NEUROLENS: SIGNAL-RELIANT EEG-TO-TEXT DECODING WITH STRUCTURED STATE SPACE ENCODING AND RELIANCE-AWARE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary EEG-to-text decoding generates the sentence a participant is reading from the neural recording alone, and underpins assistive communication and brain–computer interfaces. Progress on this task has come from stronger EEG encoders paired with a pretrained language decoder, yet once evaluation is made rigorous, with sentence-disjoint splits and no teacher forcing, absolute scores are low and it is unclear how much of each score is read from the brain signal and how much is supplied by the decoder's language prior. We identify three limitations of current systems: the encoders in use are poorly matched to short, noisy, word-level EEG sequences; the token-level objective is satisfied by fluent text whether or not the decoder reads the signal; and the contribution of the language prior is neither controlled nor measured. We propose NeuroLens, in which each component answers one of the above limitations: (1) a bidirectional S4D encoder with HiPPO-initialised long-range memory; (2) a reliance-aware objective that combines shuffled-EEG discrimination, contrastive alignment and a vocabulary prior, so that dependence on the paired recording is rewarded directly; and (3) decoder prior dampening, a tunable coefficient on decoder self-attention, together with a verification protocol that replaces the EEG input with shuffled, noise and zero signals over three seeds. On ZuCo 1.0 under a sentence-disjoint protocol without teacher forcing, NeuroLens reaches BLEU-1 of 22.44 and BLEU-4 of 1.46, improving on the previous state-of-the-art by 92% on BLEU-4, and its BLEU-4 falls by 68% when the EEG is shuffled across sentences and by 89% when it is zeroed, which shows that the outputs depend on the recorded signal. A matched-protocol ablation and a dampening sweep identify which components carry the gain and which do not. Code is available at `https://github.com/anonymous399764/AnonymousEEG2TEXT`.

1 INTRODUCTION

Open-vocabulary EEG-to-text decoding reconstructs the sentence a participant is reading from the electroencephalogram recorded during reading. EEG is the modality of practical interest because it is non-invasive, portable and resolves the millisecond-scale dynamics of language comprehension (Luck, 2014), and open-vocabulary decoding is the step from closed-set classification (Hollenstein et al., 2019) to a system that could support communication. Since Wang & Ji (2022) cast the task as sequence-to-sequence generation by replacing the text encoder of BART (Lewis et al., 2020) with a projection of word-level EEG features from ZuCo (Hollenstein et al., 2018; 2020), progress has come from better EEG representations: discrete codebooks (Duan et al., 2023), multi-view encoding (Liu et al., 2024), bidirectional pretraining (Zhou et al., 2023) and diffusion-based decoding (Jeon & Kim, 2025).

In our study, we find that a stronger encoder alone does not settle the question that matters for this task, namely how much of a reported score is decoded from the brain signal. Jo et al. (2024) showed that gains reported under teacher forcing or with sentence overlap largely vanish under a sentence-disjoint protocol without teacher forcing, and that some systems score as well on noise as on EEG. We identify three causes, illustrated in Figure 1. ① *Encoder mismatch*: word-level EEG sequences are short (at most 56 words), high-dimensional (840 features per word) and noisy; linear

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.

projections have no temporal model, recurrent encoders have no principled long-range memory, and Transformer encoders spend quadratic attention on a sequence too short and too noisy to reward it. Encoders that differ little on unigram overlap differ by up to 7× on 4-gram overlap (Figure 1c). ② *Prior-satisfiable objective*: a pretrained decoder can lower token-level cross-entropy by producing fluent text from its language prior, so nothing in standard training rewards dependence on the paired recording; even for our final model, half of the unigram overlap survives when the EEG is zeroed (Figure 1b). ③ *Unmeasured prior contribution*: the decoder's self-attention over previously generated tokens carries the prior, its share of the score is never controlled or measured, and a model whose outputs barely change when the recording is replaced would still be reported as decoding; Figure 1a shows what a model that does read the signal produces when its recording is swapped for another.

We propose NeuroLens, in which each component answers one cause (Figure 2). (1) A *bidirectional S4D encoder* (Gu et al., 2022b) models the word sequence as a diagonal state space system with HiPPO-LegS initialisation (Gu et al., 2020), which gives principled long-range memory at $\mathcal{O}(L \log L)$ cost and, in our comparison, outperforms BiLSTM, Transformer and five alternative state space encoders (cause ①). (2) A *reliance-aware objective* (RAO) adds to cross-entropy a shuffled-EEG discrimination loss that requires the model to fit its own recording better than a permuted one, symmetric InfoNCE alignment and a vocabulary prior head, together with a warm-start attention gate and a two-phase curriculum that keep the EEG pathway open while the encoder is learned (cause ②). (3) *Decoder prior dampening* (DPD) scales decoder self-attention by a coefficient α_{sa} that controls how much of the prior reaches generation, and a *verification protocol* measures what remains by replacing the EEG at test time with shuffled, noise and zero inputs over three training seeds (cause ③).

Our main contributions are summarised as follows:

- We propose NeuroLens, which pairs a bidirectional S4D encoder with a reliance-aware objective and a tunable decoder prior dampening coefficient, and to our knowledge is the first EEG-to-text system whose signal dependence is verified with input controls over multiple training seeds.
- Under sentence-disjoint evaluation without teacher forcing on ZuCo 1.0, NeuroLens attains BLEU-1 of 22.44 and BLEU-4 of 1.46 over three seeds, above every published baseline on every metric, with a margin over the strongest baseline that widens from $+2\%$ on BLEU-1 to $+92\%$ on BLEU-4.
- We quantify the prior's share: replacing the EEG with shuffled recordings lowers BLEU-4 by 68% and zeroing it by 89%, while only 2.6% of outputs are unchanged under shuffling; a matched-protocol ablation shows that the shuffled-EEG discrimination and vocabulary-prior terms each raise BLEU-4 by 16–19%, that S4D exceeds BiLSTM and Transformer encoders by 8–12%, and that partial dampening at $\alpha_{\mathrm{sa}} = 0.1$ raises BLEU-4 by 29% over no dampening.

2 Related Work

Open-vocabulary EEG-to-text decoding. Language decoding from neural signals has been demonstrated with intracranial recordings, from closed-vocabulary speech decoding (Makin et al., 2020; Moses et al., 2021) to sentence-level neuroprostheses (Willett et al., 2023; Metzger et al., 2023), and from fMRI (Tang et al., 2023). With non-invasive EEG the signal is far noisier. Early EEG work classified words or relations from ZuCo (Hollenstein et al., 2019), and Wang & Ji (2022) established open-vocabulary generation by feeding word-level EEG features to a pretrained BART decoder. DeWave (Duan et al., 2023) replaced the projection by a VQ-VAE codebook with contrastive alignment, BELT (Zhou et al., 2023) pretrained a bidirectional EEG encoder, EEG2Text (Liu et al., 2024) introduced multi-view Transformer encoding with self-supervised pretraining, NeuSpeech (Yang et al., 2024) adapted a speech recognition model to MEG, and DELTA (Jeon & Kim, 2025) replaced autoregressive decoding by a masked diffusion language model over discretised EEG tokens. Jo et al. (2024) showed that many reported gains depend on teacher forcing and sentence overlap and proposed the sentence-disjoint, free-decoding protocol we adopt, and Murad et al. (2025) re-evaluated several systems under it. All of these systems improve the encoder or the alignment and train with token-level cross-entropy alone; none verifies signal dependence with input controls, which is the evidence a reader needs after the critique of Jo et al. (2024).

State space encoders. Structured state space models represent a sequence as a linear dynamical system whose convolution kernel can be computed efficiently (Gu et al., 2022a); HiPPO initialisation gives the state a principled memory of the input history (Gu et al., 2020). S4D restricts the state

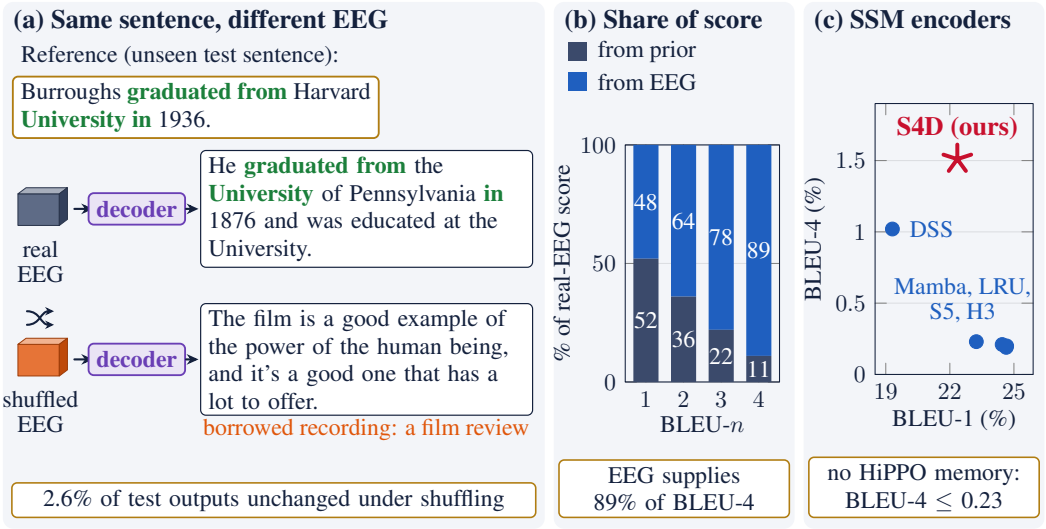


Figure 1: Signal reliance on ZuCo 1.0. (a) The same trained model decodes the same unseen test sentence from its own recording and from a recording shuffled in from another sentence; matched words in green. (b) Share of each BLEU score supplied by the language prior (the score with zero EEG) and read from the EEG (the rest), mean of three seeds. (c) Six state space encoders under identical training: unigram overlap barely separates them, 4-gram overlap does.

matrix to diagonal form (Gu et al., 2022b), DSS shows that diagonal systems suffice (Gupta et al., 2022), S5 uses a multi-input formulation with a parallel scan (Smith et al., 2023), H3 adds a shift state space for language modelling (Fu et al., 2023), LRU studies linear recurrences from a deep learning viewpoint (Orvieto et al., 2023) and Mamba adds a causal selective scan (Gu & Dao, 2023). These models were developed for long sequences such as Long Range Arena benchmarks and raw audio (Tay et al., 2021; Goel et al., 2022); we show that their long-range memory also helps on short, noisy biosignal sequences, and compare S4D with five of the alternatives under identical training.

Cross-modal alignment and reliance on the conditioning input. Contrastive alignment in the style of CLIP (Radford et al., 2021; van den Oord et al., 2018) is effective for audio (Wu et al., 2023), video (Xu et al., 2021), multi-sensor (Girdhar et al., 2023) and MEG (Défossez et al., 2023) inputs, and is most needed when one modality is much noisier than the other (Shi et al., 2022). A separate line of work studies conditional generators that produce output without reading their input: object hallucination in image captioning (Rohrbach et al., 2018), unfaithful attention (Jain & Wallace, 2019), exposure bias (Ranzato et al., 2016; Bengio et al., 2015) and degenerate decoding (Holtzman et al., 2020). Input-ablation tests (Hooker et al., 2019; DeYoung et al., 2020) and word dropout on the decoder input (Bowman et al., 2016) are the standard remedies for measuring and reducing such reliance. NeuroLens brings these tools to EEG-to-text: word dropout and self-attention dampening reduce the prior's contribution during training, and input replacement measures what remains at test time.

3 Methodology

Overview. As shown in Figure 2, NeuroLens has three parts. The *bidirectional S4D encoder* (Section 3.2) maps the word-level EEG sequence to states that replace the text-encoder output of a BART decoder. The *reliance-aware objective* (Section 3.3) trains encoder and decoder so that fitting the target from the paired recording is rewarded and fitting it from the prior is not, with a warm-start gate and a two-phase curriculum that keep the EEG pathway open. *Decoder prior dampening* (Section 3.4) scales decoder self-attention by α_{sa}, and the *verification protocol* (Section 3.4) measures the prior's remaining share at test time.

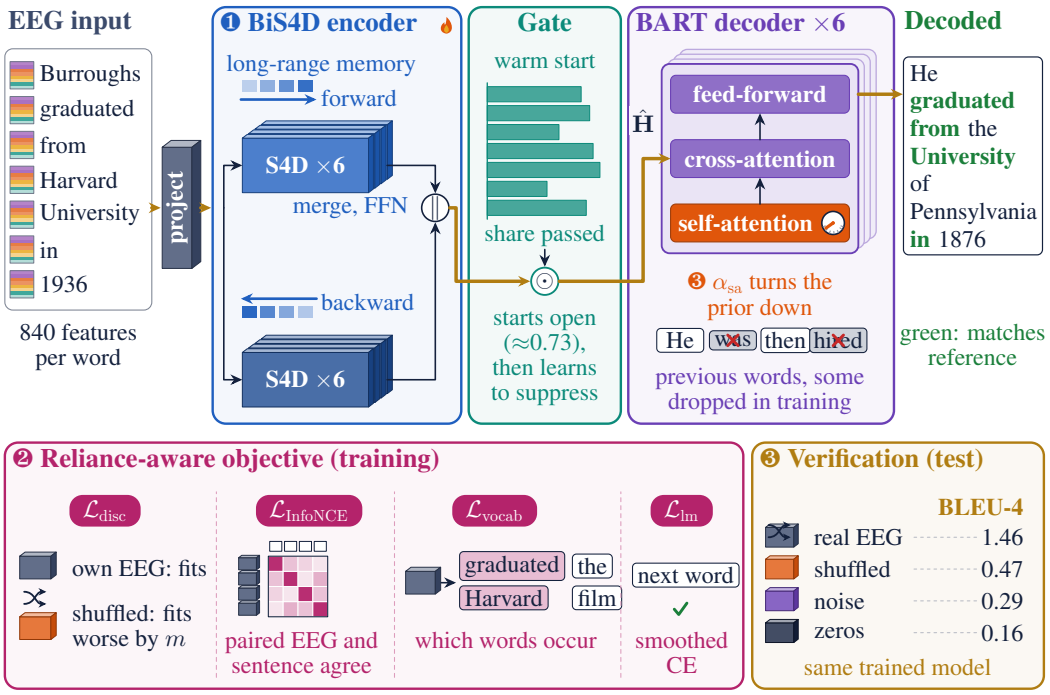


Figure 2: Overview of NeuroLens. Word-level EEG features are encoded by a bidirectional S4D stack (①), passed through a warm-start gate that sets how much of each word reaches the decoder, and read by the BART decoder through cross-attention while decoder self-attention, the path of the language prior, is scaled by α_{sa} (③). Bottom: the three terms of the reliance-aware objective (②) and the verification protocol with the measured BLEU-4 under each input replacement.

3.1 Preliminaries

Task. A sentence of L words is read while EEG is recorded. Each word yields a feature vector $\mathbf{e}_i \in \mathbb{R}^{840}$ that concatenates the power of eight frequency bands $(\theta_{1,2}, \alpha_{1,2}, \beta_{1,2}, \gamma_{1,2})$ over 105 channels, averaged over the first-pass fixation of the word (gaze-duration features). Given $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_L]$, the model generates the sentence $\mathbf{Y} = [y_1, \ldots, y_T]$ with an open vocabulary. An encoder maps $\mathbf{E}$ to states $\mathbf{H}$, and a pretrained decoder models $p(y_t \mid y_{<t}, \mathbf{H})$ through self-attention over $y_{<t}$ and cross-attention over $\mathbf{H}$.

The prior's share. Because the decoder is pretrained on text, $p(y_t \mid y_{<t})$ alone already produces fluent English, and a generated sentence can overlap the reference through the prior rather than through $\mathbf{H}$. We call the part of a score that survives when $\mathbf{H}$ carries no information about the sentence the prior's share. It can be measured from outside the model by replacing the EEG input at test time with a recording of a different sentence, with Gaussian noise, or with zeros, and it is the quantity that the objective in Section 3.3 is designed to reduce and that Section 3.4 reports.

3.2 Bidirectional S4D Encoder

Observation 1. *Word-level EEG rewards long-range memory, not attention.* With all other components fixed, S4D exceeds a Transformer encoder by 12% and a BiLSTM by 8% on BLEU-4, and among six state space encoders the four without HiPPO-style memory (Mamba, S5, H3, LRU) reach the highest BLEU-1 but the lowest BLEU-4, up to 7× below S4D (Tables 3 and 7). Unigram overlap is available from the prior; phrase-level overlap requires the encoder to carry the order of the word sequence to the decoder, which is what a HiPPO-initialised state does by construction (cause ①).

Technique. Features are projected to the working dimension $d = 512$, $\mathbf{h}_i^{(0)} = \mathrm{LayerNorm}(\mathrm{GELU}(\mathbf{W}_{\mathrm{in}}\mathbf{e}_i + \mathbf{b}_{\mathrm{in}}))$, and passed through six S4D layers. Each layer discretises the diagonal system $h'(t) = Ah(t) + Bx(t)$, $y(t) = Ch(t) + Dx(t)$ with a learned step size Δ by

the bilinear transform,

$$\bar{A} = \frac{1 + \Delta A/2}{1 - \Delta A/2}, \qquad \bar{B} = \frac{\Delta B}{1 - \Delta A/2}, \qquad \mathbf{K} = \big(C\bar{B},\, C\bar{A}\bar{B},\, \ldots,\, C\bar{A}^{L-1}\bar{B}\big), \tag{1}$$

constrains the eigenvalues of $A = \mathrm{diag}(a_1, \ldots, a_N)$ to have negative real parts, initialises them with HiPPO-LegS, $a_n = -\frac{1}{2} + i\pi(n + \frac{1}{2})$, and evaluates the convolution $y = \mathbf{K} * x$ by FFT at $\mathcal{O}(L \log L)$ per layer, against the $\mathcal{O}(L^2)$ attention of existing EEG encoders (Wang & Ji, 2022; Liu et al., 2024; Murad et al., 2025). Each layer is followed by dropout, a GLU activation, layer normalisation and a residual connection. Two independent stacks scan the word sequence forwards and backwards, and their outputs are concatenated and passed through a position-wise feed-forward sublayer to give $\mathbf{h}_{\mathrm{enc},i} \in \mathbb{R}^{768}$, the hidden width of BART-base.

Warm-start attention gate. A position-wise gate between encoder and decoder,

$$\mathbf{g}_i = \sigma(\mathbf{W}_g \mathbf{h}_{\mathrm{enc},i} + \mathbf{b}_g)\,, \qquad \hat{\mathbf{h}}_i = \mathbf{g}_i \odot \mathbf{h}_{\mathrm{enc},i} + (1 - \mathbf{g}_i) \odot \mathbf{v}_{\mathrm{null}}, \tag{2}$$

with a learnable null vector $\mathbf{v}_{\mathrm{null}}$ and bias initialised to $\mathbf{b}_g = +1$ (initial openness $\sigma(1) \approx 0.73$), passes the encoder output to the decoder from the first step and can suppress individual positions once it has evidence to do so. The gated states $\hat{\mathbf{H}}$ replace the output of BART's text encoder as keys and values of every decoder cross-attention layer. The gate is an architectural convenience rather than a source of gain: under the matched ablation protocol its removal does not lower BLEU-4 (Section 4.4).

3.3 Reliance-Aware Objective

Observation 2. *Cross-entropy alone does not reward reading the signal.* With the EEG input set to zero, the trained model still reaches a BLEU-1 of 11.67, half of its score on real EEG, and a BERTScore of 64.8 against 69.3 (Table 2): a large part of unigram and semantic overlap is obtainable from the prior without reading the recording, and token-level cross-entropy rewards it equally (cause ②). The objective must therefore reward dependence on the *paired* recording explicitly. The RAO adds three terms to the label-smoothed language modelling loss $\mathcal{L}_{\mathrm{lm}}$ ($\varepsilon = 0.1$),

$$\mathcal{L} = \mathcal{L}_{\mathrm{lm}} + \lambda_d \mathcal{L}_{\mathrm{disc}} + \lambda_c \mathcal{L}_{\mathrm{InfoNCE}} + \lambda_v \mathcal{L}_{\mathrm{vocab}}, \tag{3}$$

with $(\lambda_d, \lambda_c, \lambda_v) = (0.5, 3.0, 1.2)$.

Shuffled-EEG discrimination. The most direct term penalises the model when it fits the target as well from a mismatched recording as from its own. Within each batch the EEG inputs and their masks are permuted so that no sample keeps its recording, the decoder is run a second time on the permuted input, and the two cross-entropies are compared with a margin,

$$\mathcal{L}_{\mathrm{disc}} = \max(0,\; m - (\mathcal{L}_{\mathrm{shuffled}} - \mathcal{L}_{\mathrm{real}}))\,, \qquad m = 1.0, \tag{4}$$

where $\mathcal{L}_{\mathrm{shuffled}}$ is detached so that the gradient lowers $\mathcal{L}_{\mathrm{real}}$ rather than raising the loss on the permuted input. The term is zero once the paired recording is at least m nats more useful than a foreign one, and it is the training-time counterpart of the shuffled-EEG control in Section 3.4.

Alignment and vocabulary prior. Following CLIP (Radford et al., 2021), mean-pooled encoder states and mean-pooled states of BART's frozen text encoder are mapped by two-layer MLP heads to unit vectors $\mathbf{z}_{\mathrm{eeg}}, \mathbf{z}_{\mathrm{text}} \in \mathbb{R}^{256}$ and aligned with a symmetric InfoNCE loss,

$$\mathcal{L}_{\mathrm{InfoNCE}} = \tfrac{1}{2}\big[\mathrm{CE}(S/\tau, I) + \mathrm{CE}(S^\top/\tau, I)\big], \qquad S_{ij} = \mathbf{z}_{\mathrm{eeg},i}^\top \mathbf{z}_{\mathrm{text},j},\; \tau = 0.07, \tag{5}$$

with text embeddings detached so that the alignment does not disturb BART's representations. A bag-of-words head predicts which vocabulary items occur in the target sentence from the pooled encoder state, reusing BART's embedding matrix, and is trained with binary cross-entropy ($\mathcal{L}_{\mathrm{vocab}}$); it gives the encoder a supervision signal that does not pass through the decoder's prior. The definitions of $\mathcal{L}_{\mathrm{lm}}$ and $\mathcal{L}_{\mathrm{vocab}}$ are given in Section A.

Two-phase curriculum. If the decoder is fine-tuned before the encoder is informative, it fits the target from its prior immediately. Phase 1 (30 epochs) therefore freezes BART and trains the encoder, gate and heads (12.1M parameters) with $\mathcal{L}_{\mathrm{lm}}$, $\mathcal{L}_{\mathrm{InfoNCE}}$ and $\mathcal{L}_{\mathrm{vocab}}$ at a learning rate of 8×10^{-5} and $\alpha_{\mathrm{sa}} = 1$. Phase 2 (70 epochs) unfreezes the decoder cross-attention, feed-forward and normalisation layers (about 40M trainable parameters), activates $\mathcal{L}_{\mathrm{disc}}$ and word dropout on the decoder input, which rises linearly from 35% to 85% (Bowman et al., 2016), and optimises the full objective at 2.5×10^{-5} with α_{sa} set to its operating value. The procedure is listed in Section B.

3.4 Decoder Prior Dampening and Verification

Motivation. The prior enters generation through decoder self-attention over $y_{<t}$. Rather than leave its contribution implicit, DPD makes it a coefficient. Forward hooks scale the output of every decoder self-attention layer,

$$\mathrm{SelfAttn}_{\mathrm{out}} \leftarrow \alpha_{\mathrm{sa}} \cdot \mathrm{SelfAttn}_{\mathrm{out}}, \tag{6}$$

in Phase 2 and at inference. With $\alpha_{\mathrm{sa}} = 0$ the residual stream at step t still carries the embedding and position of y_{t-1}, so the decoder keeps first-order local context and loses only the mixing over the longer history through which the prior acts; word dropout then weakens the remaining dependence on y_{t-1}. Our main configuration uses $\alpha_{\mathrm{sa}} = 0$. The sweep in Section 4.4 shows that the coefficient should be tuned: under the matched ablation protocol $\alpha_{\mathrm{sa}} = 0.1$ gives the highest BLEU-4, 29% above no dampening, whereas full suppression and no suppression are close, so we present DPD as a regulariser of the prior's share rather than as a requirement for full suppression.

Verification protocol. Whatever the training recipe, the reader needs to know how much of a score is read from the signal. We therefore evaluate every trained model on the test set four times, with the real EEG, with EEG shuffled across samples so that each sentence receives the recording of another, with Gaussian noise of unit variance in normalised feature space, and with zeros, and we report the drop on each metric together with the fraction of outputs that are identical under real and shuffled EEG. All controls are computed with the same checkpoints, decoding settings and normalisation statistics as the main result, over three training seeds, and the same protocol is applied to a model trained without dampening so that the effect of α_{sa} on reliance can be read off directly.

4 Experiments

4.1 Settings

Data and protocol. We use ZuCo 1.0 (Hollenstein et al., 2018), which provides simultaneous EEG and eye-tracking from 12 participants over three reading paradigms, giving 12,579 word-level EEG sequences for 1,039 unique sentences. Following Jo et al. (2024), the sentences are split into train, development and test groups with no text overlap (831/103/105 sentences; 9,978/1,285/1,316 samples), so that every recording of a sentence falls in one split; per-channel normalisation statistics are computed on the training split only. All main results use greedy decoding without teacher forcing (repetition penalty 1.3, no repeated trigrams). We additionally report a subject-wise split (9/1/2 participants) and zero-shot transfer to ZuCo 2.0 (Hollenstein et al., 2020).

Baselines and metrics. R1Translator, T5Translator and BrainTranslator results are taken from Murad et al. (2025), who use the same sentence-disjoint, no-teacher-forcing protocol and correct the teacher-forcing issue identified by Jo et al. (2024); DELTA (Jeon & Kim, 2025) also uses sentence-disjoint segmentation. We report BLEU-1 to BLEU-4 (Papineni et al., 2002) via SacreBLEU (Post, 2018), ROUGE (Lin, 2004) and BERTScore F_1 (Zhang et al., 2020). Higher-order BLEU is emphasised because unigram overlap is available from the prior while phrase-level overlap is not.

Implementation. NeuroLens has 151.5M parameters (S4D dimension 512, state size 64, six layers, BART-base decoder) and is trained in PyTorch (Paszke et al., 2019) with FP16 mixed precision on a single NVIDIA RTX 4060 (8 GB), using AdamW (Loshchilov & Hutter, 2019) (weight decay 0.01, cosine schedule with 500 warm-up steps, gradient clipping at 1.0), an effective batch of 128, Gaussian feature noise ($\sigma = 0.05$) and channel dropout (2%) as augmentation, and gradient checkpointing in the decoder. The checkpoint with the best development BLEU-1 is selected. Main results are the mean and standard deviation over three training seeds (42, 99, 34); the ablation, the encoder comparison and the dampening sweep are single-seed runs (seed 42) under a fixed protocol described in Section 4.4.

4.2 Main Results

As shown in Table 1, NeuroLens outperforms all baselines on every metric. The margin over the strongest baseline, DELTA, is small on BLEU-1 ($+0.51$, or 2%) and widens with n-gram order

Table 1: Comparison with the state-of-the-art without teacher forcing (sentence-disjoint split, ZuCo 1.0). NeuroLens: mean ± standard deviation over three seeds.

Model	BLEU-N (%)				ROUGE-1 (%)			BERTScore F_1
	N=1	N=2	N=3	N=4	P	R	F	
R1Translator (2025)	12.16	3.22	1.12	0.34	12.01	9.51	9.96	—*
T5Translator (2024)	13.98	3.53	0.92	0.27	10.08	9.27	9.03	—*
BrainTranslator (2024)	14.10	2.62	0.66	0.25	9.22	11.91	9.75	—*
DELTA† (2025)	21.93	6.43	2.01	0.76	18.88	18.86	17.24	—*
NeuroLens (ours)	$22.44_{\pm 0.33}$	$7.98_{\pm 0.26}$	$2.84_{\pm 0.06}$	$1.46_{\pm 0.07}$	$19.07_{\pm 0.48}$	$19.10_{\pm 0.21}$	$18.77_{\pm 0.17}$	$69.30_{\pm 0.72}$

† DELTA uses a large diffusion language model backbone at substantially higher compute than NeuroLens. * BERTScore is not reported for the baselines, which do not release outputs under this protocol.

Table 2: EEG input controls at test time (sentence-disjoint split, greedy decoding). NeuroLens: mean over three seeds; the undampened model ($\alpha_{\text{sa}} = 1$ in training and inference) is a single seed. Percentages are relative to the real-EEG row of the same model. Identical: fraction of outputs unchanged under shuffling.

Model	EEG input	B-1	B-2	B-3	B-4	R-L F	BERT
NeuroLens	Real EEG	22.44	7.98	2.84	1.46	15.4	69.3
	Shuffled EEG	17.38 (−23%)	4.86 (−39%)	1.33 (−53%)	0.47 (−68%)	11.1	65.5
	Gaussian noise	13.58 (−39%)	3.84 (−52%)	0.98 (−65%)	0.29 (−80%)	10.5	66.1
	Zero EEG	11.67 (−48%)	2.88 (−64%)	0.64 (−77%)	0.16 (−89%)	10.6	64.8
NeuroLens, $\alpha_{\text{sa}} = 1$	Real EEG	20.30	6.02	2.23	1.00	14.2	—
	Shuffled EEG	15.98 (−21%)	3.80 (−37%)	1.11 (−50%)	0.45 (−55%)	10.5	—
	Gaussian noise	12.15 (−40%)	2.83 (−53%)	0.66 (−70%)	0.18 (−82%)	9.4	—
	Zero EEG	8.30 (−59%)	0.89 (−85%)	0.11 (−95%)	0.03 (−97%)	7.8	—
Identical outputs, real against shuffled		2.6% (NeuroLens) and 2.7% ($\alpha_{\text{sa}} = 1$)					

to +24% on BLEU-2, +41% on BLEU-3 and +92% on BLEU-4 (1.46 against 0.76), with a seed standard deviation of 0.07 on BLEU-4. A system that leans on its language prior can sustain unigram overlap while failing at phrase-level conditioning, so a margin that grows with n-gram order is the signature of phrase-level information read from the signal. DELTA uses a large diffusion language model at substantially higher compute, which indicates that capacity alone does not close this gap. Under teacher forcing, NeuroLens matches the best ROUGE-1 F and attains the best ROUGE-2 among seven systems (Section C).

4.3 Signal Reliance

Table 2 applies the verification protocol. Relative to real EEG, BLEU-1 falls by 23% when the recordings are shuffled across sentences, by 39% under Gaussian noise and by 48% under zeros, and BLEU-4 falls by 68%, 80% and 89%, from 1.46 to 0.47, 0.29 and 0.16. The drop under shuffling is the decisive one: shuffled inputs are genuine EEG with the same distribution as the real inputs, so the loss can only come from sentence-specific content, which answers the critique of Jo et al. (2024) for this model. Only 2.6% of the 1,316 test outputs are identical under real and shuffled EEG. The drops grow with n-gram order on every control, which quantifies the prior's share: about half of unigram overlap and about a tenth of 4-gram overlap survive without the signal, so higher-order BLEU is the metric on which decoding from EEG is measured. The model trained without dampening shows the same pattern, with BLEU-4 falling by 55% under shuffling, so reliance on the signal is produced by the objective and the curriculum rather than by α_{sa} alone; dampening raises the score obtained from the signal, from 1.00 to 1.46 BLEU-4 at $\alpha_{\text{sa}} = 0$ under the main protocol, and Section 4.4 examines its strength.

4.4 Ablation Study

Protocol. The ablation, the encoder comparison and the dampening sweep were run under one fixed protocol, with the training seed, the split and the decoding settings held constant and the checkpoint selected by development BERTScore, so that every row is comparable to every other. Under this protocol the full model scores 0.96 BLEU-4, against 1.46 in Table 1, where the checkpoint is selected by development BLEU-1; every row is therefore compared with the full model of the same protocol.

Table 3: Ablation under a matched protocol (sentence-disjoint split, seed 42, checkpoint selected by development BERTScore, greedy decoding). Every row is trained and selected identically; the full model under this protocol differs from Table 1 only in checkpoint selection. Δ is relative to the full model on BLEU-4.

Variant	B-1	B-2	B-3	B-4	ΔB-4	R-L F	BERT	WER
Full NeuroLens	19.96	6.26	2.21	0.96	—	12.91	69.92	1.025
Full NeuroLens, replicate run	19.36	5.74	1.84	0.76	−21%	12.54	69.81	1.016
w/o vocabulary prior $\mathcal{L}_{\text{vocab}}$	19.35	5.87	1.95	0.78	−19%	12.53	69.82	1.016
w/o shuffled-EEG discrimination $\mathcal{L}_{\text{disc}}$	19.52	5.80	1.86	0.81	−16%	12.57	69.85	1.021
w/o prior dampening ($\alpha_{\text{sa}} = 1$)	20.14	5.61	1.90	0.83	−14%	14.25	69.53	1.000
w/o word dropout	20.28	6.37	2.34	1.01	+5%	13.07	69.94	1.030
w/o InfoNCE $\mathcal{L}_{\text{InfoNCE}}$	19.97	6.32	2.34	1.04	+8%	12.93	70.18	1.026
w/o attention gate	21.28	6.87	2.51	1.11	+16%	13.77	69.99	1.021
S4D → Transformer encoder	20.10	6.20	2.08	0.84	−12%	12.72	69.97	1.031
S4D → BiLSTM	20.76	6.42	2.19	0.88	−8%	12.77	69.75	1.054
S4D → linear projection	21.56	6.99	2.46	1.00	+4%	14.01	69.35	1.132

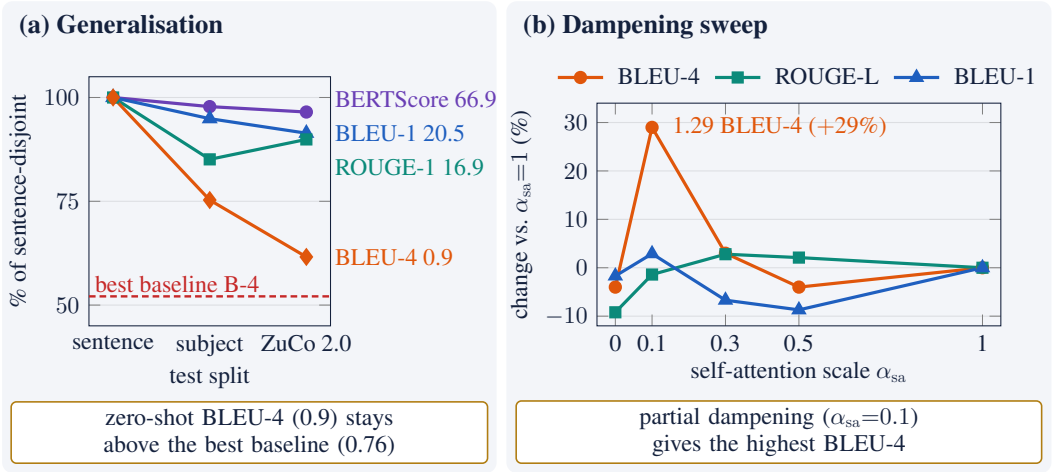
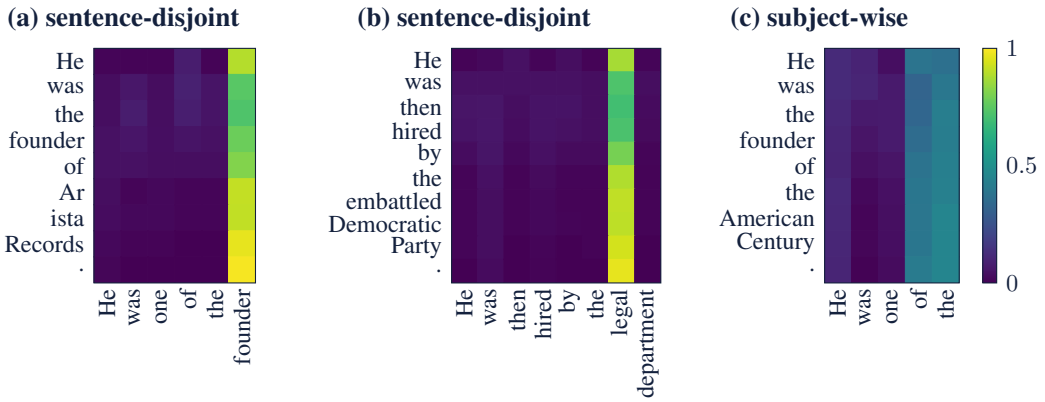


Figure 3: Generalisation and dampening strength. (a) Score retained under sentence-disjoint, subject-wise and ZuCo 2.0 zero-shot evaluation. (b) Dampening sweep under the matched protocol.

Objective terms. Table 3 shows that the two terms that act on how the decoder uses the signal each lower BLEU-4 when removed, the vocabulary prior by 19% and shuffled-EEG discrimination by 16%, with BLEU-1 to BLEU-3 and ROUGE-L falling in step, so the direction is consistent across metrics. Removing prior dampening lowers BLEU-2 to BLEU-4 by 10–14% and BERTScore by 0.4 while raising BLEU-1 and ROUGE-L, the pattern expected when the prior is restored: more unigram overlap, less phrase-level overlap. Removing InfoNCE or word dropout changes BLEU-4 by less than 0.1 and removing the attention gate raises it by 0.15; these three components are retained for training stability.

Encoder. Replacing S4D with a Transformer encoder lowers BLEU-4 by 12% and with a BiLSTM by 8%. The linear projection of Wang & Ji (2022) matches S4D on BLEU-4 and exceeds it on BLEU-1, but it has the lowest BERTScore of all variants and a word error rate of 1.13 against 1.03, which indicates longer, less faithful outputs; the state space comparison in Table 7 separates the encoders more sharply at higher n-gram order.

Dampening strength. Figure 3b and Table 6 sweep α_{sa} under the matched protocol. BLEU-4 is highest at $\alpha_{\text{sa}} = 0.1$ (1.29, 29% above no dampening and 34% above full suppression), and BLEU-1 and BLEU-2 are highest there as well, while $\alpha_{\text{sa}} = 0$, 0.5 and 1 are within 0.04 of one another. Retaining a small fraction of self-attention therefore serves the decoder better than removing it. ROUGE-L rises as self-attention is restored, which again reflects the prior's contribution to surface overlap. The main runs use $\alpha_{\text{sa}} = 0$.



target: He was one of the founder members of the Lunar Society.
decoded: **He was** the **founder of** Arista Records.

target: He was then hired by the legal department of CBS subsidiary Columbia Records.
decoded: **He was then hired by the** embattled Democratic Party.

target: He was one of the founder members of the Lunar Society.
decoded: **He was** the **founder of the** American Century.

Figure 4: Cross-attention maps under free greedy decoding (rows: decoded tokens; columns: EEG word positions). Attention concentrates on one or two EEG positions, from which the decoder reads a sentence-level summary of the recording.

4.5 Analysis

State space variants. To test whether the advantage of S4D is specific to its parameterisation, five alternative state space encoders replace it with all other components fixed (Figure 1c, Table 7). Four of them score higher on BLEU-1 but fall up to 7× behind on BLEU-4, the same signature of prior-driven unigram overlap as in Table 1. DSS, the closest counterpart, trails by 23% on BLEU-3 and 32% on BLEU-4, which indicates that HiPPO-LegS initialisation contributes beyond the diagonal formulation.

Generalisation. NeuroLens retains its scores beyond the sentence-disjoint setting (Figure 3a). With the two test participants unseen in training, BLEU-4 is 1.1, above every sentence-disjoint baseline, and zero-shot transfer to ZuCo 2.0, with 18 participants and different reading paradigms, gives a BERTScore of 66.9, a BLEU-1 of 20.5 and a BLEU-4 of 0.9, again above the best baseline (0.76).

Cross-attention behaviour. In Figure 4, attention is sharply peaked on individual EEG word positions rather than spread uniformly, and spreads over structural positions under subject-wise evaluation (Figure 4c). The attention of all output tokens concentrates on the same one or two positions: because the encoder is bidirectional, every encoder state aggregates the whole word sequence, so the decoder reads a sentence-level summary from a single salient position, in agreement with the input controls of Table 2.

Qualitative results. Table 9 gives one decoded example per error category: outputs that keep relational structure while substituting named entities and dates, a sentence decoded almost word for word until its dates diverge, and a failure on a short sentence without a semantic anchor. BERTScore values of 0.652 to 0.861 indicate sentence-level coherence even where lexical fidelity is low.

5 Conclusion

We presented NeuroLens, an EEG-to-text decoder that pairs a bidirectional S4D encoder with a reliance-aware objective and tunable decoder prior dampening, together with a protocol that measures how much of its output depends on the signal. On ZuCo 1.0 under sentence-disjoint evaluation without teacher forcing it exceeds every published baseline, by 92% on BLEU-4 over three seeds, and shuffling or zeroing the EEG removes 68% and 89% of that BLEU-4. The matched-protocol ablation shows which components carry the gain and that partial dampening is preferable to full suppression, and the gains hold for unseen participants and zero-shot on ZuCo 2.0.

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 system studied here decodes text only from recordings made with the cooperation of the reader during a reading task.

REPRODUCIBILITY STATEMENT

Section 4 specifies the sentence-disjoint split, the decoding settings and all optimisation settings, Section A defines every loss term, and Section B gives the two-phase training procedure. All main results are means over three training seeds. The code, the configuration of every reported run, the per-epoch training logs and the evaluation outputs are available in the anonymous repository linked in the abstract.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, by assisting with code implementation, and 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, 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 for readability and to format references. We have reviewed all AI-assisted work: all AI-assisted code was tested and verified by the authors, and every number in the text, tables and figures was checked against the logged results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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*, 2015.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, 2016.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7616–7633, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. doi: 10.48550/arXiv.2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1474–1487, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022a.

Albert Gu, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*, 2022.

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. doi: 10.1038/sdata.2018.291.

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 *Proceedings of the 12th Language Resources and Evaluation Conference*, 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.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

MinGyu Jeon and HyoBin Kim. DELTA: Language diffusion-based EEG-to-text architecture. *arXiv preprint arXiv:2511.21746*, 2025. doi: 10.48550/arXiv.2511.21746.

Hyeji Jo, Youngki Yang, Jinyeong Han, Yiqun Duan, Hao Xiong, and Woo Hyun Lee. Are EEG-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024. doi: 10.48550/arXiv.2405.06459.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. doi: 10.18653/v1/2020.acl-main.703.

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. doi: 10.48550/arXiv.2405.02165.

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

Steven J. Luck. *An Introduction to the Event-Related Potential Technique*. MIT Press, Cambridge, MA, 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. doi: 10.48550/arXiv.2505.13936.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135.

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.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018. doi: 10.18653/v1/W18-6319.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022. doi: 10.1609/aaai.v36i5.20476.

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.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6787–6800, 2021. doi: 10.18653/v1/2021.emnlp-main.544.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

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 Loss Definitions

The language modelling loss applies label smoothing (Szegedy et al., 2016) with $\varepsilon = 0.1$:

$$\mathcal{L}_{\text{lm}} = (1-\varepsilon)\,\text{NLL}(\hat{y}, y) + \varepsilon\Big(-\frac{1}{|V|}\sum_{v}\log p(v)\Big). \tag{7}$$

The vocabulary prior head computes $\mathbf{p}_{\text{vocab}} = \sigma(E_{\text{BART}}\,\text{GELU}(W_v\bar{\mathbf{h}}_{\text{eeg}} + b_v))$, where $E_{\text{BART}} \in \mathbb{R}^{|V|\times 768}$ is BART's shared embedding matrix, and is trained with $\mathcal{L}_{\text{vocab}} = \text{BCE}(\mathbf{p}_{\text{vocab}}, \mathbf{b}_{\text{target}})$ against the bag-of-words vector $\mathbf{b}_{\text{target}} \in \{0,1\}^{|V|}$ of the target sentence. Each projection head is $\text{Linear}(768 \to 768) \to \text{GELU} \to \text{Linear}(768 \to 256)$ followed by L2 normalisation. Table 4 lists weights and active phases.

Table 4: Loss components, weights and training phases.

Loss	Symbol	λ	Phase
Label-smoothed CE	$\mathcal{L}_{\text{lm}}$	1.0	1+2
Symmetric InfoNCE	$\mathcal{L}_{\text{InfoNCE}}$	3.0	1+2
EEG vocabulary prior	$\mathcal{L}_{\text{vocab}}$	1.2	1+2
Shuffled-EEG disc.	$\mathcal{L}_{\text{disc}}$	0.5	2 only

B Training Procedure

Algorithm 1 NeuroLens two-phase training

Require: EEG–sentence pairs $\{(\mathbf{E}, \mathbf{Y})\}$, pretrained BART, dampening coefficient α_{sa}

Phase 1 (30 epochs, $\alpha_{\text{sa}} = 1$, BART frozen)

for each batch **do**
  $\mathbf{H} \leftarrow \text{BiS4D}(\mathbf{E})$; $\hat{\mathbf{H}} \leftarrow \text{Gate}(\mathbf{H})$ ▷ Equations (1) and (2)
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{lm}} + \lambda_c\mathcal{L}_{\text{InfoNCE}} + \lambda_v\mathcal{L}_{\text{vocab}}$; update encoder, gate, heads
end for

Phase 2 (70 epochs, self-attention scaled by α_{sa}, decoder cross-attention, FFN and norms trainable)

for each batch, with word dropout p rising from 0.35 to 0.85 **do**
  $\mathcal{L}_{\text{real}} \leftarrow$ decoder loss on $(\hat{\mathbf{H}}, \mathbf{Y})$; $\mathcal{L}_{\text{shuffled}} \leftarrow$ decoder loss on permuted $\hat{\mathbf{H}}$ (detached)
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{lm}} + \lambda_d\mathcal{L}_{\text{disc}} + \lambda_c\mathcal{L}_{\text{InfoNCE}} + \lambda_v\mathcal{L}_{\text{vocab}}$ ▷ Equation (3)
  update encoder, gate, heads and unfrozen decoder layers
end for

select the checkpoint with the best development BLEU-1

Inference: greedy decoding with self-attention scaled by α_{sa}

Verification: decode the test set with real, shuffled, Gaussian-noise and zero EEG; report drops and the identical-output rate

C Teacher-Forced Comparison

Teacher forcing supplies the ground-truth token history at every step and therefore measures next-token prediction; Table 5 reports this setting for comparison with EEG-To-Text, DeWave and EEG2Text, which evaluate under it.

Table 5: Teacher-forced evaluation on the sentence-disjoint split of ZuCo 1.0.

Model	BLEU-N (%)				ROUGE-1 (%)			ROUGE-2 (%)		
	1	2	3	4	P	R	F	P	R	F
R1 (2025) w/TF	38.62	21.41	11.65	6.15	30.91	25.52	27.79	6.90	6.13	6.45
T5 (2024) w/TF	43.01	24.41	14.15	7.75	27.78	22.63	24.76	5.13	4.51	4.76
Brain (2024) w/TF	38.02	20.43	10.56	5.62	29.57	23.98	26.37	5.44	4.89	5.14
EEG-To-Text (2022) w/TF	40.10	23.10	12.50	6.80	31.70	28.80	30.10	—	—	—
DeWave (2023) w/TF	41.30	24.10	13.90	8.20	33.70	30.60	32.10	—	—	—
EEG2Text (2024) w/TF	45.20	29.10	19.70	14.10	36.90	32.00	34.20	—	—	—
NeuroLens w/TF	39.70	23.30	14.80	9.90	34.80	33.80	34.20	11.20	11.00	11.10

D Additional Results

Tables 6 and 7 list the dampening sweep and the state space encoder comparison plotted in Figure 1c, Table 8 the generalisation settings and Table 9 decoded examples.

Table 6: Dampening sweep (matched protocol, seed 42). The $\alpha_{\text{sa}} = 1$ row is an independent run of the configuration in the "w/o prior dampening" row of Table 3.

α_{sa}	B-1	B-2	B-4	R-L F
0.0 (main setting)	19.96	6.26	0.96	12.9
0.1	20.89	6.55	1.29	14.0
0.3	18.95	5.60	1.03	14.6
0.5	18.54	5.33	0.96	14.5
1.0 (none)	20.30	6.02	1.00	14.2

Table 7: State space encoder variants (seed 42, development BLEU-1 selection, $\alpha_{\text{sa}} = 0.2$ for the variants).

Encoder	B-1	B-2	B-3	B-4
S4D (ours)	22.35	7.91	2.81	1.51
DSS (2022)	19.31	5.74	2.15	1.02
Mamba-tiny (2023)	24.45	6.21	1.01	0.21
LRU-lite (2023)	23.24	6.27	1.14	0.23
S5-lite (2023)	24.66	4.94	0.97	0.20
H3-lite (2023)	24.64	5.70	0.92	0.19

Table 8: Generalisation (greedy decoding).

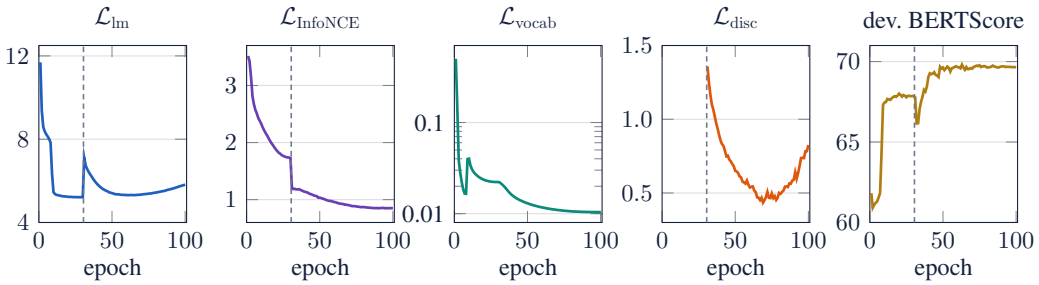
Setting	BERT	B-1	B-4	R-1 F
Sentence-disjoint	69.3	22.44	1.46	18.8
Subject-wise	67.8	21.3	1.1	16.0
ZuCo 2.0 zero-shot	66.9	20.5	0.9	16.9

Table 9: Decoded examples under free greedy decoding (sentence-disjoint split). Coloured text marks matched phrases.

Category	Ground truth	NeuroLens output	ROUGE-L	BLEU-4	BERTScore F_1
A: Correct sentiment, entity substituted	Frank Durward White (June 4, 1933 – May 21, 2003) was the Republican **governor of the U.S. state of Arkansas from** 1981 to 1983.	John Ellis "Jeb" Bush (born May 18, 1953) was a Democratic **Governor of the U.S. State of Arkansas from** 1896 to 1893.	0.511	0.309	0.861
B: Syntactic structure preserved, content wrong	President Clinton **awarded** him the **Presidential Medal of Freedom** in early 1997 for his service in the military and his political career.	He was **awarded** the **Presidential Medal of Freedom** for Best Supporting Actress in the Royal Astronomical Society.	0.450	0.195	0.747
C: High-fidelity output	**Rather than retire, Adams would go on to win election as a Democratic-Republican** to the House of Representatives beginning with the 22nd Congress, serving from March 4, 1831, until his death.	**Rather than retire, Adams would go on to win election as a Democratic-Republican** Dwight Eisenhower to the only Democrat elected to that office between 1898 and 1973.	0.525	0.410	0.804
D: Failure case	These events soon led to the American Civil War.	He was elected to the Defense Intelligence Agency and in 18, Pennsylvania.	0.191	0.036	0.652

E Training Dynamics

Figure 5 shows the four loss terms and the development BERTScore of the matched-protocol full model (seed 42); the dashed line marks the start of Phase 2 at epoch 31. The InfoNCE and vocabulary losses fall throughout training. The discrimination loss enters with Phase 2 and falls until epoch 69, and the language modelling loss falls until epoch 61; both then rise as word dropout approaches 0.85, which removes decoder-side context and forces content to be read from the EEG pathway. Development BERTScore reaches its maximum of 69.8 at epoch 63.



$\mathcal{L}_{\text{lm}}$ and $\mathcal{L}_{\text{disc}}$ rise again as word dropout nears 0.85; dev. BERTScore peaks at 69.8 (epoch 63)

Figure 5: Training loss components and development BERTScore over the two phases.