applied to EEG-to-text decoding.

We prove in Appendix E that a splitting method without brain signal and text stimuli leakage will satisfy the splitting criterion in Definition 3.1. Data leakage can be directly identified in graph $\mathcal{G}_{\mathcal{D}}$. As shown in Figure 2, if edges connected to $S_i$ are of different colors, it indicates that brain signals of $S_i$ appears in both training set and test set, which leads to brain signal leakage. Similarly, if edges connected to $T_{kj}$ are of different colors, it suggests that text segment $T_{kj}$ appears in both training set and test set, which leads to text stimuli leakage.

As a result, in the scenario of EEG signals where $(S_i, M_k, T_{kj}, F_{ijk})$ is viewed as a sample: Method (a) suffers from text stimuli leakage. Method (b) faces brain signal leakage. Method (c) is affected by leakage of both text stimuli and brain signals. Method (d) and (e) do not show any differences compared to method (c) in EEG-to-text decoding. In fMRI-to-text decoding, continuous fMRI frames and text stimuli are concatenated to form one sample. $(S_i, M_k, T_{kj}, F_{ijk})$ indicates the start point of one sample instead of the whole sample (recall Section 2.3). In this case, method (d) and (e) mean differently. Similar to method (c), method (d) and (e) face both brain signal leakage and text stimuli leakage. But for method (e) the text stimuli is slight. It only happens in the overlapping part between training samples and test samples. The situations of data leakage in different splitting methods are detailed in Table 1 and 2.

### 3.3 A Right Cross-Subject Splitting Method

We propose a right cross-subject splitting method to eliminate both brain signal leakage and text stimuli leakage. The key point is to ensure zero brain signal leakage and text stimuli leakage.

$$\mathcal{G}_{train} = \{(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, S_i \neq S'_i, T_{kj} \neq T'_{kj}\}. \tag{8}$$

Given the differences of EEG and fMRI dataset, we address them separately and propose two data splitting methods. In EEG dataset, $(S_i, M_k, T_{kj}, F_{ijk})$ forms one sample. As shown in Figure 3, our proposed splitting method consists of three steps:

- Step 1: Select $\sum_{k=1}^{K} l_k$ samples from $\mathcal{G}_\mathcal{D}$ and form a new graph $\mathcal{G}'_\mathcal{D}$ that satisfies

$$\forall (S'_i, M'_k, T'_{kj}, F'_{ijk}), (S''_i, M''_k, T''_{kj}, F''_{ijk}) \in \mathcal{G}'_\mathcal{D}, T'_{kj} \neq T''_{kj}. \tag{9}$$

- Step 2: Split $\mathcal{G}'_\mathcal{D}$ to $\mathcal{G}'_{train}$ and $\mathcal{G}'_{test}$ with a given ratio. The splitting should follow

$$\mathcal{G}'_{train} = \{(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}'_{test}, S_i \neq S'_i\}, \tag{10}$$

$$\mathcal{G}'_{test} = \{(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}'_{train}, S_i \neq S'_i\}. \tag{11}$$

- Step 3: Expand $\mathcal{G}'_{train}$ and $\mathcal{G}'_{test}$ with $\mathcal{G}'_{train\_exp}$ and $\mathcal{G}'_{test\_exp}$ separately.

$$\begin{aligned} \mathcal{G}'_{train} &\leftarrow \mathcal{G}'_{train} \cup \mathcal{G}'_{train\_exp} \\ \mathcal{G}'_{test} &\leftarrow \mathcal{G}'_{test} \cup \mathcal{G}'_{test\_exp} \end{aligned} \tag{12}$$

where $\mathcal{G}'_{train\_exp}$ and $\mathcal{G}'_{test\_exp}$ are

$$\mathcal{G}'_{train\_exp} = \{(S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_\mathcal{D} | S_i \in \mathcal{S}'_{train}, T_{kj} \in \mathcal{T}'_{train}\}, \tag{13}$$

$$\mathcal{G}'_{test\_exp} = \{(S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_\mathcal{D} | S_i \in \mathcal{S}'_{test}, T_{kj} \in \mathcal{T}'_{test}\}. \tag{14}$$

$\mathcal{S}'_{train}$, $\mathcal{T}'_{train}$, $\mathcal{S}'_{test}$, $\mathcal{T}'_{test}$ indicate subject set, text segment set in $\mathcal{G}'_{train}$ and subject set, text segment set in $\mathcal{G}'_{test}$ respectively.

Some samples are discarded in our proposed splitting method, i.e. $\mathcal{G}_\mathcal{D} \neq \mathcal{G}'_{train} \cup \mathcal{G}'_{test}$. In Appendix E, we demonstrate that it is unavoidable for some samples to be discarded in order to satisfy the cross-subject data splitting criterion.

To fMRI dataset, continuous text segments and brain signals are concatenated to form one sample $(S_i, M_k, T^*_{kj}, F^*_{ijk})$. If we follow the same splitting method as to EEG dataset, text stimuli leakage will happen in the overlapping part of two samples, when one sample is assigned to training set and the other is assigned to validation or test set. We propose a simple solution that achieves the balance between discarding as little data as possible while ensuring zero data leakage: Step 1 and Step 3 remain the same as splitting method for EEG dataset. In Step 2, $\mathcal{G}'_{train}$ and $\mathcal{G}'_{test}$ should follow

$$\mathcal{G}'_{train} = \{(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}'_{test}, S_i \neq S'_i, M_k \neq M'_k\}, \tag{15}$$

$$\mathcal{G}'_{test} = \{(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}'_{train}, S_i \neq S'_i, M_k \neq M'_k\}. \tag{16}$$

## 4 Experimental Settings

### 4.1 Implementation Detail

We test two SOTA cross-subject brain-to-text decoding models UniCoRN (Xi et al., 2023) and EEG2Text (Wang and Ji, 2022) on fMRI dataset Narratives (Nastase et al., 2021) and EEG dataset ZuCo (Hollenstein et al., 2018). Dataset details are introduced in Appendix C. Because the number of stories in ZuCo dataset is too small, and method (e) makes no difference to EEG as method (d), we only consider splitting method (a), (c), (d) for EEG. We follow the same settings of UniCoRN and EEG2Text, except all the datasets are split to the ratio of 8:1:1 for fair comparison. Details are shown in Appendix C.

### 4.2 Evaluation Metrics

**Data Leakage Metrics** We design two novel evaluation metrics **Brain Signal Leakage Rate (BSLR)** and **Text Stimuli Leakage Rate (TSLR)**

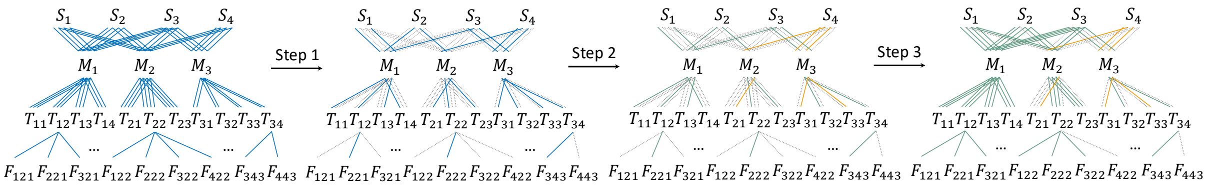


Figure 3: The detailed steps of our proposed cross-subject data splitting method. (Color printing is preferred.)

| Type | Method | Narratives | ZuCo |
|---|---|---|---|
| BSLR(%) | (a) | $\mathbf{0.00}_{\pm 0.00}$ | $\mathbf{0.00}_{\pm 0.00}$ |
| | (b) | $9.67_{\pm 4.80}$ | / |
| | (c) | $12.50_{\pm 0.04}$ | $12.50_{\pm 0.03}$ |
| | (d) | $12.80_{\pm 0.01}$ | $12.59_{\pm 0.02}$ |
| | (e) | $12.27_{\pm 0.01}$ | / |
| | (f) | $\mathbf{0.00}_{\pm 0.00}$ | $\mathbf{0.00}_{\pm 0.00}$ |
| TSLR(%) | (a) | $100.00_{\pm 0.00}$ | $22.50_{\pm 1.31}$ |
| | (b) | $\mathbf{0.00}_{\pm 0.00}$ | / |
| | (c) | $100.00_{\pm 0.00}$ | $13.07_{\pm 0.11}$ |
| | (d) | $99.82_{\pm 0.17}$ | $12.88_{\pm 0.04}$ |
| | (e) | $9.29_{\pm 0.06}$ | / |
| | (f) | $\mathbf{0.00}_{\pm 0.00}$ | $\mathbf{0.00}_{\pm 0.00}$ |

Table 3: Results of Brain Signal Leakage Rate (BSLR) and Text Stimuli Leakage Rate (TSLR). Lower is better.

to quantify two types of data leakage. Note that the situation for validation set is the same as test set, so we only consider test set in experiments. BSLR indicates the average percentage of each subject's brain signals in test set appearing in training set, which could be formulated as

$$\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \min(1, \frac{|\{F_{ijk} | F_{ijk} \in (\mathcal{G}_{test} \cap \mathcal{G}_{train})\}|}{|\{F_{ijk} | F_{ijk} \in \mathcal{G}_{train}\}|}) \tag{17}$$

where $N_{test}$ stands for the total number of subjects in test set. $|\cdot|$ stands for the cardinality of a set. Function $\min(\cdot, \cdot)$ is applied to make sure for each subject the data leakage rate is less than one.

The definition of TSLR is different for EEG signal and fMRI signal. Since $(S_i, M_k, T_{kj}, F_{ijk})$ indicates one sample in EEG dataset, definition of TSLR for EEG dataset is similar to BSLR, which measures the average percentage of certain text in test set appearing in training set.

$$\frac{1}{M_{test}} \sum_{j=1}^{M_{test}} \min(1, \frac{|\{T_{kj} | T_{kj} \in (\mathcal{G}_{test} \cap \mathcal{G}_{train})\}|}{|\{T_{kj} | T_{kj} \in \mathcal{G}_{train}\}|}) \tag{18}$$

where $M_{test}$ stands for the total number of text segments in test set. To fMRI dataset, continuous fMRI frames with corresponding text segments are concatenated as one sample. As a result, TSLR for fMRI signal is considered as the average percentage of the same text segments in test set appearing in training set, which is

$$\frac{1}{M_{test}} \sum_{j=1}^{M_{test}} \tau \frac{|\{T_{kj} | T_{kj} \in (\mathcal{G}_{test} \cap \mathcal{G}_{train})\}|}{|\mathcal{G}_{test}| \times L} \tag{19}$$

where $\tau = 0$ if $\{T_{kj} | T_{kj} \in \mathcal{G}_{test} \cap \mathcal{G}_{train}\} = \emptyset$ else

$$\tau = \min(1, \frac{|\{T_{kj} | T_{kj} \in \mathcal{G}_{train}\}|}{|\{T_{kj} | T_{kj} \in (\mathcal{G}_{test} \cap \mathcal{G}_{train})\}|}). \tag{20}$$

**Decoding Performance Metrics** Automatic evaluation metrics including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are applied to measure the decoding performance. BLEU measures the n-gram overlap between decoded content and ground truth. ROUGE-N comparing the consistency of N-grams between the decoded content and the ground truth.

## 5 Experiments and Analysis

We first conduct a data leakage verification experiment to quantify the data leakage condition of different methods with BSLR and TSLR metrics. Then we demonstrate the damage of data leakage on encoder side and decoder side. For model encoder, we analyze its validation loss under different splitting methods. For model decoder, three experiment settings are applied: (1) An additional test set that ensures zero data leakage is left out as comparison to original test set. (2) The input brain signals are randomly shuffled. (3) Training original models with more epochs and smaller learning rate.

### 5.1 Verification for Data Leakage

Experiments on BSLR and TSLR are conducted four times with different seeds. The results in Table 3 are consistent with theoretical analysis. A value of zero in BSLR and TSLR demonstrate no brain signal leakage and text stimuli leakage, while higher values suggest more significant data leakage

| Dataset | Model | Method | Original Test Set / Additional Test Set | | | |
|---|---|---|---|---|---|---|
| | | | BLEU-1 | BLEU-2 | BLEU-3 | ROUGE1-F |
| Narratives | UniCoRN | (a) | 49.56 / 18.43 | 30.49 / 1.25 | 21.07 / 0.00 | 40.65 / 16.38 |
| | | (b) | 26.37 / 23.31 | 7.50 / 5.79 | 2.48 / 1.44 | 19.62 / 18.74 |
| | | (c) | 50.24 / 16.96 | 30.83 / 0.09 | 21.23 / 0.00 | 41.01 / 15.12 |
| | | (d) | 49.63 / 17.20 | 30.29 / 1.15 | 20.85 / 0.00 | 41.03 / 15.83 |
| | | (e) | 28.94 / 21.79 | 9.39 / 4.62 | 4.07 / 1.19 | 19.49 / 18.78 |
| | | (f) | 22.83 / 21.64 | 5.69 / 4.97 | 1.43 / 1.28 | 19.04 / 18.45 |
| ZuCo | UniCoRN | (a) | 58.09 / 18.54 | 49.23 / 1.31 | 43.23 / 0.00 | 67.50 / 15.39 |
| | | (c) | 52.30 / 18.38 | 42.89 / 1.03 | 36.80 / 0.00 | 67.29 / 15.25 |
| | | (d) | 50.02 / 19.84 | 43.53 / 1.20 | 32.71 / 0.03 | 67.33 / 15.12 |
| | | (f) | 23.32 / 22.89 | 7.78 / 7.46 | 3.01 / 2.75 | 17.92 / 17.63 |
| | EEG2Text | (a) | 51.22 / 17.41 | 33.83 / 1.04 | 22.99 / 0.00 | 46.58 / 15.92 |
| | | (c) | 53.83 / 17.38 | 38.99 / 0.84 | 29.57 / 0.00 | 53.56 / 16.07 |
| | | (d) | 53.92/ 16.86 | 41.06 / 1.32 | 23.12 / 0.00 | 49.38 / 15.83 |
| | | (f) | 24.49 / 23.71 | 7.49 / 7.42 | 2.28 / 2.33 | 25.74 / 23.30 |

Table 4: Performance of brain-to-text decoding models under different splitting methods on original test set and an additional test set. The green mark denotes a method without text stimuli leakage, whereas the red mark signifies methods that have text stimuli leakage.

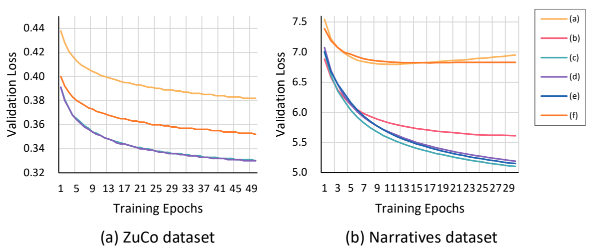


Figure 4: Validation loss of encoder under different dataset splitting methods in two datasets.

issues. Notably, only our method (f) prevents both brain signal leakage and text stimuli leakage.

### 5.2 Damage of Data Leakage to Encoder

Evaluating the encoder independently can be challenging in an end-to-end training scenario. Therefore, we primarily focus on a pre-trained encoder. Since a proper evaluation index of encoder's representation ability is missing, validation loss is applied to measure data leakage. The validation loss of encoder under different data splitting methods is shown in Figure 4. For fMRI data, the presence of brain signal leakage causes the validation loss of methods (b), (c), (d), and (e) to continuously decrease even over extended training epochs. This indicates that the encoder is actually overfitting and its representation ability is degrading. In contrast, with methods (a) and (f) that are not affected by brain signal leakage, the validation loss quickly increases after reaching its minimum within a few epochs, which aligns with the fundamental principles of machine learning. For EEG, we find validation loss keeps dropping for all methods even with very long training epochs, regardless of brain signal leakage or not. We think the poor spatial resolution of EEG signal might lead to this phenomenon.

### 5.3 Damage of Data Leakage to Decoder

**Evaluation on Additional Test Set** An additional test set that ensures zero data leakage is left out to evaluate the actual performance of brain-to-text decoding models. If the original test set is correctly split, its decoding result should be similar to that of the additional test set. From Table 4, we observe that the decoding model tends to overfit when text stimuli leakage occurs, as seen in methods (a), (c), (d), and (e) in Narratives, and methods (a) and (c) in ZuCo. The BLEU and ROUGE score is significantly lower in the additional test set. While in our proposed splitting method (f), the decoding performance of original and additional test set are similar. We also notice that methods with a high Text Stimuli Leakage Rate (TSLR), such as method (a) in Narratives, exhibit more overfitting compared to methods with a low TSLR, like method (e).

**Shuffle Input Brain Signals** We conduct a chance-level experiment to investigate whether decoding models learn language reconstruction from brain signals. Specifically, the input brain signals are randomly shuffled. Decoding performance in test set is expected to be very poor if text stimuli

| Dataset | Model | Method | Ordered Input / Shuffled Input | | | |
|---|---|---|---|---|---|---|
| | | | BLEU-1 | BLEU-2 | BLEU-3 | ROUGE1-F |
| Narratives | UniCoRN | (a) | 49.56 / 47.39 | 30.49 / 28.95 | 21.07 / 18.40 | 40.65 / 35.12 |
| | | (b) | 26.37 / 20.18 | 7.50 / 3.52 | 2.48 / 0.51 | 19.62 / 15.58 |
| | | (c) | 50.24 / 48.48 | 30.83 / 30.21 | 21.23 / 19.39 | 41.01 / 38.43 |
| | | (d) | 49.63 / 50.21 | 30.29 / 32.18 | 20.85 / 21.46 | 41.03 / 41.69 |
| | | (e) | 28.94 / 24.84 | 9.39 / 6.56 | 4.07 / 2.04 | 19.49 / 17.90 |
| | | (f) | 22.83 / 18.21 | 5.69 / 2.47 | 1.43 / 0.22 | 19.04 / 16.83 |
| ZuCo | UniCoRN | (a) | 58.09 / 59.23 | 49.23 / 51.35 | 43.23 / 44.27 | 67.50 / 68.93 |
| | | (c) | 52.30 / 50.24 | 42.89 / 37.96 | 36.80 / 30.21 | 67.29 / 63.43 |
| | | (d) | 50.02 / 51.12 | 43.53 / 40.85 | 32.71 / 28.24 | 67.33 / 64.88 |
| | | (f) | 23.32 / 19.38 | 7.78 / 2.51 | 3.01 / 0.00 | 17.92 / 15.21 |
| | EEG2Text | (a) | 51.22 / 50.63 | 33.83 / 32.19 | 22.99 / 20.63 | 46.58 / 44.70 |
| | | (c) | 53.83 / 50.33 | 38.99 / 33.42 | 29.57 / 23.19 | 53.56 / 48.78 |
| | | (d) | 53.92 / 51.46 | 41.06 / 35.87 | 23.12 / 24.75 | 49.38 / 47.42 |
| | | (f) | 24.49 / 18.72 | 7.49 / 2.01 | 2.28 / 0.00 | 25.74 / 15.36 |

Table 5: Performance of brain-to-text decoding models under different splitting methods with ordered brain signals and randomly shuffled brain signals as model input respectively.

| Dataset | Model | BLEU-N (%) | | | | ROUGE-1 (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | $N=1$ | $N=2$ | $N=3$ | $N=4$ | $R$ | $P$ | $F$ |
| Narratives | UniCoRN | **22.83** | **5.69** | **1.43** | **0.48** | **15.55** | **24.80** | **19.04** |
| ZuCo | UniCoRN | 23.32 | **7.78** | **3.01** | **1.09** | 18.47 | 20.00 | 17.92 |
| | EEG2Text | **24.49** | 7.49 | 2.28 | 0.62 | **23.98** | **23.95** | **25.74** |

Table 6: A fair benchmark for evaluating the performance of cross-subject brain-to-text decoding models.

leakage does not happen, as the shuffled input is considered as noise. However, if text stimuli in test set leaks into training set, the model will simply memorize seen text and the decoding performance is not supposed to be affected.

Results are presented in Table 5. For fMRI, we find the decoding performance of models under splitting method (a), (c), and (d) remain the same no matter the input is ordered or shuffled. Similar phenomenon is also observed in EEG dataset when it comes to splitting method (a), (c), (d). But in splitting method without text stimuli leakage, model performance with shuffled input drops significantly. This experiment demonstrates that the brain-to-text decoding task become meaningless when text stimuli leakage exists, as the Transformer block is capable of generating text that was previously encountered during the training phase.

**Longer Training Epochs with Smaller Learning Rate** According to fundamental machine learning principle, model performance in test set will first increase and then drop as the training proceeds. In this experiment, we try training models under different splitting methods with longer training epochs and smaller learning rate. If text stimuli leakage happens, the model is overfitting and its performance is supposed to keep increasing.

Results and detailed analysis are presented in Appendix F. In conclusion, the model's performance on test set continues to improve when text stimuli leakage happens, confirming that such leakage results in significant overfitting in decoding models.

### 5.4 A Fair Benchmark

We re-evaluate two SOTA models for brain-to-text decoding under our cross-subject data splitting method and release a fair benchmark. UniCoRN is tested for both fMRI and EEG decoding, EEG2Text model is tested for EEG decoding. The results are listed in Table 6. For EEG dataset, UniCoRN achieves higher results in BLEU-2,3,4 while EEG2Text is better in BLEU-1 and ROUGE-1.

## 6 Conclusion

In this paper, we focus on revealing the false dataset splitting method and its detrimental impact on cross-subject brain-to-text decoding research. We evidence that all current dataset splitting methods have data leakage problem through theoretical analysis and experiments. Such data leakage leads to model overfitting and largely exaggerates model performance, rendering model evaluation meaningless. To fix this issue, we propose a right cross-subject data splitting method. Current SOTA models are re-evaluated for further researches.

## Limitations

The limitations of this work include three aspects: (1) Although our splitting method can be applied to any natural language comprehension cognitive dataset, we only analyze cross-subject data splitting methods in fMRI and EEG dataset. We leave the investigation of other cognitive signals (e.g. ECoG, MEG, etc.) to future work. (2) Our proposed dataset splitting method meets the above requirements at the expense of discarding some data in the dataset. We recommend future datasets in this domain follow these guidelines. The division of the training set, validation set, and test set should be provided when the dataset is released. Besides, we suggest hiring new subjects with unique stimuli for the validation set and test set, which is good for testing the generalization ability of models without loss of data. (3) During experiments we find existing models rely more on a strong auto-regressive decoder to achieve good generation quality. The encoder is of limited use in all SOTA models. And we also notice in experiments that the encoder of EEG2Text keeps overfitting whether with or without brain signal leakage. We leave it as future research.

## Ethics Statement

In this paper, we introduce a new dataset splitting method to avoid data leakage for decoding brain signals to text task. Experiments are conducted on the publicly accessible cognitive datasets "Narratives" and ZuCo1.0 with the authorization from their respective maintainers. Both datasets have been de-identified by dataset providers and used for researches only.

## References

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498.

Boris Burle, Laure Spieser, Clémence Roger, Laurence Casini, Thierry Hasbroucq, and Franck Vidal. 2015. Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view. *International Journal of Psychophysiology*, 97(3):210–220.

Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. 2019. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):49.

Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. 2024. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems*, 36.

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

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. 2004. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640.

Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9:217.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

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

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. 2024. See through their minds: Learning transferable neural representation from cross-subject fmri. *arXiv preprint arXiv:2403.06361*.

Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227.

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. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific data*, 8(1):250.

Uta Noppeney and Cathy J Price. 2004. An fmri study of syntactic adaptation. *Journal of Cognitive Neuroscience*, 16(4):702–713.

Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. 2021. Decoding covert speech from eeg-a comprehensive review. *Frontiers in Neuroscience*, 15:392.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. ACL.

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

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 7047–7054. AAAI Press.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9.

Athena Vouloumanos, Kent A Kiehl, Janet F Werker, and Peter F Liddle. 2001. Detection of sounds in the auditory stream: event-related fmri evidence for differential activation to speech and nonspeech. *Journal of Cognitive Neuroscience*, 13(7):994–1005.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. 2024. Mindbridge: A cross-subject brain decoding framework. *arXiv preprint arXiv:2404.07850*.

Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 5350–5358. AAAI Press.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. 2023. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13277–13291. Association for Computational Linguistics.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2021. Towards brain-to-text generation: Neural decoding with pre-trained encoder-decoder models. In *NeurIPS 2021 AI for Science Workshop*.

## A Related Work

**Brain Signal** Brain signals can be classified into three categories: invasive, partially invasive, and non-invasive according to how close electrodes get to brain tissue. In this paper, we mainly focus on non-invasive signals EEG and fMRI. EEG signal is electrogram of the spontaneous electrical activity of the brain, with frequencies ranging from 1 Hz to 30 Hz. EEG is of high temporal resolution and relatively tolerant of subject movement, but its spatial resolution is low and it can't display active areas of the brain directly. fMRI measures brain activity by detecting changes of blood flow. Blood flow of a specific region increases when this brain area is in use. The spatial resolution of fMRI is measured by the size of voxel, which is a three-dimensional rectangular cuboid ranging from 3mm to 5mm (Vouloumanos et al., 2001; Noppeney and Price, 2004). Unlike EEG which samples brain signals continuously, fMRI samples based on a fixed time interval named TR, usually at second level.

**Brain-to-text Decoding** Previous research on brain-to-text decoding (Herff et al., 2015; Anumanchipalli et al., 2019; Zou et al., 2021; Moses et al., 2021; Défossez et al., 2023) mainly focused on word-level decoding in a restricted vocabulary with hundreds of words (Panachakel and Ramakrishnan, 2021). These models typically apply recurrent neural network or long short-term memory (Hochreiter and Schmidhuber, 1997) network to build mapping between brain signals and words in vocabulary. Despite relatively good accuracy, these methods fail to generalize to unseen words. Some progress (Sun et al., 2019) has been made by expanding word-level decoding to sentence-level through encoder-decoder framework or using less noisy ECoG data (Burle et al., 2015; Anumanchipalli et al., 2019). However, these models struggle to generate accurate and fluent sentences limited by decoder ability. Wang and Ji (2022) introduced the first open vocabulary EEG-to-text decoding model by leveraging the power of pre-trained language models. Xi et al. (2023) improved the model design and proposed a unified framework for decoding both fMRI and EEG signals.

## B Brain-to-text Decoding Models

**UniCoRN** UniCoRN provides a unified encoder-decoder framework for EEG and fMRI to text decoding. The training of UniCoRN follows a three-stage manner. The fMRI encoder is first pre-trained with a cognitive signal reconstruction task to capture spatial feature via a 3D-CNN module. Then a Transformer encoder is stacked into the fMRI encoder to capture temporal connections. Finally BART is fine-tuned to translate fMRI representation into natural language in the generation stage.

**EEG2Text** EEG2Text treats each EEG feature sequence as an encoded sentence by the human brain. An additional encoder is then trained to map the embedding from the human brain to the embedding from the pretrained BART. EEG feature vectors are used directly as initial word embeddings to feed into the model.

## C Implementation Details

We apply the "Narratives" (Nastase et al., 2021) dataset for fMRI-to-text decoding and the ZuCo (Hollenstein et al., 2018) dataset for EEG-to-text decoding in experiments. The "Narratives" dataset contains fMRI data from 345 subjects listening to 27 diverse stories. Since the data collection process involves different machines, we only consider fMRI data with $64 \times 64 \times 27$ voxels. The ZuCo dataset includes 12 healthy adult native English speakers reading English text for 4 to 6 hours. It contains simultaneous EEG and Eye-tracking data. The reading tasks include Normal Reading (NR) and Task-specific Reading (TSR) extracted from movie views and Wikipedia. Both datasets are split into training, validation, and test set with a ratio of 80%, 10%, 10% in all experiments.

We perform the same filtering steps to "Narratives" dataset as UniCoRN paper (Xi et al., 2023) and the same filtering steps to ZuCo1.0 as EEG2Text paper (Wang and Ji, 2022). In BSLR and TSLR calculation, the number of four different seeds are set as $1, 2, 3, 4$ respectively. In signal reconstruction task for encoder of UniCoRN, the batch size of EEG and fMRI data is 512 and 320 respectively. The learning rate is set as 1e-4 and 1e-3 separately as the author claimed in the original paper. In the fair benchmark, for fMRI data, encoder of UniCoRN is trained through 1e-4 learning rate and decaying to 1e-6 finally for 30 training epochs. Decoder is trained through 1e-4 learning rate and decaying to 1e-6 finally for 10 training epochs with 90 batch size. Sample length $L$ is set as 10 for all experiments related to fMRI. For EEG data, EEG2Text model is trained with 1e-6 learning rate for 80 epochs. UniCoRN model is trained with the same settings as fMRI data.

## D Cross-Subject Data Splitting in Practice

We present the pseudo-code of two dataset splitting methods for EEG and fMRI signal. We only consider a bipartite graph $\mathcal{G}_1 = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ instead of a 4-*partite* graph in real practice. For EEG signal, $\mathcal{U} = \{S_i\}_{i=1}^{N}$, $\mathcal{V} = \{T_j\}_{j=1}^{M}$. While for fMRI signal, $\mathcal{U} = \{S_i\}_{i=1}^{N}$, $\mathcal{V} = \{M_k\}_{k=1}^{K}$. $\mathcal{E}$ is the edge between node in $\mathcal{U}$ and node in $\mathcal{V}$. $N, M, K$ indicate the total number of subjects, text segments and stories. We assert $M > N$ for EEG dataset and $K < N$ for fMRI dataset, so $e = (u, v) \in \mathcal{E}$ exists for every $v \in \mathcal{V}$, as each text segment or story is listened by at least one subject. As shown in step 1 of Figure 3, first we pick one edge for each node $v \in \mathcal{V}$ and build a new bipartite graph $\mathcal{G}_2 = (\mathcal{U}, \mathcal{V}, \mathcal{E}')$. Then following step 2, we split graph $\mathcal{G}_2$ by subject $\mathcal{U}$ with the given splitting ratio and form three disjoint graphs $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$. In step 3, we extend each graph $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$ by adding edges without data leakage.

The main difference of splitting methods for EEG and fMRI lies in how $\mathcal{G}_2$ is generated. We always choose the side with fewer nodes in bipartite graph $\mathcal{G}_1$ to generate $\mathcal{G}_2$. Specifically, in Algorithm 1 where we assert $|\mathcal{U}| < |\mathcal{V}|$, the adjacency matrix is initialized as $M \times N$. In Algorithm 2 where $|\mathcal{V}| < |\mathcal{U}|$, the adjacency matrix is initialized as $N \times K$. All assertions are based on real cognitive datasets. One more thing to notice is that in Line 14 of both pseudo-code, the loop indicates extending training set, validation set, and test set respectively. So the names of variable should be alternated in the repeat loop and the displayed part in pseudo-code is a case example of extending training set. We write it in this way for simplicity of expression.

## E Supplementary Proof

**Why a method without brain signal leakage and text stimuli leakage must satisfy cross-subject brain-to-text decoding criterion** Training set $\mathcal{G}_{train}$ without brain signal leakage and text stimuli leakage is formatted in

$$\begin{aligned}\mathcal{G}_{train} = \{&(S_i, M_k, T_{kj}, F_{ijk})| \\ &\forall (S_i', M_k', T_{kj}', F_{ijk}') \in \mathcal{G}_{test}, \\ &S_i \neq S_i', T_{kj} \neq T_{kj}'\} \\ &= \mathcal{S}_{train} \otimes \mathcal{M} \otimes \mathcal{T}_{train} \otimes \mathcal{F}\end{aligned} \tag{21}$$

where $\mathcal{S}_{train} = \{S_i | \forall S'_i \in \mathcal{S}_{test}, S_i \neq S'_i\}, \mathcal{T}_{train} = \{T_{kj} | \forall T'_{kj} \in \mathcal{T}_{test}, T_{kj} \neq T'_{kj}\}$. Since $F_{ijk} \in \mathcal{F}$ indicates brain signal of subject $S_i$ stimulated by text segment $T_{kj}$, and given the definition of operator $\otimes$, $\mathcal{F}$ is determined when $\mathcal{S}$ and $\mathcal{T}$ are specified, which is

$$\begin{aligned} \mathcal{F} &= \{F_{ijk} | i \in I, kj \in J\}, \\ I &= \{i | S_i \in \mathcal{S}_{train}\}, \\ J &= \{kj | T_{kj} \in \mathcal{T}_{train}\}. \end{aligned} \tag{22}$$

$\mathcal{F}$ can also be written as $\mathcal{F} = \{F_{ijk} | i \in I\}$, $I = \{i | \forall j, \forall k, F_{ijk} \notin \mathcal{F}_{test}\}$, which is equal to Definition 3.1.

**Why the proposed splitting method satisfy zero data leakage** Take the splitting method for EEG signal as example, the training set and test set after step 1 and step 2 already satisfy

$$\begin{aligned} \mathcal{G}_{train} = \{&(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, \\ &T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, S_i \neq S'_i, T_{kj} \neq T'_{kj}\} \end{aligned} \tag{23}$$

$$\begin{aligned} \mathcal{G}_{test} = \{&(S_i, M_k, T_{kj}, F_{ijk}) | \forall (S'_i, M'_k, \\ &T'_{kj}, F'_{ijk}) \in \mathcal{G}_{train}, S_i \neq S'_i, T_{kj} \neq T'_{kj}\} \end{aligned} \tag{24}$$

So we only need to prove expanded graph $\mathcal{G}'_{train\_exp}$ and $\mathcal{G}'_{test\_exp}$ satisfy zero data leakage, which is obvious from Equation 13 and 14.

**Why we must discard samples to ensure no data leakage** If $\mathcal{G}_{train} \cup \mathcal{G}_{test} = \mathcal{G}_{\mathcal{D}}$, suppose $\forall (S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_{train}$, $(S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}, S_i \neq S'_i, T_{kj} \neq T'_{kj}$. For $f(\mathcal{E}) = (M_k, T_{kj})$, $f(\mathcal{E}') = (M'_k, T'_{kj})$, $T_{kj} \neq T'_{kj}$:

- If $M_k = M'_k$, then there must exist a subject $S_i = S'_i$ such that he is stimulated by the whole stories.
- If $M_k \neq M'_k$, then there must exist a subject $S_i = S'_i$ such that he is stimulated by two different stories.

As a result, if $\mathcal{G}_{train} \cup \mathcal{G}_{test} = \mathcal{G}_{\mathcal{D}}$, then $\exists (S_i, M_k, T_{kj}, F_{ijk}) \in \mathcal{G}_{train}$, $(S'_i, M'_k, T'_{kj}, F'_{ijk}) \in \mathcal{G}_{test}$, $s.t.\ S_i = S'_i$ or $T_{kj} = T'_{kj}$. Some samples must be discarded to ensure no data leakage.

## F Supplementary Experiment

Results and analysis on experiments on longer training epochs with smaller learning rate is supplemented. If evaluation indicators keep improving as training epochs increase, we believe part of the test set is leaked into training set and the model is overfitting. For fMRI signal, we test five current dataset splitting methods under different training settings. As shown in Table 7, we test two kinds of UniCoRN models. One is UniCoRN with hyper-parameters claimed in the original paper, and the other is UniCoRN* whose encoder is randomly initialized. Besides, UniCoRN* is trained with longer epochs and smaller learning rate. In method (a), (c), (d), due to text stimuli leakage, if we reduce the learning rate and extend training epochs, UniCoRN* performs much better than UniCoRN and its performance keeps rising with longer training epochs. As to method (b) and (e) with no text stimuli leakage, changing training epochs or learning rates makes no obvious difference to model performance. For EEG signal, the conclusion is similar as shown in Table 8. For method (a) and (c) with text stimuli leakage, model performance keeps rising with longer training epochs. For method (d) without text stimuli leakage, both models reach optimal performance after the first few rounds of training epochs.

| Model | Epoch+lr+Method | BLEU-N (%) | | | | ROUGE-1 (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | $N = 1$ | $N = 2$ | $N = 3$ | $N = 4$ | $F$ | $P$ | $R$ |
| UniCoRN | 10+1e-3+(a) | 49.56 | 30.49 | 21.07 | 15.49 | 44.83 | 50.41 | 40.65 |
| | 10+1e-3+(b) | 26.37 | 7.50 | 2.48 | 0.99 | 22.28 | 25.99 | 19.62 |
| | 10+1e-3+(c) | 50.24 | 30.83 | 21.23 | 15.60 | 44.68 | 49.44 | 41.01 |
| | 10+1e-3+(d) | 49.63 | 30.29 | 20.85 | 15.32 | 45.06 | 50.47 | 41.03 |
| | 10+1e-3+(e) | 28.94 | 9.39 | 4.07 | 1.53 | 21.68 | 24.64 | 19.49 |
| UniCoRN* | 20+1e-4+(a) | 50.19 | 34.25 | 25.98 | 21.00 | 46.59 | 50.36 | 43.62 |
| | 30+1e-4+(a) | 55.46 | 40.99 | 32.85 | 27.56 | 52.08 | 55.02 | 49.68 |
| | 20+1e-4+(b) | 25.91 | 8.80 | 3.84 | 1.66 | 20.65 | 27.74 | 16.57 |
| | 30+1e-4+(b) | 25.91 | 8.80 | 3.84 | 1.66 | 20.65 | 27.74 | 16.57 |
| | 20+1e-4+(c) | 72.44 | 60.84 | 53.35 | 47.88 | 70.52 | 74.10 | 67.53 |
| | 30+1e-4+(c) | 72.82 | 61.42 | 53.95 | 48.44 | 71.24 | 74.41 | 68.57 |
| | 20+1e-4+(d) | 65.31 | 51.02 | 42.54 | 36.72 | 62.76 | 67.09 | 59.29 |
| | 30+1e-4+(d) | 66.56 | 53.00 | 44.75 | 39.02 | 63.89 | 67.51 | 60.95 |
| | 20+1e-4+(e) | 32.15 | 12.34 | 5.57 | 2.45 | 24.28 | 30.43 | 20.35 |
| | 30+1e-4+(e) | 32.15 | 12.34 | 5.57 | 2.45 | 24.28 | 30.43 | 20.35 |

Table 7: Generation quality of UniCoRN model for fMRI under different training settings. Here UniCoRN* indicates the encoder of UniCoRN is randomly initialized instead of pre-trained through signal reconstruction task.

| Model | Epoch+lr+Method | BLEU-N (%) | | | | ROUGE-1 (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | $N = 1$ | $N = 2$ | $N = 3$ | $N = 4$ | $F$ | $P$ | $R$ |
| UniCoRN | 50+1e-4+(a) | 58.09 | 49.23 | 43.23 | 38.43 | 63.88 | 61.12 | 67.50 |
| | 80+1e-4+(a) | 60.88 | 50.52 | 43.42 | 37.84 | 65.17 | 61.16 | 70.72 |
| | 50+1e-4+(c) | 52.30 | 42.89 | 36.80 | 32.17 | 57.39 | 51.09 | 67.29 |
| | 80+1e-4+(c) | 60.78 | 55.92 | 53.18 | 51.10 | 84.64 | 63.16 | 71.50 |
| | 50+1e-4+(d) | 22.90 | 7.36 | 2.71 | 0.95 | 17.73 | 19.90 | 17.33 |
| | 80+1e-4+(d) | 22.90 | 7.36 | 2.71 | 0.95 | 17.73 | 19.90 | 17.33 |
| EEG2Text | 50+1e-4+(a) | 51.22 | 33.83 | 22.99 | 16.05 | 46.40 | 46.85 | 46.58 |
| | 80+1e-4+(a) | 63.32 | 52.52 | 45.19 | 39.50 | 65.96 | 64.74 | 68.01 |
| | 50+1e-4+(c) | 53.83 | 38.99 | 29.57 | 23.01 | 53.64 | 54.19 | 53.56 |
| | 80+1e-4+(c) | 65.42 | 57.56 | 52.56 | 48.60 | 73.00 | 69.99 | 77.01 |
| | 50+1e-4+(d) | 23.92 | 8.16 | 3.21 | 1.20 | 20.78 | 19.96 | 23.89 |
| | 80+1e-4+(d) | 23.92 | 8.16 | 3.21 | 1.20 | 20.78 | 19.96 | 23.89 |

Table 8: Generation quality of UniCoRN and EEG2Text model for EEG under different training settings.

**Algorithm 1:** Dataset splitting method for EEG signal

**Initialize:** Bipartite graph $\mathcal{G}_1 = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, $\mathcal{G}_2 = (\mathcal{U}, \mathcal{V}, \mathcal{E}')$ where $\mathcal{U} = \{S_i\}_{i=1}^{N}$ and $\mathcal{V} = \{T_j\}_{j=1}^{M}$, Adjacency matrix $A_1$ of $\mathcal{G}_1$ where $A_1[i][j] = 1$ if node $i$ and node $j$ is connected else $A_1[i][j] = 0$, Adjacency matrix $A_2$ of $\mathcal{G}_2$ where $A_2[i][j] = 0$, Array $C$ where $len(C) = len(\mathcal{U})$ and $C[i] = 0$;

**for** $u \leftarrow U_1$ **to** $U_N$ **do**
  $C_{copy} \leftarrow C$;
  **for** $v \leftarrow A_1[u][0]$ **to** $A_1[u][M]$ **do**
    **if** $v = 0$ **then**
      $C_{copy}[v.index] \leftarrow \infty$;
  $Minimum = \min(C_{copy})$;
  $A_2[u][Minimum.index] \leftarrow 1$;
  $C[Minimum.index] \leftarrow C[Minimum.index] + 1$;   `// Make degree of nodes balanced`
Split by subjects $\mathcal{U}$ according to default ratio;
$\mathcal{G}_2 = \mathcal{G}_{train} \cup \mathcal{G}_{val} \cup \mathcal{G}_{test}$, $\mathcal{U}_{train} \cap \mathcal{U}_{val} \cap \mathcal{U}_{test} = \emptyset$, $\mathcal{V}_{train} \cap \mathcal{V}_{val} \cap \mathcal{V}_{test} = \emptyset$;
**repeat**   `// To three sets respectively, below is for training set`
  **for** $u$ **in** $\mathcal{U}$ **do**
    **for** $v$ **in** $\mathcal{V}$ **do**
      **if** $e = (u, v) \in \mathcal{E}$ **and** $e = (u, v) \notin \mathcal{E}'_{train}$ **and** $u \notin \mathcal{U}_{val} \cup \mathcal{U}_{test}$ **then**
        $\mathcal{E}'_{train} \leftarrow \mathcal{E}'_{train} \cup \{e\}$;
**until** $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$ are all extended;
**return** $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$;

**Algorithm 2:** Dataset splitting method for fMRI signal

**Initialize:** Bipartite graph $\mathcal{G}_1 = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, $\mathcal{G}_2 = (\mathcal{U}, \mathcal{V}, \mathcal{E}')$ where $\mathcal{U} = \{S_i\}_{i=1}^{N}$, $\mathcal{V} = \{M_k\}_{k=1}^{K}$, Adjacency matrix $A_1$ of $\mathcal{G}_1$ where $A_1[i][j] = 1$ if node $i$ and node $j$ is connected else $A_1[i][j] = 0$, Adjacency matrix $A_2$ of $\mathcal{G}_2$ where $A_2[i][j] = 0$, Array $C$ where $len(C) = len(\mathcal{V})$ and $C[i] = 0$;

**for** $v \leftarrow V_1$ **to** $V_K$ **do**
  $C_{copy} \leftarrow C$;
  **for** $u \leftarrow A_1[v][0]$ **to** $A_1[v][K]$ **do**
    **if** $u = 0$ **then**
      $C_{copy}[u.index] \leftarrow \infty$;
  $Minimum = \min(C_{copy})$;
  $A_2[v][Minimum.index] \leftarrow 1$;
  $C[Minimum.index] \leftarrow C[Minimum.index] + 1$;   `// Make degree of nodes balanced`
Split by tasks $\mathcal{V}$ according to default ratio;
$\mathcal{G}_2 = \mathcal{G}_{train} \cup \mathcal{G}_{val} \cup \mathcal{G}_{test}$, $\mathcal{U}_{train} \cap \mathcal{U}_{val} \cap \mathcal{U}_{test} = \emptyset$, $\mathcal{V}_{train} \cap \mathcal{V}_{val} \cap \mathcal{V}_{test} = \emptyset$;
**repeat**   `// To three sets respectively, below is for training set`
  **for** $v$ **in** $\mathcal{V}$ **do**
    **for** $u$ **in** $\mathcal{U}$ **do**
      **if** $e = (u, v) \in \mathcal{E}$ **and** $e = (u, v) \notin \mathcal{E}'_{train}$ **and** $v \notin \mathcal{V}_{val} \cup \mathcal{V}_{test}$ **then**
        $\mathcal{E}'_{train} \leftarrow \mathcal{E}'_{train} \cup \{e\}$;
**until** $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$ are all extended;
**return** $\mathcal{G}_{train}, \mathcal{G}_{val}, \mathcal{G}_{test}$;