SPECTRALOOM: SPECTRAL-ATTENTIVE MULTI-SCALE NEURAL DECODING FOR OPEN-VOCABULARY EEG-TO-TEXT TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary EEG-to-text decoding generates the sentence a participant reads from the electroencephalogram recorded at each fixated word, and is a step towards non-invasive brain–computer interfaces for communication. Existing decoders flatten the eight frequency bands of every word event into one vector, model temporal context at a single scale, and connect the EEG encoder to a pretrained language model through a fixed linear projection. We propose SpectraLoom, in which each component answers one of these limitations: (1) *Spectral Band Attention* (SBA) computes a word-specific gate over the eight band slices with 315 parameters and retains all 840 input features; (2) a *multi-scale Conv1D–BiLSTM encoder* combines parallel convolutions over three, five and seven neighbouring word events with sentence-wide bidirectional recurrence; and (3) a *Cross-Attention Bridge* (CAB) refines every word-aligned EEG state by attending over the whole sentence before it enters the BART encoder. We evaluate five systems trained under one protocol, with a sentence-grouped split, identical decoding and three seeds, on ZuCo 1.0 and zero-shot on ZuCo 2.0. SpectraLoom attains the highest ROUGE-L and the lowest word error rate on both corpora (16.80 and 97.55 in domain, 17.65 and 92.70 zero-shot) with 44% fewer parameters than T5-large, and it attains the highest teacher-forced BLEU-1 to BLEU-4. Over three seeds, removing or simplifying any component lowers BLEU-1 to BLEU-4 and ROUGE-L on both corpora, and a word-independent band gate costs 4.87 BLEU-1. Across three seeds, its BLEU-1 varies with a standard deviation of 0.74, against 3.35 for BrainTranslator with the same BART-large backbone. Code is available at https://github.com/anonymoususer98700/spectraloom.

1 INTRODUCTION

Decoding language from brain activity has advanced rapidly with invasive recordings, from speech neuroprostheses (Moses et al., 2021; Willett et al., 2023; Metzger et al., 2023) to sentence decoding from cortical activity (Makin et al., 2020), and with non-invasive fMRI and MEG (Tang et al., 2023; Défossez et al., 2023). Electroencephalography (EEG) is the non-invasive modality of practical interest: it is portable, inexpensive and resolves the millisecond dynamics of reading (Luck, 2014). In open-vocabulary EEG-to-text decoding, word-level EEG features recorded during natural reading (Hollenstein et al., 2018; 2020) are mapped into the input space of a pretrained sequence-to-sequence model that generates the sentence (Wang & Ji, 2022). Subsequent systems have introduced discrete codebooks (Duan et al., 2023), multi-view encoders with self-supervised pretraining (Liu et al., 2024), contrastive bootstrapping (Zhou et al., 2023), recurrent encoders (Murad et al., 2025) and diffusion decoders (Jeon & Kim, 2025), and Jo et al. (2024) established that such systems must be evaluated without teacher forcing and against input controls.

In our study, we find that the design of current decoders leaves three properties of the word-level EEG signal unused. ① *Spectral blindness*: every word event carries band power in eight frequency bands, and oscillatory activity in different bands is associated with different aspects of language comprehension (Bastiaansen & Hagoort, 2015; Weiss & Mueller, 2003; Jensen et al., 2007; Klimesch, 2012), yet existing encoders concatenate the bands into one 840-dimensional vector and apply the same transformation to every word. ② *Single-scale context*: a fixation sequence carries structure

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

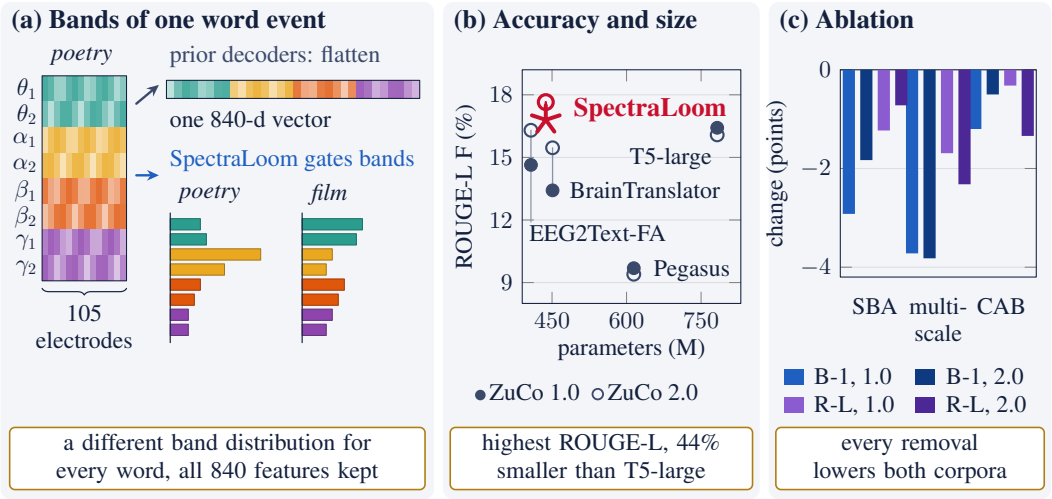


Figure 1: Motivation and evidence on ZuCo: (a) band structure of a word event, (b) ROUGE-L against model size, (c) change in BLEU-1 (B-1) and ROUGE-L (R-L) on ZuCo 1.0 and ZuCo 2.0 when one component is removed (three seeds).

at several spans, from the current word to its neighbours to the whole sentence, whereas existing encoders apply either a Transformer (Wang & Ji, 2022) or a recurrent network (Murad et al., 2025) at one scale. ③ *Rigid interface*: the encoder output reaches the pretrained model through a fixed linear projection, so each word-aligned state enters the language model without access to the rest of the sentence.

We propose SpectraLoom, in which each component answers one of the above limitations (Figure 2). (1) *Spectral Band Attention* (SBA) reshapes each word event into eight band slices, scores each slice with a shared query and gates the slices with a word-specific softmax distribution, adding 315 parameters and retaining all 840 features (addresses limitation ①). (2) A *multi-scale Conv1D–BiLSTM encoder* applies parallel convolutions over three, five and seven word events, in the manner of Inception branches (Szegedy et al., 2015), followed by a three-layer bidirectional LSTM with sinusoidal positions (addresses limitation ②). (3) A *Cross-Attention Bridge* (CAB) projects each encoder state into a query that attends over the whole EEG sequence, following learned-query interfaces in multimodal models (Jaegle et al., 2021; Alayrac et al., 2022; Li et al., 2023), and passes the refined states through the BART encoder to the decoder (addresses limitation ③). We evaluate SpectraLoom against four locally trained systems under one protocol, with a sentence-grouped split, identical decoding and three seeds (Figure 1).

Our main contributions are summarised as follows:

- We propose SpectraLoom, an EEG-to-text decoder that combines word-specific spectral band gating, multi-scale temporal encoding and an EEG-conditioned cross-attention interface to a pretrained BART model.
- Under a matched protocol with a sentence-grouped split, identical decoding and three seeds, SpectraLoom attains the highest ROUGE-L and the lowest word error rate among five systems on ZuCo 1.0 and zero-shot on ZuCo 2.0, with 44% fewer parameters than T5-large and the highest teacher-forced BLEU-1 to BLEU-4, and each of its three components raises BLEU-1 to BLEU-4 and ROUGE-L on both corpora over three seeds.
- We show that SpectraLoom trains an EEG-conditioned BART reproducibly: with the same BART-large backbone, its seed standard deviation is 0.74 BLEU-1 and 0.99 ROUGE-L in domain, against 3.35 and 2.54 for BrainTranslator, whose EEG front end is a Transformer with a linear projection.

2 Related Work

Language decoding from neural signals. Invasive recordings support closed- and open-vocabulary decoding at high accuracy (Makin et al., 2020; Moses et al., 2021; Willett et al., 2023; Metzger

et al., 2023), and non-invasive fMRI and MEG recover semantic content and perceived speech (Tang et al., 2023; Défossez et al., 2023; Yang et al., 2024). With EEG, early work on ZuCo classified words and relations from word-level features (Hollenstein et al., 2019), and Wang & Ji (2022) introduced open-vocabulary generation by projecting word-level EEG into BART (Lewis et al., 2020). DeWave (Duan et al., 2023) added a vector-quantised codebook, EEG2Text (Liu et al., 2024) a multi-view Transformer with EEG pretraining, BELT (Zhou et al., 2023) contrastive bootstrapping, R1 (Murad et al., 2025) a recurrent encoder, and DELTA (Jeon & Kim, 2025) residual vector quantisation with a diffusion language model. Jo et al. (2024) showed that scores obtained with teacher forcing overstate decoding quality and that input controls are required. All of these systems treat the eight bands of a word event as one feature vector and connect the encoder to the language model through a linear map; SpectraLoom models the band structure explicitly and replaces the linear map with a sentence-level attention interface, and Section B reports the input controls of Jo et al. (2024) as training-time interventions.

Frequency-selective EEG representations. Band-limited oscillations are associated with lexical retrieval, syntactic and semantic unification, attention and memory during language comprehension (Bastiaansen & Hagoort, 2015; Weiss & Mueller, 2003; Jensen et al., 2007; Klimesch, 2012), and band-specific features are discriminative in EEG emotion recognition, where critical bands and channels differ across tasks (Zheng & Lu, 2015; Song et al., 2020). Deep EEG decoders learn spectral filters implicitly through temporal convolutions (Schirrmeister et al., 2017; Lawhern et al., 2018; Song et al., 2023), and self-supervised EEG encoders learn from raw signals (Kostas et al., 2021); reviews of the field note that most models consume band features as an undifferentiated vector (Craik et al., 2019). Channel attention in vision, such as squeeze-and-excitation and CBAM (Hu et al., 2018; Woo et al., 2018), reweights feature groups from their own statistics. SBA transfers this idea to the band axis of word-level EEG, with a single shared query so that the gate is word-specific but band-comparable, and a scale factor that makes a uniform gate an identity.

Multi-scale encoders and learned-query interfaces. Parallel convolutions with different receptive fields capture complementary structure in images and speech (Szegedy et al., 2015; Baevski et al., 2020; Hsu et al., 2021), and bidirectional recurrent networks (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997; Graves et al., 2013) model sequential biosignals. For connecting a perceptual encoder to a language model, attention with learned or conditioned queries is the dominant design: Perceiver and Perceiver IO read inputs through latent queries (Jaegle et al., 2021; 2022), Flamingo resamples visual features before a frozen language model (Alayrac et al., 2022), BLIP-2 uses a query Transformer (Li et al., 2023), and cross-modal attention aligns unsynchronised modalities (Xu et al., 2015; Tsai et al., 2019). SpectraLoom combines multi-scale convolution with recurrence for word-level EEG and derives the bridge queries from the EEG states themselves, so that every memory slot stays aligned with a word event.

3 METHODOLOGY

Overview. As shown in Figure 2, SpectraLoom transforms the word-level EEG sequence of a sentence in three stages before it enters BART-large. SBA (Section 3.2) gates the band slices of each word, the multi-scale Conv1D–BiLSTM encoder (Section 3.3) contextualises the gated sequence locally and across the sentence, and the CAB (Section 3.4) refines each word-aligned state with sentence-level attention. The refined sequence replaces the token embeddings of BART's encoder, and the decoder generates the sentence autoregressively.

3.1 PRELIMINARIES

Task. A participant reads a sentence while EEG and eye movements are recorded. For each of the $L \leq 56$ fixated words, band power over the gaze duration is computed in $B = 8$ frequency bands (two sub-bands each of theta, alpha, beta and gamma) at $E = 105$ electrodes, giving $\mathbf{x}_i \in \mathbb{R}^{BE}$ with $BE = 840$. Given $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$, the model generates the subword sequence $\mathbf{Y} = [y_1, \ldots, y_T]$ of the sentence,

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{t=1}^{T} p\big(y_t \mid y_{<t}, f_\theta(\mathbf{X})\big), \tag{1}$$

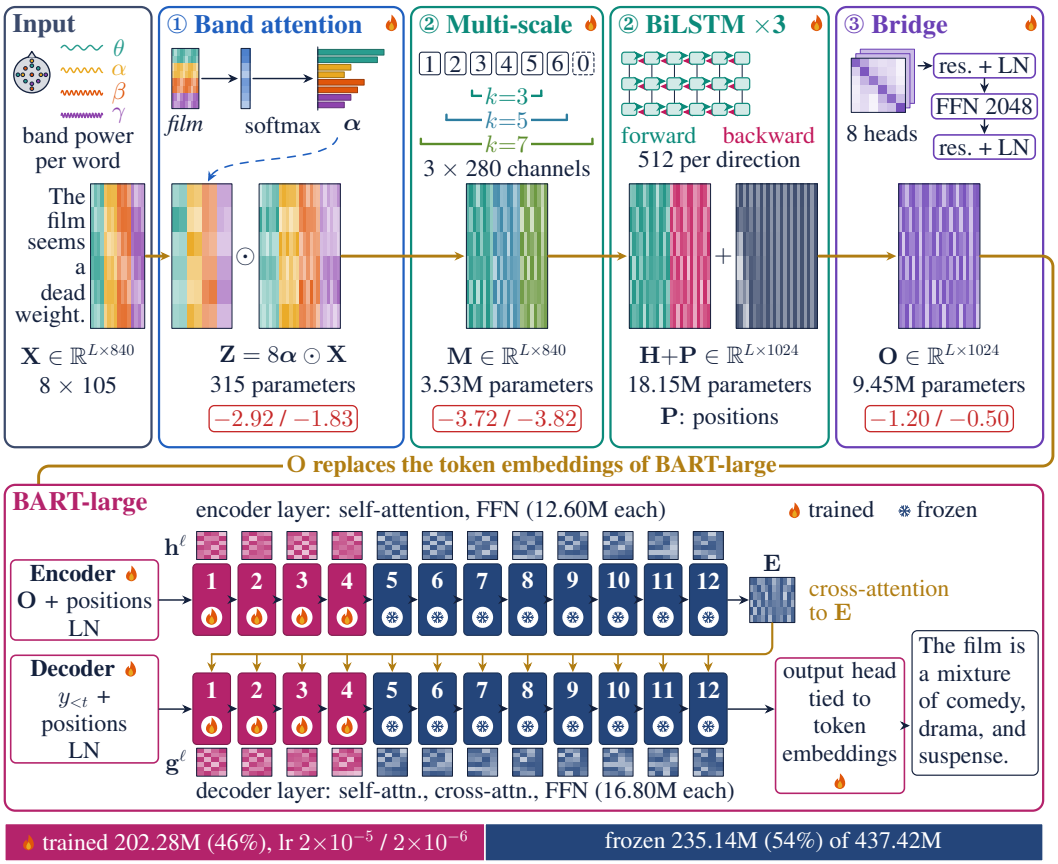


Figure 2: Architecture of SpectraLoom: intermediate tensors at every stage, trained and frozen components. Red boxes give the BLEU-1 change on ZuCo 1.0 / ZuCo 2.0 when the module is removed (three seeds).

where f_θ is the SpectraLoom encoder and the conditional is computed by BART-large (Lewis et al., 2020). The word boundaries come from the eye tracker, so the input is a sequence of word events of known length.

Normalisation. Features are standardised per dimension with statistics computed on the training split only, $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \boldsymbol{\mu})/(\boldsymbol{\sigma} + 10^{-8})$, and passed through a layer normalisation (Ba et al., 2016); during training, Gaussian noise with standard deviation 0.02 is added. Padded positions are masked after every operation that could mix them with valid positions.

3.2 Spectral Band Attention

Observation 1. *Band-specific gating improves phrase-level overlap.* Replacing the learned gate by a uniform gate, with every other component fixed, lowers BLEU-2 from 4.61 to 3.04 (−34%), BLEU-1 by 2.92, ROUGE-L by 1.23 and zero-shot BLEU-1 by 1.83 over three seeds, and a static gate that learns one weight per band for all words lowers BLEU-1 by 4.87 (Table 3). Uniform or word-independent treatment of the bands discards information that a word-specific weighting recovers (limitation ①).

Technique. SBA reshapes each normalised word vector into its band slices, $\mathbf{X}_i^{\text{band}} \in \mathbb{R}^{B\times E}$, normalises every slice with a shared layer normalisation to remove the difference in power scale between low and high bands, and scores it with a shared query $\mathbf{w}_q \in \mathbb{R}^E$:

$$s_{i,b} = \mathbf{w}_q^\top \,\mathrm{LN}\big(\mathbf{X}_{i,b}^{\text{band}}\big), \qquad \alpha_{i,b} = \frac{\exp(s_{i,b})}{\sum_{b'=1}^{B}\exp(s_{i,b'})}, \qquad \mathbf{z}_{i,b} = B\,\alpha_{i,b}\,\mathbf{X}_{i,b}^{\text{band}}. \tag{2}$$

The gated slices are concatenated back to $\mathbf{z}_i \in \mathbb{R}^{840}$. Because the query is shared, the scores of different bands are computed in the same space and the softmax yields a word-specific distribution over bands; because the gate multiplies the un-normalised slices, the band identity and electrode layout are preserved. The factor B makes a uniform distribution an identity map, so the gate redistributes emphasis around a mean multiplier of one instead of shrinking the input, and the uniform-gate ablation receives exactly the original features. SBA adds 315 parameters: the query (105) and the scale and bias of the layer normalisation (210).

Design choices. Three choices distinguish SBA from generic feature attention (Hu et al., 2018; Woo et al., 2018). First, the query is shared across bands, so a score measures the same electrode pattern in every band and the resulting weights are comparable across bands; a separate query per band would let each band define its own scale and would remove this comparability. Second, the softmax makes the bands compete for a fixed budget at every word, which forces a relative allocation instead of an unconstrained rescaling that later layers could absorb. Third, scores are computed on band-normalised slices, so that a band with high absolute power, typically theta, does not dominate the allocation, while the gate multiplies the slices before normalisation and therefore leaves the band amplitudes that carry information intact. The cost is $\mathcal{O}(LBE)$ per sentence, negligible beside the encoder.

3.3 Multi-Scale Conv1D–BiLSTM Encoder

Observation 2. *Multi-scale context supports transfer under distribution shift.* Replacing the parallel convolutions by a normalised linear layer of the same width lowers BLEU-3 by 36%, BLEU-4 by 45% and ROUGE-L by 1.69 in domain, and lowers zero-shot BLEU-1 by 3.82 and ROUGE-L by 2.32, the largest zero-shot loss after the static band gate (Table 3). Structure at several spans of the fixation sequence generalises across participants and sessions better than per-word features alone (limitation ②).

Technique. Three parallel one-dimensional convolutions with kernel sizes $k \in \{3, 5, 7\}$ and 280 output channels each are applied along the word axis with same-padding,

$$\mathbf{c}_i^{(k)} = \mathrm{GELU}\Big(\sum_{j=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \mathbf{W}_j^{(k)} \mathbf{z}_{i+j} + \mathbf{b}^{(k)} \Big), \qquad \mathbf{m}_i = \mathrm{Drop}\Big(\mathrm{LN}\big[\mathbf{c}_i^{(3)}; \mathbf{c}_i^{(5)}; \mathbf{c}_i^{(7)}\big] \Big) \in \mathbb{R}^{840}, \tag{3}$$

so that the three branches cover the neighbouring one, two and three words on each side and the concatenation keeps the working width at 840. GELU activations (Hendrycks & Gimpel, 2016) and dropout (Srivastava et al., 2014) with rate 0.3 are used throughout. The sequence $\mathbf{M}$ is then encoded by a three-layer bidirectional LSTM (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) with 512 units per direction, run on packed sequences so that padding does not enter the recurrence,

$$\mathbf{h}_i = \big[\overrightarrow{\mathrm{LSTM}}(\mathbf{M})_i \,;\, \overleftarrow{\mathrm{LSTM}}(\mathbf{M})_i\big] \in \mathbb{R}^{1024}, \qquad \hat{\mathbf{h}}_i = \mathrm{Drop}\big(\mathrm{Drop}(\mathrm{LN}(\mathbf{h}_i)) + \mathbf{p}_i\big), \tag{4}$$

where $\mathbf{p}_i$ is the sinusoidal position encoding of word i (Vaswani et al., 2017). Input-to-hidden weights use Xavier initialisation and recurrent weights orthogonal initialisation (Saxe et al., 2014). The convolutions capture local co-activation across neighbouring fixations, and the recurrence integrates both directions over the full sentence. The two stages are complementary: a convolution with kernel k sees a fixed window of k words with position-specific weights, which suits short phrases whose internal order matters, whereas the recurrence carries a state of unbounded span, which suits sentence-level context. Both are linear in L, with cost $\mathcal{O}(L \sum_k k\, d^2)$ for the convolutions and $\mathcal{O}(L\, d_h^2)$ for the recurrence, and the sinusoidal positions restore the absolute word index that packing and recurrence leave implicit before the bridge applies order-agnostic attention.

3.4 Cross-Attention Bridge

Observation 3. *A sentence-level interface improves transfer.* Replacing the bridge by a normalised linear projection to the decoder width lowers BLEU-1 by 1.20 in domain, and zero-shot ROUGE-L by 1.34 and BLEU-4 by 0.39 (Table 3). A word-aligned state that can read the rest of the sentence before it reaches the language model carries information that a position-wise projection does not (limitation ③).

Technique. Each encoder state is projected to a query in the decoder width $d = 1024$, $\mathbf{q}_i = \mathbf{W}_q \hat{\mathbf{h}}_i + \mathbf{b}_q$, and all queries attend over the valid encoder states with eight heads,

$$\mathbf{a}_i = \mathrm{MHA}\big(\mathbf{q}_i, \hat{\mathbf{H}}, \hat{\mathbf{H}}\big), \qquad \mathbf{u}_i = \mathrm{LN}(\mathbf{q}_i + \mathbf{a}_i), \qquad \mathbf{o}_i = \mathrm{LN}\big(\mathbf{u}_i + \mathrm{Drop}(\mathrm{FFN}(\mathbf{u}_i))\big), \quad (5)$$

where MHA is multi-head attention (Vaswani et al., 2017) with a key-padding mask, the residual connections follow He et al. (2016), and the feed-forward network expands to $2d$ with a GELU. Deriving the queries from the EEG states, rather than using sample-independent learned queries (Jaegle et al., 2021; Li et al., 2023), keeps one memory slot per word event and conditions every slot on its own recording. The refined sequence $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_L]$ is passed to BART-large as input embeddings; the BART encoder processes it, and the decoder attends to the encoder output at every layer. Padded slots are set to zero.

Why conditioned queries. A bank of free learned queries, as in Perceiver or BLIP-2 (Jaegle et al., 2021; Li et al., 2023), compresses an input into a fixed number of slots and is well suited to long visual inputs, but for sentences of at most 56 words it removes the one-to-one correspondence between memory slots and word events, and a free query that ignores its input can let the decoder fall back on its language prior. Deriving each query from its own encoder state keeps the correspondence, conditions every slot on the recording, and still lets each slot gather sentence-level context through attention. The bridge costs $\mathcal{O}(L^2 d)$, which is small for $L \leq 56$, and accounts for 9.45M of the 31.13M EEG-module parameters, against 3.53M for the convolutions and 18.15M for the recurrence.

3.5 Training and Decoding

All EEG modules are trained jointly with the token embeddings and the first four encoder and decoder layers of BART-large; the remaining eight layers of each stack are frozen, so that 202.28M of the 437.42M parameters are trained (Figure 2). The objective is the token-level cross-entropy of Equation (1). AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01 uses a learning rate of 2×10^{-5} for the EEG modules and 2×10^{-6} for the unfrozen BART parameters, with a 10% linear warm-up followed by cosine decay, gradient clipping at 1.0 and an effective batch of 64. Training runs for at most 30 epochs with early stopping on the development loss (patience 5), and the best development checkpoint is evaluated. Sentences are generated with beam search (five beams, at most 32 tokens, length penalty 1.4, repetition penalty 1.5, no repeated bigrams), with settings chosen on the ZuCo 1.0 development split; the repetition penalty and the bigram constraint counter the degenerate repetition to which beam search is prone (Holtzman et al., 2020). The procedure is listed in Section A.

4 Experiments

4.1 Settings

Data and protocol. We train on the three reading tasks of ZuCo 1.0 (sentiment, normal and task-specific reading) (Hollenstein et al., 2018) and evaluate on its test split and zero-shot on the task-specific reading of ZuCo 2.0 (Hollenstein et al., 2020), which has different participants and recording sessions. The split groups all readings of a normalised sentence into one partition, so that no sentence occurs in two partitions, and removes from training every ZuCo 1.0 sentence that also occurs in ZuCo 2.0. After EEG validity filtering, the partitions contain 8,978 training, 1,105 development and 1,127 test trials, and 6,629 zero-shot trials. Normalisation statistics are computed on the training split only.

Baselines and metrics. We train four systems under the identical protocol: BrainTranslator (Wang & Ji, 2022), which feeds a six-layer Transformer encoder into BART-large; the same EEG encoder with T5-large (Raffel et al., 2020) and Pegasus-XSum (Zhang et al., 2020) as the language model; and EEG2Text-FA, our adaptation of EEG2Text (Liu et al., 2024) to word-level features, which encodes the band and electrode structure of each word event with a convolution and a six-layer Transformer and feeds BART-large. All five systems use the same split, epoch budget, early stopping and decoding settings, and seeds 42, 100 and 312. We report corpus BLEU-1 to BLEU-4 (Papineni et al., 2002) via SacreBLEU (Post, 2018), ROUGE-L F_1 (Lin, 2004), chrF (Popović, 2015), word error rate (WER)

Table 1: Comparison under the matched protocol (free five-beam decoding, three seeds, mean ± sample standard deviation). ZuCo 2.0 is evaluated zero-shot with the ZuCo 1.0 checkpoints. Params: total registered parameters (M). Uniq.: distinct generated sentences (%).

	Model	Params	B-1	B-2	B-3	B-4	R-L F	chrF	WER↓	Uniq.
ZuCo 1.0	BrainTranslator (2022)	451.06	$15.63_{\pm 3.35}$	$4.73_{\pm 1.02}$	$1.43_{\pm 0.54}$	$0.43_{\pm 0.31}$	$13.42_{\pm 2.54}$	$17.09_{\pm 0.96}$	$119.16_{\pm 7.61}$	0.80
	T5-large (2020)	782.44	$18.42_{\pm 0.32}$	$4.58_{\pm 0.46}$	$1.93_{\pm 0.33}$	$0.98_{\pm 0.23}$	$16.42_{\pm 0.28}$	$14.58_{\pm 0.43}$	$101.20_{\pm 0.73}$	13.64
	Pegasus-XSum (2020)	614.52	$6.38_{\pm 3.33}$	$1.06_{\pm 1.69}$	$0.10_{\pm 0.15}$	$0.03_{\pm 0.04}$	$9.69_{\pm 3.04}$	$11.09_{\pm 0.77}$	$97.95_{\pm 0.21}$	0.12
	EEG2Text-FA (2024)	407.54	$14.75_{\pm 3.85}$	$2.61_{\pm 1.34}$	$0.62_{\pm 0.69}$	$0.14_{\pm 0.13}$	$14.64_{\pm 1.78}$	$12.16_{\pm 2.10}$	$102.00_{\pm 2.99}$	4.41
	SpectraLoom (ours)	437.42	$17.76_{\pm 0.74}$	$4.61_{\pm 0.13}$	$1.29_{\pm 0.09}$	$0.38_{\pm 0.11}$	$16.80_{\pm 0.99}$	$14.65_{\pm 0.72}$	$97.55_{\pm 1.70}$	6.68
ZuCo 2.0	BrainTranslator (2022)	451.06	$18.32_{\pm 2.25}$	$6.49_{\pm 1.05}$	$2.74_{\pm 0.71}$	$1.28_{\pm 0.57}$	$15.46_{\pm 2.30}$	$16.74_{\pm 0.55}$	$107.87_{\pm 6.45}$	0.12
	T5-large (2020)	782.44	$16.43_{\pm 0.86}$	$3.90_{\pm 0.40}$	$1.57_{\pm 0.19}$	$0.77_{\pm 0.11}$	$16.08_{\pm 0.26}$	$14.36_{\pm 0.67}$	$95.96_{\pm 1.27}$	5.50
	Pegasus-XSum (2020)	614.52	$4.90_{\pm 3.00}$	$0.78_{\pm 1.30}$	$0.13_{\pm 0.22}$	$0.02_{\pm 0.04}$	$9.40_{\pm 2.92}$	$10.02_{\pm 0.64}$	$95.76_{\pm 1.07}$	0.02
	EEG2Text-FA (2024)	407.54	$14.19_{\pm 5.15}$	$3.36_{\pm 2.22}$	$0.85_{\pm 1.14}$	$0.34_{\pm 0.51}$	$16.30_{\pm 2.58}$	$11.58_{\pm 2.55}$	$94.54_{\pm 1.54}$	1.92
	SpectraLoom (ours)	437.42	$16.33_{\pm 0.60}$	$4.64_{\pm 0.65}$	$1.61_{\pm 0.52}$	$0.69_{\pm 0.27}$	$17.65_{\pm 0.79}$	$13.24_{\pm 0.98}$	$92.70_{\pm 1.92}$	1.16

Table 2: Teacher-forced evaluation on the ZuCo 1.0 test split (three seeds, mean ± sample standard deviation).

Model	B-1	B-2	B-3	B-4	R-L F	chrF	WER↓
BrainTranslator	$38.64_{\pm 0.27}$	$21.34_{\pm 0.19}$	$12.25_{\pm 0.13}$	$6.96_{\pm 0.07}$	$33.65_{\pm 0.44}$	$29.82_{\pm 0.35}$	$68.66_{\pm 0.39}$
T5-large	$37.61_{\pm 0.17}$	$20.31_{\pm 0.15}$	$10.96_{\pm 0.10}$	$6.05_{\pm 0.09}$	$35.24_{\pm 0.13}$	$28.44_{\pm 0.08}$	$68.82_{\pm 0.15}$
Pegasus-XSum	$37.34_{\pm 0.32}$	$19.60_{\pm 0.25}$	$10.36_{\pm 0.35}$	$5.69_{\pm 0.38}$	$31.72_{\pm 0.18}$	$27.50_{\pm 0.08}$	$70.59_{\pm 0.33}$
EEG2Text-FA	$37.76_{\pm 2.25}$	$20.62_{\pm 2.27}$	$11.74_{\pm 1.76}$	$6.70_{\pm 1.22}$	$32.91_{\pm 1.98}$	$28.77_{\pm 2.19}$	$69.63_{\pm 2.47}$
SpectraLoom	$39.03_{\pm 0.78}$	$21.64_{\pm 0.71}$	$12.37_{\pm 0.47}$	$7.04_{\pm 0.31}$	$34.29_{\pm 1.17}$	$30.03_{\pm 0.57}$	$68.16_{\pm 0.85}$

and the fraction of distinct generated sentences, all on the percentage scale, as mean and sample standard deviation over seeds.

Implementation. Models are implemented in PyTorch (Paszke et al., 2019) with Hugging Face Transformers (Wolf et al., 2020) and trained in BF16 on a single NVIDIA RTX 5080 (16 GB), with a physical batch of 8 and eight accumulation steps. Deterministic algorithms are enabled and all random generators are seeded. The ablation uses all three seeds and the input controls of Section B use seed 312, under the identical protocol.

4.2 Main Results

As shown in Table 1, SpectraLoom attains the highest ROUGE-L and the lowest WER of the five systems on both corpora. In domain, its ROUGE-L of 16.80 exceeds T5-large by 0.38 and BrainTranslator by 3.38, and its WER of 97.55 is 3.65 points below T5-large and 21.61 below BrainTranslator. Zero-shot on ZuCo 2.0, both leads hold and the ROUGE-L margin over T5-large grows to 1.57: ROUGE-L 17.65 against 16.08 for T5-large and 15.46 for BrainTranslator, and WER 92.70 against 95.96 and 107.87. SpectraLoom reaches these scores with 437.42M parameters, 44% fewer than T5-large and 3% fewer than BrainTranslator. T5-large attains the highest in-domain BLEU-1, BLEU-3 and BLEU-4 and the most diverse outputs, and BrainTranslator the highest chrF and zero-shot BLEU, with a seed standard deviation of 3.35 BLEU-1 against 0.74 for SpectraLoom. EEG2Text-FA, which shares the BART-large decoder, trails SpectraLoom by 3.01 BLEU-1 and 2.16 ROUGE-L in domain and by 2.14 BLEU-1 and 1.34 ROUGE-L zero-shot. Pegasus-XSum produces one or two distinct sentences per seed and scores lowest on the overlap metrics.

SpectraLoom transfers without loss of phrase-level overlap: from ZuCo 1.0 to ZuCo 2.0, its BLEU-4 rises from 0.38 to 0.69, its ROUGE-L from 16.80 to 17.65 and its WER falls from 97.55 to 92.70, whereas T5-large loses 1.99 BLEU-1 and 0.21 BLEU-4 under the same shift.

Teacher-forced evaluation. Teacher forcing supplies the reference prefix and takes the most likely next token at every position, so it measures conditional next-token prediction separately from free generation (Jo et al., 2024; Ranzato et al., 2016; Bengio et al., 2015). On the same test split and checkpoints (Table 2), SpectraLoom attains the highest BLEU-1 to BLEU-4 (39.03, 21.64, 12.37 and 7.04), the highest chrF (30.03) and the lowest WER (68.16) of the five systems, and T5-large the highest ROUGE-L. Its advantage over BrainTranslator and EEG2Text-FA, which use the same BART-large backbone, holds on every metric.

Table 3: Component ablation (three seeds, mean ± sample standard deviation). Each row replaces one component with all others fixed; the static gate learns one weight per band shared by every word.

	ZuCo 1.0				ZuCo 2.0 zero-shot		
Variant	B-1	B-2	B-4	R-L	B-1	B-4	R-L
SpectraLoom	$17.76_{\pm 0.74}$	$4.61_{\pm 0.13}$	$0.38_{\pm 0.11}$	$16.80_{\pm 0.99}$	$16.33_{\pm 0.60}$	$0.69_{\pm 0.27}$	$17.65_{\pm 0.79}$
Uniform SBA gate	$14.83_{\pm 2.78}$	$3.04_{\pm 0.89}$	$0.18_{\pm 0.16}$	$15.56_{\pm 0.42}$	$14.50_{\pm 3.72}$	$0.42_{\pm 0.31}$	$16.93_{\pm 1.81}$
Static SBA gate	$12.88_{\pm 3.87}$	$2.30_{\pm 1.47}$	$0.09_{\pm 0.08}$	$14.39_{\pm 2.13}$	$11.48_{\pm 5.27}$	$0.21_{\pm 0.34}$	$14.56_{\pm 3.47}$
w/o multi-scale	$14.04_{\pm 4.11}$	$3.19_{\pm 1.51}$	$0.21_{\pm 0.14}$	$15.11_{\pm 0.92}$	$12.51_{\pm 4.75}$	$0.46_{\pm 0.44}$	$15.32_{\pm 2.71}$
w/o CAB (linear)	$16.56_{\pm 0.43}$	$4.08_{\pm 0.36}$	$0.26_{\pm 0.20}$	$16.48_{\pm 1.88}$	$15.83_{\pm 1.39}$	$0.30_{\pm 0.25}$	$16.30_{\pm 1.35}$

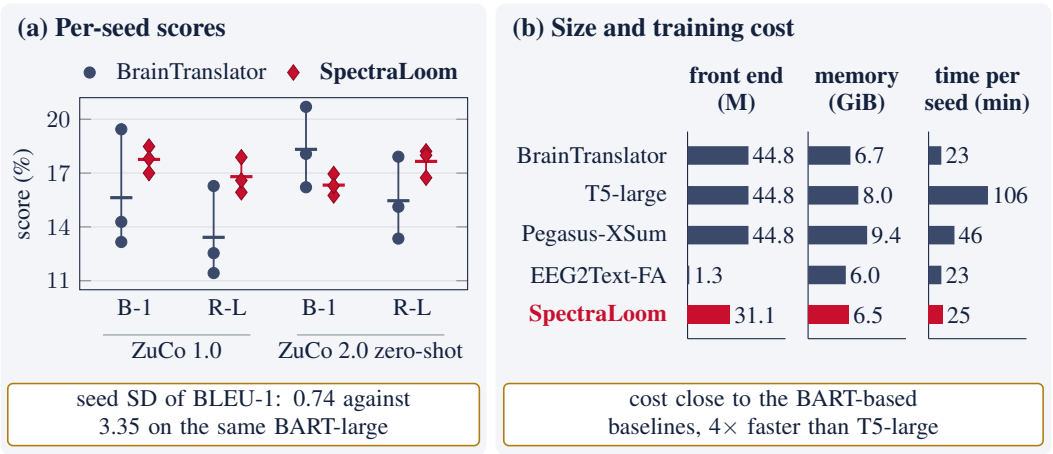


Figure 3: Stability and cost. (a) Per-seed scores on the same BART-large backbone. (b) Size and training cost.

4.3 Ablation Study

Table 3 replaces one component at a time under all three seeds, with the replaced module substituted by a normalised alternative of the same output width. Every variant falls below the full model on BLEU-1 to BLEU-4 and ROUGE-L on both corpora, so each component contributes in domain and under transfer. Band gating matters most when it is word-specific: a uniform gate lowers BLEU-1 by 2.92 in domain and 1.83 zero-shot, and a static gate, which learns eight band weights shared by every word, lowers it by 4.87 and 4.85, the largest losses of any variant. Removing the multi-scale convolutions lowers BLEU-1 by 3.72 and 3.82 and ROUGE-L by 1.69 and 2.32, which identifies local context over neighbouring words as the component that carries phrase structure and transfers. Removing the bridge lowers BLEU-1 by 1.20 in domain, and zero-shot ROUGE-L by 1.34 and BLEU-4 by 0.39, so the bridge acts mainly on transfer.

4.4 Analysis

Stability across seeds. SpectraLoom is markedly more stable than BrainTranslator, which shares its BART-large backbone and differs only in the EEG front end (Figure 3a): its seed standard deviation is 0.74 BLEU-1 and 0.99 ROUGE-L in domain and 0.60 and 0.79 zero-shot, against 3.35, 2.54, 2.25 and 2.30 for BrainTranslator, whose per-seed BLEU-1 ranges from 13.16 to 19.44 in domain, and 3.85, 1.78, 5.15 and 2.58 for EEG2Text-FA, which also shares the backbone. With the same language model, the spectral gate, multi-scale encoder and bridge therefore train an EEG-conditioned BART both to a lower word error rate and more reproducibly than the Transformer front ends of both baselines.

Efficiency. SpectraLoom's EEG front end has 31.13M parameters, against 44.77M for the shared Transformer encoder of BrainTranslator, T5-large and Pegasus-XSum, which includes an unused registered layer of 6.27M, and 1.25M for the compact front end of EEG2Text-FA. It trains with 6.48 GiB of peak memory, against 5.97 to 9.35 GiB for the baselines, and in 25.3 minutes per seed, against 22.7 to 106.1 minutes (Figure 3b); training times include early stopping, after 12 to 17 epochs

for SpectraLoom and 30 for T5-large. SBA itself adds 315 parameters, so spectral selectivity is obtained at negligible cost.

5 Conclusion

We presented SpectraLoom, an EEG-to-text decoder that gates the frequency bands of every word event, encodes the gated sequence at several temporal scales and across the sentence, and connects it to BART through an EEG-conditioned cross-attention bridge. Under a matched protocol with three seeds, SpectraLoom attains the highest ROUGE-L and the lowest word error rate among five systems on ZuCo 1.0 and zero-shot on ZuCo 2.0, with 44% fewer parameters than T5-large, and each of its components improves both corpora. With the same language model, it also trains more reproducibly than two alternative front ends, with a seed standard deviation of 0.74 BLEU-1 against 3.35 and 3.85.

Ethics statement

This work uses the publicly released ZuCo 1.0 and ZuCo 2.0 corpora (Hollenstein et al., 2018; 2020), which were recorded from adult participants who gave informed consent, as described in the original publications; no new data were collected. All experiments operate on de-identified word-level features. Decoding language from brain signals concerns mental privacy, and the systems studied here decode text only from recordings made with the cooperation of the reader during a reading task.

Reproducibility statement

The code, data splits, training scripts and logged outputs of every run are available in the anonymous repository linked in the abstract. Section A gives the training and evaluation procedure, Section C lists all hyperparameters, and the splits are fixed by a sentence-grouped manifest with split seed 2026; all results use seeds 42, 100 and 312.

AI use statement

In this work, we used generative AI tools to interpret results, by drafting and revising the text that reports them. We have not used generative AI tools to design the methodology or experiments or to implement methods, and generating synthetic data, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to edit parts of the paper, to suggest its structure and to format references. We have reviewed all AI-assisted work: every number in the text, tables and figures was checked against the logged experiment outputs, and every figure against the model implementation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Marcel Bastiaansen and Peter Hagoort. Frequency-based segregation of syntactic and semantic unification during online sentence level language comprehension. *Journal of Cognitive Neuroscience*, 27(11):2095–2107, 2015.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Alexander Craik, Yongtao He, and Jose L. Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, 16(3):031001, 2019.

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 Wang, and Chin-Teng Lin. DeWave: Discrete EEG waves encoding for brain dynamics to text translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Alex Graves, Abdel-Rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013. doi: 10.1109/ICASSP.2013.6638947.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

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

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

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pp. 538–549, 2019. doi: 10.18653/v1/K19-1050.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proc. 12th Language Resources and Evaluation Conference (LREC)*, pp. 138–146, 2020.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Language Processing*, 29:3451–3460, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proc. 38th Int. Conf. Machine Learning (ICML)*, pp. 4651–4664, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

Ole Jensen, Joern Kaiser, and Jean-Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, 30(7):317–324, 2007.

Mingyu Jeon and Hyobin Kim. DELTA: Language diffusion-based EEG-to-text architecture. *arXiv preprint arXiv:2511.21746*, 2025.

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

Wolfgang Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606–617, 2012.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. 58th Annual Meeting of the ACL*, pp. 7871–7880, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

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

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. EEG2TEXT: Open vocabulary EEG-to-text decoding with EEG pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, 2nd edition, 2014.

Joseph G. Makin, David A. Moses, and Edward F. Chang. Machine translation of cortical activity to text with an encoder-decoder framework. *Nature Neuroscience*, 23(4):575–582, 2020. doi: 10.1038/s41593-020-0608-8.

Sean L. Metzger, Kaylo T. Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael A. Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K. Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620:1037–1046, 2023.

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, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Saydul Akbar Murad, Ashim Dahal, and Nick Rahimi. EEG-to-text translation: A model for deciphering human brain activity. *arXiv preprint arXiv:2505.13936*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the ACL*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Maja Popović. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

Matt Post. A call for clarity in reporting BLEU scores. In *Proc. 3rd Conf. Machine Translation (WMT)*, pp. 186–191, 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

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

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020. doi: 10.1109/TAFFC.2018.2817622.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, et al. Going deeper with convolutions. In *Proc. IEEE/CVF CVPR*, pp. 1–9, 2015.

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.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, et al. Multimodal transformer for unaligned multimodal language sequences. In *Proc. ACL*, pp. 6558–6569, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

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

Sabine Weiss and Horst M. Mueller. The contribution of EEG coherence to the investigation of language. *Brain and Language*, 85(2):325–343, 2003.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. *Nature*, 620:1031–1036, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. 32nd Int. Conf. Machine Learning (ICML)*, pp. 2048–2057, 2015.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proc. 37th Int. Conf. Machine Learning (ICML)*, pp. 11328–11339, 2020.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

Jiaqi Zhou, Yiqun Duan, Yu Zou, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. BELT: Bootstrapping electroencephalography-to-language decoding and zero-shot sentiment classification by natural language supervision. *arXiv preprint arXiv:2309.12056*, 2023.

A Training Procedure

Algorithm 1 SpectraLoom training and evaluation

Require: sentence-grouped manifest, seed s, pretrained BART-large

fit per-feature normalisation on the ZuCo 1.0 training split
initialise SBA, convolutions, BiLSTM (Xavier and orthogonal) and CAB; seed all generators
unfreeze BART token embeddings and the first four encoder and decoder layers
for epoch $= 1$ to 30 **do**
  for each batch **do**
    $\mathbf{Z} \leftarrow \mathrm{SBA}(\mathrm{LN}(\tilde{\mathbf{X}}) + \boldsymbol{\eta})$ ▷ Equation (2)
    $\hat{\mathbf{H}} \leftarrow \mathrm{BiLSTM}(\mathrm{Conv}_{3,5,7}(\mathbf{Z})) + \mathbf{P}$ ▷ Equations (3) and (4)
    $\mathbf{O} \leftarrow \mathrm{CAB}(\hat{\mathbf{H}})$; loss $\leftarrow$ cross-entropy of BART on $(\mathbf{O}, \mathbf{Y})$ ▷ Equations (1) and (5)
    accumulate gradients; update every 8 batches (EEG modules 2×10^{-5}, BART 2×10^{-6})
  end for
  evaluate development loss; keep the best checkpoint; stop after 5 epochs without improvement
end for
decode test and ZuCo 2.0 with five-beam search and the development-selected settings

B Input Controls

Table 4 reports the input controls recommended by Jo et al. (2024), applied as training-time interventions in the manner of remove-and-retrain evaluation (Hooker et al., 2019). SpectraLoom is retrained and evaluated with every EEG value replaced by Gaussian noise or by zeros with the valid-position mask retained, or by zeros without the mask.

Table 4: Input controls (seed 312). SpectraLoom is retrained and evaluated with the input replaced.

	ZuCo 1.0			ZuCo 2.0				
Training input	B-1	B-4	R-L	B-1	B-4	R-L	WER↓	Uniq.
Real EEG	17.80	0.41	16.59	16.29	0.73	16.74	94.92	0.85
Gaussian noise†	16.90	0.38	15.46	16.64	0.88	16.98	93.08	1.30
Zero values†	18.45	0.75	16.53	18.64	1.04	18.35	93.63	0.24
Zero values, no mask	14.71	0.29	15.60	16.18	0.92	19.33	91.37	0.02

† valid-position mask retained.

C Training Hyperparameters

Table 5 summarises the configuration shared by all runs.

Table 5: Training and decoding configuration.

Setting	Value
Input	840 gaze-duration features (8 bands $\times$ 105 electrodes), $L_{\max} = 56$
SBA	shared query $\mathbb{R}^{105}$, shared LayerNorm, gate scale $B = 8$ (315 parameters)
Convolutions	kernels 3, 5, 7; 280 channels each; GELU; LayerNorm; dropout 0.3
BiLSTM	3 layers, 512 units per direction, packed sequences, sinusoidal positions
CAB	EEG-derived queries, 8 heads, FFN $1024 \rightarrow 2048 \rightarrow 1024$, dropout 0.3
Language model	BART-large; embeddings and first 4 encoder and decoder layers trainable
Optimiser	AdamW, weight decay 0.01; LR 2×10^{-5} (EEG) and 2×10^{-6} (BART)
Schedule	10% linear warm-up, cosine decay; gradient clipping 1.0
Batch	8×8 accumulation $= 64$; BF16
Stopping	at most 30 epochs; patience 5 on development loss
Decoding	5 beams, max 32 tokens, length pen. 1.4, repetition pen. 1.5, no repeated bigrams
Seeds	42, 100, 312 (split seed 2026)