GLANCE: GLOBAL–LOCAL ALIGNMENT OF NEURAL AND CONTEXTUAL EMBEDDINGS FOR BRAIN–LANGUAGE SENTENCE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying heard language from neural activity requires models that align distributed, time-varying brain signals with linguistic representations. We introduce GLANCE (Global–Local Alignment of Neural and Contextual Embeddings), a framework for sentence retrieval from stereo-electroencephalography (sEEG) recorded during naturalistic movie viewing. GLANCE models within-lead temporal dynamics using recurrent encoders, combines information across leads with spatial attention, and integrates temporal context into a global sentence representation. A complementary local pathway compares time-resolved neural features with contextual word embeddings from each candidate transcript. Both pathways operate in embedding space, combining whole-sentence similarity with word-level evidence without requiring word-onset supervision or a fixed word-to-time alignment. GLANCE is trained with a multi-positive contrastive objective that accounts for repeated transcripts, lightweight adaptation of frozen text features, and paired full and partially masked neural views. We evaluate the framework on 13 recordings from four Brain Treebank participants, with separate models trained and tested within each participant–movie recording. Under 100-way retrieval, GLANCE achieves 26.2% subject-averaged Recall@10 and exceeds the 10% chance level in every recording. Combining global and local scores generally improves retrieval over global-only scoring across recordings. These findings demonstrate the value of joint sentence-level and word-level matching for retrieving heard sentences from naturalistic intracranial activity.

1 INTRODUCTION

Can neural activity recorded during naturalistic listening identify which sentence a person heard? Recent approaches connect intracranial recordings to pretrained speech and language representations, enabling speech retrieval and reconstruction of perceived language (Evanson et al., 2025; Shams et al., 2025). Naturalistic listening presents a demanding setting for this task: utterances vary in wording, duration, and context, while neural responses unfold across time and recording sites. Brain Treebank (Wang et al., 2024), which pairs stereo-electroencephalography (sEEG) during movie viewing with aligned dialogue transcripts, provides an opportunity to study sentence identification within ongoing conversations and audiovisual context.

We formulate the problem as sentence retrieval. Given the neural activity associated with a heard sentence, a model ranks its transcript among alternative candidates. This formulation evaluates neural–text matching through an explicit candidate set and a known chance level, without requiring text generation. DEEPER and its successor BPR learn cross-modal representations for passage retrieval from scalp EEG during reading, using word-aligned neural features derived from eye-tracking fixations (McGuire & Moshfeghi, 2024; 2025). Our setting instead uses sentence windows from sEEG during movie viewing. SSENSE addresses this setting by encoding individual electrode spectrograms, averaging their representations, and aligning the result with frozen CLIP text embeddings (Liu, 2025). Although its spectrogram encoder captures time–frequency patterns, its final matching score compares one pooled neural representation with one sentence embedding. This leaves an opportunity to retain time-resolved neural features and use them to assess individual candidate words alongside the sentence as a whole.

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

The central question is how to combine these two levels of evidence. A sentence embedding provides a representation of the complete utterance, while contextual word embeddings allow individual words to contribute explicitly to the matching score. BrainMosaic combines semantic-unit prediction with sentence-level supervision to guide language reconstruction (Li et al., 2026), illustrating the relevance of modeling multiple linguistic scales. Here, we investigate their combination directly within retrieval: each candidate is scored both by its overall compatibility with the neural window and by the compatibility of its words with time-resolved neural features. The challenge is to obtain useful word-level evidence without prescribing when each word must align with the neural response.

We introduce **GLANCE (Global–Local Alignment of Neural and Contextual Embeddings)**, a framework for brain–language alignment that combines sentence-level and contextual word-level matching through a shared neural encoder to retrieve heard sentences from sEEG. GLANCE processes superlet time–frequency representations of sEEG, organizing contacts by their recording leads. Lead-specific recurrent networks model temporal activity within each lead, and a spatial Transformer integrates their outputs at each time bin. The resulting sequence supports two pathways. A global pathway summarizes the neural window for sentence-level alignment. A local pathway preserves time-resolved features and compares them with the contextual word embeddings of each candidate transcript. Soft pooling over neural time bins, followed by averaging over candidate words, produces a local score that is combined with the global similarity. This matching mechanism requires neither word-onset supervision nor a fixed word-to-time alignment.

Sentence and word representations are supplied by frozen EmbeddingGemma, with a residual sentence adapter and a learned word projection adapting the cached features for neural matching. We train GLANCE with a symmetric multi-positive contrastive objective that treats repeated transcripts as positives. A second, partially masked neural view is aligned with the same sentence targets and encouraged to agree with the full-view representation. Together, these objectives train the shared encoder to support sentence retrieval using both global alignment and local matching evidence.

We evaluate GLANCE on 13 recordings from four Brain Treebank participants, training and testing a separate model within each participant–movie recording. Under a common 100-way retrieval protocol, mean Recall@10 reaches 26.2% with subjects weighted equally, and performance exceeds the 10% chance level in every recording. Across recordings, combining global and local scores generally improves retrieval over global-only scoring. We further assess the framework through matched baseline comparisons and component-removal ablations examining the local pathway and temporal masking.

Our contributions are:

- GLANCE, a framework for sentence retrieval from sEEG that combines lead-wise spatiotemporal encoding with sentence-level alignment and contextual word-to-time matching, without word-level timing supervision.
- An evaluation across four participants and 13 movie recordings under a common 100-way retrieval protocol, complemented by pathway analyses across recordings and baseline comparisons and component ablations on one recording.

2 Related work

Language decoding and retrieval. Pretrained speech and language representations have enabled several approaches to decoding perceived language from neural activity. EEG-to-text models generate words with a pretrained sequence decoder (Wang & Ji, 2022; Duan et al., 2023), Tang et al. (2023) reconstruct continuous language from fMRI with a language-model prior, and Neuro2Semantic maps intracranial recordings to sentence embeddings and reconstructs text through embedding inversion (Shams et al., 2025). Retrieval approaches instead identify the corresponding stimulus among alternatives: Défossez et al. (2023) align MEG and EEG with pretrained speech representations to retrieve heard speech segments, while Evanson et al. (2025) learn neural–speech alignment from extended naturalistic sEEG recordings. These formulations address related but distinct objectives: reconstructing linguistic content and identifying it within a candidate set. Evaluation also requires care. In the EEG-to-text systems they examined, Jo et al. (2024) found that teacher forcing inflated reported performance and that noise inputs could yield scores comparable to

EEG. Our study uses sentence retrieval to evaluate neural–text compatibility directly, with explicit candidate sets and a defined chance level.

From reading EEG to heard-sentence retrieval. DEEPER and its subsequent BPR formulation use contrastive dual encoders to retrieve text passages from scalp EEG recorded during reading (McGuire & Moshfeghi, 2024; 2025). Their neural inputs comprise sequences of word-level band-power features obtained through eye-tracking fixations. BPR additionally adapts a frozen text encoder with a trainable layer. In our listening setting, the model receives a complete sentence window, and word-level matching operates without fixation-segmented neural inputs. Brain Treebank provides a setting for studying this problem through sEEG recorded during movie viewing and aligned dialogue transcripts (Wang et al., 2024); Neuroprobe builds on the corpus to benchmark decoding of stimulus properties (Zahorodnii et al., 2025). Closest to our retrieval task, SSENSE encodes each electrode's superlet spectrogram with a shared convolutional network, averages electrode embeddings with equal weight, and aligns the result with frozen CLIP text features (Liu, 2025). Its encoder captures time–frequency patterns, but its retrieval score compares one pooled neural representation with one sentence embedding. This leaves open how time-resolved evidence for individual candidate words can contribute alongside sentence-level similarity. We adapt SSENSE, DEEPER, BPR, and the sentence-embedding pathway of BrainMosaic to our protocol as baselines (Section 4.1).

Representations of intracranial activity. The spatial organization of sEEG introduces a further modeling consideration: recording sites are sparse, and their number and placement depend on clinical needs, so there is no shared montage across participants (Chau et al., 2025). BrainBERT learns contextual representations of intracranial time–frequency signals through masked self-supervised training (Wang et al., 2023), while Population Transformer learns to aggregate temporal embeddings across variable electrode ensembles (Chau et al., 2025). Related architectures explicitly separate temporal and spatial processing. The sEEGnificant model applies temporal attention within electrodes before combining information across electrodes using their three-dimensional coordinates (Mentzelopoulos et al., 2024). Du-IN groups channels from selected brain regions into regional representations for masked pretraining and spoken-word classification (Zheng et al., 2024). These approaches establish the value of considering how recording sites are represented and combined. GLANCE uses recording leads as the grouping unit, with temporal modeling within leads and attention across leads, to produce a shared sequence for sentence-level and word-level retrieval.

Global and local cross-modal matching. Sentence-level alignment measures the overall compatibility between a neural recording and a candidate transcript, but does not explicitly score the neural evidence for individual candidate words. VL-JEPA provides a related precedent for predicting text representations from another modality and comparing them in an embedding space (Chen et al., 2026). GLANCE likewise matches neural and textual representations in embedding space, combining whole-sentence similarity with contextual word-level matching. Related work uses multiple linguistic scales for different purposes. BrainMosaic combines semantic-unit prediction with sentence-level supervision to guide language-model reconstruction (Li et al., 2026); our local pathway instead contributes candidate-specific word evidence directly to the retrieval score. The matching mechanism draws on contextual late interaction in ColBERT (Khattab & Zaharia, 2020) and fine-grained cross-modal alignment in FILIP and X-CLIP (Yao et al., 2022; Ma et al., 2022). In GLANCE, contextual candidate word embeddings interact with time-resolved neural embeddings. Because their temporal correspondence is not prescribed, each word's compatibility is pooled softly over time, without word-onset supervision or a discrete word-to-time assignment. GLANCE combines this local compatibility with whole-sentence alignment through a shared neural encoder, allowing word-level supervision to participate in representation learning as well as candidate ranking.

3 METHODS

3.1 RETRIEVING HEARD SENTENCES

When a person hears a sentence during a movie, the neural response unfolds over time and across sEEG leads, each carrying multiple contacts along an electrode trajectory. Recovering which sentence was heard therefore poses two problems at once: combining activity distributed over leads and time, and distinguishing transcripts that may share words or meaning. We address these with

a lead-structured neural encoder and complementary sentence-level and word-level matching pathways (Fig. 1).

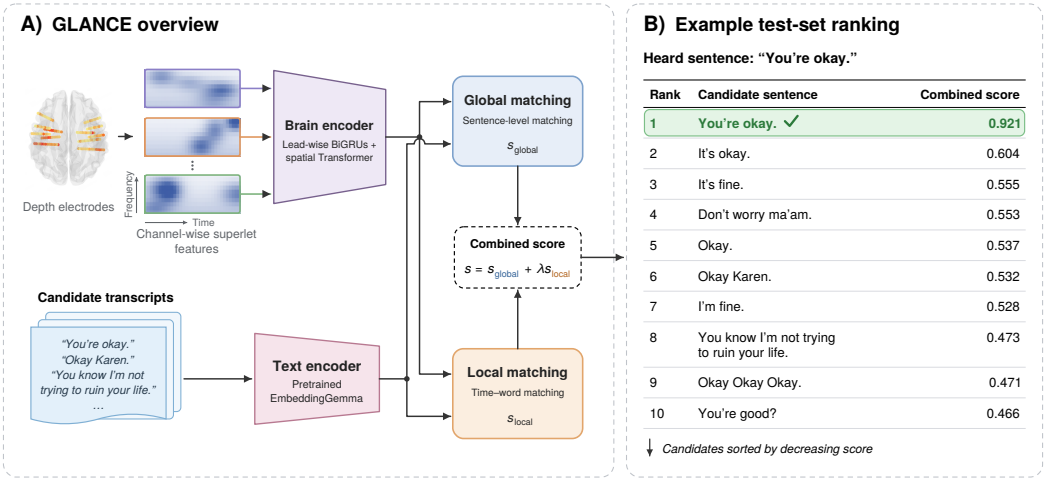


Figure 1: **Retrieving a heard sentence from sEEG.** A lead-structured brain encoder and frozen transcript encoder feed complementary global and local matching pathways. Their scores rank candidate transcripts.

For sentence i, let $\mathbf{X}_i \in \mathbb{R}^{C\times F\times T}$ denote sEEG time–frequency features from C channels, F frequencies, and T time bins spanning sentence onset through one second after its offset; let s_i denote the heard sentence's transcript. Given $\mathbf{X}_i$, the model scores and ranks candidate transcripts s_j. Retrieval provides a controlled test with an explicit candidate set and chance baseline. The model receives sEEG features and candidate transcript embeddings, but no movie audio.

3.2 Lead-structured neural encoder

Contacts on the same lead are spatially close and tend to covary, whereas different leads often sample different regions and differ in number and placement across participants. We therefore use leads as the encoder's spatial units (Fig. 2, left). We group channels by lead labels and feed each lead's channel–frequency features over time to a lead-specific bidirectional GRU. A learned lead embedding is added to each output state. At each time bin, a spatial Transformer attends over the lead states and a learned spatial [CLS] token. Its [CLS] output is the shared neural representation $\mathbf{u}_{i,t} \in \mathbb{R}^{256}$; write $\mathbf{U}_i = (\mathbf{u}_{i,1}, \ldots, \mathbf{u}_{i,T})$.

Two pathways branch from $\mathbf{U}_i$ (Fig. 2, right). The *global pathway* asks which sentence the window resembles as a whole. It prepends a learned temporal [CLS] token, adds learned positional embeddings, applies a temporal Transformer, and projects its [CLS] output to a normalized brain vector $\mathbf{b}_i \in \mathbb{R}^{768}$. The *local pathway* preserves time-resolved evidence: a temporal convolution, GELU, and linear projection produce normalized vectors $\mathbf{p}_{i,t} \in \mathbb{R}^{128}$. Both pathways share the lead and spatial encoder.

3.3 Transcript representations

Frozen EmbeddingGemma supplies a sentence embedding and contextual word embeddings for each candidate transcript; these features are precomputed and cached. A trainable residual adapter maps the sentence embedding to a normalized vector $\widetilde{\mathbf{y}}_j \in \mathbb{R}^{768}$. Its output layer is initialized to zero, so the adapter initially preserves the normalized pretrained embedding. A separate trainable linear projection maps each contextual word embedding to a normalized vector $\mathbf{w}_{j,k} \in \mathbb{R}^{128}$ for word k of candidate j.

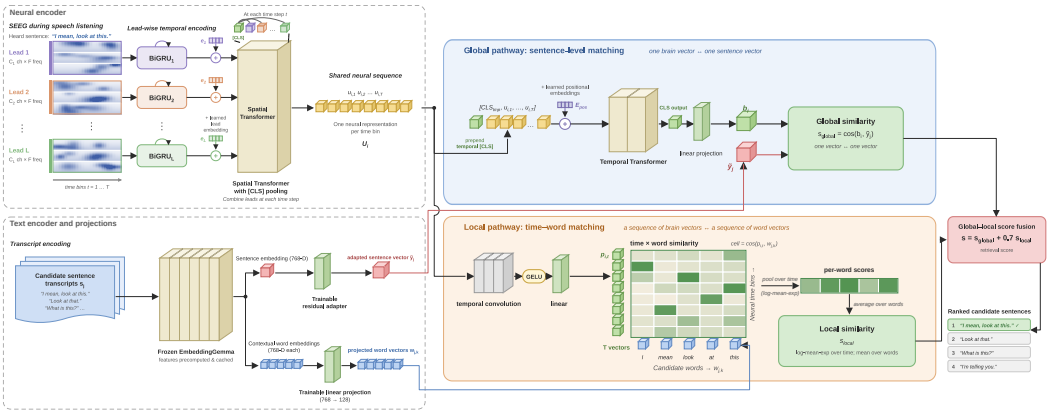


Figure 2: **Neural encoding and global–local matching.** Lead-wise temporal encoding and spatial pooling produce one neural vector per time bin. The global pathway compares a summary of the neural window with a candidate sentence; the local pathway compares time-bin vectors with contextual candidate-word vectors before pooling. Their scores are combined for retrieval.

3.4 Global–Local Matching

All vectors used for matching have unit norm. The global score is $s_{ij}^{\text{global}} = \langle \mathbf{b}_i, \widetilde{\mathbf{y}}_j \rangle$. Locally, we compare each candidate word with every recorded time bin using $r_{ijtk} = \langle \mathbf{p}_{i,t}, \mathbf{w}_{j,k} \rangle$. For the set $\mathcal{T}_i$ of recorded bins and M_j words in candidate j, log-mean-exp pooling over time followed by averaging over words gives $a_{ijk} = \eta \log\big(|\mathcal{T}_i|^{-1} \sum_{t \in \mathcal{T}_i} e^{r_{ijtk}/\eta}\big)$ and $s_{ij}^{\text{local}} = M_j^{-1} \sum_{k=1}^{M_j} a_{ijk}$, with $\eta = 0.10$. This score requires neither word-onset labels nor a fixed word-to-time alignment. The final retrieval score is $s_{ij} = s_{ij}^{\text{global}} + \lambda s_{ij}^{\text{local}}$, with $\lambda = 0.7$. Score-level fusion also lets us evaluate the two pathways separately.

3.5 Training and Evaluation

Dialogue contains repeated lines, so trials with identical transcripts should be positives for one another. We use symmetric multi-positive InfoNCE. In a minibatch of size B, $P(i)$ contains the indices of all transcripts matching s_i after Unicode and whitespace normalization. For a brain–transcript score matrix Z, define

$$\ell_i(Z) = \log \sum_{j=1}^{B} e^{Z_{ij}/\tau} - \log \sum_{j \in P(i)} e^{Z_{ij}/\tau},$$
$$\mathcal{L}_{\text{sym}}(Z) = \frac{1}{2B} \sum_{i=1}^{B} \big[\ell_i(Z) + \ell_i(Z^\top)\big], \quad \tau = 0.07. \tag{1}$$

We score the full window with the combined score matrix S. During training, we also encode a second view in which a contiguous four-bin block is replaced by learned, lead-specific mask tokens for 50% of examples. This view gives a global brain vector $\widehat{\mathbf{b}}_i$ and a global-only score matrix $\widehat{S}^{\text{global}}$ against the same adapted sentence vectors. The objective is

$$\mathcal{L} = \tfrac{1}{2}\mathcal{L}_{\text{sym}}(S) + \tfrac{1}{2}\mathcal{L}_{\text{sym}}(\widehat{S}^{\text{global}}) + \frac{0.1}{B} \sum_{i=1}^{B} \big[1 - \langle \widehat{\mathbf{b}}_i, \text{sg}(\mathbf{b}_i) \rangle\big], \tag{2}$$

where sg stops gradients through the full-view target. The local pathway learns from full windows; masking is applied only in training. We optimize the brain network, sentence adapter, and word projector with AdamW while keeping EmbeddingGemma frozen, and select the model by validation Recall@10 of the combined score.

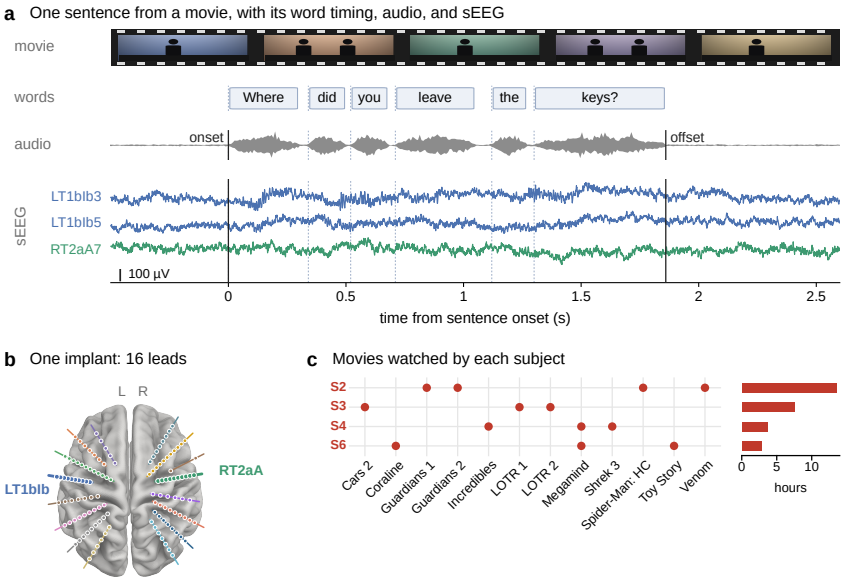


Figure 3: Brain Treebank. (a) One sentence from a movie, with its hand-corrected word onsets and offsets, the audio, and sEEG from three contacts on two leads. (b) Sample subject with 16 leads implanted (c) Movies watched by each subject subjects and their total viewing time (Wang et al., 2024); red marks the subjects and the 13 movie viewings used in our experiments.

At test time, each query is ranked against a deterministic 100-candidate bank comprising its correct transcript identity and 99 distinct distractor identities from the same held-out split. We use 30 banks per query and report Recall@1, Recall@10, and mean reciprocal rank, averaged over queries and banks, for global, local, and combined scores. Distinct identities prevent repeated lines from appearing twice in one bank.

4 Experiments and Results

We evaluate whether the model can retrieve heard sentences from sEEG across multiple participants and movies, how it compares with prior approaches, and what the global and local pathways contribute.

4.1 Setup

Data. We use Brain Treebank (Wang et al., 2024), a public corpus of stereo-EEG recordings collected from patients with epilepsy while they watched feature films during clinical monitoring (Figure 3). The corpus provides manually corrected dialogue transcripts and word-level timing annotations. We study 13 recordings from four subjects (2, 3, 4, and 6), with multiple movies per subject. Each retained sentence of at most 4 s gives one trial: the sEEG from sentence onset to 1 s after offset, converted to superlet magnitudes (Moca et al., 2021) at 50 frequencies between 1 and 200 Hz and 20 time bins per second. Contacts are grouped by electrode lead. Per-recording statistics and full preprocessing details are given in Appendix A.

Experimental setup. We train and test a separate model for each of the 13 recordings, so each model keeps the contacts and leads of its subject. We do not test transfer to new subjects or movies. Within a recording, trials are split at random into 70% training, 10% validation, and 20% test, with repeated lines kept in the same split, and we repeat this for three seeds. Text targets are frozen EmbeddingGemma embeddings of each sentence and of its words (Vera et al., 2025). Models are trained with AdamW, and we keep the checkpoint with the best validation R@10 for the combined score.

Because recordings differ in the number of available test transcripts, we use 100-way retrieval with sampled candidate sets to keep the number of candidates and chance levels consistent across recordings. Each test query ranks its true sentence against 99 other distinct sentences from the test split, over 30 fixed candidate sets per query. Chance is therefore 1% for R@1, 10% for R@10, and approximately 0.052 for MRR. Unless noted, results use the combined score S and are averaged over seeds. Appendix D lists all hyperparameters.

Baselines. We compare four prior brain-to-language methods and a linear reference on Subject 2, Spider-Man: Homecoming, using the same data splits, candidate sets, and retrieval metrics. SSENSE (Liu, 2025) retrieves sentences from Brain Treebank sEEG by encoding each electrode separately, averaging the electrode embeddings, and aligning the result with a frozen CLIP text encoder. DEEPER and its successor BPR (McGuire & Moshfeghi, 2024; 2025) are dual encoders for passage retrieval from EEG. BrainMosaic (Li et al., 2026) predicts concept words and a sentence embedding from EEG or sEEG; we use its sentence embedding to rank candidate transcripts. Ridge regression from neural features to sentence embeddings provides a linear reference. Appendix E describes the implementations and adaptations to our retrieval task.

4.2 Results

Results across subjects and movies. Table 2 reports GLANCE's performance across 13 recordings from four subjects. Mean combined R@10 exceeds the 10% chance level in every recording, ranging from 22.5% for Subject 6 watching Coraline to 34.4% for Subject 2 watching Spider-Man: Homecoming. Averaging movies within each subject and then weighting subjects equally gives 26.2% R@10, 6.0% R@1, and 0.131 MRR. Performance also varies between participants watching the same movie: Megamind yields 26.3% R@10 for Subject 4 and 22.8% for Subject 6. Thus, retrieval remains above chance across the evaluated participants and movies, with variation in performance under recording-specific training.

Comparison with prior models. Table 1 compares GLANCE with SSENSE, DEEPER, BPR, BrainMosaic, and ridge regression on Subject 2 watching *Spider-Man: Homecoming*, the recording with the most retained trials, using the same data splits, candidate sets, and retrieval metrics for every model. GLANCE achieves the highest performance on all three metrics, reaching 10.0% R@1, 34.4% R@10, and 0.181 MRR. BPR is the strongest baseline, followed by SSENSE, which is the closest comparison in task and recording modality, using electrode-wise spectrogram encoding and mean pooling for sentence retrieval from sEEG. DEEPER and the sentence-embedding variant of BrainMosaic were originally developed for EEG passage retrieval during reading and for semantic-unit decoding and sentence reconstruction, respectively, and perform below SSENSE under this protocol. Ridge regression also exceeds chance at R@10, indicating that a linear mapping from time-averaged neural features captures some retrieval-relevant information. GLANCE's stronger performance across all three metrics supports the effectiveness of its learned neural–text alignment under this protocol. Baseline implementations and adaptations are detailed in Appendix E.

Table 1: Sentence retrieval on Subject 2, *Spider-Man: Homecoming*, using 100-way candidate sets. Values are mean ± SD over three seeds; recall is reported in %. Chance performance is 1% R@1, 10% R@10, and 0.052 MRR.

Model	R@1	R@10	MRR
BrainMosaic (Li et al., 2026)	2.27 ± 0.99	19.03 ± 0.79	0.086 ± 0.01
BPR (McGuire & Moshfeghi, 2025)	5.64 ± 3.04	28.29 ± 2.14	0.137 ± 0.02
Ridge regression	3.64 ± 0.09	16.68 ± 0.03	0.097 ± 5e-4
DEEPER (McGuire & Moshfeghi, 2024)	3.031 ± 0.41	21.18 ± 1.48	0.098 ± 4e-3
SSENSE (Liu, 2025)	2.58 ± 0.32	26.26 ± 1.0	0.105 ± 4e-3
GLANCE	**10.0** ± 0.8	**34.4** ± 1.9	**0.181** ± 7e-3

Ablation. We examine two design choices: matching at both the sentence and word levels, and training with a partially masked second view. We first evaluate the contribution of each matching score within the full model, then retrain variants to assess the effects of removing the local pathway or temporal masking.

Table 2: 100-way sentence retrieval, with one model per recording. Mean ± sample SD across available runs for 3 seeds. Recall is reported in %. R@1 and MRR use the combined score; local and global R@10 evaluate the two pathways of the same trained model separately.

			R@10			
Subj.	Movie	R@1	Local	Global	Combined	MRR
2	Spider-Man: Homecoming	10.0 ± 0.8	27.0 ± 2.1	33.2 ± 1.6	34.4 ± 1.9	0.181 ± 7e-3
2	Venom	6.2 ± 0.4	17.5 ± 3.0	25.9 ± 1.4	26.0 ± 2.1	0.135 ± 6e-3
2	Guardians 1	5.1 ± 2.0	20.8 ± 3.6	24.7 ± 1.0	27.9 ± 2.1	0.126 ± 2e-2
2	Guardians 2	5.4 ± 1.3	20.9 ± 1.4	26.3 ± 2.3	28.5 ± 0.8	0.130 ± 1e-2
3	Cars 2	6.2 ± 0.5	20.5 ± 1.6	26.9 ± 1.1	28.6 ± 1.1	0.136 ± 7e-3
3	Lord of the Rings 1	5.8 ± 2.5	19.0 ± 1.3	22.6 ± 1.7	22.9 ± 4.0	0.122 ± 3e-2
3	Lord of the Rings 2	4.8 ± 1.3	17.3 ± 2.7	25.6 ± 1.9	25.5 ± 2.0	0.123 ± 1e-2
4	Shrek the Third	5.9 ± 1.4	16.9 ± 1.1	22.5 ± 1.6	22.8 ± 2.1	0.122 ± 1e-2
4	The Incredibles	7.0 ± 2.4	22.6 ± 2.7	26.7 ± 6.1	29.5 ± 5.3	0.152 ± 3e-2
4	Megamind	6.6 ± 1.8	20.0 ± 1.6	25.4 ± 2.1	26.3 ± 2.3	0.138 ± 2e-2
6	Coraline	3.4 ± 2.0	13.3 ± 1.6	23.3 ± 1.2	22.5 ± 1.9	0.105 ± 2e-2
6	Megamind	5.6 ± 1.5	20.6 ± 1.1	21.2 ± 4.0	22.8 ± 3.6	0.117 ± 2e-2
6	Toy Story	6.1 ± 1.6	19.3 ± 0.3	25.4 ± 3.2	26.3 ± 1.8	0.131 ± 1e-2
Mean over subjects		6.0	19.5	25.2	26.2	0.131
Chance		1.0	10.0	10.0	10.0	0.052

Table 3: Ablation study

Model	R@1	R@10	MRR
Full model	10.0 ± 0.8	34.4 ± 1.9	0.181 ± 7e-3
No Local	9.3 ± 0.7	30.8 ± 1.0	0.168 ± 7e-3
No Masking	9.4 ± 1.73	31.6 ± 2.45	0.171 ± 2e-2

Table 2 reports the two matching pathways separately for each trained model. Scored alone, the local pathway is weaker, with subject-averaged R@10 of 19.5% compared with 25.2% for the global pathway. Yet adding it to the global score raises R@10 to 26.2%. The combined score improves on global-only scoring in 11 of 13 recordings, with decreases of less than one percentage point in the remaining two. Thus, word-level matching provides useful complementary evidence despite its lower standalone performance.

Table 3 reports retrained variants on Subject 2 watching Spider-Man: Homecoming, averaged over three seeds. Removing the local pathway reduces R@10 from 34.4% to 30.8%, a decrease of 3.6 percentage points. Disabling temporal masking reduces R@10 to 31.6%, a decrease of 2.8 percentage points. Both variants also have lower mean R@1 and MRR than the full model. These comparisons support the contributions of the local pathway and temporal masking to retrieval performance in this setting.

Analysis of retrieval errors. We examine incorrect top-ranked predictions on the test trials of Subject 2 watching Spider-Man: Homecoming. For each error, we measure the selected transcript's percentile within the candidate comparison pool on several properties (Figure 4). Random selection would yield an average near the 50th percentile. Incorrect predictions fall on average at the 61st percentile of text similarity to the heard sentence, measured by cosine similarity between frozen EmbeddingGemma embeddings (95% bootstrap CI: 59–63). The average remains at the 61st percentile when the comparison is restricted to candidates occurring more than 10 minutes from the query in the film, suggesting that nearby temporal context alone does not explain the preference. Duration similarity, temporal proximity, and shared speaker identity show smaller departures from the random reference, at the 55th, 52nd, and 51st percentiles, respectively. Thus, even when retrieval fails, the selected transcripts tend to be more similar to the target in text-embedding space than randomly selected distractors. To complement this analysis of individual errors, Appendix C

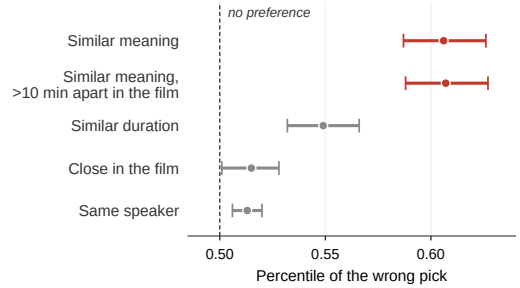


Figure 4: What the model's wrong picks share with the heard sentence (Subject 2, test set). When the top-ranked of 100 candidates is wrong, each row gives its percentile among the other 99 candidates on one property; 0.5 means no preference. Meaning is the cosine between frozen EmbeddingGemma embeddings. The second row counts only candidates heard more than 10 minutes apart in the film. Bars are 95% bootstrap intervals over sentences.

examines alignment between brain and text representations averaged within text-defined sentence groups (Figure 6) and retrieval performance under neural feature-window truncation (Figure 5).

5 Discussion

This work investigates whether time–frequency representations of intracranial activity can support retrieval of sentences heard during naturalistic movie viewing. GLANCE achieves above-chance retrieval in every evaluated recording, providing evidence across participants and movies that superlet features contain information useful for distinguishing heard transcripts from alternative candidates. In the baseline comparison, GLANCE outperforms the evaluated neural models and a linear reference, supporting the effectiveness of its learned neural–text alignment. Together, these findings support sentence retrieval in embedding space as a useful approach to studying the relationship between intracranial activity and heard language within ongoing audiovisual experience. Analyses of the learned embedding space point in the same direction: the model's wrong choices tend to be sentences with similar meaning, brain and text representations agree at the level of sentence types, and retrieval improves when longer neural feature windows are used.

GLANCE combines spatiotemporal encoding with matching at two linguistic scales. The global pathway evaluates the candidate sentence as a whole, while the local pathway allows its contextual words to contribute directly to the retrieval score without requiring word-aligned neural inputs. Combining these scores improves retrieval over global-only scoring in most recordings, suggesting that word-level compatibility supplies useful evidence beyond the global sentence match. Component-removal experiments on the baseline-comparison recording provide additional support: the full model outperforms variants trained without the local pathway or without temporal masking. These results support the contributions of both word-level matching and masked-view training in the evaluated setting.

GLANCE is evaluated with recording-specific training, so transfer to unseen participants or movies remains untested. The current evaluation also assumes known sentence boundaries and ranks a provided set of candidate transcripts. Future work could investigate joint training across recordings and adaptation across electrode layouts to reduce recording-specific supervision requirements. Extending the framework to continuous neural streams and larger transcript collections would assess sentence identification under less constrained retrieval conditions. Another direction is to extend retrieval to paraphrased or multilingual candidate transcripts, testing whether the learned alignment can identify linguistic content beyond the wording of the heard sentence.

ETHICS STATEMENT

This work analyzes the publicly released Brain Treebank dataset. The original recordings were collected from patients with epilepsy at Boston Children's Hospital under approval from the Boston

Children's Hospital/Harvard institutional review board, with informed consent. The released recordings are anonymized, and we use them in accordance with the dataset's CC BY 4.0 license. We do not collect new data or report participant identities. We acknowledge the ethical considerations associated with research involving human intracranial recordings and follow the dataset's usage and privacy requirements.

REPRODUCIBILITY STATEMENT

Code for model implementation, training, and evaluation is available at `https://github.com/anonymresearcher027/GLANCE`.

DISCLOSURE OF AI USE

We used generative AI tools to assist with refining methodological ideas, developing and debugging code, and discussing interpretations of experimental results. We also used these tools to summarize and identify relevant literature, draft and edit manuscript text, format tables, and create and refine scientific figures. The authors take responsibility for the final content of this work, including all AI-assisted text, code, figures, and scientific claims.

REFERENCES

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations*, 2025.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, 2026.

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

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Chang, Yu-Cheng Wang, and Chin-Teng Lin. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Linnea Evanson, Mingfang Zhang, Hubert Banville, Saarang Panchavati, Pierre Bourdillon, and Jean-Rémi King. From minutes to days: Scaling intracranial speech decoding with supervised pretraining. *arXiv preprint arXiv:2512.15830*, 2025.

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

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.

Jiahe Li, Junru Chen, Fanqi Shen, Jialan Yang, Jada Li, Zhizhang Yuan, Baowen Cheng, Meng Li, and Yang Yang. Assembling the mind's mosaic: Towards EEG semantic intent decoding. In *International Conference on Learning Representations (ICLR)*, 2026.

Yijun Liu. sEEG-based encoding for sentence retrieval: A contrastive learning approach to brain-language alignment. *arXiv preprint arXiv:2504.14468*, 2025.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM International Conference on Multimedia*, 2022.

Niall McGuire and Yashar Moshfeghi. DEEPER: Dense electroencephalography passage retrieval. *arXiv preprint arXiv:2412.06695v1*, 2024.

Niall McGuire and Yashar Moshfeghi. Towards brain passage retrieval: An investigation of EEG query representations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. doi: 10.1145/3726302.3730097.

Georgios Mentzelopoulos, Evangelos Chatzipantazis, Ashwin G. Ramayya, Michelle J. Hedlund, Vivek P. Buch, Kostas Daniilidis, Konrad P. Kording, and Flavia Vitale. Neural decoding from stereotactic EEG: Accounting for electrode variability across subjects. In *Advances in Neural Information Processing Systems*, 2024.

Vasile V. Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C. Mureşan. Time-frequency super-resolution with superlets. *Nature Communications*, 12:337, 2021.

Siavash Shams, Richard Antonello, Gavin Mischler, Stephan Bickel, Ashesh Mehta, and Nima Mesgarani. Neuro2Semantic: A transfer learning framework for semantic reconstruction of continuous language from human intracranial EEG. In *Interspeech*, 2025.

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

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations (ICLR)*, 2023.

Christopher Wang, Adam Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R Madsen, Scellig Stone, Gabriel Kreiman, et al. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. *Advances in Neural Information Processing Systems*, 37:96505–96540, 2024.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *AAAI Conference on Artificial Intelligence*, 2022.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

Andrii Zahorodnii, Christopher Wang, Bennett Stankovits, Charikleia Moraitaki, Geeling Chau, Andrei Barbu, Boris Katz, and Ila R. Fiete. Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli. *arXiv preprint arXiv:2509.21671*, 2025.

Hui Zheng, Hai-Teng Wang, Wei-Bang Jiang, Zhong-Tao Chen, Li He, Pei-Yang Lin, Peng-Hu Wei, Guo-Guang Zhao, and Yun-Zhe Liu. Du-IN: Discrete units-guided mask modeling for decoding speech from intracranial neural signals. In *Advances in Neural Information Processing Systems*, 2024.

A Dataset details

We describe the Brain Treebank data (Wang et al., 2024) used in our experiments, together with our preprocessing, sample construction, and data splits.

A.1 BRAIN TREEBANK RECORDINGS

Participants. We use 13 participant–movie recordings from Subjects 2, 3, 4, and 6 of Brain Treebank, comprising one male and three female participants aged 9–18 years. These participants were patients with epilepsy at Boston Children's Hospital who underwent depth-electrode implantation for clinical seizure localization. Electrode type, number, and placement were determined solely by clinical needs. The original data collection was approved by the Boston Children's Hospital/Harvard institutional review board and conducted with informed consent; the released recordings are anonymized.

Stimuli and viewing. Our evaluation includes four movies for Subject 2 and three each for Subjects 3, 4, and 6. Participants chose which movies to watch. Movies were re-encoded at 23.976 frames per second and presented using a custom MATLAB player on a 15.4-inch laptop positioned 60–100 cm from the participant. Participants could adjust the volume and pause playback; the experimenter also paused the movie during interruptions or when the participant became distracted. Participants did not speak during viewing and heard no speech other than the movie dialogue.

Acquisition and synchronization. Each depth probe (Ad-Tech) carries 6 to 16 contacts, 0.8 mm in diameter and 2 mm long, spaced 1.5 mm apart, recording intracranial field potentials. Subjects had 12 to 18 probes and 106 to 246 contacts. Signals were recorded with XLTEK and BioLogic hardware at 2,048 Hz, and data recorded during or right after seizures were discarded. During the movie, a trigger box sent a 15-ms, 80-mV pulse to a separate recording channel every 100 ms, logged with both wall-clock and movie time; bursts of 10, 8, 9, and 11 pulses marked movie start, pause, resume, and end. An automated tool used these triggers to align movie and neural data.

Annotations and localization. Transcripts of each movie's audio were corrected by hand. Word onsets were annotated manually on spectrograms of the audio track, word offsets were corrected, and every sentence was parsed into Universal Dependencies with manual correction of part-of-speech tags and dependency relations. Speaker identity and scene labels were also annotated. Contacts were localized by co-registering post-operative imaging to the pre-operative MRI and assigned to Desikan–Killiany regions. The release flags contacts with long periods of static signal as corrupted (114 contacts in total). Brain Treebank is released under CC BY 4.0.

A.2 PREPROCESSING

Each subject–movie recording is processed on its own.

Contacts. We exclude contacts flagged as corrupted in the release and contacts that stay noisy after filtering. Contacts are grouped by lead using the electrode labels and localization files of the release. Table 5 lists the leads of each Subject.

Signal. Starting from the 2,048-Hz recordings, we remove line noise with zero-phase notch filters at the 60-Hz harmonics from 60 to 480 Hz, apply zero-phase Butterworth high-pass (1 Hz) and low-pass (500 Hz) filters, and downsample by two to 1,024 Hz. At each time point we then estimate a common signal for each lead as the median over its valid contacts and subtract it from each of those contacts; corrupted contacts do not enter the median.

Sentences and time–frequency features. We drop sentences longer than 4 s. For each remaining sentence we extract an 8.5-s window from 1.5 s before to 7 s after sentence onset and compute adaptive superlet magnitudes (Moca et al., 2021) over it, with 50 linearly spaced frequencies from 1 to 200 Hz, orders from 2 to 12, and a base of 3 cycles, using zero extension at the window edges. The magnitudes are downsampled in time by 50, to almost 20Hz, and stored as 32-bit floats.

Model window. We crop the time–frequency map from sentence onset to 1 s after sentence offset. Since sentences are at most 4 s long, the crop is at most 5 s, and we right-pad it with zeros to a fixed 5-s grid of 103 bins. A binary mask marks the valid bins. Each sample is therefore a $C \times 50 \times 103$ tensor with a 103-bin mask.

Normalization. For each split, we estimate a mean and a standard deviation for every contact–frequency pair from the valid bins of the training sentences only, with a floor of 10^{-6} on the variance, and apply them unchanged to the training, validation, and test sets. Padded bins stay at zero.

A.3 Recordings used and lead structure

Table 4 summarizes the 13 participant–movie recordings, including sentence exclusions, retained samples, and lead counts. Table 5 lists the lead labels by participant and hemisphere. Retained contacts are grouped by lead, with each group processed by a separate bidirectional GRU. The number of retained contacts varies across recordings.

Table 4: The 13 recordings used in our experiments, from transcribed sentences to model samples. Sentences in the movie are from Brain Treebank; the other columns are after our preprocessing.

Subj.	Movie	Sentences	Over 4 s	Other	Samples	Leads
2	Spider-Man: Homecoming	2,196	78	14	2,104	16
	Venom	1,379	93	49	1,237	16
	Guardians of the Galaxy	1,174	141	110	923	16
	Guardians Vol. 2	1,290	177	35	1,078	16
3	Cars 2	2,051	100	1	1,950	12
	Lord of the Rings 1	1,515	264	318	933	12
	Lord of the Rings 2	1,717	281	3	1,433	12
4	Shrek the Third	1,064	146	3	915	15
	The Incredibles	1,516	128	372	1,016	15
	Megamind	1,472	135	9	1,328	15
6	Coraline	998	83	443	472	12
	Megamind	1,472	135	653	684	12
	Toy Story	1,320	52	5	1,263	12

A.4 Language targets and word timing

Each sentence is encoded with frozen EmbeddingGemma (`google/embeddinggemma-300m`) (Vera et al., 2025), using the sentence-similarity prompt "`task: sentence similarity | query:`" followed by the transcript. This produces one ℓ_2-normalized, 768-dimensional vector per sentence. For each split, we subtract the mean of the training-sentence embeddings from all sentence embeddings and renormalize them. Contextual word embeddings are extracted using the same encoder. Padded word positions are masked and excluded from local-score aggregation.

Sentence identities are defined by Unicode NFC normalization and whitespace collapsing, with case and punctuation preserved. During training, all examples in a minibatch with the same identity are treated as positives in both directions of the contrastive loss.

For analysis, the current word-target pipeline uses lexical-word masks only and does not retain or use released word onset or offset timestamps. These lexical-word masks are not used to supervise GLANCE's local matching or prescribe word-to-time assignments.

A.5 Splits and evaluation

For each recording and each of three seeds, trials are randomly partitioned into 70% training, 10% validation, and 20% test, with all occurrences of the same transcript identity assigned to the same split. The exception is Subject 6 watching Coraline, where limited data required a 50/25/25 split to retain enough distinct test identities for 100-way retrieval. Normalization statistics and embedding means are computed from the training split only. For GLANCE, we select the checkpoint with the highest validation R@10 of the combined score. Each test query is evaluated against 30 candidate sets generated with fixed seeds, each containing its correct transcript identity and 99 distinct distractor identities sampled from the test split. Within each recording, all compared models, including baselines and ablation variants, use the same splits and candidate sets.

Table 5: Leads of Subjects 2, 3, 4, and 6, grouped by hemisphere. Labels follow the Brain Treebank release.

Subject	Hemisphere	Leads
2	Left	LT1bIb, LT1c, LT2aA, LT2b, LT3a, LT3bHa, LT3cHb, LT3d
2	Right	RT1aIa, RT1bIb, RT1c, RT2aA, RT2b, RT2c, RT3aHa, RT3bHb
3	Left	F2Ia, F3c, F3b, F3aOF, F3d, T1aIc, P2a, T1b, T1cIe, P2b, O1aIb, O1bId
4	Left	LF1aCaOF, LF1bCb, LF1cCc, LF3aOFa, LF3bIa, LF3cIc, LT1aIb, LT1bId, LT1cCd, LT2aA, LT2bHb, LT3Ha
4	Right	RF1aCaOF, RF1bCb, RF1cCc
6	Left	F3aOFa, F3bOFb, F3cOFc, T2A, T1Id, F2aCa, F2c, F2dCb, F3dIb, P2Ie, F2bIa, F2eIc

Global, local, and combined scores are evaluated using the same candidate sets and the trained sentence adapter and word projection, as applicable. Candidates are ranked by descending score. Recall@K is the fraction of query–candidate-set pairs in which the correct transcript ranks among the top K; mean reciprocal rank averages the reciprocal of its rank over the same pairs.

B Additional methodological details

This section supplements the model and objective in Section 3. Signal preprocessing, window construction, and train-split normalization are described in Appendix A.2; the notation below follows the main Methods.

B.1 Lead inputs and variable-length windows

Contacts are grouped by the lead labels and ordered by contact number. At each recorded time bin, the channel–frequency features of lead ℓ form one vector of length $C_\ell F$, where C_ℓ is the number of retained contacts on that lead. Each lead has a separate bidirectional GRU with 128 hidden units in each direction. A learned lead embedding is added to each 256-dimensional GRU output before spatial attention. The number of leads and contacts per lead follow each recording's retained channels; lead-specific parameters are learned for that recording.

The preprocessing in Appendix A.2 pads the model input to a common duration. We retain the observed length of each window and use packed GRU sequences so that padding is not processed as recorded activity. The spatial [CLS] output is set to zero at padded bins; these bins are masked as keys and values in the global temporal Transformer and excluded from local score pooling. Artificial masking during training does not alter which bins are observed.

B.2 Neural and transcript encoders

The spatial Transformer has one layer, four attention heads, a 512-dimensional feed-forward sublayer, and a learned spatial [CLS] token. Applied across leads independently at each time bin, its [CLS] output is $\mathbf{u}_{i,t} \in \mathbb{R}^{256}$. The global branch prepends a separate learned temporal [CLS] token to $\mathbf{U}_i$, adds learned positional embeddings, and applies two temporal Transformer layers with eight heads and 1,024-dimensional feed-forward sublayers. A linear map projects the temporal [CLS] output from 256 to 768 dimensions to form the normalized $\mathbf{b}_i$. Both Transformers use dropout of 0.1.

The local branch takes $\mathbf{U}_i$ directly. Its temporal convolution has kernel width five, maps 256 to 128 channels, and uses two-bin padding on each side to preserve the time axis. GELU and a bias-free $128 \to 128$ linear map yield normalized $\mathbf{p}_{i,t}$. A separate bias-free $768 \to 128$ linear map produces normalized contextual word vectors $\mathbf{w}_{j,k}$. Padded word positions are excluded when averaging over candidate words.

Frozen EmbeddingGemma provides cached, centered 768-dimensional sentence and contextual word features, as described in Appendix A. For a centered sentence feature $\mathbf{y}_j$, the trainable residual adapter computes $\mathbf{z}_j = \mathrm{norm}(\mathbf{y}_j)$ and $\widetilde{\mathbf{y}}_j = \mathrm{norm}(\mathbf{z}_j + 0.05\, f_{64}(\mathbf{z}_j))$, where f_{64} is a $768 \to 64$ linear layer, GELU, and a $64 \to 768$ linear layer. The final layer's weights and bias are initialized to zero, so the adapter initially returns $\mathbf{z}_j$.

B.3 Masked-view training and optimization

Each training example has a full view and a second view encoded by the same network. In the second view, temporal masking is applied with probability 0.5. When applied, it covers one contiguous block of four observed bins. The start is sampled uniformly from valid start positions. At the selected bins, each lead's entire channel–frequency input is replaced by its own learned mask vector before the GRU. All leads share the selected time positions. The second view still passes through the global branch when no artificial mask is selected.

The full view contributes symmetric multi-positive InfoNCE using the combined score; the second contributes symmetric multi-positive InfoNCE using its global score only. The normalized full-view brain vector is detached within the cosine-consistency term, but remains trainable through the full-view retrieval loss. The full-view retrieval, second-view retrieval, and consistency weights are 0.5, 0.5, and 0.1. Retrieval temperature is 0.07, local log-mean-exp temperature is 0.10, and the local score coefficient is $\lambda = 0.70$.

All active neural modules start from random initialization; the sentence adapter starts in the identity configuration above. We train jointly with batch size 64 and AdamW (weight decay 0.01). Gradients of the trainable brain network and sentence adapter are clipped separately to norm 1.0. We evaluate validation retrieval every five epochs and select the checkpoint with the highest combined Recall@10; test results do not influence selection. Evaluation uses unmasked windows.

C Representation analysis

We conduct two complementary analyses on the test trials of Subject 2 watching *Spider-Man: Homecoming*: retrieval under neural feature-window truncation and alignment between brain and text representations at the sentence-group level.

Effect of neural feature-window length. We evaluate the trained model using progressively longer prefixes of each neural feature window, from sentence onset to t seconds afterward. Later bins are treated as padding, and the model is not retrained. Global-score R@10 increases from 15.2% at 0.25 s to 26.2% at 1 s with further improvement for longer feature windows (Figure 5). Because every window contains at least 1 s of recorded activity, all queries have equal retained duration at each cutoff up to 1 s. Differences in retained duration between queries therefore cannot serve as a retrieval cue at these cutoffs.

The results show that retrieval benefits from longer prefixes of the neural features in this evaluation. This is an offline input-truncation analysis: it assesses the trained model's sensitivity to feature-window length rather than establishing when linguistic information first becomes available in the neural signal.

Sentence groups. Single-trial brain embeddings are noisy, so we also compare the brain and text spaces at the level of groups of sentences. We use all test trials of Subject 2 and average repeated lines. We cluster these sentences into six groups with k-means on their frozen EmbeddingGemma embeddings; the number of groups was fixed before the analysis. We average each group in each space and compute the cosine between every brain group mean and every text group mean. The test statistic is the diagonal gap: the mean cosine of matching groups minus the mean cosine of mismatched ones, compared with 1,000 shuffles of the group labels. We name the groups after reading their sentences, for example "The guy in the chair." (descriptions), "Of course." (short replies), "No put that down."(negations), "Come on man look." (commands), "I gotta check my calendar first." (first-person statements), and "Does she know?" (questions).

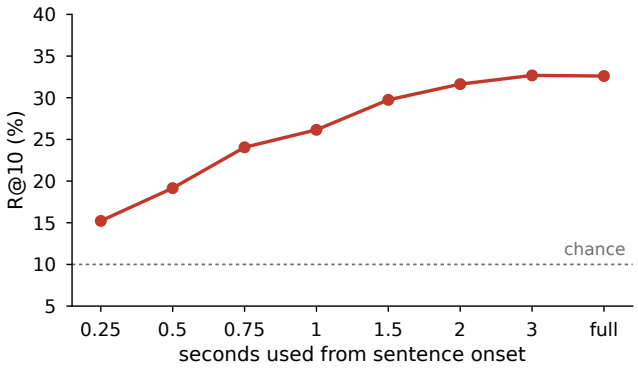


Figure 5: Effect of neural feature-window length on retrieval for Subject 2 watching *Spider-Man: Homecoming*. Global-score R@10 is evaluated on all test trials using feature-window prefixes ending at the indicated times after sentence onset. Later bins are treated as padding, without retraining. "Full" denotes the complete window, from sentence onset to 1 s after offset. At each cutoff up to 1 s, retained duration is identical across queries. The dotted line indicates 10% chance performance.

Results. For five of six groups, the brain mean has its highest cosine similarity with the corresponding text-group mean. The diagonal gap is 0.44, while the network at its initial weights matches one of six groups (Figure 6). Using a linear rather than binned adjustment for sentence duration and word count gives a gap of 0.46, with five of six groups matched. Across ten clustering initializations, the diagonal gap ranges from 0.36 to 0.45. The clusters have limited separation, with a silhouette score of 0.079, while agreement across clustering initializations has an adjusted Rand index of 0.68. The strongest off-diagonal similarities link short replies and first-person statements, which also form the most similar pair of text-group means (cosine 0.38). The questions group instead matches most closely to the negations group. Overall, these descriptive results indicate correspondence between brain and text representations averaged within text-defined sentence groups.

D Implementation details

Table 6 lists the settings used for every recording. The only part that changes across recordings is the number of leads and the number of contacts on each, which set the number and input size of the lead GRUs.

E Baseline implementations

We compare GLANCE with adaptations of SSENSE, DEEPER, BPR, and BrainMosaic, together with a ridge-regression reference, on Subject 2 watching *Spider-Man: Homecoming*. This recording contains the largest number of retained sentence trials. The comparison uses the 100-way sentence-retrieval protocol described in Section 4.1, with identical data splits and candidate sets across models. Model selection uses validation data, and test retrieval is evaluated using the selected models.

SSENSE. Our implementation adapts the electrode-wise encoding and averaging architecture of Liu (2025) to Subject 2. A shared ResNet-18, initialized from random weights, processes each electrode's superlet spectrogram independently. We modify its input convolution to accept a single-channel time–frequency map and replace its final layer with a projection to the 768-dimensional text-embedding space. For each sentence, we average the electrode embeddings and ℓ_2-normalize the resulting neural vector. We use centered EmbeddingGemma sentence embeddings as text targets and ℓ_2-normalize them as well. Each electrode contributes a 50×103 spectrogram. Training uses a one-directional neural-to-text InfoNCE loss with in-batch negatives and temperature 0.07, optimized with Adam at a learning rate of 5×10^{-4}.

BPR. We implement the BPR architecture of McGuire & Moshfeghi (2025) as a sentence-level sEEG-to-text retrieval baseline. Each neural time bin contains the concatenated $126 \times 50 = 6300$

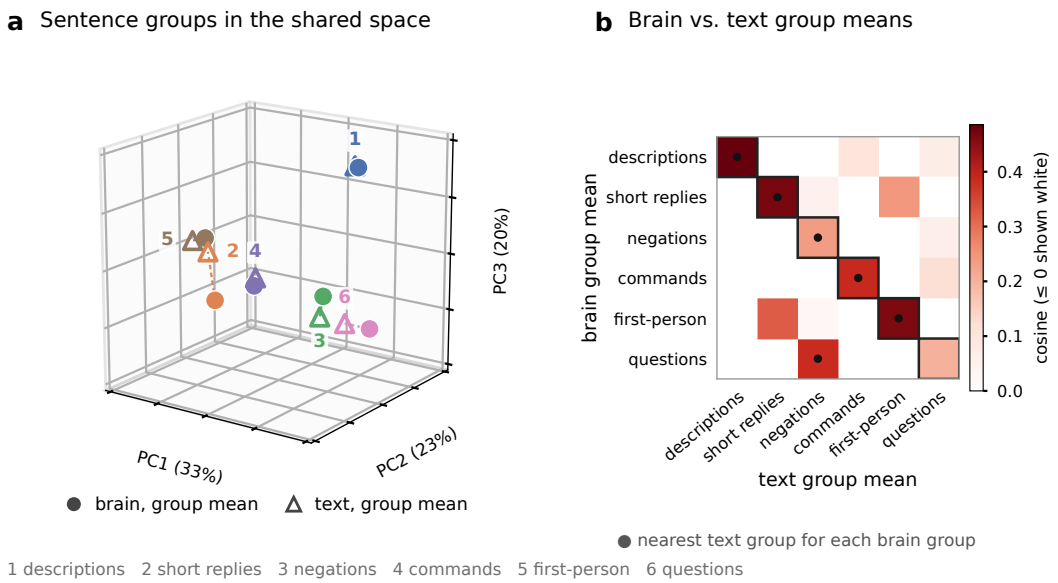


Figure 6: Alignment of text-defined sentence groups in brain and text embedding spaces for Subject 2 watching Spider-Man: Homecoming. Comparisons use residual embeddings after regression on binned sentence duration and word count. (a) Mean brain embeddings (circles) and mean text embeddings (triangles), projected onto the first three principal components of the text-group means, explaining 76% of their variance. Brain means are rescaled by a common factor for visualization; cosine comparisons use the unscaled embeddings. (b) Cosine similarity between brain-group means (rows) and text-group means (columns). Values at or below zero are shown in white. Dots identify the highest-similarity text group for each brain group; five of six matches lie on the diagonal. The diagonal-gap statistic is significant under the label-permutation test ($p = 0.001$).

Table 6: Model and training settings, shared by all recordings.

Component	Setting
Lead encoder	one bidirectional GRU per lead, 128 units per direction ($d = 256$) input: contacts on the lead $\times$ 50 frequencies per time bin
Spatial Transformer	1 layer, 4 heads, feed-forward 512, dropout 0.1, learned summary token
Temporal Transformer	2 layers, 8 heads, feed-forward 1024, dropout 0.1, learned summary token and positional embeddings; linear map to 768
Local branch	Conv1d $256 \to 128$ (kernel 5), GELU, linear $128 \to 128$; word projection $768 \to 128$
Text adapter	bottleneck 64, residual scale $\alpha = 0.05$, output layer set to zero
Scores and losses	temperature $\tau = 0.07$, local weight $\lambda = 0.7$, pooling softness $\eta = 0.1$, consistency weight $\beta = 0.1$
Masking	one block of 4 time bins, probability 0.5, learned mask vector per lead
Optimizer	AdamW, weight decay 0.01, gradient-norm clip 1.0, batch 64
Learning rates	lead GRUs, spatial Transformer, mask vectors: 3×10^{-4} temporal Transformer and projection: 5×10^{-5} summary token and positional embeddings: 2×10^{-5} local branch and word projection: 3×10^{-4}; adapter: 10^{-5}
Schedule	constant learning rates, validation every 5 epochs
Model selection	best validation R@10 of the combined score
Evaluation	100-way ranking, 30 candidate sets per test query

contact–frequency features, which a learned projection maps to 512 dimensions. The neural encoder uses a learned summary token, learned temporal positional embeddings, and three Transformer layers with eight attention heads and hidden width 512. Its summary representation is projected to 768 dimensions and ℓ_2-normalized. The text encoder consists of frozen DistilBERT-base-uncased followed by one trainable Transformer adaptation layer. We retain BPR's residual combination with

the layer-normalized DistilBERT representations and use the ℓ_2-normalized adapted [CLS] representation for retrieval.

The primary condition pairs complete neural sentence windows with their corresponding transcripts, retaining BPR's encoders and loss while replacing its inverse-cloze training task with full-sentence alignment. Training uses one-directional neural-to-text InfoNCE with temperature 0.07 and uniformity regularization weighted by 0.1. Repeated normalized sentences are excluded from one another as false negatives. After an initial run at the paper's reported learning rate of 10^{-6}, we evaluated 10^{-5}, 3×10^{-5}, 10^{-4}, and 3×10^{-4}. The primary condition uses 10^{-4}, selected by validation MRR without reference to test performance. Optimization uses AdamW with weight decay 0.1, batch size 128, 10 epochs of linear warmup followed by linear learning-rate decay, gradient-norm clipping at 1.0, and early stopping with patience five.

We additionally evaluate a sentence-level adaptation of BPR's inverse-cloze task (ICT). A contiguous span containing approximately 30% of the sentence words is selected using the available word timestamps, and the corresponding sEEG time bins form the neural query. The selected words are removed from the paired transcript with probability 0.9. This adapts ICT to individual sentences rather than longer passages and is reported separately from the full-sentence training condition used in the main comparison. Under the same fixed 100-way candidate-bank evaluation, the ICT variant achieves mean R@1 of 2.35%, R@10 of 17.36%, and MRR of 0.0806 across three seeds.

DEEPER. We adapt DEEPER (McGuire & Moshfeghi, 2024) to sentence retrieval from superlet features. Each neural time bin contains the concatenated values from 126 contacts and 50 frequencies, yielding 6300 input features. A learned linear projection maps these features to width 512. Learned temporal positional embeddings are added, a learned summary token is prepended, and padded time bins are masked. The neural encoder is a six-layer, eight-head Transformer with feed-forward width 2048 and dropout 0.3. Its summary representation is projected to 768 dimensions and ℓ_2-normalized. The text encoder is frozen DistilBERT, using its [CLS] representation; only the neural encoder and projection layers are trained.

Training uses one-directional neural-to-text InfoNCE with temperature 0.07, in-batch negatives, and five BM25-selected hard negatives per training query. We use a sentence-level adaptation of inverse-cloze training: a contiguous span of $\max(1, \lfloor 0.3N \rfloor)$ words is sampled from a sentence with N aligned words, and the neural bins overlapping that span form the query. The selected words are removed from the paired transcript with probability 0.9, unless removal would leave an empty transcript. Word timestamps are used to construct these training queries.

Optimization uses AdamW with learning rate 10^{-4}, weight decay 0.01, linear warmup over the first 10% of training steps, linear decay thereafter, and gradient-norm clipping at 1.0. The checkpoint is selected by validation MRR using complete neural sentence windows. The main test evaluation likewise uses complete sentence windows to rank candidate transcripts under the common 100-way retrieval protocol; evaluation with inverse-cloze queries is reported separately.

BrainMosaic. We implement a retrieval adaptation of BrainMosaic (Li et al., 2026). The neural encoder combines a ModernTCN backbone with a Transformer encoder–decoder and learned query slots. We retain all 50 frequency features per contact by flattening the contact and frequency dimensions, giving an input of shape $[\text{contacts} \times \text{frequencies}, \text{time}]$. The padded sentence windows are processed with temporal patch size and stride five, producing 20 temporal tokens. We use 13 query slots: one for the global sentence embedding and 12 for semantic-unit embeddings. The semantic-unit slot count is set to the maximum number of units per sentence across the training, validation, and test splits.

Our adaptation uses frozen EmbeddingGemma to provide centered 256-dimensional representations for sentences and semantic units. Semantic units are constructed using the released clustering procedure, with embedding centering based on the training split. Training runs for 50 epochs using a cosine-alignment loss and a semantic-unit InfoNCE loss weighted by 0.2.

At evaluation, we rank candidate transcripts by cosine similarity between the predicted global sentence embedding and their text embeddings, using the same 30 fixed 100-way candidate banks as the other models. Sentence-level alignment is trained with a cosine loss, whereas the contrastive term

operates on semantic-unit slots. The reported results characterize this adaptation under our retrieval protocol.

Ridge regression. As a linear reference, we average the normalized neural features over the valid time bins of each neural window, yielding a fixed-dimensional feature vector. We then fit an L_2-regularized multi-output Ridge regression mapping these neural features to frozen 768-dimensional Gemma sentence embeddings. Regression parameters are fitted on the training split, while the regularization strength is selected using validation retrieval performance. Predicted embeddings are L_2-normalized before cosine-based retrieval scoring.