On Length Bias in EEG-to-Text Decoding

Iakovos Tenedios
Computer and Information Sciences
NeuraSearch Lab, University of Strathclyde
Glasgow
iakovos.tenedios@strath.ac.uk

Yashar Moshfeghi
Computer and Information Sciences
NeuraSearch Lab, University of Strathclyde
Glasgow
yashar.moshfeghi@strath.ac.uk

Abstract

Sentence-level EEG-to-text decoding models typically pad variable-duration neural signals to a fixed-length window, creating a structural confound where segment duration correlates with sentence identity independently of neural content, a problem we term length bias. Despite being acknowledged in prior work, the prevalence, magnitude, and architectural dependence of this effect have never been systematically quantified. We introduce length-matched noise baselines that isolate this confound by replacing only the true signal region with Gaussian noise while preserving zero-padding, and apply them across three English listening EEG datasets under a strict subject-held-out protocol. Evaluating ten sentence classifiers and three generation models based on the current state-of-the-art, we show that reported decoding scores are substantially inflated by length cues across datasets, with the degree of inflation varying markedly by architecture. Among the classifiers evaluated, BIOT is the only one whose accuracy under length-matched noise falls to chance on every dataset. A systematic decomposition of its components identifies Short-Time Fourier Transform (STFT) tokenisation as the only component whose presence determines near-complete length invariance. We transfer this finding into two new generation variants that largely eliminate encoder-side length bias, enabling a controlled analysis of where residual bias originates. A classification-to-generation evaluation further suggests that leading generation models operate largely as sentence retrieval systems, with a linear classifier recovering nearly all of their reported BLEU-4 scores. Together, these results establish that length bias is a pervasive and previously unquantified failure mode in EEG-to-text evaluation, and provide both diagnostic tools and architectural directions for future work.[1]

1 Introduction

Decoding speech from EEG holds significant promise for a communication prosthesis in impaired populations such as amyotrophic lateral sclerosis (ALS) patients [36], where the ability to communicate through brain signals alone could restore critical quality of life. EEG recorded during passive listening provides a naturally labelled, abundant signal that serves as the practical foundation for building such systems, avoiding the data scarcity that limits imagined speech research [25]. Realising this potential, however, requires that reported decoding performance reflects genuine neural signal processing and not spurious correlates of experimental design.

Recent open-vocabulary EEG-to-text decoding models report striking BLEU and ROUGE scores on listening EEG benchmarks [21, 35]. However, a structural confound arises in sentence-level pipelines. Since neural signals of varying duration are zero-padded to a fixed window, models may exploit the correlation between segment length and sentence identity rather than decoding from neural content, as illistrated in Figure 1. This concern is not incidental but architectural, as it arises from the design

[1] Our code can be found here.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

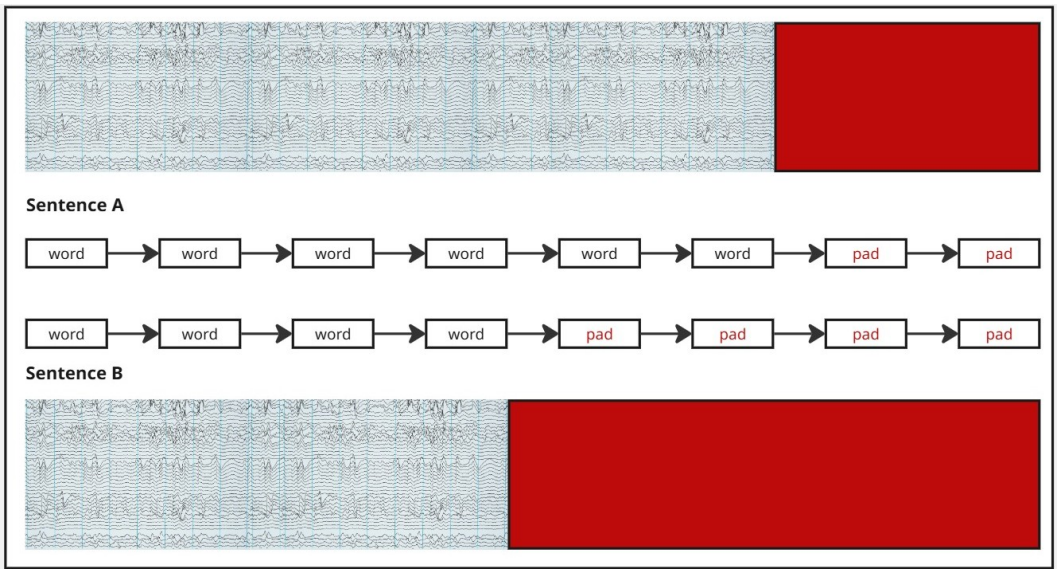

Figure 1: Illustration of length bias in sentence-level EEG decoding. EEG segments corresponding to individual sentences vary in duration and are zero-padded (Shaded Red Area) to a fixed window length.

of every current-generation pipeline in which padded EEG is aligned to a spectrogram-based decoder whose input inherently encodes duration information. Despite being noted in past research [21], the prevalence and magnitude of this effect have never been systematically quantified. Generative systems can reach strong scores while relying on structural confounds rather than the intended content, so auditing what such systems rely on is necessary. This paper is a concrete instance of such an audit.

A second, related question is whether current generation models perform genuine sequence-to-sequence decoding or implicitly operate as classifiers over a fixed stimulus set. When trained on a closed sentence vocabulary, an autoregressive decoder may learn to retrieve the most likely transcript from cues like segment length rather than decoding open-vocabulary text. Sentence classification over the same stimulus set provides a direct probe of this hypothesis as, if a generation model's scores can be reproduced by classifying the EEG segment and retrieving the corresponding transcript, the generation pipeline adds little value beyond retrieval.

We address these questions with four contributions. **First**, we introduce *length-matched noise* baselines that replace only the true signal region with Gaussian noise while preserving zero-padding, measuring the explicit contribution of segment length to reported performance. Applied across a multi-dataset English listening EEG benchmark spanning Brennan [6], HBN [2], and NeuraSearch (our 28-subject dataset), these baselines reveal that a substantial portion of generation scores is attributable to length cues rather than neural decoding.

Second, we identify BIOT [34] as the only classifier in our sentence classification benchmark whose accuracy under length-matched noise falls to chance on every dataset and conduct a systematic decomposition of its three architectural components, STFT tokenisation, channel embeddings, and sinusoidal positional encodings, to isolate the source of this invariance. We verify experimentally that STFT tokenisation is the only component in our decomposition whose presence determines length invariance, as zero-valued padding windows produce exactly zero STFT coefficients, making all fully zero-padded frames inert and confining residual length information largely to a single boundary frame per sequence.

Third, motivated by this finding, we introduce **BrainECHO w/ Full BIOT** and **BrainECHO w/ STFT**, two variants of BrainECHO in which the time-domain convolutional frontend is replaced with STFT-based tokenisers. Both variants deal effectively with explicit length bias from the EEG encoder stage while retaining the three-stage alignment-to-Whisper pipeline, enabling a controlled ablation of how much length bias in generation originates from the EEG encoder versus Whisper's spectrogram processing.

Fourth, we introduce a *classification-to-generation evaluation* experiment that quantifies how much of each generation model's scores can be replicated by retrieving the predicted sentence from a trained classifier. This evaluation directly tests the retrieval hypothesis and reveals whether generation pipelines add value beyond sentence-level identification.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 4 describes the evaluation design, including datasets, models, and the BIOT decomposition. Section 5 presents results. Section 6 provides discussion and conclusions.

2 Related Work

Before open-vocabulary generation, EEG-based speech decoding was framed as simple classification over speech segments, sets of words, or phonemes [7, 8, 30]. Wang and Ji [32] introduced the open-vocabulary EEG-to-text task, feeding word-level EEG rhythm features into a pretrained BART model. While achieving high reported scores, this and subsequent BART-based approaches [4, 10, 11, 32, 33] rely on teacher forcing during evaluation, wherein the model receives ground-truth tokens as input at each decoding step. Jo et al. [16, 17] demonstrated that under this evaluation regime, models trained on random noise achieve comparable performance to those trained on actual EEG. This motivated a non-teacher-forcing evaluation for meaningful assessment, which we adopt throughout.

Yang et al. [35] proposed NeuSpeech, treating raw MEG signals as a specialised form of speech and feeding them directly into a pretrained Whisper model for end-to-end decoding without teacher forcing. Building on this, Li et al. [21] proposed BrainECHO, a three-stage training paradigm for Whisper fine-tuning that achieves current state-of-the-art performance on sentence-level listening EEG. Critically, sentence-level decoding requires long EEG segments exceeding 12 seconds, as shorter segments of 3–4 seconds are insufficient to capture sentence semantics [21], a requirement that makes padding-induced length bias particularly acute. A related but distinct question is whether models fine-tuned on a closed stimulus vocabulary perform genuine decoding or implicitly retrieve sentences from cues like segment duration, yet neither concern has been explored with more controlled baselines.

Length bias is a domain-specific form of shortcut learning, where a model exploits a feature that correlates with the label instead of the intended signal [12, 19]. Applied domains offer parallel examples, such as a pneumonia classifier that recognised the hospital system producing each radiograph rather than the pathology [37]. The general principle is known, yet the open problem in each field is to identify which shortcut operates, measure its contribution, find the architectural property enabling it, and remove it, as this work does for padding-induced duration. Addressing length bias matters beyond EEG, as the same shortcut plausibly affects other variable-length neural decoding, such as sentence decoding from MEG, fNIRS, or fMRI signals, and any pipeline that pads variable-length inputs before classifying over a label inventory whose members differ in duration. Across modalities, auditing generative models for reliance on spurious structure is a growing concern, and the present analysis offers a concrete diagnostic for one such structure.

Patch tokenisation of continuous signals descends from the Vision Transformer [9]. In audio, the Audio Spectrogram Transformer processes spectrogram patches with a pure transformer [13], and LaBraM tokenises EEG in a spectral, frequency-aware space [14]. BIOT [34], on which our proposed variants build, belongs to the same family. These models adopted spectral tokenisation for its representational quality, and the length invariance it confers under zero padding has gone unremarked, even though it proves central to evaluation validity here.

The unquantified contribution of length bias to reported scores, the untested retrieval hypothesis, and the uncharacterised vulnerability of different model families to this confound represent critical gaps in the current evaluation framework. No prior work has introduced length-controlled baselines to isolate the length confound, examined whether generation scores can be reproduced by sentence classification alone, nor characterised which architectures are most susceptible. We directly address all three as length-matched noise baselines quantify the explicit contribution of segment duration across architectures, a systematic decomposition identifies the architectural properties that confer length invariance, and a classification-to-generation evaluation tests whether generation pipelines add value beyond simple retrieval.

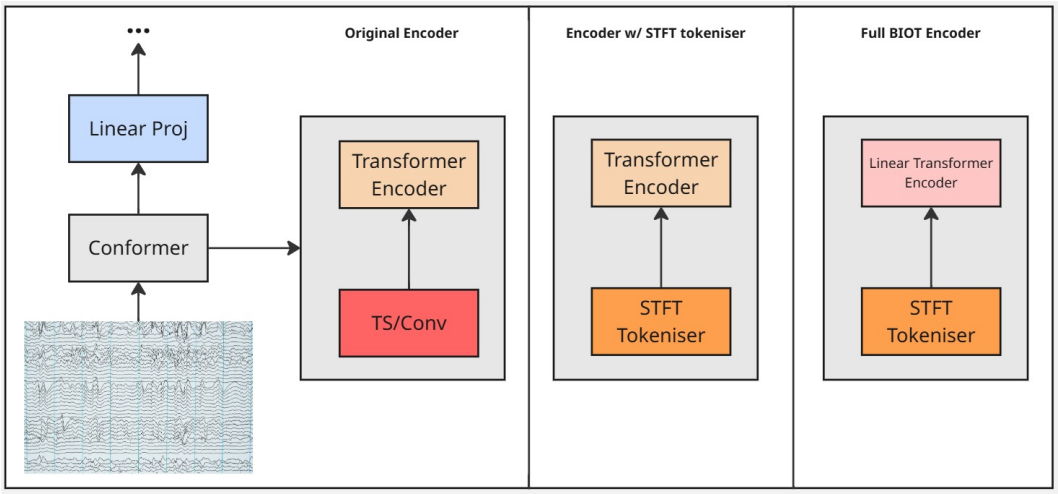

Figure 2: The three EEG encoder configurations evaluated in this work. **Left:** BrainECHO's original time-domain `TS_Conv` frontend, which operates directly on the raw padded signal and is susceptible to length bias. **Centre:** BrainECHO w/ STFT, replacing `TS_Conv` with an STFT tokeniser while retaining BrainECHO's standard Transformer encoder. **Right:** BrainECHO w/ Full BIOT, additionally replacing the Transformer with BIOT's LinearAttentionTransformer.

3 Approach

Length bias in sentence-level EEG decoding arises from the time-domain frontend, where convolutional patchers operating directly on raw EEG can exploit the discontinuity at the zero-padding boundary as a shortcut to sentence identity, independently of neural content. We address this architecturally by replacing the time-domain frontend with an STFT-based tokeniser, which confers near-complete length invariance by construction. For a zero-padded signal, fully zero-valued windows produce exactly zero STFT coefficients, making all padded frames inert to the encoder. Residual length information is confined to weak cues, mainly the boundary frame where real signal and padding meet, which is a structurally unavoidable and minimal source. This property is verified numerically in Section 4.3 and Appendix E. We introduce two encoder variants that integrate this tokeniser into the BrainECHO framework [21], the current state-of-the-art for sentence-level listening EEG decoding, which outperforms prior BART-based [10, 32] and earlier Whisper-based approaches [35]. Both variants retain BrainECHO's three-stage training pipeline (VQVAE pre-training, EEG-to-audio alignment, Whisper fine-tuning) and differ only in their EEG encoder frontend, enabling a controlled isolation of encoder-side length bias from any confound introduced by the downstream decoder. Figure 2 illustrates the three encoder configurations.

BrainECHO w/ STFT. The temporal-spatial convolutional frontend (`TS_Conv`) of BrainECHO is replaced with an STFT tokeniser. Each EEG channel is converted to a magnitude spectrogram via STFT ($n_{\text{fft}} = 200$, hop $= 100$, `centre=False` at 200 Hz), projecting the frequency axis ($n_{\text{freq}} = 101$) to $d_{\text{model}} = 256$ via a linear layer. Learned per-channel identity embeddings and sinusoidal positional encodings are added, and the resulting token sequence is processed by BrainECHO's own standard Transformer encoder (4-layer `nn.TransformerEncoderLayer`, $d_{\text{model}} = 256$, 8 heads). This variant isolates the contribution of STFT tokenisation from the choice of transformer architecture.

BrainECHO w/ Full BIOT. This variant additionally replaces BrainECHO's Transformer encoder with BIOT's 4-layer LinearAttentionTransformer [34], retaining the same STFT tokeniser and channel/positional embeddings. This incorporates the full BIOT encoder and enables a direct comparison of transformer architectures under identical tokenisation. In both variants, temporal pooling reduces the token sequence to match the downstream alignment target, and all subsequent stages are identical to BrainECHO.

Table 1: Summary of the three benchmark datasets. Split sizes are given as number of subjects in training / validation / test sets.

Dataset	Subjects	Channels	Sentences	Stimulus
Brennan [6]	49	61	139	Audiobook (earphones)
HBN [2]	123	129	171	Movie watching (speakers)
NeuraSearch	28	64	48	Isolated sentences (speakers)

4 Experimental Design

The experimental design addresses the three gaps identified above through four complementary components: a multi-dataset benchmark with length-controlled noise conditions, a sentence classification benchmark spanning established and generation-derived architectures, a systematic BIOT decomposition, and a classification-to-generation evaluation.

4.1 Datasets and Splits

Prior EEG-to-text work evaluates on a single EEG dataset [10, 21, 32], making it hard to assess whether results generalise across recording conditions, subject populations, or stimulus types. Existing English datasets capture either continuous audiobook listening [6] or incidental audio from video watching [2], representing only a narrow slice of the listening EEG landscape. We span three English datasets with substantially different recording contexts to enable systematic comparison of length-bias patterns and decoding strategies across paradigms. This work focuses on English datasets.

Brennan [6] contains EEG recordings from 49 participants who listened passively to a 12.4-minute audiobook recording of the first chapter of *Alice's Adventures in Wonderland* via earphones. Data were recorded with 61 active electrodes using a Brain Products actiCHamp amplifier at 500 Hz. This dataset represents a naturalistic continuous narrative listening paradigm with a fixed set of approximately 139 sentences.

HBN [2] is drawn from Release 1 of the Healthy Brain Network EEG dataset, which contains recordings from 136 participants aged 5–21 years during naturalistic movie watching. We retain 123 subjects for whom complete data are available across all four movie stimuli (*Despicable Me*, *Diary of a Wimpy Kid*, *Fun with Fractals*, and *The Present*) and resting state. EEG was recorded with 129 channels at 500 Hz. Because the audio stimuli are not distributed with the dataset for fair use reasons, we do not include them directly, and instead, we release code to download and align the audio with the EEG recordings.

NeuraSearch is a new dataset collected by our lab, recorded from 28 participants (12 female, 14 male, 2 non-binary; mean age 26.1, SD 6.9). EEG was recorded at 512 Hz using a g.HIamp biosignal amplifier (62 active channels, earlobe reference). The stimulus set comprises 48 sentences. Sentence length characteristics are $\mu = 7.4$ word tokens, $\sigma = 0.8$, and the combined vocabulary spans 265 unique word types. Sentences were listened to by participants with randomised trial presentation order, each sentence appeared three times per session with a fixed inter-trial interval, yielding 144 trials, extracted as listening-period segments and padded to a fixed length of 1000 samples. The study received ethical approval from the Ethics Committee of the Department of Computer and Information Sciences of the University of Strathclyde. [2]

The padding fraction varies substantially across datasets, with a mean of 64.2% and an SD of 20.4% on Brennan, 84.7% and 12.0% on HBN, and 36.7% and 12.5% on NeuraSearch. Full statistics are reported in Appendix A.

All datasets were preprocessed by notch filtering (60 Hz), bandpass filtering (0.5–99 Hz), resampling to 200 Hz, robust scaling, and per-participant normalisation. In all three datasets, splits are performed strictly at the subject level with no participant appearing in more than one split. This ensures that all reported results reflect subject-independent generalisation, the most stringent and clinically relevant evaluation setting [25]. Table 1 summarises the key properties of each dataset.

[2] Data can be made available upon request.

Data Splits. We adopt subject-level splits [25] of approximately 70/10/20% for training, validation, and test sets across all three datasets, with no participant appearing in more than one partition. This ensures all reported results reflect generalisation to unseen individuals.

Evaluation Metrics. Consistent with previous research [21, 32], for generation models we report BLEU-1 through BLEU-4 [26], word error rate (WER), and ROUGE-1, ROUGE-2 [22]. For classification mo,dels we report top-1 accuracy and macro-averaged F1 score, alongside the chance level for each dataset ($1/N$ where N is the number of sentences).

4.2 Models

Noise Baselines. To verify that models decode genuine neural signals rather than exploiting dataset artifacts, we evaluate three baselines alongside every model. *Gaussian noise* replaces the entire EEG input with samples drawn from the test signal distribution. *Length-matched Gaussian noise* replaces only the true signal region (up to the input mask boundary) with Gaussian noise, leaving padding unchanged. This baseline specifically targeting the concern, that models may exploit EEG segment length or padding position to infer sentence identity. *Random sentence selection* applies to generation models only, where model output is discarded, and a sentence is drawn uniformly at random from the corpus, providing a corpus-level lower bound on lexical overlap metrics. Full implementation details are provided in Appendix B.3.

4.2.1 Generation Models

To quantify the contribution of the EEG encoder frontend to length bias, we evaluate three generation models built on the BrainECHO framework, differing only in their EEG encoder: **BrainECHO** [21] with its original TS_Conv frontend, **BrainECHO w/ STFT**, and **BrainECHO w/ Full BIOT**, both described in Section 3 and depicted in Figure 2. All three models are evaluated without teacher forcing using beam search (beam size 5, repetition penalty 5.0, no-repeat nn n-gram size 2). Implementation details are provided in Appendix F.

4.2.2 Classification Models

Established EEG Classifiers To characterise architectural vulnerability to length bias across model families, we evaluate seven established deep learning classifiers drawn from the `braindecode` library [5], applied end-to-end directly to raw EEG epochs. **EEGNet** [20] uses depthwise and separable convolutions to learn frequency-specific spatial filters with minimal parameters, making it a strong compact baseline across BCI paradigms. **ShallowConvNet** and **DeepConvNet** [27] are FBCSP-inspired convolutional networks of two and five convolutional layers, respectively, jointly proposed as end-to-end alternatives to hand-crafted spectral features. **EEGConformer** [28] combines a shallow CNN front-end that extracts local spatio-temporal features with a transformer encoder that captures long-range temporal dependencies, bridging the inductive biases of convolutional and attention-based architectures. **TIDNet** [18] performs explicit temporal-spatial decomposition via separate temporal filter banks and spatial projection layers, and was specifically designed to improve cross-subject generalisation. **ATCNet** [3] processes parallel overlapping temporal windows through a multi-head self-attention block followed by a temporal convolutional network, allowing it to attend to key temporal segments while modelling sequential dependencies. **BIOT** [34] is a biosignal foundation model that tokenises each EEG channel into fixed-length segments and processes them with a transformer augmented with channel and relative position embeddings, designed for cross-dataset and cross-modality generalisation. All models receive the same preprocessed EEG epochs used in the generation experiments.

Frozen encoder probes. To assess whether representations learned during generation pre-training encode sentence-discriminative information independently of length cues, we attach a two-layer classification head (LayerNorm + Linear) to the frozen encoders of all three generation models. Each encoder is initialised from its Stage 2 alignment checkpoint with all parameters frozen; only the classification head is trained. For BrainECHO, the EEG representation is extracted from the Transformer output prior to VQ-VAE quantisation and mean-pooled over the sequence dimension. Details are provided in Appendix B.

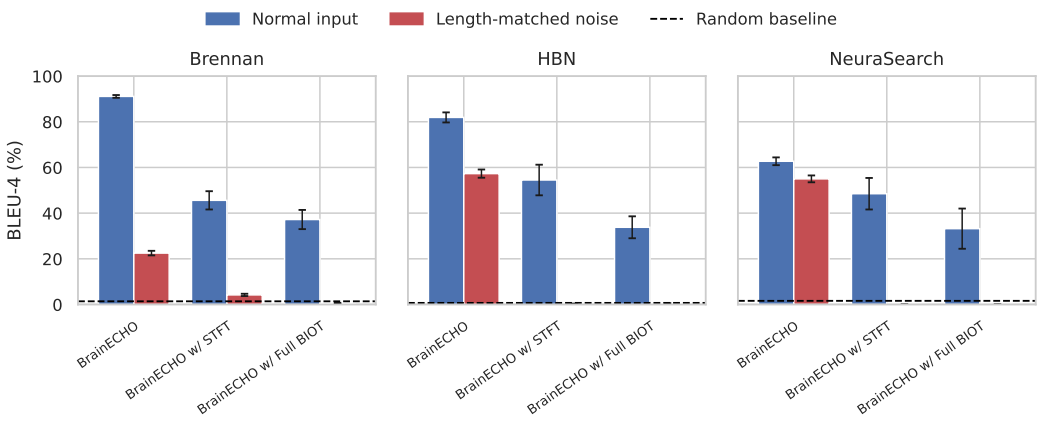

Figure 3: Generation results across all datasets, shown as BLEU-4 with paired bars per model for normal input and length-matched noise. Bars are subject means and error bars subject SDs, and the dashed line marks the random baseline. Values reported in Table 4 in Appendix D, which reports BLEU-1 through BLEU-4, ROUGE, and WER together with all noise conditions.

4.3 BIOT Decomposition Study

To identify which component is responsible for BIOT's length invariance, we ablate its three components namely STFT tokenisation, channel identity embeddings, and sinusoidal positional encodings, in all combinations, replacing STFT with a Conv1d patcher of equivalent stride when removed. The patcher uses kernel size 200 and stride 100, which match the STFT window and hop length exactly, so both frontends give an identical frame count and the same receptive field and temporal resolution. The only remaining difference between the two frontends is the frequency-domain transform itself. Eight variants are trained end-to-end on all three datasets and evaluated under normal, length-matched noise, and input noise conditions. The Conv1d patcher reads 200 raw samples per frame, which is more information than the 101 STFT magnitude bins, yet the variants without STFT still retain well-above-chance length-noise accuracy on Brennan and HBN, with `no_stft` at 58.8% on Brennan, while every variant that keeps STFT falls to near chance. STFT tokenisation is the only component in our decomposition whose presence determines length invariance. Full results and numerical verification of the zero-coefficient property are provided in Appendix E.

4.4 Classification-to-Generation Evaluation

To test whether Whisper-based generation operates as a retrieval system over the closed sentence vocabulary, we construct a *classification-to-generation (Cls2Gen) baseline*. For each test sample, the trained classification model predicts a sentence class, and the corresponding transcript is retrieved and scored with the same generation metrics (BLEU, ROUGE, WER) used for the generation models. This retrieval score represents the generation performance achievable by a perfect classifier without any sequence-to-sequence decoding, which is an upper bound on what sentence identification alone can achieve.

The Cls2Gen baseline is also evaluated under the length-matched noise condition, directly quantifying how much of the classifier's retrieval performance is attributable to length cues. A small gap between normal and length-noise Cls2Gen scores indicates that the classifier (and by extension any generation model sharing similar features) relies primarily on EEG content. Only the normal condition is reported in Figure 3. Full methodological details and results are given in Appendix G.

5 Results

Figure 3 reports generation results and Figure 4 reports classification results across all three datasets under subject-held-out evaluation, with the full metric tables reproduced in Appendices C and D.

Length bias is pervasive but architecture-dependent. All ten classifiers exceed chance on real EEG, and input noise collapses all to chance, confirming genuine signal sensitivity. Length-matched noise reveals substantial and variable length exploitation across architectures. Among end-to-end classifiers, DeepConvNet shows the most extreme dependence, with length-noise accuracy exceeding real-EEG accuracy on Brennan (80.8% vs 75.7%). On NeuraSearch the collapse has two levels, beginning with the dataset itself, whose length-matched, low variance stimuli leave little length to exploit, so most models fall toward chance. A set of architectures with convolutional frontends then preserves enough temporal resolution to exploit the residual boundary, with the same mechanism as BrainECHO's temporal-spatial convolution, and the model-by-model numbers and the confusion analysis appear in Appendix C. For generation, length-matched noise accounts for 22.5 of 91.1 BLEU-4 points on Brennan, 57.3 of 81.9 on HBN, and 55.0 of 62.7 on NeuraSearch, indicating that between 25% and 88% of reported scores are attributable to segment duration rather than neural content. This variation across datasets is itself a finding as single-dataset evaluation conceals how severely length bias inflates reported performance.

STFT tokenisation is the only component in our decomposition whose presence determines length invariance, and the invariance it confers is near-complete rather than absolute. BIOT is the only classifier whose length-noise accuracy collapses to chance on every dataset while maintaining above-chance real-EEG accuracy. All decomposition variants that remove STFT, using the stride-matched Conv1d patcher of Section 4.3, retain well-above-chance length-noise accuracy regardless of whether other components are present, with `no_stft` at 58.8% on Brennan. All variants retaining STFT collapse to near chance under length noise. Full decomposition results and the numerical verification of the zero-coefficient property are reported in Appendix E.

Encoder-side length bias transfers to generation. BrainECHO w/ Full BIOT and BrainECHO w/ STFT largely eliminate encoder-side explicit length bias, with length-noise generation scores falling near or below the random baseline on all datasets. BrainECHO achieves the highest raw scores across all three datasets, while its frozen encoder probe achieves 85.5% classification accuracy on Brennan with low length sensitivity (6.6% under length noise). BrainECHO w/ Full BIOT and BrainECHO w/ STFT match standalone BIOT classification accuracy on NeuraSearch (65.7% and 66.9%) while performing below it on Brennan and HBN, consistent with their STFT-based encoders collapsing to chance under length noise, with the two variants differing only in the transformer block (Table 7 in Appendix F). Confusion matrix analysis of BrainECHO w/ Full BIOT reveals systematic sentence-pair confusions consistent across subjects on all three datasets. The ten most confused pairs and full confusion matrices are reported in Appendix J.

Generation models operate largely as retrieval systems. The Cls2Gen evaluation scores 89.1 BLEU-4 on Brennan, just below BrainECHO's real-EEG score, indicating that nearly all generation performance on that dataset is reproducible by sentence retrieval alone without sequence-to-sequence decoding. On HBN and NeuraSearch, Cls2Gen scores for BrainECHO w/ Full BIOT and BrainECHO w/ STFT exceed their actual generation scores, revealing that once encoder-side length bias is removed, Whisper's spectrogram processing becomes the primary remaining bottleneck rather than the EEG encoder. Full Cls2Gen results across all models and noise conditions are reported in Appendix G.

6 Discussion and Conclusion

Length bias is an existing and previously unquantified evaluation confound. Sentence-level EEG decoding pipelines pad variable-duration signals to a fixed window, creating a structural confound that has been noted [21, 35] but never systematically measured. Our length-matched noise baselines make this explicit, with between 25% and 88% of reported BLEU-4 scores across three datasets and two model families attributable to segment duration rather than neural content, and with the degree of inflation varying substantially across datasets. Brennan's high padding variability (σ=20.4%) makes duration a highly discriminative per-sentence cue whereas NeuraSearch's shorter, more uniform stimuli reduce but do not eliminate this effect. That this pattern is consistent across datasets with different recording contexts, stimulus types, and subject populations confirms that length bias is a sentence-level evaluation problem, not an artifact of any single dataset or model. Length bias repeats the pattern of the teacher-forcing flaw, which spread through several BART-based pipelines [4, 10, 31] before Jo et al. identified it [16, 17], and it remains active, since NEURAL-VOX [15] evaluates on the same listening benchmarks, including Brennan, with the same

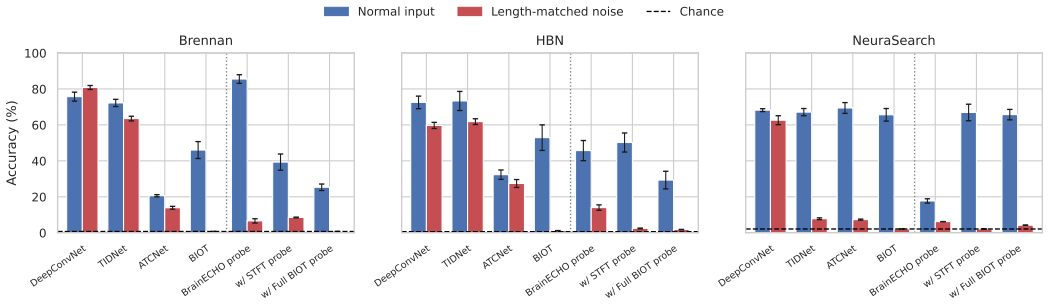

Figure 4: Sentence classification accuracy across all datasets, shown as paired bars for normal input and length-matched noise. End-to-end classifiers appear left of the dotted line and frozen encoder probes right of it. Bars are subject means, error bars subject SDs, and the dashed line marks chance (0.7% Brennan, 0.6% HBN, 2.1% NeuraSearch). Full accuracy and F1 results for all ten models are in Table 3 in Appendix C.

fixed-window padding and BrainECHO as a baseline without accounting for length bias, so present head-to-head comparisons inherit it. A diagnostic separating decoding from length-driven retrieval therefore becomes more needed, as audio and text reconstruction claims grow.

STFT tokenisation is the architectural remedy. A systematic ablation of the components of BIOT [34] identifies STFT tokenisation as the only component in our decomposition whose presence determines length invariance. A window that lies entirely inside the zero padding gives exactly zero STFT coefficients, a property that follows from the transform and is verified numerically in Appendix E, so the invariance is near-complete rather than absolute, with two weak cues remaining, the single boundary frame that straddles signal and padding, and the number of all-zero frames, which still changes with the amount of padding. This may possibly explain why BrainECHO w/ STFT keeps a small above-chance accuracy under length-matched noise on Brennan (8.5% against 0.7% chance), whose padding is large and highly variable, while the variants fall to chance on NeuraSearch, whose padding is low and almost uniform. This finding transfers directly to generation since our two proposed variants, BrainECHO w/ STFT and BrainECHO w/ Full BIOT, largely eliminate encoder-side length bias across all three datasets, providing the first generation models in this setting whose scores are not substantially inflated by padding artifacts. The Brennan advantage of the original time-domain frontend, at 85.5% frozen encoder accuracy against 39.3% for the STFT variant, stays dataset specific and is not confound-free, with 6.6% under length-matched noise against a 0.7% chance floor and a BLEU-4 fall from 91.1 to 22.5 under length-matched noise.

Robustness and performance trade-off. The STFT variants score below the original BrainECHO, and a length-stratified evaluation probes whether the removed information was spurious or useful. Within ten narrow bins of real signal length, where duration carries little information, BrainECHO w/ STFT BLEU-1 rises from 19.0 in the shortest HBN bin to 77.1 in the longest, and from 22.5 to 67.3 across Brennan bins 2 to 10, whereas a model leaning on length would score roughly flat (Figure 6 in Appendix H). We read this as strong supporting evidence, and claim no rise on NeuraSearch, whose sentences are length-matched by design. The Cls2Gen comparison (Figure 3 and Appendix G) gives the key contrast, with BrainECHO's NeuraSearch Cls2Gen at 18.3 against 62.7 for generation, BrainECHO w/ STFT's Cls2Gen exceeding BrainECHO's on HBN and NeuraSearch, and the large performance gap belonging to Brennan alone, where length is most exploitable. Performance falls, but length-matched noise sensitivity falls to near chance, and better methods may recover the loss while retaining robustness. Length may also have shaped training.

Current generation pipelines operate largely as retrieval systems. Because Cls2Gen nearly matches or exceeds generation once encoder-side length bias is removed (Section 5), generation scores in this closed-vocabulary setting do not reliably measure open-vocabulary decoding ability. A qualitative inspection of generated outputs (Appendix I) corroborates this, with verbatim sentence repetitions across different test inputs on Brennan consistent with retrieval rather than sequence generation. Retrieval is chiefly a property of the benchmarks rather than the models, and it forms an

axis separate from length bias, since the STFT variants remove most of the encoder-side bias while no encoder change removes closed-inventory retrieval. Under normalised matching, BrainECHO's Brennan outputs equal training-inventory sentences for 92.2% of the 1,112 test samples, produced as only 134 distinct strings, and because every test sentence also appears in training over 48 to 166 classes, inventory retrieval is the sensible default (Table 11 in Appendix H). BLEU and ROUGE were designed for open-vocabulary generation, while over a small closed set classification accuracy measures the same quantity more directly, and cannot separate retrieval from generation, so a task should earn the open-vocabulary label only by demonstration, by holding out sentences, holding out vocabulary, or adding a Cls2Gen retrieval baseline beside the generation score. We recommend held-out-sentence evaluation alongside held-out subjects, with a Cls2Gen-style baseline as standard practice. Beyond EEG, the combination of length-matched noise and Cls2Gen generalises as an auditing template for generative systems that may substitute structure for content.

Limitations. Four limitations should be noted. First, the closed-vocabulary setting increases the likelihood that generation models behave as retrieval systems, and our Cls2Gen analysis is a diagnostic rather than definitive verification that no open-vocabulary decoding occurs. Second, the length-matched noise baseline should be understood as a practical diagnostic for non-neural structure rather than a complete causal disentanglement, as residual interactions between noise statistics, boundary effects, and model architecture may contribute to observed performance. Third, our attribution of residual length sensitivity to decoder-side spectrogram processing is inferential. Directly ablating the decoder's positional encodings or padding strategy would be required to confirm this mechanism. Fourth, despite spanning three datasets, the benchmark remains constrained by small subject counts and fixed sentence inventories relative to modern speech benchmarks.

Future Directions. Removing structural confounds before interpreting any score comes first, with the decoder-side ablation above as the most important next step. Beyond it, deeper per-subject sampling, held-out-sentence evaluation, larger and more diverse stimulus inventories, transfer across subjects and datasets, multilingual extension [1, 24, 29], and confound baselines reported by default will yield scores that are likely lower, and that is progress. The choice of noise distribution defines what the baseline controls for, as Gaussian noise removes all structure beyond duration, whereas a surrogate built by phase scrambling or by matching the spectrum would retain spectral content while destroying temporal structure. Biological signal baselines, like ocular activity, while destroying neural content would require reconstructing and reinjecting ocular contributions across the scalp channels, and thus, such surrogates are valuable but hard to build. The present listening data involve no reading-related eye movements, and surrogate baselines supplement rather than replace strict experimental protocols.

Broader impact. The motivating application is assistive communication for people whose speech is severely impaired, such as in ALS, and that goal is served only when reported results reflect genuine neural decoding. Inflated scores risk encouraging premature clinical expectations or deployment before the evidence supports the decisions those expectations imply.

Conclusion. We have identified length bias as a primary and previously unquantified confound in sentence-level EEG-to-text evaluation, demonstrated that STFT tokenisation provides a principled architectural remedy, and shown that leading generation models operate largely as retrieval systems over closed stimulus sets. The combination of length-matched noise baselines, near length-invariant encoder variants, and classification-to-generation analysis constitutes a methodological template for rigorous evaluation of EEG-to-text frameworks going forward. More broadly, generative AI systems need audits of this kind, so that reported capability reflects real competence.

Acknowledgments and Disclosure of Funding

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0011).

References

[1] Bernd Accou, Lies Bollens, Marlies Gillis, Wendy Verheijen, Hugo Van Hamme, and Tom Francart. Sparrkulee: A speech-evoked auditory response repository from ku leuven, containing the eeg of 85 participants. *Data*, 9(8), 2024.

[2] Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Jennifer Andersen, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Henry Butler, Victoria Castagna, Nicolas Camacho, Elisha Chan, Danielle Citera, Jon Clucas, Samantha Cohen, Sarah Dufek, Megan Eaves, Brian Fradera, Judith Gardner, Natalie Grant-Villegas, Gabriella Green, Camille Gregory, Emily Hart, Shana Harris, Megan Horton, Danielle Kahn, Katherine Kabotyanski, Bernard Karmel, Simon P. Kelly, Kayla Kleinman, Bonhwang Koo, Eliza Kramer, Elizabeth Lennon, Catherine Lord, Ginny Mantello, Amy Margolis, Kathleen R. Merikangas, Judith Milham, Giuseppe Minniti, Rebecca Neuhaus, Alexandra Levine, Yael Osman, Lucas C. Parra, Ken R. Pugh, Amy Racanello, Anita Restrepo, Tian Saltzman, Batya Septimus, Russell Tobe, Rachel Waltz, Anna Williams, Anna Yeo, Francisco X. Castellanos, Arno Klein, Tomas Paus, Bennett L. Leventhal, R. Cameron Craddock, Harold S. Koplewicz, and Michael P. Milham. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1):170181, December 2017.

[3] Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. *IEEE transactions on industrial informatics*, 19(2):2249–2258, 2022.

[4] Hamza Amrani, Daniela Micucci, and Paolo Napoletano. Deep representation learning for open vocabulary electroencephalography-to-text decoding. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[5] Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, Joseph Paillard, Bruna Junqueira Lopes, Sara Sedlar, Thomas Moreau, Sylvain Chevallier, Alexandre Gramfort, and Robin Tibor Schirrmeister. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models.

[6] Jonathan R. Brennan and John T. Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1):e0207741, January 2019.

[7] Lopez-Bernal D., Balderas D., Ponce P., and Molina A. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in Human Neuroscience*, 16((Lopez-Bernal, Balderas, Ponce, Molina) Tecnologico de Monterrey, National Department of Research, Mexico City, Mexico):867281, 2022.

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

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] 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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[11] Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883, 2023.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[13] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[14] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

[15] Zhishuo Jin, Dongdong Li, Qin Zhou, and Zhe Wang. NEURAL-VOX: NEURal auditory language decoding for voice and text reconstruction. *Neural Networks*, 204:109221, December 2026.

[16] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are eeg-to-text models working? (arXiv:2405.06459), October 2024. arXiv:2405.06459 [cs].

[17] Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Evaluating eeg-to-text models through noise-based performance analysis. *Scientific Reports*, 16(1):350, December 2025.

[18] Demetres Kostas and Frank Rudzicz. Thinker invariance: enabling deep neural networks for bci across more people. *Journal of Neural Engineering*, 17(5):056008, 2020.

[19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[20] 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.

[21] Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. Brainecho: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, page 2762–2778, Vienna, Austria, 2025. Association for Computational Linguistics.

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

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

[24] Mohammad Momenian, Zhengwu Ma, Shuyi Wu, Chengcheng Wang, Jonathan Brennan, John Hale, Lars Meyer, and Jixing Li. Le petit prince hong kong (lpphk): Naturalistic fmri and eeg data from older cantonese speakers. *Scientific Data*, 11(1):992, 2024.

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

[26] Kishore Papineni. Machine translation evaluation: N-grams to the rescue. In *LREC*, 2002.

[27] 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.

[28] 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, 2022.

[29] Carlos Valle, Carolina Mendez-Orellana, Christian Herff, and Maria Rodriguez-Fernandez. Identification of perceived sentences using deep neural networks in eeg. *Journal of Neural Engineering*, 21(5):056044, October 2024.

[30] Lee W., Seong J.J., Ozlu B., Shim B.S., Marakhimov A., and Lee S. Biosignal sensors and deep learning-based speech recognition: A review. *Sensors (Basel, Switzerland)*, 21(4), 2021.

[31] Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7278–7292, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[32] Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5350–5358, 2022.

[33] Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13277–13291, 2023.

[34] Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

[35] 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:2403.01748), 2024. arXiv:2403.01748 [cs].

[36] Collin York, Christopher Olm, Ashley Boller, Leo McCluskey, Lauren Elman, Jenna Haley, Emily Seltzer, Lama Chahine, John Woo, Katya Rascovsky, et al. Action verb comprehension in amyotrophic lateral sclerosis and parkinson's disease. *Journal of neurology*, 261(6):1073–1079, 2014.

[37] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018.

[38] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

A Padding Statistics and Signal Length Distributions

Table 2 reports per-dataset statistics of the padding fraction, defined as the proportion of each fixed-length EEG window occupied by zero-padding rather than real signal. Figure 5 shows the full distribution of padding fractions across all samples in each dataset.

Table 2: Padding fraction statistics (%) per dataset. Padding fraction is defined as $1 - \text{real signal length}/\text{window length}$.

Dataset	N	Mean	Median	SD	Min	Max	Q25	Q75
Brennan	6,811	64.2	69.8	20.4	5.7	93.5	52.3	81.1
HBN	21,033	84.7	89.0	12.0	45.1	98.2	80.3	93.4
NeuraSearch	3,960	36.7	37.0	12.5	5.6	60.0	28.6	46.0

Brennan has the highest padding variability of the three datasets (σ = 20.4%), spanning nearly the full range from minimally padded to almost entirely padded samples. This variability is what makes segment duration a discriminative cue per sentence: classifiers and generation models can learn to associate specific duration ranges with specific sentences from the stimulus set. HBN has the highest mean padding fraction (84.7%) but lower variability (σ = 12.0%), concentrating most samples in a high-padding regime. NeuraSearch has the lowest mean padding fraction (36.7%) and similarly

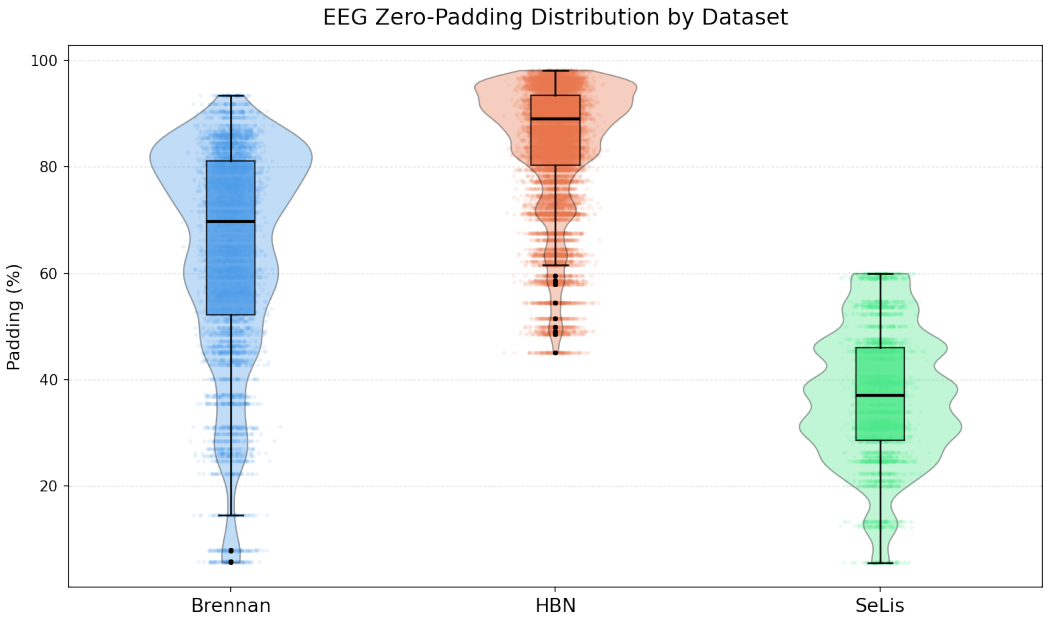

Figure 5: Distribution of padding fractions (%) across all samples in each dataset.

low variability (σ = 12.5%), reflecting its shorter, more uniform isolated sentences. The interaction between mean padding fraction and its variability across sentences may possibly determine how exploitable segment length is as a classification cue in each dataset.

B Classification Evaluation: Full Model Descriptions and Hyperparameters

All classification models were trained using the AdamW optimiser [23] with learning rate 1×10^{-3}, weight decay 10^{-4}, and a cosine annealing learning rate schedule. Training ran for up to 50 epochs with early stopping applied when validation accuracy did not improve for 10 consecutive epochs. Batch size was 32 for all models. The same subject-disjoint splits used for the generation benchmark were applied here where test subjects were held out entirely during training and validation. Reported metrics are macro-averaged accuracy and F1 computed over the held-out test set, with mean and standard deviation reported across test subjects.

B.1 Braindecode Baselines

All seven braindecode models were implemented using the `braindecode` library (v1.12+) and trained end-to-end from random initialisation on the full EEG epoch (no hand-crafted feature extraction).

EEGNet [20]. EEGNet is a compact CNN that encapsulates core EEG feature extraction concepts using three sequential operations: (1) a temporal convolution that learns a data-driven filter bank across the time axis; (2) a depthwise convolution applied per filter that learns electrode-specific spatial weights; and (3) a separable convolution that summarises temporal patterns per feature map before mixing across maps. This three-stage design reduces parameters dramatically compared to standard CNNs while retaining the ability to extract frequency-specific spatial patterns. It is included as the primary compact CNN baseline, having demonstrated strong generalisation across motor imagery, P300, and error-related negativity paradigms. *Parameters used:* `drop_prob=0.25`, `final_conv_length="auto"`.

ShallowConvNet [27]. ShallowConvNet is a two-stage CNN inspired by Filter Bank Common Spatial Patterns (FBCSP): a temporal convolution learns broadband frequency filters followed by a spatial convolution that collapses the electrode dimension. Square activation and average pooling are then applied to compute approximate band power features, and a linear classifier produces the

output. The shallow architecture is deliberately interpretable, with learned filters often recovering physiologically meaningful frequency bands. *Parameters used:* `final_conv_length="auto"`.

DeepConvNet [27]. DeepConvNet uses five convolutional stages (compared to ShallowConvNet's two), interleaved with batch normalisation, ELU activations, max pooling, and dropout. The first stage applies a temporal-then-spatial factored convolution; subsequent stages apply joint spatio-temporal convolutions to build increasingly abstract representations. It was proposed alongside ShallowConvNet as a deeper end-to-end alternative, trading interpretability for representational capacity. *Parameters used:* `final_conv_length="auto"`.

EEGConformer [28]. EEGConformer combines a CNN front-end with a transformer encoder. The CNN module applies a temporal convolution (forming a filter bank) followed by a depth-wise spatial convolution, producing a sequence of local spatio-temporal feature vectors. These are passed to a multi-head self-attention transformer encoder, which captures long-range temporal dependencies beyond the pooling horizon of the convolutional layers. A fully connected classifier head produces the final output. EEGConformer is included to test whether global temporal context improves sentence classification relative to purely local CNN baselines. *Parameters used:* `n_filters_time=40`, `filter_time_length=25`, `drop_prob=0.5`, `att_depth=6`, `att_heads=10`, `att_drop_prob=0.5`, `final_fc_length="auto"`.

TIDNet [18]. TIDNet (Thinker-Invariant Decomposition Network) separates temporal and spatial feature learning into two explicitly decoupled streams. A bank of temporal filters with shared weights first extracts oscillatory patterns independently at each electrode. A spatial projection layer then combines electrode signals for each temporal filter. A bottleneck module reduces dimensionality before a temporal summary network produces the final representation. TIDNet was specifically designed to improve generalisation across subjects ("thinker invariance"), which is directly relevant to the subject-held-out evaluation protocol used here. *Parameters used:* `s_growth=24`, `t_filters=32`, `drop_prob=0.4`, `pooling=15`, `temp_layers=2`, `spat_layers=1`, `temp_span=0.05`, `bottleneck=3`, `summary=n_times//15`.

ATCNet [3]. ATCNet processes the EEG signal through three sequential modules. First, a convolutional block (temporal convolution followed by depthwise spatial convolution with average pooling) encodes low-level spatio-temporal information into a sequence of high-level temporal representations. This sequence is divided into multiple overlapping windows, each processed by a multi-head self-attention block that highlights the most informative temporal segments. Each attended window is then passed through a temporal convolutional network (TCN) using dilated causal convolutions to model sequential dependencies. The outputs of all windows are concatenated and classified. ATCNet is included as an attention-augmented temporal model that explicitly segments the temporal dimension, complementing the global attention of EEGConformer. *Parameters used:* `n_windows=5`, `att_head_dim=8`, `att_num_heads=2`, `tcn_depth=2`, `tcn_kernel_size=4`, `tcn_drop_prob=0.3`.

BIOT [34]. BIOT is a biosignal transformer designed for cross-data learning. Each EEG channel is independently tokenised into fixed-length temporal segments, treating each segment as a "word" in a biosignal "sentence". Channel embeddings and sinusoidal relative position embeddings are added to each token to preserve spatial and temporal context across the variable-format inputs. A standard transformer encoder processes the resulting sequence of tokens. BIOT was pre-trained on large corpora of EEG and ECG signals before fine-tuning on downstream tasks; we use it without its pre-trained weights (trained from scratch) to isolate the architectural contribution, which is directly comparable to the other braindecode baselines. *Parameters used:* default braindecode implementation parameters.

B.2 Frozen Encoder Probes

The frozen encoder experiments are designed as linear probing evaluations: the encoder weights are fixed at their pre-trained values and only a two-layer classification head (LayerNorm + Linear) is trained on the labelled EEG data. This experimental design isolates the question of *representation quality*: does the encoder, trained solely to align EEG with audio representations, already encode

information sufficient to discriminate between sentences without any classification-specific fine-tuning?

BrainECHO. The BrainECHO encoder (Stages 1–2 from [21]) maps raw EEG to a latent space aligned with mel-spectrogram representations via a VQ-VAE. For classification, the output of the transformer module is extracted prior to the VQ-VAE quantisation stage, mean-pooled over the sequence dimension, and passed to the classification head. This representation is 256-dimensional and captures the aligned EEG–audio features before the information-reducing discretisation step. The Stage 2 alignment checkpoint is used. *Feature dim:* 256. *Head:* LayerNorm(256) $\rightarrow$ Linear(256, N).

BrainECHO w/ Full BIOT. The BrainECHO w/ Full BIOT encoder uses BIOT's LinearAttention-Transformer following STFT tokenisation. The token sequence $(B, C \cdot t_s, 256)$ is mean-pooled over the sequence dimension to produce a 256-dimensional representation, which is passed to the same two-layer classification head. The Stage 2 alignment checkpoint is used. *Feature dim:* 256. *Head:* LayerNorm(256) $\rightarrow$ Linear(256, N).

BrainECHO w/ STFT. Identical to BrainECHO w/ Full BIOT but using BrainECHO's standard transformer encoder following STFT tokenisation. The Stage 2 alignment checkpoint is used. *Feature dim:* 256. *Head:* LayerNorm(256) $\rightarrow$ Linear(256, N).

B.3 Noise Conditions and Evaluation Protocol

Two additional evaluation conditions are reported for each model to assess robustness and verify that models are using EEG signal content rather than label-correlated artifacts:

Input Noise: The entire EEG input is replaced with independently sampled Gaussian noise $\mathcal{N}(\mu EEG, \sigma EEG)$ of the same shape, using the trained model's weights without modification. A model that achieves above-chance accuracy under this condition has learned to exploit biases in the training data (e.g. class-imbalanced batches or dataset statistics) rather than EEG signal content.

Length-Matched Noise: Only the real-signal portion of the EEG (identified by the attention mask, which marks non-padded timepoints) is replaced with Gaussian noise, while the zero-padding region is left unchanged. This condition tests whether models exploit the signal–padding boundary as a shortcut rather than genuine EEG patterns.

For generation models, a **Random** baseline replaces model output with a sentence drawn uniformly at random from the corpus, providing a lower bound on lexical overlap metrics. The **Classification-to-Generation (Cls2Gen)** condition retrieves the transcript of the classifier's predicted class and scores it as if it were generation output; full details are in Appendix G.

C Full Classification Results

Full classification results are presented in Table 3

Why most models collapse on NeuraSearch. NeuraSearch combines a low mean padding fraction (36.7%) and low padding variance (12.5%) with sentences that are length-matched by design (mean 7.4 tokens, SD 0.8). Length is therefore barely exploitable, which is the primary, dataset level cause of the collapse, and models that lean heavily on length on Brennan and HBN fall toward chance, with EEGNet at 10.4% and TIDNet at 7.8% under length-matched noise. On top of this dataset level effect, three architectures retain a secondary exploitation of the residual boundary. DeepConvNet scores 62.6% under length-matched noise against 68.2% under normal input, EEGConformer scores 60.3% against 65.2%, and ShallowConvNet scores 31.5% against 65.9%. What these three models have in common is a convolutional stage that still resolves time finely enough to locate where signal meets padding even when length variance is small, which is the same mechanism attributed to BrainECHO's temporal-spatial convolution. This is an architecture specific effect rather than a convolution-wide one, since EEGNet, ATCNet, and TIDNet are also convolutional and still collapse. The confusion analysis in Appendix J completes the picture, since the most confused NeuraSearch pairs reflect EEG representational similarity rather than duration.

Table 3: Full sentence classification results across all ten models and three datasets. Values are subject mean ± subject SD. *Chance* is $1/N$ classes. *+Input Noise* replaces the entire EEG with Gaussian noise; *+Len. Noise* replaces only the true signal region. Blue rows denote STFT-based length-invariant architectures.

Model	**Brennan**		**HBN**		**NeuraSearch**	
	Acc.	*F1*	*Acc.*	*F1*	*Acc.*	*F1*
Chance	0.7	0.7	0.6	0.6	2.1	2.1
Braindecode end-to-end classifiers						
EEGNet	65.6 ±26.3	61.4 ±28.7	69.3 ±7.4	63.5 ±8.0	62.6 ±16.5	56.7 ±17.8
+ Input Noise	0.6 ±0.2	0.0 ±0.1	0.4 ±0.5	0.0 ±0.0	2.8 ±0.4	0.2 ±0.0
+ Len. Noise	50.0 ±2.5	40.8 ±2.0	42.2 ±2.4	33.3 ±2.3	10.4 ±1.1	4.6 ±0.5
ShallowConvNet	29.4 ±3.7	19.6 ±3.0	18.5 ±1.9	10.7 ±1.7	65.9 ±6.5	59.9 ±6.3
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.7 ±0.7	0.2 ±0.1
+ Len. Noise	21.0 ±1.7	11.4 ±1.6	18.6 ±0.7	10.0 ±0.7	31.5 ±1.2	22.0 ±1.6
DeepConvNet	75.7 ±2.5	72.1 ±3.0	72.5 ±3.5	65.6 ±4.1	68.2 ±0.8	60.1 ±0.9
+ Input Noise	0.5 ±0.3	0.0 ±0.0	0.6 ±0.3	0.0 ±0.0	2.2 ±0.3	0.2 ±0.3
+ Len. Noise	80.8 ±1.1	75.8 ±1.4	59.7 ±1.7	50.6 ±1.8	62.6 ±2.5	57.6 ±1.8
EEGConformer	58.3 ±1.2	47.5 ±1.6	52.6 ±2.3	43.2 ±2.0	65.2 ±3.5	57.3 ±3.6
+ Input Noise	0.5 ±0.6	0.0 ±0.1	0.8 ±0.6	0.1 ±0.1	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	50.7 ±2.2	40.2 ±2.1	22.2 ±1.5	13.3 ±1.2	60.3 ±2.0	52.6 ±1.9
TIDNet	72.2 ±2.0	67.1 ±1.7	73.3 ±5.3	68.0 ±5.6	67.1 ±2.0	62.5 ±2.1
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.5	0.0 ±0.1	2.1 ±0.4	0.1 ±0.1
+ Len. Noise	63.5 ±1.3	54.0 ±1.7	61.8 ±1.6	53.8 ±1.7	7.8 ±0.5	1.9 ±0.3
ATCNet	20.6 ±0.6	12.2 ±0.7	32.3 ±2.6	22.8 ±2.3	69.4 ±3.0	63.4 ±3.3
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	13.9 ±0.8	6.5 ±0.8	27.4 ±2.2	18.2 ±1.8	7.3 ±0.3	3.7 ±0.1
BIOT	46.0 ±4.7	38.1 ±4.7	52.9 ±7.1	46.5 ±6.9	65.6 ±3.5	58.9 ±3.7
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	0.7 ±0.0	0.0 ±0.0	1.2 ±0.0	0.1 ±0.0	2.1 ±0.0	0.1 ±0.0
Frozen encoder probes						
BrainECHO	85.5 ±2.4	81.5 ±2.5	45.7 ±5.6	38.9 ±5.2	17.7 ±1.2	11.1 ±0.9
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	6.6 ±1.2	3.0 ±0.7	14.0 ±1.5	7.5 ±1.2	6.2 ±0.0	2.5 ±0.0
BrainECHO w/ STFT	39.3 ±4.5	30.8 ±4.7	50.2 ±5.3	43.5 ±5.5	66.9 ±4.6	60.2 ±4.3
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	8.5 ±0.2	2.9 ±0.2	2.4 ±0.2	0.7 ±0.2	2.1 ±0.0	0.1 ±0.0
BrainECHO w/ Full BIOT	25.3 ±1.8	16.3 ±1.1	29.3 ±4.9	22.4 ±4.0	65.7 ±2.9	59.3 ±3.1
+ Input Noise	0.7 ±0.0	0.0 ±0.0	0.6 ±0.0	0.0 ±0.0	2.1 ±0.0	0.1 ±0.0
+ Len. Noise	0.7 ±0.0	0.0 ±0.0	1.8 ±0.0	0.1 ±0.0	4.2 ±0.0	0.3 ±0.0

D Full Generation Results

Complete generation results can be found in Table 4

E BIOT Decomposition: Full Results and Mathematical Verification

E.1 Decomposition Results by Dataset

Table 5 summarises the classification accuracy of all eight BIOT decomposition variants under the normal and length-matched noise conditions across all three datasets. A variant is considered *length-biased* if its length-noise accuracy is substantially above chance.

The variants without STFT replace the STFT tokeniser with a Conv1d with kernel size 200 and stride 100, applied per channel to the raw signal, while all other components stay unchanged.

The key finding is that `no_stft` and all variants without STFT achieve well-above-chance accuracy under length-matched noise on Brennan and HBN, regardless of whether channel tokens or positional encodings are present. All variants that include STFT collapse to near-chance under length-matched noise. The NeuraSearch exception (all variants collapse to chance under length-matched noise) may be attributed to NeuraSearch's shorter and less variable sentences, which may not provide sufficient length signal for shortcuts to emerge even without STFT.

E.2 Numerical Verification of STFT Length Invariance

For each test sample, we apply STFT to (A) the full padded signal $[x_{\text{real}} \mid \mathbf{0}]$ and (B) the real signal x_{real} only, and compare the resulting spectrograms frame-by-frame over the real-signal frames. Since STFT is linear and operates on non-overlapping windows (centre=False), a window w composed entirely of zeros satisfies $\text{STFT}(w) = \mathbf{0}$ exactly. The maximum absolute difference between spectrograms A and B over real frames is 0.00 across all three datasets ($N = 197$ Brennan, $N = 120$ HBN, $N = 200$ NeuraSearch), confirming invariance. Eighty HBN samples with real signal shorter than $n_\text{fft} = 200$ samples are excluded as they cannot be tokenised. This comparison excludes the single boundary frame that may straddle the real-signal and padding regions and such a frame carries partial length information and represents a minimal but structurally unavoidable residual cue.

F Generation Baseline Implementation Details

F.1 BrainECHO

BrainECHO [21] follows a three-stage decoupled training strategy designed to prevent the pre-trained language model from dominating training and suppressing the EEG signal.

Stage 1 Mel-VQVAE Pre-training. A Vector-Quantized Variational Autoencoder (VQVAE) is pre-trained to reconstruct log Mel spectrograms extracted from the audio stimuli of each dataset. The VQVAE consists of a convolutional encoder, a learned codebook of size 2048, and a convolutional decoder. The Mel spectrograms are computed using the Whisper processor at 16 kHz with 80 mel bins. Training uses the Adam optimiser with learning rate 2×10^{-4} for 400 epochs with batch size 16.

Stage 2 Brain-Audio Alignment. A Conformer-based EEG encoder (4 Transformer layers, $d_{\text{model}} = 256$, FFN dimension 1024) with temporal-spatial convolutions is trained to map raw EEG signals to the VQVAE latent space. The alignment loss is a mean squared error between the EEG embedding and the quantized Mel spectrogram embedding for the corresponding sentence. Training uses AdamW with learning rate 1×10^{-4} for 40 epochs with batch size 16. A 5% noise augmentation is applied by randomly replacing one sample per batch with Gaussian noise paired with a silence Mel spectrogram.

Stage 3 Whisper Fine-tuning. The Whisper-base (English, `openai/whisper-base.en`) encoder is added and fine-tuned jointly with the EEG encoder using AdaLoRA [38] applied to the attention projection layers (`k_proj`, `q_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`). The EEG encoder frontend modules (`TS_Conv`, `Transformer`, `pos_embed_e`) are frozen. Initial rank $r = 12$, target rank $r = 4$. Training uses AdamW with learning rate 1×10^{-4} for 40 epochs, batch size 16. Early stopping patience of 4 epochs.

F.2 BrainECHO w/ STFT

BrainECHO w/ STFT is a variant of BrainECHO in which the temporal-spatial convolutional frontend (`TS_Conv`) is replaced with an STFT-based tokenizer, while all other components inlcuding the standard `TransformerEncoder`, alignment loss, Mel-VQVAE, and Whisper fine-tuning stages, remain identical to the original BrainECHO. The motivation is to isolate the contribution of frequency-domain tokenization to length invariance, independently of any changes to the transformer architecture.

EEG Encoder STFT Tokenizer. Each EEG channel $c \in \{1, \ldots, C\}$ is processed independently. The raw time-domain signal $\mathbf{x}_c \in \mathbb{R}^T$ is passed through a Short-Time Fourier Transform (STFT) with $n_{\text{fft}} = 200$, hop_length $= 100$, and `center=False`, yielding a magnitude spectrogram $\mathbf{S}_c \in \mathbb{R}^{n_{\text{freq}} \times t_s}$, where $n_{\text{freq}} = n_{\text{fft}}/2 + 1 = 101$ and t_s is the number of STFT frames. The frequency axis is projected to $d_{\text{model}} = 256$ via a linear layer (`PatchFrequencyEmbedding`), producing a token sequence $\mathbf{T}_c \in \mathbb{R}^{t_s \times d_{\text{model}}}$. Learned per-channel identity embeddings (channel tokens) $\mathbf{e}_c \in \mathbb{R}^{d_{\text{model}}}$

and sinusoidal positional encodings are added to each token. The token sequences for all channels are concatenated along the sequence dimension, yielding $\mathbf{E} \in \mathbb{R}^{(C \cdot t_s) \times d_{\text{model}}}$.

Length Invariance. The STFT is a linear operation applied to non-overlapping windows. All frames falling entirely within the zero-padded region carry exactly zero frequency energy. A single boundary frame may straddle the signal/padding boundary, partially encoding sequence length but this is structurally unavoidable with fixed-length STFT framing and represents a minimal residual length cue compared to other encoders.

Temporal Pooling. Since the STFT produces $C \cdot t_s$ tokens (compared to BrainECHO's fixed mel_len$/r = 300$ temporal steps after `TS_Conv`), an `AdaptiveAvgPool1d` layer reduces the token sequence to $2 \times$ mel_len$/r = 600$ tokens along the temporal dimension before being passed to `fc_eeg` and `conv_out`, which then produce the final 300 spatial-temporal tokens matching the downstream Mel-VQVAE latent format. All remaining downstream layers (`fc_eeg`, `conv_out`, `mid_fc`, `ln`) and training stages are identical to BrainECHO.

Stage 3 Frozen Modules. During Whisper fine-tuning, the STFT tokenizer components (`patch_embedding`, `channel_tokens`, `positional_encoding`), the `TransformerEncoder`, and the `temporal_pool` layer are frozen. AdaLoRA is applied to the Whisper encoder with the same configuration as BrainECHO.

F.3 BrainECHO w/ Full BIOT

BrainECHO w/ Full BIOT replaces BrainECHO's entire EEG encoder frontend. Both the `TS_Conv` tokenizer and the standard `TransformerEncoder` are replaced with the STFT-based encoder from BIOT [34]. This variant therefore tests whether BIOT's linear attention transformer architecture provides additional benefit over BrainECHO's standard transformer when combined with STFT tokenization. Stages 1 and 3, and all downstream layers from `fc_eeg` onwards, are identical to BrainECHO.

EEG Encoder BIOT STFT Encoder. The STFT tokenization procedure is identical to BrainECHO w/ STFT (same n_{fft}, hop_length, channel tokens, and positional encodings), producing $\mathbf{E} \in \mathbb{R}^{(C \cdot t_s) \times d_{\text{model}}}$. The token sequence is then processed by BIOT's `LinearAttentionTransformer` (4 layers, $d_{\text{model}} = 256$, 8 heads) rather than BrainECHO's standard `TransformerEncoder`. BIOT's transformer uses linear attention with feed-forward networks implemented as `SwiGLU` gated units, which reduces the quadratic attention complexity over the long $(C \cdot t_s)$ sequence from $O((C \cdot t_s)^2)$ to $O(C \cdot t_s)$, an important efficiency consideration given the $C \cdot t_s = 62 \times 23 = 1{,}426$ token sequence for Brennan at $T = 2{,}400$ samples.

Temporal Pooling. Identical to BrainECHO w/ STFT: `AdaptiveAvgPool1d` to $2 \times$mel_len$/r = 600$ tokens.

Stage 3 Frozen Modules. During Whisper fine-tuning, the `biot_encoder` and `temporal_pool` are frozen. AdaLoRA is applied to the Whisper encoder with the same configuration as BrainECHO.

Comparison with BrainECHO w/ STFT. The three variants form a controlled ablation over tokenizer and transformer architecture choices:

F.4 Parameter Counts

Parameter counts on Brennan. On Brennan, the two STFT-based variants differ by about 3,000 parameters in total, all of it in the transformer block (3,159,040 against 3,155,968). Both share one STFT tokeniser of 41,984 parameters and identical downstream stages, so the only architectural difference is the encoder, where one uses standard full quadratic self-attention and the other uses BIOT's linear-attention transformer over the same tokens. The linear attention design buys efficiency at the cost of expressive power, which is the most plausible explanation for the consistent gap between the two variants. Every variant is trained on the same splits with the same epochs and sample counts, so sample count is not a confound.

G Classification-to-Generation Evaluation: Methodological Details

G.1 Method

For each test sample we proceed as follows:

1. Run the trained classification model on the preprocessed EEG to obtain a predicted sentence class $\hat{y} \in \{0, \ldots, N-1\}$.
2. Retrieve the target string associated with $\hat{y}$ from the test split files via the `target_string` field stored per sample.
3. Treat the retrieved string as the model's predicted text and the ground truth `target_string` as the reference.
4. Compute BLEU-1 through BLEU-4, WER, and ROUGE-1/2/L, with per-subject grouping and mean $\pm$ SD.

When $\hat{y} = y$ the retrieved text is identical to the reference, contributing a perfect score. When $\hat{y} \neq y$ the score reflects lexical overlap between two distinct sentences, which is near zero for semantically distinct stimuli. The overall retrieval BLEU-4 is therefore $\approx$ accuracy $\times\, 1.0 + (1 - \text{accuracy}) \times \epsilon$.

G.2 Noise Conditions

The Cls2Gen baseline is evaluated under three conditions: **Normal** (standard inference), **Input noise** (entire EEG replaced with $\mathcal{N}(\mu EEG, \sigma EEG)$), and **Length noise** (real signal region only replaced with $\mathcal{N}(\mu EEG, \sigma EEG)$, padding preserved). A large normal-to-LN gap indicates the classifier exploits sentence duration with a small gap indicating reliance on EEG content.

G.3 Classification-to-Generation Full Results

Tables 8, 9, and 10 report the full Cls2Gen results for all ten classifiers across Brennan, HBN, and NeuraSearch respectively. For each model, the predicted sentence class is used to retrieve the corresponding transcript, which is then scored with the same generation metrics applied to the generation models.

Brennan. On Brennan, BrainECHO's frozen encoder achieves the highest Cls2Gen BLEU-4 of any classifier (89.1%), which is within 2.0 points of BrainECHO's actual Whisper generation score (91.1%). This near-identical performance between a simple linear classification head and a full autoregressive Whisper decoder is the strongest evidence that BrainECHO's generation pipeline is operating as a retrieval system on this dataset. Its length-noise Cls2Gen score drops substantially (6.5% BLEU-4), confirming that BrainECHO's encoder representations retain some length sensitivity but are not solely reliant on it. BrainECHO w/ Full BIOT and BrainECHO w/ STFT both collapse to chance under length noise (1.7% and 10.4% BLEU-4 respectively), with BrainECHO w/ STFT's residual reflecting the small encoder-side length sensitivity noted in the main classification results. Among end-to-end classifiers, BIOT collapses entirely to chance under length noise (0.3% BLEU-4), with its normal Cls2Gen score (54.4%) considerably below BrainECHO's, consistent with its moderate classification accuracy on Brennan.

HBN. On HBN, BrainECHO's Whisper generation score (81.9% BLEU-4) substantially exceeds its Cls2Gen score (53.9%). However, BrainECHO's Cls2Gen length-noise score (10.6%) is considerably lower than its generation length-noise score (57.3%), indicating that Whisper's residual length dependence on HBN originates primarily from the spectrogram input rather than the encoder representations alone. Among end-to-end classifiers, DeepConvNet achieves the highest Cls2Gen BLEU-4 (83.8%) but with extreme length sensitivity (73.9% under length noise), confirming that its performance is predominantly length-driven. TIDNet achieves similarly high scores (82.7%) with comparable length sensitivity (74.9%). BIOT again collapses entirely to chance under length noise (0.9% BLEU-4) while maintaining above-chance normal performance (64.9%). BrainECHO w/ Full BIOT and BrainECHO w/ STFT both show near-zero length-noise Cls2Gen scores (1.9% and 3.2% respectively), confirming full encoder-side length invariance, while their normal Cls2Gen scores (43.1% and 62.0%) bracket the standalone BIOT result.

NeuraSearch. In NeuraSearch the Cls2Gen scores of BrainECHO w/ Full BIOT (65.8%) and BrainECHO w/ STFT (66.6%) both substantially exceed their respective Whisper generation scores (33.2% and 48.5%), and are competitive with the best end-to-end classifiers on this dataset (ATCNet 69.6%, BIOT 65.3%). This demonstrates that the sentence-discriminative information is present in the STFT-based encoder representations, recoverable by a linear head, but Whisper's decoder cannot exploit it effectively without the length cue. BrainECHO's frozen encoder performs poorly on NeuraSearch (18.3% Cls2Gen BLEU-4), consistent with its Stage 2 representations being poorly suited to the smaller, more controlled sentence set when length information is limited. BIOT's length-noise Cls2Gen score collapses fully to chance (2.3% BLEU-4) while its normal score (65.3%) is among the highest of any model, reinforcing that BIOT's discriminative performance on NeuraSearch can be attributed to EEG content rather than segment duration.

H Exact Training String Matches and Length-Stratified Results

This appendix reports two analyses. The first measures how often each generation model's output equals a sentence from the training inventory. The second bins test samples by real signal length and scores each model inside each bin, so that duration carries little information within a bin.

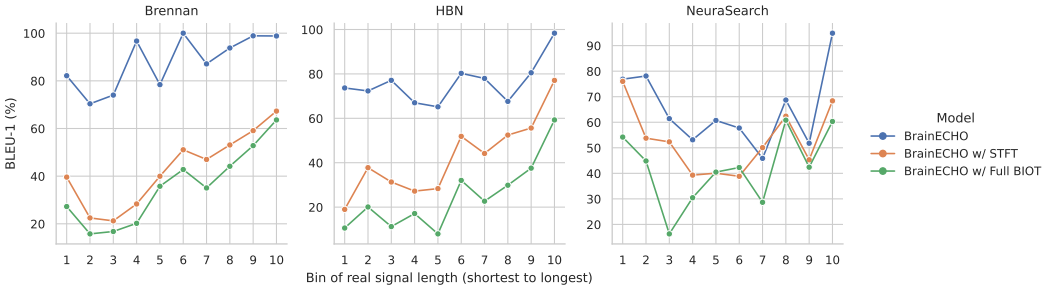

Figure 6: Length-stratified BLEU-1 (0 to 100 scale) by decile of real signal length (samples at the 200 Hz sampling rate), one line per model per dataset. Bins 1 to 10 run from the shortest to the longest real signal length. HBN excludes the story *The Present*. BLEU-4 is zero in the two shortest HBN bins for every model because those segments are too short for 4-gram overlap, which is why the main text argues with BLEU-1.

I Qualitative Examples of Generated Outputs

Tables 13, 14, and 15 present curated outputs from all three generation models on Brennan, HBN, and NeuraSearch respectively. Rows are selected to illustrate four qualitatively distinct patterns. First (i), all models produce verbatim correct output, confirming that decoding above chance is achievable. Second (ii), BrainECHO is correct while the STFT variants produce incorrect output, reflecting the cost of suppressing the encoder-side length cue exploited by BrainECHO. Third (iii), all models fail, but via distinct mechanisms, with BrainECHO substituting a different corpus sentence verbatim [CS], BrainECHO w/ Full BIOT generating novel text absent from the stimulus corpus entirely [H], and BrainECHO w/ STFT typically lying between the two. Fourth (iv), BrainECHO w/ Full BIOT or BrainECHO w/ STFT producing output that preserves the correct syntactic frame or phrasal fragment with only a single substitutable error, and outputs of this type are candidates for recovery by a post-hoc language model corrector that could exploit contextual and linguistic priors, a capability structurally unavailable to any retrieval-based approach. HBN additionally exhibits a fifth pattern unique to its multi-stimulus design, (v) cross-movie contamination [XC], in which the predicted sentence originates from a different stimulus film than the target, revealing that models treat the combined multi-movie corpus as a single retrieval inventory rather than partitioning by stimulus context.

Italics denote verbatim correct output (up to minor transcription normalisation such as capitalisation). In near-correct outputs, the differing word(s) are shown in **bold**. [CS] marks a corpus substitution (a different stimulus sentence retrieved verbatim); [H] marks open-vocabulary hallucination (text

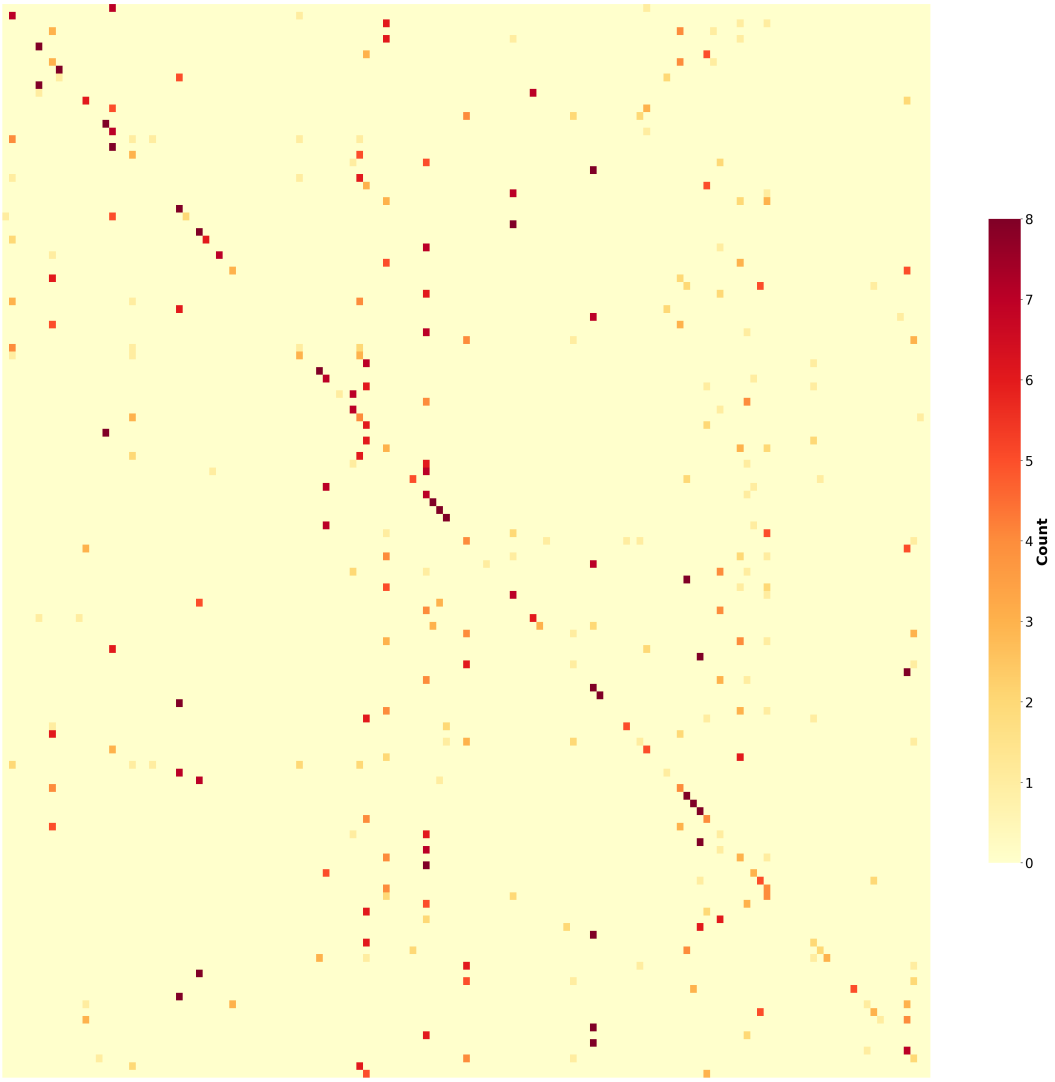

Figure 7: Sentence-level confusion matrix for BrainECHO w/ Full BIOT on Brennan (139 sentences). Off-diagonal mass is concentrated in a small number of systematically confused pairs rather than distributed uniformly.

not present in the stimulus set); [XC] marks cross-movie contamination (sentence retrieved from a different stimulus film).

J Confusion Matrix Analysis

Figures 7, 8, and 9 show the sentence-level confusion matrices for BrainECHO w/ Full BIOT's frozen encoder with linear classification head across all three datasets. Tables 16, 17, and 18 list the ten most confused sentence pairs per dataset, reporting the number of times the true sentence was predicted as the confused alternative across all test subjects. The key observation across all three datasets is that the most frequent misclassifications are *systematic* as the same sentence pairs are confused repeatedly across held-out subjects, indicating that these errors reflect structural ambiguity in the learned EEG representations rather than random decoder noise.

Brennan. Confusion counts are uniformly low (maximum 8 across all pairs), consistent with Brennan's large sentence set (139 classes) and the genuinely discriminative EEG responses elicited by a naturalistic audiobook. The confused pairs largely involve sentences of comparable length,

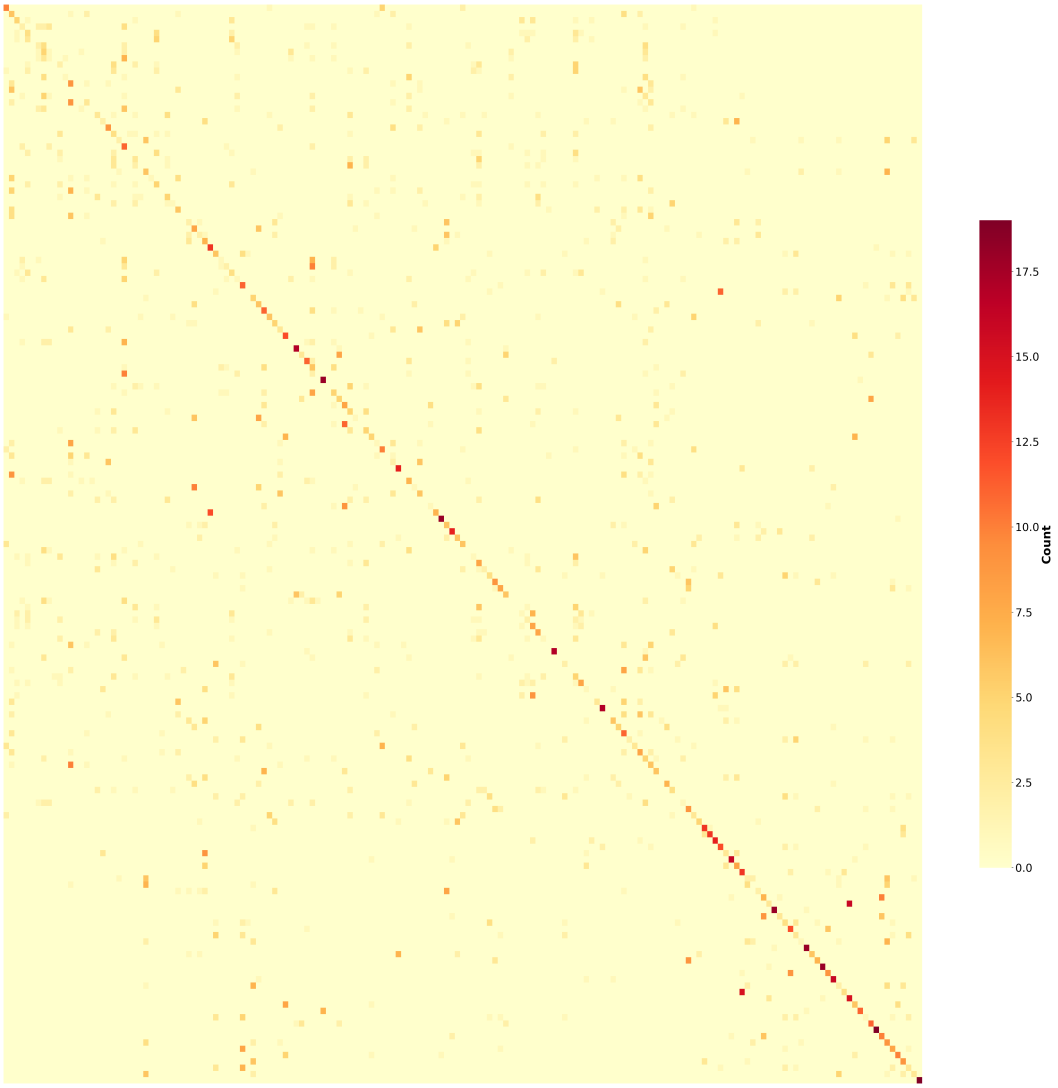

Figure 8: Sentence-level confusion matrix for BrainECHO w/ Full BIOT on HBN (171 sentences). Confusions cluster among very short sentences, consistent with their high padding fraction and limited real-signal content.

suggesting that even for a less length-prone encoder, sentences with similar temporal structure produce overlapping EEG representations.

HBN. The most confused pairs on HBN are dominated by very short sentences like "Manny", "Fine", "Keep reading", "Good night, kittens", which share high padding fractions and minimal real-signal content. These confusions reflect a genuine decoding ceiling as when the real EEG segment is only a few hundred samples long within a 2400-sample window, the encoder has limited signal to distinguish between sentences of similar duration and prosodic structure. Confusion counts are higher than Brennan (up to 16), consistent with HBN's more extreme padding regime.

NeuraSearch. Several of the most confused pairs on NeuraSearch cross the biased/unbiased sentence boundary like for example, "He eats quickly, then washes dirty dishes" (unbiased) confused with "Older athletes should retire instead of competing" (biased). Since NeuraSearch sentences are explicitly length-matched across conditions, these cross-condition confusions confirm that BrainECHO w/ Full BIOT's errors on this dataset are driven by EEG representational similarity rather than segment duration. The confusion counts are higher than Brennan (up to 18 across 6 test subjects), reflecting

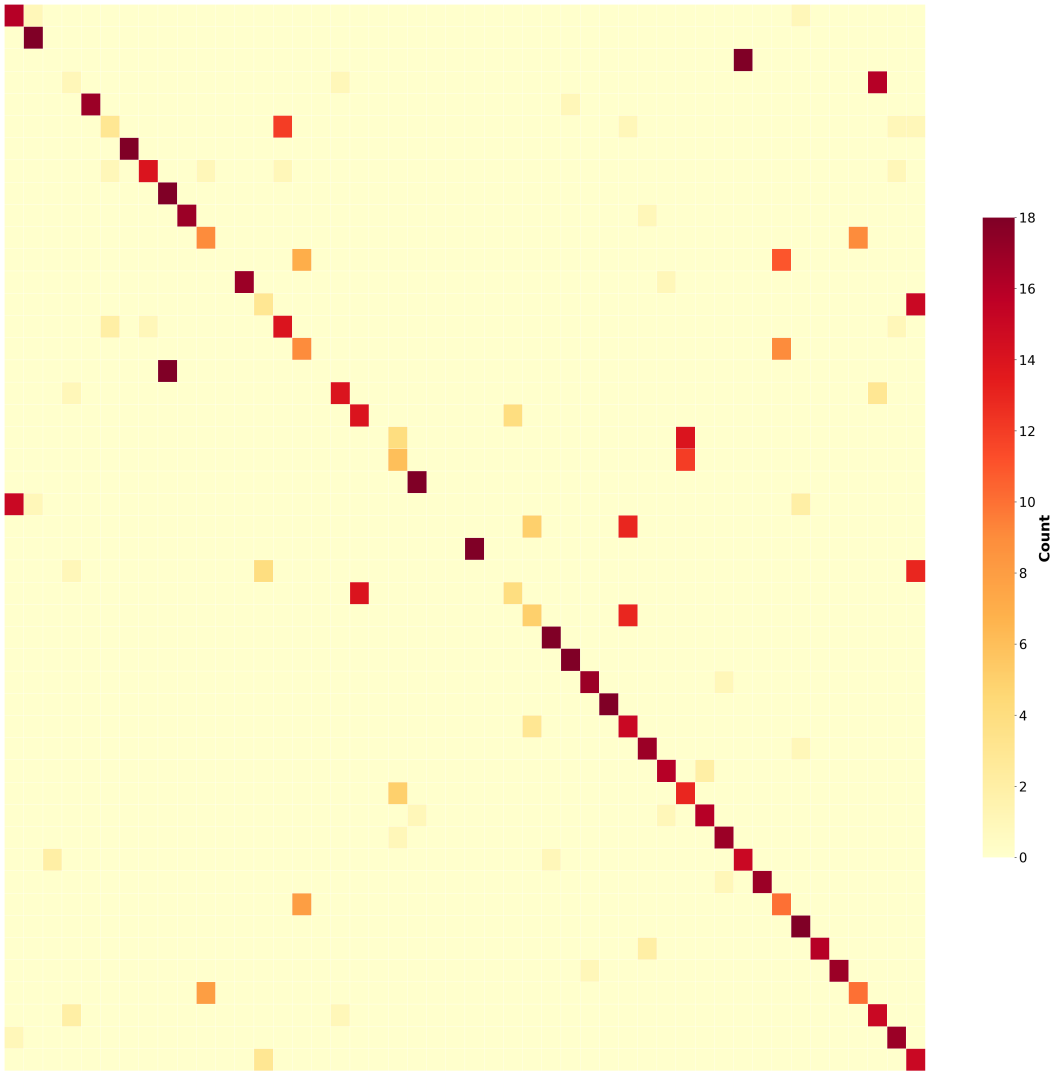

Figure 9: Sentence-level confusion matrix for BrainECHO w/ Full BIOT on NeuraSearch (48 sentences). Despite sentences being length-matched by design, systematic confused pairs persist, confirming that confusions are driven by EEG representational similarity rather than duration.

the smaller sentence set and the greater decoding difficulty per sentence given NeuraSearch's shorter stimulus durations.

K Compute Details

All training and evaluation reported in this paper ran on a single NVIDIA RTX A6000 GPU with 48 GB of memory. As indicative timings for the BrainECHO stages, the VQVAE pre-training, brain-audio alignment, and Whisper fine-tuning stages took about 30 minutes, 1.5 hours, and 1 hour on Brennan, about 40 minutes, 5 hours, and 8 hours on HBN, and about 2 minutes, 15 minutes, and 1 hour on NeuraSearch. These timings exclude preprocessing and evaluation.

Table 4: Full generation results including Gaussian noise conditions across all datasets. Values are subject mean ± subject SD. *+Noise* replaces the entire EEG input with Gaussian noise drawn from the test signal distribution. *+Len-Noise* replaces only the true signal region. *+Cls2Gen* retrieves the predicted sentence transcript from a linear classifier. Blue rows denote STFT-based length-invariant variants.

	BLEU-N (%) ↑				ROUGE (%) ↑		WER (%) ↓
Model	*N=1*	*N=2*	*N=3*	*N=4*	*R-1*	*R-2*	*WER*
Brennan							
Random	8.7 ±0.5	3.1 ±0.7	1.9 ±0.7	1.4 ±0.7	10.2 ±0.5	1.4 ±0.5	127.5 ±2.8
BrainECHO	**92.1** ±0.5	**91.5** ±0.5	**91.3** ±0.6	**91.1** ±0.6	**89.7** ±0.4	**88.3** ±0.5	**8.8** ±0.5
+ Cls2Gen	89.7 ±1.6	89.2 ±1.7	89.1 ±1.8	89.1 ±1.8	87.2 ±2.1	85.6 ±2.3	11.7 ±1.9
+ Noise	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	1.1 ±0.1	0.0 ±0.0	99.9 ±0.1
+ Len-Noise	28.6 ±1.1	25.4 ±0.9	23.7 ±0.9	22.5 ±1.0	47.1 ±1.3	40.0 ±1.3	74.9 ±0.9
BrainECHO w/ STFT	50.2 ±3.4	47.1 ±3.8	46.0 ±3.9	45.6 ±4.0	46.0 ±2.4	38.8 ±3.2	59.5 ±4.4
+ Cls2Gen	50.1 ±5.2	47.2 ±5.8	46.3 ±6.0	45.9 ±6.1	46.1 ±3.9	39.7 ±4.5	59.8 ±6.6
+ Noise	3.2 ±0.2	0.6 ±0.1	0.0 ±0.0	0.0 ±0.0	7.6 ±0.4	0.3 ±0.1	99.8 ±0.3
+ Len-Noise	12.7 ±0.8	7.2 ±0.6	5.3 ±0.5	4.2 ±0.5	15.5 ±0.9	5.6 ±0.8	103.1 ±1.1
BrainECHO w/ Full BIOT	43.1 ±3.4	39.5 ±3.8	38.0 ±4.1	37.2 ±4.2	38.9 ±2.8	30.7 ±3.6	68.0 ±4.5
+ Cls2Gen	37.3 ±2.4	33.5 ±2.6	32.4 ±2.6	31.9 ±2.6	33.7 ±1.8	25.7 ±1.8	74.6 ±3.0
+ Noise	0.4 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	3.3 ±0.0	0.2 ±0.0	98.7 ±0.0
+ Len-Noise	5.6 ±0.3	1.9 ±0.2	1.0 ±0.1	0.5 ±0.2	9.6 ±0.4	1.8 ±0.4	113.5 ±0.9
HBN							
Random	3.0 ±0.9	1.3 ±1.0	0.9 ±1.0	0.7 ±0.9	3.5 ±0.9	0.7 ±0.5	137.2 ±3.8
BrainECHO	**81.8** ±2.5	**81.7** ±2.4	**81.8** ±2.3	**81.9** ±2.2	**76.8** ±3.6	**68.2** ±3.1	**21.8** ±3.1
+ Cls2Gen	53.8 ±5.4	53.3 ±5.6	53.4 ±5.7	53.9 ±5.7	48.1 ±5.5	40.7 ±5.1	55.4 ±6.7
+ Noise	3.2 ±0.5	0.4 ±0.4	0.1 ±0.2	0.0 ±0.0	3.8 ±0.6	0.3 ±0.2	136.7 ±5.0
+ Len-Noise	60.6 ±1.6	59.3 ±1.6	58.3 ±1.7	57.3 ±1.8	60.7 ±1.8	52.4 ±1.7	47.2 ±2.4
BrainECHO w/ STFT	54.2 ±5.6	53.7 ±6.0	54.0 ±6.3	54.5 ±6.7	44.9 ±4.3	40.4 ±4.4	55.9 ±7.5
+ Cls2Gen	61.3 ±5.2	61.1 ±5.4	61.4 ±5.5	62.0 ±5.6	53.1 ±5.3	47.0 ±5.0	46.2 ±6.7
+ Noise	3.6 ±0.2	0.9 ±0.2	0.2 ±0.2	0.0 ±0.0	6.3 ±0.5	0.5 ±0.2	267.5 ±8.8
+ Len-Noise	3.9 ±0.6	1.5 ±0.5	0.6 ±0.5	0.1 ±0.3	5.6 ±0.7	1.3 ±0.6	121.7 ±3.5
BrainECHO w/ Full BIOT	34.8 ±3.7	33.6 ±4.1	33.5 ±4.5	33.8 ±4.8	27.9 ±2.6	23.8 ±2.9	81.5 ±6.5
+ Cls2Gen	42.3 ±4.6	41.8 ±4.7	42.3 ±4.8	43.1 ±4.9	32.8 ±4.7	27.7 ±4.5	68.0 ±5.7
+ Noise	3.2 ±0.4	0.7 ±0.3	0.2 ±0.3	0.0 ±0.0	4.4 ±0.6	0.3 ±0.2	143.7 ±4.8
+ Len-Noise	4.4 ±0.4	1.7 ±0.3	0.8 ±0.2	0.2 ±0.4	6.5 ±0.5	1.2 ±0.2	154.9 ±4.6
NeuraSearch							
Random	3.1 ±0.7	2.0 ±0.8	1.7 ±0.8	1.6 ±0.9	4.2 ±0.8	2.0 ±0.7	103.6 ±1.1
BrainECHO	**65.1** ±1.6	**64.1** ±1.6	**63.4** ±1.7	**62.7** ±1.7	**66.7** ±1.6	**64.8** ±1.5	**38.1** ±1.9
+ Cls2Gen	20.4 ±0.9	19.1 ±1.0	18.6 ±1.0	18.3 ±1.0	21.9 ±1.3	19.0 ±1.3	86.2 ±1.4
+ Noise	1.8 ±0.2	0.3 ±0.1	0.0 ±0.0	0.0 ±0.0	3.0 ±0.2	0.2 ±0.1	234.1 ±7.1
+ Len-Noise	59.3 ±1.3	57.6 ±1.4	56.4 ±1.4	55.0 ±1.5	62.7 ±1.2	59.9 ±1.3	44.9 ±1.7
BrainECHO w/ STFT	52.0 ±6.6	50.7 ±6.8	49.6 ±6.9	48.5 ±6.9	54.9 ±5.6	52.4 ±5.9	54.7 ±9.5
+ Cls2Gen	67.1 ±4.5	66.8 ±4.6	66.7 ±4.6	66.6 ±4.6	68.1 ±4.5	67.2 ±4.6	34.6 ±4.9
+ Noise	0.1 ±0.1	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.1 ±0.1	0.0 ±0.0	101.3 ±0.7
+ Len-Noise	2.3 ±0.4	1.0 ±0.2	0.0 ±0.0	0.0 ±0.0	2.2 ±0.2	0.9 ±0.1	120.7 ±2.3
BrainECHO w/ Full BIOT	41.7 ±7.8	38.4 ±8.4	35.6 ±8.7	33.2 ±8.8	47.9 ±6.3	41.5 ±7.2	70.4 ±12.4
+ Cls2Gen	66.2 ±2.6	65.9 ±2.6	65.8 ±2.6	65.8 ±2.6	66.5 ±2.9	65.8 ±2.9	35.8 ±2.9
+ Noise	1.5 ±0.3	0.4 ±0.3	0.0 ±0.0	0.0 ±0.0	1.5 ±0.3	0.2 ±0.1	131.5 ±6.2
+ Len-Noise	1.0 ±0.2	0.1 ±0.2	0.0 ±0.0	0.0 ±0.0	1.0 ±0.2	0.1 ±0.0	130.1 ±4.7

Table 5: BIOT decomposition accuracy (%) under normal and length-matched noise (+LN) conditions. Chance levels are 0.72% on Brennan, 0.58% on HBN, and 2.08% on NeuraSearch.

	Brennan		**HBN**		**NeuraSearch**	
Variant	Normal	+LN	Normal	+LN	Normal	+LN
`full`	46.7	0.7	48.6	2.3	66.6	2.1
`no_pos_enc`	54.4	1.3	56.7	1.5	68.5	2.1
`no_chan_tok`	53.1	2.4	40.5	4.1	65.9	2.1
`no_stft`	73.0	58.8	70.9	45.8	35.8	2.1
`no_pos_no_chan`	60.2	5.4	61.3	1.8	68.1	2.1
`no_pos_no_stft`	78.2	60.2	73.2	24.3	65.6	2.9
`no_chan_no_stft`	78.8	59.9	71.2	38.0	57.2	2.1
`no_all`	77.3	60.5	72.6	35.8	67.1	4.2

Model	**Tokenizer**	**Transformer**
BrainECHO	`TS_Conv` (time domain)	Standard `TransformerEncoder`
BrainECHO w/ STFT	STFT	Standard `TransformerEncoder`
BrainECHO w/ Full BIOT	STFT	BIOT `LinearAttentionTransformer`

Table 6: EEG encoder architecture comparison across BrainECHO variants.

Table 7: Parameter counts on Brennan. The tokeniser column covers the tokenisation frontend, the transformer block column covers the encoder transformer, and the downstream column covers the alignment and decoding stages shared by all variants.

Model	**Total**	**Tokeniser**	**Transformer block**	**Downstream**
BrainECHO	43,727,085	2,211,456	3,159,040	36,023,172
BrainECHO w/ STFT	41,557,613	41,984	3,159,040	36,023,172
BrainECHO w/ Full BIOT	41,554,541	41,984	3,155,968	36,023,172

Table 8: Cls2Gen results on Brennan (10 classifiers). Each classifier predicts a sentence class from held-out test EEG; the predicted sentence text is then scored with standard generation metrics, bridging classification and generation evaluation. Values are subject mean ± subject SD. Bold denotes best main-row result per metric across all models.

	BLEU-N (%) ↑				ROUGE (%) ↑		WER (%) ↓
Model	*N=1*	*N=2*	*N=3*	*N=4*	*R-1*	*R-2*	*WER*
Braindecode end-to-end classifiers							
EEGNet	73.6 ±22.0	72.2 ±23.3	71.9 ±23.7	71.8 ±23.9	69.4 ±23.4	65.8 ±26.2	31.6 ±26.6
+ Input Noise	7.5 ±0.4	2.5 ±0.7	1.6 ±0.7	1.3 ±0.6	11.3 ±0.4	1.1 ±0.4	240.5 ±1.8
+ Len. Noise	58.2 ±2.4	55.7 ±2.7	54.9 ±2.8	54.7 ±2.9	54.3 ±1.7	48.4 ±2.3	49.4 ±2.8
ShallowConvNet	40.7 ±4.2	37.4 ±4.7	36.4 ±4.9	36.1 ±5.0	37.2 ±3.3	29.9 ±3.7	69.6 ±5.1
+ Input Noise	9.0 ±0.0	3.1 ±0.0	1.9 ±0.0	1.5 ±0.0	13.8 ±0.0	1.3 ±0.0	208.7 ±0.0
+ Len. Noise	32.2 ±1.4	28.2 ±1.5	26.8 ±1.6	26.2 ±1.6	30.0 ±1.2	22.2 ±1.3	81.7 ±1.8
DeepConvNet	80.5 ±3.1	79.5 ±3.3	79.3 ±3.4	79.2 ±3.4	78.4 ±2.3	75.9 ±2.5	22.9 ±3.6
+ Input Noise	8.3 ±0.5	2.7 ±0.9	1.5 ±0.9	1.1 ±0.9	12.9 ±0.5	1.1 ±0.5	229.0 ±1.7
+ Len. Noise	85.2 ±1.3	84.4 ±1.4	84.2 ±1.5	84.2 ±1.5	82.8 ±1.1	80.9 ±1.3	17.4 ±1.5
EEGConformer	68.5 ±1.4	66.9 ±1.6	66.6 ±1.7	66.5 ±1.8	62.7 ±1.0	58.5 ±1.2	37.5 ±1.9
+ Input Noise	7.5 ±0.3	2.7 ±0.6	1.7 ±0.7	1.4 ±0.6	11.5 ±0.3	1.3 ±0.4	241.9 ±1.3
+ Len. Noise	59.8 ±1.5	57.8 ±1.6	57.2 ±1.7	57.0 ±1.8	55.4 ±1.0	50.3 ±1.3	47.8 ±1.8
TIDNet	78.7 ±1.3	77.7 ±1.3	77.4 ±1.4	77.4 ±1.4	75.0 ±1.8	72.4 ±2.0	25.0 ±1.5
+ Input Noise	7.8 ±0.0	2.5 ±0.0	1.4 ±0.0	1.1 ±0.0	13.0 ±0.0	1.3 ±0.0	101.2 ±0.0
+ Len. Noise	70.6 ±1.8	69.2 ±2.0	68.8 ±2.0	68.7 ±2.1	66.9 ±1.8	63.4 ±1.9	34.9 ±2.0
ATCNet	34.3 ±0.8	30.4 ±0.8	29.3 ±0.9	28.9 ±0.9	29.7 ±0.6	20.9 ±0.6	78.6 ±1.2
+ Input Noise	9.0 ±0.0	3.1 ±0.0	1.9 ±0.0	1.5 ±0.0	13.8 ±0.0	1.3 ±0.0	208.7 ±0.0
+ Len. Noise	25.0 ±0.9	20.1 ±1.0	18.7 ±1.0	18.0 ±1.0	23.4 ±0.5	14.4 ±0.6	89.2 ±1.3
BIOT	57.5 ±4.4	55.2 ±4.8	54.6 ±5.0	54.4 ±5.1	52.0 ±3.9	46.4 ±4.7	50.8 ±5.5
+ Input Noise	1.3 ±0.0	0.3 ±0.0	0.1 ±0.0	0.1 ±0.0	9.0 ±0.0	0.7 ±0.0	95.8 ±0.0
+ Len. Noise	6.4 ±0.0	1.3 ±0.0	0.5 ±0.0	0.3 ±0.0	9.4 ±0.0	0.9 ±0.0	185.8 ±0.3
Frozen Encoder + Linear Classification Head							
BrainECHO	**89.7** ±1.6	**89.2** ±1.7	**89.1** ±1.8	**89.1** ±1.8	**87.2** ±2.1	**85.6** ±2.3	**11.7** ±1.9
+ Input Noise	9.0 ±0.0	3.1 ±0.0	1.9 ±0.0	1.5 ±0.0	13.8 ±0.0	1.3 ±0.0	208.7 ±0.0
+ Len. Noise	15.5 ±0.9	9.1 ±1.1	7.3 ±1.1	6.5 ±1.1	18.0 ±1.0	6.8 ±1.2	109.7 ±1.6
BrainECHO w/ Full BIOT	37.3 ±2.4	33.5 ±2.6	32.4 ±2.6	31.9 ±2.6	33.7 ±1.8	25.7 ±1.8	74.6 ±3.0
+ Input Noise	8.9 ±0.0	3.4 ±0.0	2.0 ±0.0	1.6 ±0.0	12.6 ±0.0	1.6 ±0.0	178.2 ±0.0
+ Len. Noise	9.0 ±0.0	3.4 ±0.0	2.1 ±0.0	1.7 ±0.0	12.6 ±0.0	1.3 ±0.0	154.4 ±0.1
BrainECHO w/ STFT	50.1 ±5.2	47.2 ±5.8	46.3 ±6.0	45.9 ±6.1	46.1 ±3.9	39.7 ±4.5	59.8 ±6.6
+ Input Noise	7.1 ±0.0	2.4 ±0.0	1.6 ±0.0	1.3 ±0.0	10.7 ±0.0	0.9 ±0.0	179.5 ±0.0
+ Len. Noise	18.7 ±0.1	12.8 ±0.1	11.1 ±0.2	10.4 ±0.2	19.6 ±0.2	8.9 ±0.2	100.6 ±0.3

Table 9: Cls2Gen results on HBN (10 classifiers). Each classifier predicts a sentence class from held-out test EEG; the predicted sentence text is then scored with standard generation metrics, bridging classification and generation evaluation. Values are subject mean ± subject SD. Bold denotes best main-row result per metric across all models.

	BLEU-N (%) ↑				ROUGE (%) ↑		WER (%) ↓
Model	*N=1*	*N=2*	*N=3*	*N=4*	*R-1*	*R-2*	*WER*
Braindecode end-to-end classifiers							
EEGNet	77.4 ±5.4	77.5 ±5.4	78.0 ±5.3	78.5 ±5.3	71.0 ±7.1	62.7 ±6.2	26.7 ±6.5
+ Input Noise	3.4 ±0.5	1.6 ±0.6	1.0 ±0.6	0.8 ±0.6	5.7 ±0.6	1.3 ±0.5	235.2 ±5.1
+ Len. Noise	58.0 ±2.9	58.2 ±3.0	59.1 ±3.0	60.3 ±3.1	46.3 ±2.4	39.9 ±2.4	49.5 ±3.2
ShallowConvNet	31.9 ±2.9	31.0 ±3.2	31.2 ±3.4	31.8 ±3.6	22.8 ±1.8	16.9 ±1.9	82.4 ±3.6
+ Input Noise	0.1 ±0.0	0.1 ±0.0	0.0 ±0.0	0.0 ±0.0	1.6 ±0.0	1.0 ±0.0	101.4 ±0.0
+ Len. Noise	31.9 ±1.0	30.9 ±1.1	31.1 ±1.1	31.8 ±1.2	23.8 ±0.6	18.4 ±0.6	82.7 ±1.2
DeepConvNet	**82.4** ±2.8	**82.7** ±2.9	**83.2** ±3.0	**83.8** ±3.0	74.6 ±3.4	**66.7** ±3.1	**21.3** ±3.5
+ Input Noise	3.5 ±0.9	1.3 ±0.7	0.6 ±0.8	0.4 ±0.8	3.6 ±0.7	0.7 ±0.4	150.1 ±6.1
+ Len. Noise	71.7 ±1.3	72.3 ±1.2	73.0 ±1.2	73.9 ±1.2	62.2 ±1.5	55.6 ±1.3	33.5 ±1.4
EEGConformer	66.2 ±2.3	66.5 ±2.4	67.3 ±2.4	68.2 ±2.4	56.2 ±2.4	50.0 ±2.2	40.3 ±3.0
+ Input Noise	3.4 ±0.7	1.3 ±0.8	0.7 ±0.8	0.6 ±0.7	4.5 ±0.8	0.7 ±0.5	155.8 ±3.1
+ Len. Noise	29.9 ±1.7	27.9 ±1.8	27.3 ±1.8	27.3 ±1.9	26.8 ±1.6	20.5 ±1.5	85.0 ±2.2
TIDNet	81.4 ±4.4	81.6 ±4.5	82.1 ±4.5	82.7 ±4.5	**74.9** ±5.1	66.7 ±4.9	22.3 ±5.2
+ Input Noise	1.9 ±0.8	0.8 ±0.6	0.5 ±0.6	0.2 ±0.7	2.2 ±0.4	0.5 ±0.4	126.4 ±4.1
+ Len. Noise	73.2 ±2.0	73.5 ±2.0	74.2 ±1.9	74.9 ±1.9	64.1 ±2.8	58.0 ±2.7	32.4 ±2.6
ATCNet	40.3 ±2.8	38.9 ±2.9	38.6 ±3.0	38.6 ±3.1	36.2 ±2.8	29.3 ±2.7	72.3 ±3.6
+ Input Noise	3.6 ±0.0	1.6 ±0.0	1.2 ±0.0	1.0 ±0.0	6.6 ±0.0	0.9 ±0.0	349.5 ±0.0
+ Len. Noise	36.0 ±1.6	35.0 ±1.6	35.1 ±1.7	35.4 ±1.7	30.3 ±1.7	24.7 ±1.6	74.2 ±1.9
BIOT	64.1 ±6.6	63.9 ±7.0	64.3 ±7.2	64.9 ±7.3	56.3 ±6.6	49.7 ±6.4	43.3 ±8.0
+ Input Noise	1.7 ±0.0	0.7 ±0.0	0.4 ±0.0	0.0 ±0.0	5.7 ±0.0	1.5 ±0.0	102.4 ±0.4
+ Len. Noise	1.0 ±0.0	0.7 ±0.0	0.8 ±0.0	0.9 ±0.0	4.8 ±0.1	1.9 ±0.0	100.9 ±0.1
Frozen Encoder + Linear Classification Head							
BrainECHO	53.8 ±5.4	53.3 ±5.6	53.4 ±5.7	53.9 ±5.7	48.1 ±5.5	40.7 ±5.1	55.4 ±6.7
+ Input Noise	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	1.2 ±0.0	0.0 ±0.0	99.9 ±0.0
+ Len. Noise	14.0 ±1.8	11.8 ±1.9	11.0 ±2.1	10.6 ±2.2	19.4 ±1.7	12.9 ±1.3	96.7 ±2.1
BrainECHO w/ Full BIOT	42.3 ±4.6	41.8 ±4.7	42.3 ±4.8	43.1 ±4.9	32.8 ±4.7	27.7 ±4.5	68.0 ±5.7
+ Input Noise	4.7 ±0.0	2.2 ±0.0	1.5 ±0.0	1.2 ±0.0	6.1 ±0.0	1.1 ±0.0	250.3 ±0.0
+ Len. Noise	5.9 ±0.0	3.3 ±0.0	2.3 ±0.0	1.9 ±0.0	7.1 ±0.0	1.7 ±0.0	182.4 ±0.7
BrainECHO w/ STFT	61.3 ±5.2	61.1 ±5.4	61.4 ±5.5	62.0 ±5.6	53.1 ±5.3	47.0 ±5.0	46.2 ±6.7
+ Input Noise	1.5 ±0.0	0.8 ±0.0	0.6 ±0.0	0.5 ±0.0	5.3 ±0.0	1.9 ±0.0	108.5 ±0.0
+ Len. Noise	2.7 ±0.2	2.6 ±0.1	2.7 ±0.1	3.2 ±0.1	4.0 ±0.3	1.8 ±0.1	102.1 ±0.4

Table 10: Cls2Gen results on NeuraSearch (10 classifiers). Each classifier predicts a sentence class from held-out test EEG; the predicted sentence text is then scored with standard generation metrics, bridging classification and generation evaluation. Values are subject mean ± subject SD. Bold denotes best main-row result per metric across all models.

	BLEU-N (%) ↑				ROUGE (%) ↑		WER (%) ↓
Model	*N=1*	*N=2*	*N=3*	*N=4*	*R-1*	*R-2*	*WER*
Braindecode end-to-end classifiers							
EEGNet	63.1 ±16.3	62.8 ±16.4	62.6 ±16.5	62.5 ±16.5	63.8 ±16.1	62.9 ±16.4	39.1 ±17.3
+ Input Noise	2.4 ±0.4	2.1 ±0.5	2.0 ±0.5	2.0 ±0.5	2.8 ±0.5	1.9 ±0.5	104.3 ±0.7
+ Len. Noise	11.0 ±1.5	10.2 ±1.6	9.9 ±1.6	9.8 ±1.6	12.1 ±1.5	10.1 ±1.6	92.2 ±1.5
ShallowConvNet	66.4 ±6.5	66.1 ±6.6	66.0 ±6.6	65.9 ±6.5	66.8 ±6.2	65.9 ±6.5	35.5 ±6.7
+ Input Noise	4.0 ±0.5	2.7 ±0.5	2.4 ±0.5	2.2 ±0.4	7.3 ±0.5	4.4 ±0.5	97.5 ±0.6
+ Len. Noise	32.6 ±1.0	31.8 ±1.1	31.7 ±1.1	31.7 ±1.2	34.2 ±1.0	32.1 ±1.1	69.5 ±1.0
DeepConvNet	68.9 ±0.9	68.5 ±0.9	68.3 ±0.9	68.2 ±0.9	69.5 ±0.9	68.4 ±0.8	33.0 ±0.9
+ Input Noise	3.9 ±0.5	2.7 ±0.6	2.4 ±0.6	2.2 ±0.6	4.1 ±0.6	1.9 ±0.5	106.1 ±0.6
+ Len. Noise	62.2 ±3.9	61.8 ±3.9	61.7 ±3.9	61.6 ±3.9	62.8 ±3.8	61.6 ±3.8	40.2 ±4.1
EEGConformer	65.5 ±3.7	65.1 ±3.7	64.9 ±3.7	64.7 ±3.7	66.7 ±3.5	65.6 ±3.6	36.3 ±3.9
+ Input Noise	4.2 ±0.0	2.9 ±0.0	2.6 ±0.0	2.5 ±0.0	4.2 ±0.0	2.1 ±0.0	105.9 ±0.0
+ Len. Noise	59.0 ±2.0	58.5 ±2.1	58.3 ±2.1	58.2 ±2.2	60.1 ±1.8	58.7 ±2.0	43.4 ±2.1
TIDNet	67.4 ±2.0	67.0 ±2.1	66.9 ±2.1	66.8 ±2.1	68.3 ±1.9	67.4 ±2.0	34.4 ±2.1
+ Input Noise	4.5 ±0.1	3.0 ±0.0	2.6 ±0.0	2.4 ±0.0	8.2 ±0.2	4.3 ±0.1	98.6 ±0.1
+ Len. Noise	8.8 ±0.4	7.9 ±0.4	7.6 ±0.4	7.4 ±0.4	10.3 ±0.3	8.0 ±0.4	97.0 ±0.4
ATCNet	**69.9** ±3.1	**69.7** ±3.2	**69.6** ±3.2	**69.6** ±3.2	**70.1** ±2.9	**69.5** ±2.9	**31.9** ±3.1
+ Input Noise	4.2 ±0.0	2.9 ±0.0	2.6 ±0.0	2.5 ±0.0	4.2 ±0.0	2.1 ±0.0	105.9 ±0.0
+ Len. Noise	10.2 ±0.3	8.6 ±0.2	8.0 ±0.2	7.7 ±0.2	11.2 ±0.3	8.3 ±0.3	96.4 ±0.5
BIOT	65.9 ±3.5	65.5 ±3.6	65.3 ±3.7	65.3 ±3.7	67.0 ±3.3	66.0 ±3.4	35.9 ±3.8
+ Input Noise	5.6 ±0.0	3.4 ±0.0	2.9 ±0.0	2.7 ±0.0	8.5 ±0.0	4.1 ±0.0	115.2 ±0.0
+ Len. Noise	3.9 ±0.0	2.8 ±0.0	2.5 ±0.0	2.3 ±0.0	4.1 ±0.0	2.1 ±0.0	98.5 ±0.1
Frozen Encoder + Linear Classification Head							
BrainECHO	20.4 ±0.9	19.1 ±1.0	18.6 ±1.0	18.3 ±1.0	21.9 ±1.3	19.0 ±1.3	86.2 ±1.4
+ Input Noise	2.8 ±0.0	1.9 ±0.0	1.7 ±0.0	1.6 ±0.0	4.0 ±0.0	2.1 ±0.0	96.9 ±0.0
+ Len. Noise	10.1 ±0.2	7.9 ±0.1	7.2 ±0.0	6.8 ±0.0	11.6 ±0.1	7.6 ±0.0	95.8 ±0.2
BrainECHO w/ Full BIOT	66.2 ±2.6	65.9 ±2.6	65.8 ±2.6	65.8 ±2.6	66.5 ±2.9	65.8 ±2.9	35.8 ±2.9
+ Input Noise	2.6 ±0.0	2.3 ±0.0	2.2 ±0.0	2.2 ±0.0	3.1 ±0.0	2.1 ±0.0	106.5 ±0.0
+ Len. Noise	5.8 ±0.3	4.9 ±0.3	4.5 ±0.3	4.3 ±0.3	6.6 ±0.3	4.4 ±0.3	98.3 ±0.6
BrainECHO w/ STFT	67.1 ±4.5	66.8 ±4.6	66.7 ±4.6	66.6 ±4.6	68.1 ±4.5	67.2 ±4.6	34.6 ±4.9
+ Input Noise	3.9 ±0.0	2.8 ±0.0	2.5 ±0.0	2.3 ±0.0	4.1 ±0.0	2.1 ±0.0	98.0 ±0.0
+ Len. Noise	3.9 ±0.0	2.8 ±0.0	2.5 ±0.0	2.3 ±0.0	4.1 ±0.0	2.1 ±0.0	98.0 ±0.0

Table 11: Exact training string match rates for generation outputs. Raw denotes verbatim string equality between the output and a training-split sentence. Normalised counts a match when the output equals a training sentence after lowercasing, trimming, collapsing whitespace, and removing terminal punctuation, using the same normalisation to check inventory membership. The NeuraSearch raw rate is depressed by formatting differences between generated and stored sentences rather than by model behaviour, which is why the normalised rate is the headline figure in the main text.

Dataset	**Model**	**Raw (%)**	**Norm. (%)**	**Distinct**
Brennan	BrainECHO	89.9	92.2	134
Brennan	BrainECHO w/ STFT	78.4	80.7	193
Brennan	BrainECHO w/ Full BIOT	48.7	52.3	455
HBN	BrainECHO	86.5	86.5	384
HBN	BrainECHO w/ STFT	85.3	85.7	370
HBN	BrainECHO w/ Full BIOT	82.5	83.4	513
NeuraSearch	BrainECHO	26.5	77.9	106
NeuraSearch	BrainECHO w/ STFT	22.5	62.4	240
NeuraSearch	BrainECHO w/ Full BIOT	5.6	34.6	487

Table 12: Pooled and length-neutral (bin-averaged) BLEU-1 (0 to 100 scale) from a separate rescoring of the outputs behind Figure 6.

Dataset	Model	Pooled	Length neutral
Brennan	BrainECHO	92.1	88.0
Brennan	BrainECHO w/ STFT	50.3	42.9
Brennan	BrainECHO w/ Full BIOT	43.2	35.4
HBN	BrainECHO	81.9	76.0
HBN	BrainECHO w/ STFT	54.3	42.5
HBN	BrainECHO w/ Full BIOT	34.7	24.8
NeuraSearch	BrainECHO	65.1	64.9
NeuraSearch	BrainECHO w/ STFT	51.8	52.6
NeuraSearch	BrainECHO w/ Full BIOT	41.3	42.1

Table 13: Curated generation examples on Brennan (Alice in Wonderland audiobook, 139 sentences). Rows are ordered by qualitative pattern. Long sentences are truncated (. . .).

Target	BrainECHO	BrainECHO w/ Full BIOT	BrainECHO w/ STFT
Nor did Alice think it so very much out of the way to hear the rabbit say to itself, O dear, O dear, I shall be late!	*Nor did Alice think it so very much out of the way to hear the rabbit say to itself, O dear, o dear, I shall be late!*	*Nor did Alice think it so very much out of the way to hear the rabbit say to itself, O dear, o dear, I shall be late!*	*Nor did Alice think it so very much out of the way to hear the rabbit say to itself, O dear, o dear, I shall be late!*
It was labeled orange marmalade.	*It was labeled orange marmalade.*	It was **made with** orange marmalade.	*It was labeled orange marmalade.*
So suddenly.	*So suddenly.*	Dinah was the cat. [CS]	*So suddenly.*
Then she looked at the sides of the well and noticed that they were filled with cupboards and bookshelves.	*Then she looked at the sides of the well and noticed that they were filled with cupboards and bookshelves.*	She looked down upon a little three-legged table, all made of solid glass. [CS]	**Suddenly** she looked at the sides of the well and noticed that they were filled with cupboards and bookshelves.
that she'd never before seen a rabbit with either a waistcoat pocket or a watch to take out of it, and burning with curiosity, she ran across the field after it (. . .)	*that she'd never before seen a rabbit with either a waistcoat pocket or a watch to take out of it. And burning with curiosity, she ran across the field after it (. . .)*	that the rabbit actually took a look at what was going to happen every day, and he ran across the field. But her light was flattered out of her mind. [H]	But they were all locked, and when Alice had been all the way down one side and up the other trying every door, she walked sadly down the middle, wondering how she was ever to get out again. [CS]
Either the well was very deep, or she fell very slowly.	but then I wonder what latitude or longitude I've got to. [CS]	Once or twice she peeped into the book her sister was reading. [CS]	But I shall have to ask them what the name of the country is, you know. [CS]
Here and there she saw maps and pictures hung upon pegs.	which certainly was not here before, said Alice. [CS]	Please, ma'am, is this New Zealand or Australia? [CS]	*Here and there she saw maps and pictures hung upon pegs.*
First, she tried to look down and make out what she was coming to.	And she was quite surprised to find that she remained the same size. [CS]	There was nothing else to do, so Alice soon began talking again. [CS]	**First,** she was quite surprised to find that she remained the same size.

Table 14: Curated generation examples on HBN (movie watching, four films: a bedtime scene, a school drama, a nature documentary, and *Fun with Fractals*). Rows 5–8 illustrate cross-movie contamination [XC]: the model retrieves a sentence from a different film than the one containing the target. Rows 2–3 show a reversed-retrieval pair in which BrainECHO predicts each sentence as the other's output, consistent with length-based confusion between two short, same-duration stimuli.

Target	**BrainECHO**	**BrainECHO w/ Full BIOT**	**BrainECHO w/ STFT**
I mean it.	*I mean it.*	*I mean it.*	*I mean it.*
Do you actually like this?	Keep reading. [CS]	A butt can't be cute. [CS]	*Do you actually like this?*
Keep reading.	Do you actually like this? [CS]	*Keep reading.*	Do you actually like this? [CS]
Then their mother came out and said, time for kittens to go to bed.	*Then their mother came out and said, time for kittens to go to bed.*	*Then their mother came out and said, time for kittens to go to bed.*	**Then there are other main monsters, so** time for kittens to go to bed.
And without a bedtime story, we'll just keep getting up and bugging you.	*And without a bedtime story, we'll just keep getting up and bugging you.*	If we keep going, we can make more and more black triangles at each step. [XC]	Now let's zoom in even more and look at this, it looks like the fern again. [XC]
Wow, this is garbage.	He's just trying to make kids like us feel bad. [CS]	We already have 27 black triangles. [XC]	*Wow, this is garbage.*
Their mother smiled and said with a purr, fine.	*Their mother smiled and said with a purr, fine.*	The physical appearance of the please looks like the whole fern, but we can create more and more black triangles at each step. [H]	*Their mother smiled and said with a purr, fine.*
Mommy, we're not tired at all.	*Mommy, we're not tired at all.*	*Mommy, we're not tired at all.*	You'll be dead or homeschooled by the end of the year anyway. [XC]

Table 15: Curated generation examples on NeuraSearch (48 isolated sentences, controlled paradigm). Unlike Brennan, sentences share syntactic frames, making subject-level substitution errors particularly diagnostic.

Target	**BrainECHO**	**BrainECHO w/ Full BIOT**	**BrainECHO w/ STFT**
Not all engineers work on complex technical projects.	*Not all engineers work on complex technical projects.*	*Not all engineers work on complex technical projects.*	*Not all engineers work on complex technical projects.*
Assistants organize schedules and manage office tasks.	*Assistants organize schedules and manage office tasks.*	**Assets** organize schedules and manage office tasks.	**My systems** organize schedules and manage office tasks.
Teenagers don't bother preparing healthy breakfast meals.	Musicians practice instruments regularly to improve their skills. [CS]	**Luciges** don't bother preparing healthy breakfast meals.	**Managers** don't bother preparing healthy breakfast meals.
Senior citizens waste time on trivial hobbies.	Some programmers don't enjoy debugging code all day. [CS]	So close, we're still holding on to it all day. [H]	Some programmers don't enjoy debugging code all day. [CS]
Latino janitors must clean office buildings every night.	I take a shower, feeling refreshed and clean. [CS]	I need a **janitor** to read an interesting book. [H]	I take a shower, feeling refreshed and clean. [CS]
Asian students study mathematics constantly exceeding all expectations.	*Asian students study mathematics constantly exceeding all expectations.*	**Aspects** constantly exceeding all expectations.	**The mathematics** constantly exceeding all expectations.
He prepares dinner with fresh herbs.	*He prepares dinner with fresh herbs.*	**I don't always do them** with fresh herbs.	I don't usually read newspapers in the morning. [CS]
Real men don't spend time folding laundry.	**Children** don't spend time folding laundry.	I'm so glad you guys weren't doing all day long. [H]	Some programmers don't enjoy debugging code all day. [CS]

Table 16: Top 10 confused sentence pairs for BrainECHO w/ Full BIOT on Brennan. Count = number of times the true sentence was predicted as the confused alternative across all test subjects.

Rank	True sentence	Predicted sentence	Count
1	But when the rabbit actually took a watch out of its waistcoat pocket and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind.	as well as she could, for the hot day made her feel very sleepy and stupid, whether the pleasure of making a daisy chain would be worth the trouble of getting up and picking the daisies, when suddenly	8
2	for she had plenty of time as she went down to look about her and to wonder what was going to happen next.	that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.	8
3	Then she looked at the sides of the well and noticed that they were filled with cupboards and bookshelves.	She knelt down and looked along the passage into the loveliest garden you ever saw.	8
4	How brave they'll all think me at home!	But the rabbit was no longer to be seen.	8
5	Dinah was the cat.	So suddenly.	8
6	Away went Alice like the wind, and was just in time to hear it say, as it turned a corner, Oh, my ears and whiskers, how late it's getting!	for she had read several nice little histories about children who'd gotten burnt and eaten up by wild beasts and other unpleasant things.	8
7	However, on the second time round, she came upon a low curtain she had not noticed before, and behind it was a little door, about 15 inches high.	and that if you cut your finger very deeply with a knife it usually bleeds, and she had never forgotten that if you drink much from a bottle marked poison,	8
8	Alice opened the door and found it led into a small passage.	She ate a little bit, and said anxiously to herself,	8
9	And even if my head would go through, thought poor Alice, it would be of very little use without my shoulders.	She did not like to drop the jar for fear of killing somebody, so she managed to put it into one of the cupboards as she fell past it.	8
10	It had, in fact, a sort of mixed flavor of cherry tart, custard, pineapple, roast turkey, toffee, and hot buttered toast.	and that if you cut your finger very deeply with a knife it usually bleeds, and she had never forgotten that if you drink much from a bottle marked poison,	8

Table 17: Top 10 confused sentence pairs for BrainECHO w/ Full BIOT on HBN.

Rank	True sentence	Predicted sentence	Count
1	But actually, if you look closer, each of the small flowers is made up of even smaller portions.	We can mark the center of each side of the triangle and connect the dots to make a new upside-down triangle.	16
2	Let's start with the black triangle.	Here's what a whole fern looks like.	15
3	Have a great day, sweetie pie!	Then their mother came out and said, time for kittens to go to bed.	12
4	Mommy, we're not tired at all.	Today, let's have some fun with fractals.	11
5	Though while you sleep,	Good night, kittens.	11
6	Keep reading.	I don't like this book.	10
7	This is going on forever.	It is still no.	10
8	Are you sure you don't want me to walk you in?	Okay, let's get this over with.	10
9	Manny.	Fine.	10
10	At first sight, if we see this bunch of flowers, we can see that the bunch is made up of many small flowers.	You'll notice that each time we remove an upside down triangle, we get three new black triangles.	10

Table 18: Top 10 confused sentence pairs for BrainECHO w/ Full BIOT on NeuraSearch.

Rank	True sentence	Predicted sentence	Count
1	He eats quickly, then washes dirty dishes.	Older athletes should retire instead of competing.	18
2	He prepares dinner with fresh herbs.	I don't usually read newspapers in the morning.	18
3	I fold the clothes from the laundry basket.	Indian workers dominate all technology jobs.	16
4	Some programmers don't enjoy debugging code all day.	Senior citizens waste time on trivial hobbies.	15
5	I get up early on weekend mornings.	She brews coffee in the kitchen.	15
6	Artists express creativity through various visual mediums.	Young adults skip basic hygiene prioritizing screen time.	14
7	Husbands help out occasionally with household cleaning tasks.	Not all scientists conduct experiments in laboratory settings.	14
8	Musicians practice instruments regularly to improve their skills.	Teenagers don't bother preparing healthy breakfast meals.	13
9	Real men don't spend time folding laundry.	Senior citizens waste time on trivial hobbies.	13
10	The wife manages all childcare while working full-time.	Teenagers don't bother preparing healthy breakfast meals.	13

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The four contributions stated in the introduction (length-matched noise baselines, BIOT decomposition, two new generation variants, classification-to-generation evaluation) are all delivered and evidenced in Sections 4 and 5. The abstract accurately characterises the scope as sentence-level listening EEG across three English datasets under subject-held-out evaluation.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5 contains an explicit Limitations paragraph addressing four points: the closed-vocabulary evaluation setting, the diagnostic rather than causal nature of the length-matched noise baseline, the inferential attribution of residual bias to the decoder, and dataset scale constraints relative to modern speech benchmarks.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: The paper makes no formal theoretical claims. The STFT zero-coefficient property is verified numerically rather than proved formally, with full verification details in Appendix B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?

   Answer: [Yes]

   Justification: All model architectures are described in Section 3 and Appendices C and D. Preprocessing pipelines, hyperparameters, optimisers, training schedules, data splits, and evaluation metrics are detailed in Section 4 and the appendix. Code is available at the anonymous repository linked in the abstract footnote.

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

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

   Answer: [Yes]

   Justification: Code is released anonymously via the repository linked in the abstract footnote, including preprocessing scripts, model implementations, and evaluation pipelines. Brennan and HBN are publicly available datasets; preprocessing code and instructions to align HBN audio are included. NeuraSearch will be released publicly at a later stage and is available upon request during review.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits (70/10/20 subject-level), optimisers, learning rates, batch sizes, early stopping criteria, and beam search parameters are reported in Section 4 and Appendices C–E. All classification and generation hyperparameters are fully specified.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: All results in Tables 1 and 2 and all appendix tables are reported as mean $\pm$ standard deviation computed across held-out test subjects, capturing inter-subject variability as the primary source of variance. This is stated in Section 4.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute details including GPU type, memory, and approximate training times per model per dataset are reported in Appendix K.

9. **Code of ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research involves human EEG data collected with institutional ethical approval (stated in Section 4). Participants provided informed consent. No deceptive practices were used. The anonymised submission preserves participant privacy.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Broader impacts are discussed in Section 6. Positive impacts include enabling communication for individuals with ALS and other motor impairments. Negative impacts discussed include the risk of overreliance on inflated benchmark scores leading to premature deployment of unreliable systems in clinical contexts.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The released assets (EEG decoding models and diagnostic baselines) do not present meaningful risks of misuse. The NeuraSearch dataset contains anonymised EEG recordings with no personally identifiable information.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited and are licenses explicitly mentioned?

    Answer: [Yes]

    Justification: All datasets (Brennan, HBN), models (BrainECHO, BIOT, NeuSpeech, braindecode classifiers), and evaluation libraries are cited with their original papers throughout. The braindecode library (BSD licence) and HuggingFace Transformers (Apache 2.0) are used in accordance with their terms. Stimuli of HBN from Youtube not explicitly provided following fair usage practises.

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [Yes]

    Justification: The NeuraSearch dataset is documented in Section 4 including recording hardware, sampling rate, stimulus design, trial structure, preprocessing pipeline, and ethical approval. The two new model variants (BrainECHO w/ STFT and BrainECHO w/ Full BIOT) are fully documented in Section 3 and Appendix C. Code for all new assets is included in the anonymous repository.

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

    Question: For research with human subjects, does the paper include full details about instructions given to participants and compensation?

    Answer: [Yes]

    Justification: The NeuraSearch dataset was collected with participants recruited at the University of Strathclyde. Participants were compensated in accordance with institutional guidelines. Full experimental protocol details are provided in Section 4. Participant instructions are available upon request alongside the dataset.

15. **Institutional review board (IRB) approvals**

    Question: Does the paper describe potential risks to study participants and whether IRB approvals were obtained?

    Answer: [Yes]

    Justification: Ethical approval for the NeuraSearch data collection was obtained from the relevant institutional review board (stated in Section 4, anonymised for review). EEG recording presents minimal risk to participants; no invasive procedures were used.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods?

    Answer: [N/A]

    Justification: LLMs are not a component of the core methodology. Whisper is used as a pretrained ASR decoder following prior work, which is standard in this research area and does not constitute a novel LLM contribution.