SENSE: Semantic Neural Speech Synthesis from Brain Dynamics via Spatial Graph Encoding

Jisoo Park[1*] **Seonghak Lee**[2] **Hyojin Park**[3] **Junseok Kwon**[1,2]
[1]Department of Artificial Intelligence, Chung-Ang University, Korea
[2]Department of Computer Science and Engineering, Chung-Ang University, Korea
[3]School of Psychology, Centre for Human Brain Health (CHBH), University of Birmingham, UK
susiehome@cau.ac.kr

Abstract

Reconstructing speech from non-invasive brain signals offers a promising pathway for restoring communication in individuals who are cognitively intact but unable to speak. Existing EEG-to-speech approaches formulate this task as *acoustic reconstruction*, optimizing waveform fidelity while ignoring whether the generated speech preserves high-level semantic content. In this work, we revisit this formulation and argue that EEG signals carry not only acoustic but also semantic information. We identify two key limitations of prior methods: (1) the neglect of spatial relationships between EEG electrodes, and (2) the failure to exploit the semantic structure of the N400 paradigm, where congruent and incongruent trials reflect distinct semantic processing. We propose **SENSE** (**S**emantic-**E**EG **N**eural **S**peech Synth**E**sis), which combines a graph-based EEG encoder over electrode geometry with EEG Semantic Conditioning (ESC), aligning EEG to a pretrained semantic space using only congruent trials. On the N400 dataset, SENSE consistently outperforms prior methods on both acoustic and semantic metrics, and model-internal channel attribution suggests distributed reliance on auditory, sensorimotor, and centro-parietal regions, consistent with known speech-perception neuroscience. In the unseen-subject setting, SENSE trained on only two subjects already surpasses the strongest baseline trained on all eighteen subjects in word error rate, and matches it on acoustic metrics with as few as eight subjects. [1]

1 Introduction

Restoring speech communication from brain activity is a central goal of brain-computer interface (BCI) research, with direct implications for individuals who are cognitively intact but unable to speak due to conditions such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome [1, 2]. Beyond speech, neural decoding has enabled motor intention decoding [3, 4], visual stimulus reconstruction [5, 6], and decoding of high-level semantic representations [7, 8]. State-of-the-art results, however, largely rely on invasive or high-fidelity modalities such as electrocorticography (ECoG) and functional MRI (fMRI), which are limited by surgical requirements or low temporal resolution [9]. Among non-invasive alternatives, electroencephalography (EEG) is particularly attractive due to its portability, low cost, and millisecond-level temporal resolution [10–12].

Despite EEG's well-known limitations of noise, artifact susceptibility, and indirect reflection of neural sources [12, 14, 15], recent work has shown that speech can be reconstructed from EEG in *perceived speech* settings, where neural activity recorded during passive listening is paired with the speech signal. Early approaches predict acoustic features and then synthesize speech from them [16], while

*Work done during a research visit to the University of Birmingham.
[1]Code and audio samples are available at https://jisoo-o.github.io/website/projects/SENSE/.

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

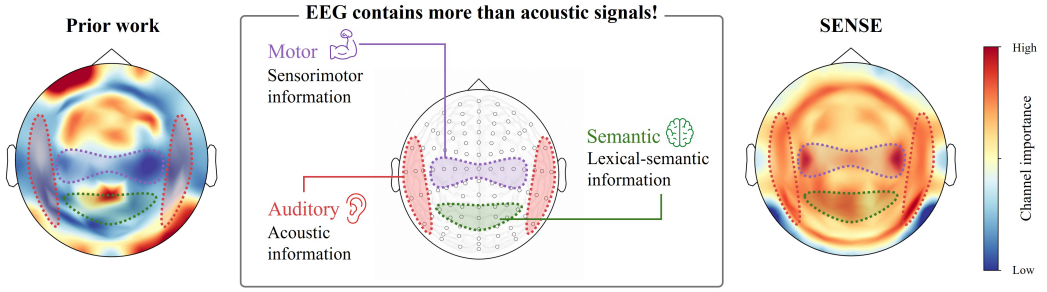


Figure 1: **Motivation for the proposed SENSE framework in EEG-to-Speech Generation.** EEG during speech perception carries acoustic, sensorimotor, and lexical-semantic information distributed across the scalp (center). Prior work [13] captures channel patterns without clear anatomical structure (left); SENSE recovers attribution aligned with all three functional regions (right) via *EEG Semantic Conditioning* and *Graph-based EEG Encoding*.

more recent methods directly generate waveforms from EEG end-to-end [13, 17]. These models are trained with waveform- or spectral-level objectives and evaluated using acoustic measures such as Mel Cepstral Distortion and Mel-Correlation [13, 17]. While these metrics capture signal fidelity, they offer limited insight into whether the reconstructed speech preserves semantic content. As a result, EEG-to-speech generation has been largely framed as an acoustic reconstruction problem, leaving open whether semantic information in EEG is being captured at all.

These observations suggest that the acoustic-only framing is incomplete. First, FESDE [17] reports that using all EEG channels outperforms restricting input to the temporal and parietal regions associated with auditory processing [18]. If the task were purely acoustic, auditory-related regions should suffice; the improvement from broader channel usage suggests that EEG-to-speech generation leverages information beyond auditory processing, though the underlying mechanism remains unclear. Second, existing models overlook the data's structure. The standard benchmark, the N400 corpus [19], contains sentences with semantically congruent or incongruent endings designed to elicit the N400 event-related potential, a well-established index of lexical and conceptual semantic processing [20, 21]. However, prior models [13, 17] treat congruent and incongruent trials identically across all training objectives, missing the opportunity to use the congruency distinction for semantic alignment. Taken together, EEG-to-speech generation involves both acoustic and semantic information, and that the paradigm's semantic structure could be used to ground EEG representations in linguistic meaning.

We address these gaps along two axes. First, for semantic grounding, we align EEG representations with a pretrained CLIP space [22], restricted to congruent trials so that the alignment target reflects the brain's coherent processing of the target word. CLIP is well-suited to this setting because the N400 paradigm fixes the sentence frame and varies only the final word, concentrating semantic differences at the sentence-final position; its sentence-level embedding therefore captures variations that directly correspond to the semantic content being processed by the brain. Although CLIP is not trained on neural signals, prior work has shown that its embedding space captures structure also represented in the brain [7, 8, 23]. Second, for spatial structure, we model EEG as a graph defined by electrode geometry, in contrast to existing EEG-to-speech encoders that treat channels as unordered feature dimensions [13, 16, 17, 24]. Graph-based EEG models have been explored for classification tasks such as emotion recognition [25, 26], but to our knowledge, have not been applied to the EEG-to-speech generation. Together, these design choices yield representations that jointly capture neural activity across multiple functionally distinct regions and high-level semantic content, enabling the model to leverage information beyond auditory regions, as we confirm in Section 4.4.

We propose **SENSE** (**S**emantic-**E**EG **N**eural **S**peech Synth**E**sis), a framework for EEG-to-speech generation that integrates spatial modeling with paradigm-aware semantic conditioning. SENSE includes a graph-based EEG encoder that reflects electrode geometry and an EEG Semantic Conditioning (ESC) module that aligns EEG representations with a pretrained semantic space using congruent trials, incorporating semantic information into speech generation. Figure 1 illustrates the motivation of our work. We summarize our contributions as follows:

- We reformulate EEG-to-speech generation to incorporate both acoustic and semantic information, showing that semantic structure benefits both training and evaluation.
- We propose **SENSE**, a framework that combines a graph-based EEG encoder with EEG Semantic Conditioning (ESC), leveraging the congruency structure of N400-style corpora to align EEG representations with a pretrained semantic space.
- We show that SENSE improves both reconstruction quality and semantic consistency, achieving strong gains in unseen-subject generalization, and surpassing the strongest baseline trained on all 18 subjects in word error rate even when trained on only two subjects.

2 Related Work

Neural Speech and Language Decoding. Decoding speech and language from brain activity has advanced across both invasive and non-invasive modalities. Intracortical and ECoG-based approaches achieve strong performance, including continuous speech synthesis [27], high-throughput neuroprostheses [1, 2, 28], and low-resource ECoG-to-speech synthesis via wav2vec 2.0 [29] transfer learning [30], but all require invasive recordings [9]. Non-invasive alternatives are safer but operate on weaker signals: fMRI enables language reconstruction with language-model priors [7], while MEG and EEG offer high temporal resolution [31]. Several recent non-invasive works target *text* rather than waveforms (via contrastive retrieval against wav2vec 2.0 [31], discrete EEG codex with pretrained BART [32], or Whisper-based MEG-to-text decoding [33]) and thus differ from our task in either modality or output. Within EEG-to-speech *waveform generation*, early work predicted acoustic features [16], while recent end-to-end models reconstruct waveforms directly [13, 17]; these are the comparable systems on the N400 corpus [19] and serve as our baselines. All formulate the task as acoustic reconstruction, despite evidence that using all channels improves performance beyond auditory regions [17, 18]. In contrast, we treat EEG-to-speech generation as involving both acoustic and semantic information, and incorporate semantic structure into both training and evaluation.

Semantic Representations from Brain Activity. Semantic representations capture high-level meaning beyond sensory details. Prior work shows that such information is encoded in brain activity: semantic organization across cortex has been mapped during natural speech [23], and conceptual meaning can be decoded from fMRI [8]. The N400 component further indicates that semantic processing is accessible at the EEG level [20, 21]. More recently, brain signals have been aligned with pretrained embedding spaces: fMRI and EEG have been mapped into the CLIP [22] for visual retrieval and reconstruction [5, 34–37], and MEG/EEG have been aligned with wav2vec 2.0 [29] to identify corresponding speech segments from brain activity [31]. These results suggest that pretrained embedding spaces, despite not being trained on neural data, can serve as effective targets for extracting semantically meaningful information from brain signals via alignment. Unlike alignment-based approaches focused on retrieval or identification, we use semantic alignment as a conditioning mechanism within a generative pipeline that directly synthesizes speech waveforms from EEG.

EEG Representation Learning. EEG representation learning is challenging due to low signal-to-noise ratio and inter-subject variability, and limited data [12, 15, 38]. Convolutional models such as EEGNet [39] treat channels as feature dimensions processed by temporal and spatial filters and existing EEG-to-speech models follow this convention without explicitly modeling spatial relationships [13, 16, 17, 24]. Graph-based approaches model electrode geometry with learned or biologically motivated adjacencies [25, 26], but are primarily applied to classification tasks such as emotion recognition. Large-scale self-supervised pretraining has also been explored to improve generalization [40, 41]. In contrast, we apply graph-based spatial modeling to a generative setting rather than classification, and incorporate structure not through within-modality pretraining but via alignment with a semantic space learned from large-scale multimodal data.

3 Method

Figure 2 illustrates the overall pipeline. Each training example consists of an $\mathbf{X} \in \mathbb{R}^{128\times T}$, the corresponding sentence transcript c, and the recorded speech audio y. EEG is encoded by the EEG Module (Section 3.1) into a latent representation $\mathbf{M} \in \mathbb{R}^{d\times T'}$ (d=192), which is then mapped to the prior of a VITS speech decoder [42], comprising a normalizing flow and a HiFi-GAN vocoder [43]. The decoder is supervised using y through the standard VITS training objectives. During training, $\mathbf{M}$

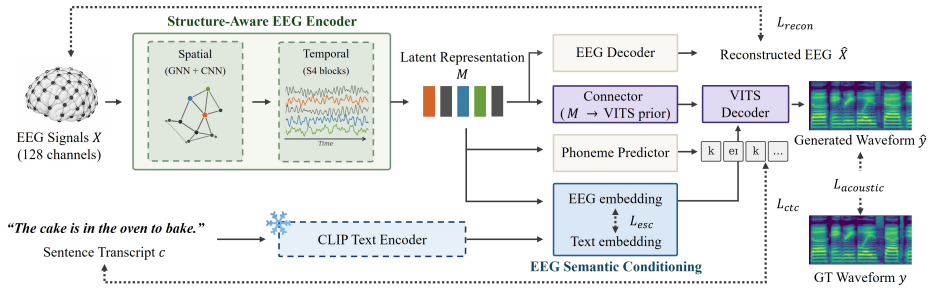


Figure 2: **SENSE training pipeline.** EEG signal $\mathbf{X}$ is encoded by a structure-aware EEG encoder into a latent $\mathbf{M}$. $\mathbf{M}$ is mapped to the VITS prior via a connector module and conditioned at the decoder using an EEG embedding. The EEG embedding is aligned with a frozen CLIP text embedding (ESC, congruent trials only). Solid arrows indicate the forward data flow, while dashed arrows represent loss supervision, connecting model predictions to their corresponding training targets.

is further guided by two supervisory signals derived from c: (1) phoneme sequence supervision, where phoneme targets are obtained by tokenizing c (Section 3.2), and (2) EEG Semantic Conditioning, where c is used as input to a frozen CLIP model (Section 3.3). At inference time, only EEG is required. A semantic conditioning vector $\mathbf{g}_{\text{eeg}} \in \mathbb{R}^d$, derived from $\mathbf{M}$, is used to directly condition the decoder.

SENSE addresses two limitations identified in Section 1. First, prior EEG-to-speech models treat EEG channels as unordered feature dimensions, ignoring their spatial structure; we address this with a graph-based encoder defined over electrode geometry. Second, prior models use the N400 corpus in a paradigm-agnostic manner, treating congruent and incongruent trials identically; we address this with EEG Semantic Conditioning (ESC), which restricts semantic alignment to congruent trials.

3.1 Structure-Aware EEG Encoder

Prior EEG-to-speech encoders are typically trained solely with a self-reconstruction or phoneme objectives [13, 17, 24], producing representations optimized primarily for acoustic fidelity. In contrast, SENSE jointly supervises the EEG module with both phonemic and semantic objectives, yielding a latent representation that captures acoustic structure, phonetic content, and high-level semantic meaning. The module consists of a spatial encoder that exploits electrode geometry, a temporal encoder that models long-range dependencies, and two auxiliary components used only during training.

Spatial Encoder. EEG electrodes are arranged according to standard scalp positions, and signals originating from spatially proximate cortical sources exhibit strong correlations [14]. Conventional 1D CNN encoders overlook this spatial structure by treating channels independently. To address this limitation, we employ a graph-based spatial encoder that explicitly models electrode topology. The encoder integrates three components: a learnable channel gate, a graph convolution over electrode geometry, and a parallel CNN skip connection.

• *Channel attention gate:* A learnable per-channel sigmoid gate, $\mathbf{X}_{\text{att}} = \mathbf{X} \odot \sigma(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^{128}$ is a learnable parameter and σ denotes the sigmoid function, reweights each electrode prior to spatial mixing, with a positive bias on auditory-related electrodes [18] as a warm-start.

• *Graph convolution:* The 128 electrodes are represented as graph nodes with 3D coordinates defined by the BioSemi montage [44]. We construct a fixed adjacency $\mathbf{A} \in \mathbb{R}^{128\times128}$ using a Gaussian kernel over pairwise Euclidean distances. At each time step, node features are updated through two layers of residual graph convolution: $\mathbf{h}_i^{(\ell)} = \mathbf{h}_i^{(\ell-1)} + f^{(\ell)}\left(\sum_j A_{ij}\, \mathbf{h}_j^{(\ell-1)}\right)$, $\ell \in \{1, 2\}$, where each $f^{(\ell)}$ is a learnable feedforward block. The resulting node representations are then averaged over all 128 nodes and projected to dimension d, yielding $\mathbf{H}_{\text{GNN}} \in \mathbb{R}^{d\times T}$. We adopt a fixed anatomy-derived adjacency rather than learning the graph structure from data [25, 45]; with only 7,200 training pairs, joint optimization with the downstream task is prone to overfitting.

• *CNN skip connection:* Since the graph convolution involves aggregation over all 128 nodes, it can attenuate channel-specific information. To mitigate this, we introduce a parallel CNN branch applied to $\mathbf{X}_{\text{att}}$, which captures local channel interactions through learned convolutions, producing $\mathbf{H}_{\text{CNN}} \in \mathbb{R}^{d\times T}$. The outputs of two branches are combined via a learnable scalar gate $g \in \mathbb{R}$:

$\mathbf{H} = \mathbf{H}_{\text{GNN}} + \sigma(g) \cdot \tilde{\mathbf{H}}_{\text{CNN}}$, where $\tilde{\mathbf{H}}_{\text{CNN}}$ denotes $\mathbf{H}_{\text{CNN}}$ rescaled to match the variance of $\mathbf{H}_{\text{GNN}}$. The gate g is initialized to favor the graph branch during the early training.

Temporal Encoder. $\mathbf{H}$ is temporally downsampled and processed by four S4 blocks [46], yielding $\mathbf{M} \in \mathbb{R}^{d \times T'}$. We use S4 because EEG sequences remain long after downsampling; FESDE [17] adopts the same backbone for fair comparison.

Auxiliary Components. Following FESDE [17], we attach two auxiliary components used only during training. An *EEG decoder* (lightweight transposed convolutions) reconstructs the input signal from $\mathbf{M}$ under a length-masked cosine loss $\mathcal{L}_{\text{recon}}$ in (1). A *connector*, consisting of a self-attention encoder followed by a linear projection, maps $\mathbf{M}$ into the mean and log-variance of the VITS prior. A stop-gradient between the EEG module and the connector prevents VITS prior-matching gradients from dominating the EEG encoder, so $\mathbf{M}$ is shaped jointly by reconstruction, phoneme, and semantic objectives. The EEG decoder is discarded at inference; the connector remains. Implementation details are in Appendix A.

3.2 Phoneme Sequence Supervision

Mel-spectrogram reconstruction provides holistic acoustic supervision but does not explicitly encourage $\mathbf{M}$ to encode categorical phonemic structure. Prior work shows that auxiliary phoneme prediction significantly improves EEG-to-speech performance [13]. We adopt this approach with a Conformer block [47] followed by an LSTM attention decoder over a 51-symbol IPA vocabulary, trained with a CTC loss [48] against the phoneme sequence tokenized from c. The predictor is discarded at inference. Implementation details are in Appendix A.3.

3.3 EEG Semantic Conditioning

Phoneme supervision encourages $\mathbf{M}$ to encode fine-grained phonetic structure, but leaves a mismatch: the decoder is conditioned on text during training, yet no such signal is available at inference. We resolve this with EEG Semantic Conditioning (ESC), which learns to map the EEG latent into the decoder's conditioning space and replaces text-based conditioning at inference. Unlike prior brain-to-CLIP alignment work [5, 34, 36] targeting retrieval or reconstruction, ESC is designed for generative conditioning via two choices: a *congruency restriction* and a *conditioning bridge* that exposes $\mathbf{g}_{\text{eeg}}$ to the decoder during training so it remains in-distribution at inference.

Conditioning-Space Alignment. The speech decoder conditions its flow and HiFi-GAN vocoder via a learned projection $W_{\text{clip}} \colon \mathbf{c}_{\text{text}} \in \mathbb{R}^{512} \to \mathbf{g}_{\text{clip}} \in \mathbb{R}^{d}$ of CLIP text embeddings, where $\mathbf{c}_{\text{text}}$ is the output of a frozen CLIP ViT-B/32 encoder applied to the sentence c. To enable text-free conditioning, we learn an additional linear map $W_{\text{eeg}} \in \mathbb{R}^{d \times d}$ that projects the temporal mean of $\mathbf{M}$ into the same conditioning space, yielding $\mathbf{g}_{\text{eeg}} \in \mathbb{R}^{d}$. The alignment target is $\mathbf{g}_{\text{clip}} = \text{sg}[W_{\text{clip}}(\mathbf{c}_{\text{text}})]$, where $\text{sg}[\cdot]$ denotes stop-gradient; this prevents $\mathcal{L}_{\text{esc}}$ from pulling W_{clip} toward the EEG-reachable subspace and degrading conditioning quality, so W_{eeg} targets a fixed space shaped solely by acoustic objectives.

Congruency-Aware Cosine Loss. The N400 corpus contains congruent and incongruent sentence-final words. Aligning the EEG representation to semantic content is meaningful only for congruent trials, where neural responses reflect coherent semantic processing of the target word [20, 21]. Let $\mathcal{C}$ denote the subset of congruent indices in the batch. The ESC loss minimizes cosine distance in the conditioning space restricted to $\mathcal{C}$: $\mathcal{L}_{\text{esc}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \left(1 - \frac{\mathbf{g}_{\text{eeg},i}^{\top} \mathbf{g}_{\text{clip},i}}{\|\mathbf{g}_{\text{eeg},i}\| \, \|\mathbf{g}_{\text{clip},i}\|} \right)$. All other training objectives in Section 3.4 use both congruent and incongruent trials, so incongruent trials contribute to acoustic and phonemic learning without injecting noise into semantic alignment.

Conditioning Bridge. To ensure that the decoder accepts $\mathbf{g}_{\text{eeg}}$ at inference, we expose it during training. The decoder is conditioned on three sources, each sampled with roughly equal probability: $\mathbf{g}_{\text{eeg}}$, $W_{\text{clip}}(\mathbf{c}_{\text{text}})$, and a learnable null embedding (analogous to unconditional dropout in classifier-free guidance training [49], though no guided sampling is performed at inference). When conditioning on $\mathbf{g}_{\text{eeg}}$, gradients are blocked at both $\mathbf{M}$ and W_{eeg}, so acoustic gradients do not leak back into the EEG encoder and W_{eeg} is optimized solely by $\mathcal{L}_{\text{esc}}$. This ensures $\mathbf{g}_{\text{eeg}}$ is in-distribution at inference and the decoder remains compatible with text-based conditioning during training.

Table 1: **Comparison of end-to-end EEG-to-speech models** across three evaluation conditions.

	Model	MCD↓	Mel-Corr↑	STOI↑	WER↓	BERT↑	CLIP-Sim↑
Both	FESDE [17]	11.80 ±0.12	19.87 ±1.97	0.2785	1.2186	0.0381	0.7557
	FE-Phoneme [13]	10.52 ±0.08	28.29 ±1.34	0.3979	1.1610	0.0428	0.7509
	SENSE	**10.44** ±0.10	**31.18** ±0.60	**0.3996**	**1.0186**	**0.0852**	**0.7654**
Audio	FESDE [17]	11.71 ±0.04	19.14 ±0.47	0.2792	1.2125	0.0289	**0.7601**
	FE-Phoneme [13]	10.73 ±0.06	26.53 ±0.30	0.3918	1.1360	0.0602	0.7554
	SENSE	**10.44** ±0.01	**28.50** ±0.24	**0.4039**	**1.0375**	**0.0959**	**0.7601**
Subject	FESDE [17]	11.73 ±0.04	18.70 ±0.29	0.2744	1.2171	0.0269	0.7587
	FE-Phoneme [13]	10.65 ±0.02	26.90 ±0.08	0.3806	1.1231	0.0564	0.7555
	SENSE	**9.928** ±0.03	**32.07** ±0.14	**0.4126**	**0.9948**	**0.1088**	**0.7647**

3.4 Training Objective and Inference

Three Complementary Supervisory Signals. SENSE supervises $\mathbf{M}$ at three levels of abstraction. EEG self-reconstruction ($\mathcal{L}_{\text{recon}}$) preserves low-level signal fidelity, phoneme supervision ($\mathcal{L}_{\text{ctc}}$) imposes discrete phonetic structure, and semantic conditioning ($\mathcal{L}_{\text{esc}}$) injects high-level meaning by aligning the latent with a pretrained semantic space. These signals operate at distinct granularities (raw signal, phonemes, sentence-level meaning) and are complementary: $\mathbf{M}$ must simultaneously support EEG reconstruction, phoneme prediction, and semantic alignment.

Total Loss. We define the total loss as:

$$\mathcal{L} = \underbrace{\lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{fm}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}}}_{\mathcal{L}_{\text{acoustic (VITS)}}} + \underbrace{\lambda_{\text{recon}}\mathcal{L}_{\text{recon}}}_{\text{EEG self-recon}} + \underbrace{\lambda_{\text{ctc}}\mathcal{L}_{\text{ctc}}}_{\text{phoneme}} + \underbrace{\lambda_{\text{esc}}\mathcal{L}_{\text{esc}}}_{\text{semantic}}, \tag{1}$$

with $\{\lambda_{\text{mel}}, \lambda_{\text{kl}}, \lambda_{\text{recon}}, \lambda_{\text{ctc}}, \lambda_{\text{esc}}\}=\{45, 1.0, 1.0, 0.3, 0.5\}$. $\mathcal{L}_{\text{mel}}$ (mel-spectrogram L1 reconstruction), $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{fm}}$, and $\mathcal{L}_{\text{kl}}$ follow VITS [42]. $\mathcal{L}_{\text{recon}}$ is the cosine reconstruction loss applied to the EEG decoder output (Section 3.1), $\mathcal{L}_{\text{ctc}}$ is the phoneme prediction loss of FE-Phoneme [13] (Section 3.2), and $\mathcal{L}_{\text{esc}}$ is the congruency-restricted alignment loss defined in Section 3.3.

Inference. The EEG decoder, phoneme predictor, and CLIP encoder are discarded. Raw EEG is processed by the EEG encoder to produce $\mathbf{M}$, which is mapped to $\mathbf{g}_{\text{eeg}}$ via W_{eeg} and to the VITS prior via the Connector. The decoder synthesizes the waveform conditioned on $\mathbf{g}_{\text{eeg}}$, without text or guided sampling; the null embedding from training is unused at inference and serves only as regularization.

4 Experiments

Dataset. We evaluated on the N400 EEG corpus [19], which comprises 24 subjects, each with 440 sentences, recorded using 128-channel EEG at 512 Hz. Following [13, 17], we excluded four unreliable subjects, leaving 20 subjects total, and used a standard preprocessing pipeline; speech was downsampled to 22,050 Hz with 80-band mel-spectrogram targets (Appendix A.1). The training set contains 7,200 EEG-speech pairs from 18 subjects $\times$ 400 sentences. We evaluated performance under three test conditions following [17]: *Unseen audio* (training subjects, held-out sentences; 720 pairs), *Unseen subject* (held-out subjects sub-23, sub-24 with training sentences; 800 pairs), and *Unseen both* (held-out subjects with held-out sentences; 80 pairs). For cross-subject scaling experiments, we used a nested subset structure $k{=}2 \subset k{=}4 \subset k{=}8 \subset k{=}18$ (Appendix A.5).

Baselines and Metrics. We compared SENSE against FESDE [17] and FE-Phoneme [13], the comparable existing end-to-end EEG-to-speech models on this corpus, retrained using their official implementations. Acoustic fidelity was measured by MCD, Mel-Correlation, and STOI; linguistic content by WER (via Whisper-large [50]); rescaled BERTScore (BERT) [51], and CLIP-Sim [22]. WER may exceed 1 due to insertion errors. Since incongruent trials are designed to violate the sentence-final semantic expectation, semantic metrics were computed only on congruent trials; acoustic metrics used all trials.

Training Setup. SENSE was trained for 200k iterations until validation performance plateaus, with loss weights $\lambda_{\text{ctc}} = 0.3$ and $\lambda_{\text{esc}} = 0.5$ selected via the sensitivity analysis in Appendix A.13. Baselines were trained for 100k iterations as specified in their official implementations and original papers, with extended training providing no further gains. All results were averaged over five seeds. Implementation details are in Appendix A.

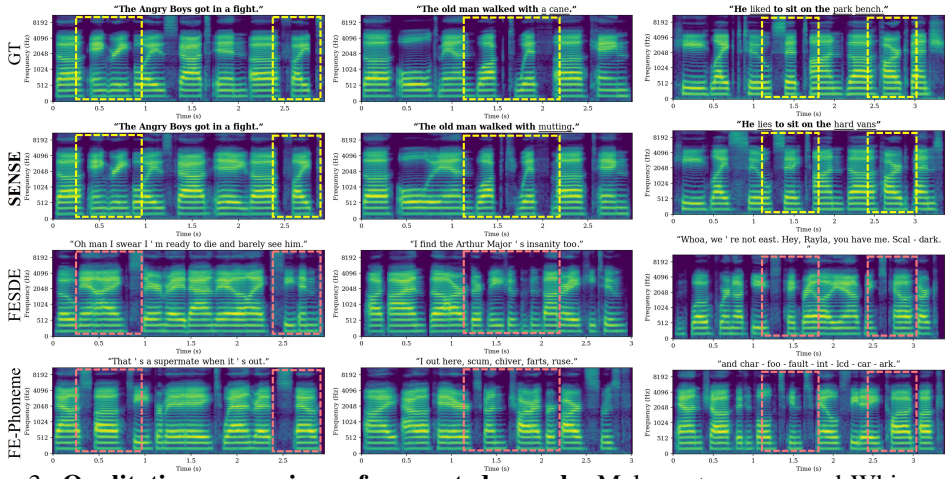


Figure 3: **Qualitative comparison of generated speech.** Mel-spectrograms and Whisper-large transcriptions for ground truth, SENSE, FESDE, and FE-Phoneme. Words matching the ground truth are shown in **bold**, while semantically or phonetically similar words are underlined.

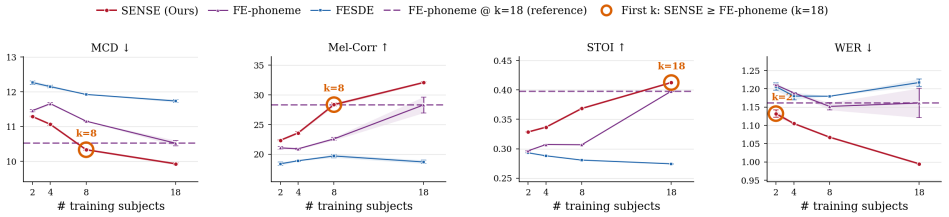


Figure 4: **Scaling with the number of training subjects.** Models are trained on $k \in \{2, 4, 8, 18\}$ subjects and evaluated on the held-out pair (sub-23, sub-24). Shaded regions: ± 1 std over 5 seeds. Dashed line: FE-Phoneme at k=18 (best baseline). Orange circles: the smallest k at which SENSE first matches or exceeds this reference.

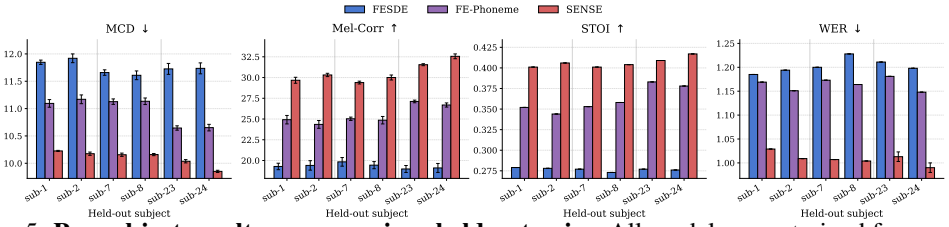


Figure 5: **Per-subject results across various held-out pairs.** All models are retrained from scratch on the remaining 18 subjects for each pair. Error bars: ± 1 std over 5 seeds.

4.1 Main Results

SENSE outperforms both baselines across all splits and metrics (Table 1), with the largest gains on the *unseen subject* split, indicating that it learns representations that generalize across subjects rather than relying on subject-specific cues. Absolute WER values near 1 nonetheless show that intelligibility-level reconstruction from EEG remains an open challenge. BERTScore and CLIP-Sim differences are smaller, as these sentence-level metrics are less sensitive to fine-grained lexical differences than WER. Qualitative examples in Figure 3 show that SENSE produces speech that is both acoustically consistent and semantically meaningful, while baselines often fail to recover lexical content. *For additional qualitative evidence, please refer to the **audio samples** in the supplementary video and Appendix A.8.*

4.2 Cross-Subject Generalization

We evaluated cross-subject generalization in two complementary ways: (1) scaling the number of training subjects, and (2) varying which subjects are held out. Figure 4 shows performance as the number of training subjects k increases, evaluated on a fixed held-out pair (sub-23, sub-24). Our SENSE consistently outperforms both baselines, with the largest gains in the low-data regime. Notably, SENSE already surpasses FE-Phoneme's k=18 WER at k=2, and matches or exceeds it on MCD, Mel-Correlation, and STOI by k=8. Since evaluation is conducted on unseen subjects, these

Table 2: **Ablation study of the proposed SENSE components**. Each row corresponds to removing one component from the full model. w shuffled ESC uses the same ESC loss and conditioning bridge as SENSE, but randomly permutes CLIP targets within each batch.

Model	Ch.Att	GNN	CNN Skip	CTC	ESC	MCD↓	Mel-Corr↑	STOI↑	WER↓	BERT↑	CLIP-Sim↑
SENSE	✓	✓	✓	✓	✓	**9.93** ±0.03	**32.07** ±0.14	**0.4126**	**0.9948**	**0.1088**	**0.7647**
w/o ESC	✓	✓	✓	✓	-	10.34 ±0.03	29.43 ±0.27	0.4016	1.0323	0.0763	0.7601
w shuffled ESC	✓	✓	✓	✓	✓ (shuf)	10.38 ±0.04	29.25 ±0.28	0.4002	1.0385	0.0721	0.7595
w/o CTC	✓	✓	✓	-	✓	10.50 ±0.01	27.24 ±0.27	0.3721	1.0488	0.0588	0.7624
w/o CNN Skip	✓	✓	-	✓	✓	10.40 ±0.01	29.64 ±0.25	0.3970	1.0306	0.0724	0.7579
w/o GNN	✓	-	-	✓	✓	10.11 ±0.02	29.54 ±0.22	0.3951	1.0355	0.0788	0.7577
w/o Ch.Att	-	✓	✓	✓	✓	10.24 ±0.01	29.08 ±0.16	0.4046	1.0170	0.0926	0.7639

Table 3: **Effect of adding ESC to baseline encoders.** We incorporate ESC into FESDE and FE-Phoneme without other changes. While both baselines improve, they remain consistently below SENSE across all metrics, indicating that ESC complements rather than replaces the SENSE encoder.

Backbone	ESC	MCD ↓	Mel-Corr ↑	STOI ↑	WER ↓	BERT ↑	CLIP-sim ↑
FESDE [17]	✓	10.93 ±0.03	26.27 ±0.19	0.363	1.084	0.045	0.7500
FE-Phoneme [13]	✓	10.04 ±0.02	31.17 ±0.21	0.411	1.102	0.070	0.7544
SENSE	✓	**9.93** ±0.03	**32.07** ±0.14	**0.413**	**0.995**	**0.109**	**0.7647**

Table 4: **Performance across different graph adjacency structures.** *BioSemi topo*: anatomy-based adjacency derived from electrode geometry (used in SENSE). *Fully connected*: uniform adjacency. *Random*: fixed random adjacency. *GAT*: learned edge attention via Graph Attention Network.

Model	Graph Topology	CNN Skip	MCD↓	Mel-Corr↑	STOI↑	WER↓
SENSE	Biosemi topo	✓	**9.93** ±0.03	**32.07** ±0.14	**0.4126**	**0.9948**
GNN (topo, no skip)	Biosemi topo	–	10.40 ±0.02	29.64 ±0.25	0.3970	1.0306
GNN (GAT attn. topo)	Biosemi topo	–	10.21 ±0.03	29.10 ±0.10	0.3835	1.0183
GNN (fully-conn.)	Fully-conn.	–	10.96 ±0.05	26.17 ±0.27	0.3684	1.0945
GNN (random)	Random	–	10.67 ±0.03	27.15 ±0.18	0.3803	1.0806

Table 5: **Performance across different semantic encoders** in the ESC module.

Semantic Encoder	MCD ↓	Mel-Corr ↑	STOI ↑	WER ↓	BERT ↑
SBERT [52]	10.06 ±0.01	30.25 ±0.09	0.399	1.011	0.094
CLAP [53]	10.02 ±0.04	29.62 ±0.21	0.388	1.013	0.092
CLIP [22]	**9.93 ±0.03**	**32.07 ±0.14**	**0.413**	**0.995**	**0.109**

results reflect cross-subject generalization rather than in-distribution scaling, suggesting that SENSE learns representations driven by stimulus content rather than subject-specific patterns.

To test whether SENSE's advantage depends on the choice of held-out subjects, we retrained all models from scratch using two additional held-out pairs, (sub-1, sub-2) and (sub-7, sub-8). Figure 5 reports per-subject performance across the resulting six held-out subjects. SENSE outperforms both baselines on every metric and every subject, with consistent relative gains over FE-Phoneme. Although absolute performance varies modestly across subjects, the relative ranking of the methods remains consistent. Full per-subject results are provided in Appendix A.14.

4.3 Ablation Study

Table 2 shows that all components of the proposed SENSE contribute to overall performance. Removing ESC degrades semantic metrics, while removing CTC causes an even larger increase in WER, indicating that phoneme-level supervision also preserves intelligibility (Table 2). Spatial components (GNN, CNN skip, channel attention) provide consistent gains, complementing both acoustic and semantic modeling. Adding ESC alone to baseline encoders improves them but still falls short of SENSE (Table 3), showing that semantic conditioning is insufficient without our encoder design. To verify that ESC exploits genuine EEG-semantic correspondence rather than acting as generic regularization, we trained SENSE with CLIP targets randomly shuffled within each batch (Table 2, *w shuffled ESC* row). Shuffled targets yield performance comparable to (and slightly below) the w/o ESC baseline, confirming that ESC's gain depends on correct semantic targets rather than on auxiliary supervision alone.

We compared four graph adjacency schemes (Table 4). The anatomy-based BioSemi topology outperforms learnable GAT, fully-connected, and random adjacencies, indicating that local distance-based connectivity is the relevant inductive bias. Jointly learning the graph from only 7,200 pairs is under-constrained, while removing structure entirely is more harmful. Adding the CNN skip on top of the BioSemi GNN further improves performance, confirming complementarity between graph-based spatial mixing and channel-wise feature aggregation.

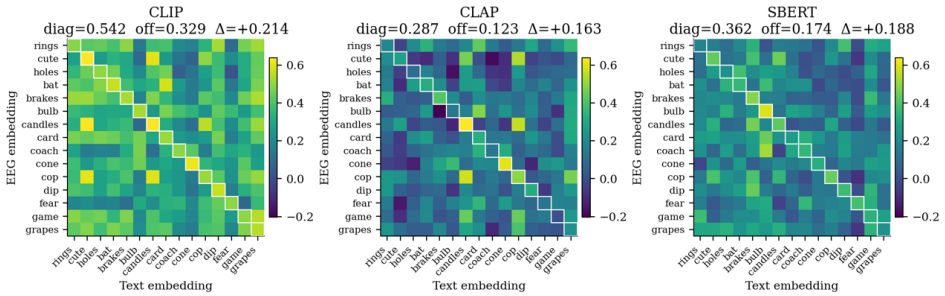
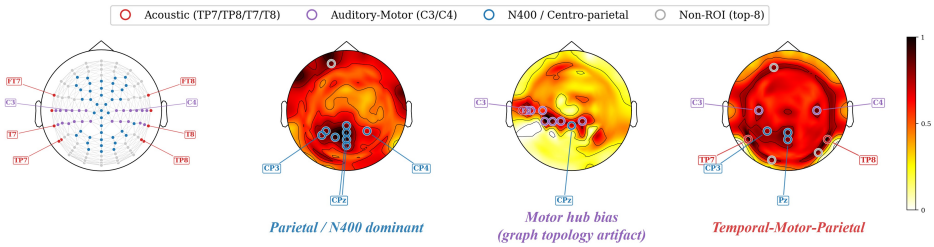


Figure 6: **EEG–text alignment across semantic encoders.** Cosine similarity matrices between mean EEG embeddings (rows) and text embeddings (columns) for matched words. Stronger word-specific alignment appears as a brighter diagonal. Δ is the gap between diagonal and off-diagonal means.



(a) Fixed Anatomical Graph (b) w/o GNN (c) GNN (no skip gate) (d) GNN with Skip

Figure 7: **ESC attribution across encoder variants.** (a) Anatomical graph used in SENSE. (b–d) Per-channel gradient magnitude of EEG-CLIP alignment, normalized per model. (b) Without GNN: centro-parietal concentration consistent with N400. (c) GNN without skip: bias toward central hub electrodes, a graph-topology artifact. (d) SENSE: tripartite pattern over temporal-auditory (TP7/TP8), motor (C3/C4), and centro-parietal (CPz). Circles: top-8 channels per model; colors: functional ROI.

We also evaluated three pretrained semantic encoders for the ESC alignment target (Table 5): CLIP, CLAP [53], and SBERT [52]. CLIP outperforms CLAP and SBERT on all metrics. CLAP, despite being trained for audio–language alignment, performs worse than text-only SBERT, possibly because the N400 stimuli often involve concrete, image-able nouns (*chair*, *dog*, *ants*) for which CLIP's image–text pretraining provides more discriminative semantic structure. Figure 6 corroborates this: CLIP shows the brightest diagonal and the largest gap between diagonal and off-diagonal values.

4.4 Channel Attribution Analysis

We complemented the quantitative results with two analyses of the model's internal behavior, examining whether the learned representations align with our motivation that EEG during speech perception encodes information beyond auditory regions. The channel attention gate (Section 3.1) is initialized with a positive bias toward auditory-related electrodes and zero elsewhere; no prior is placed on sensorimotor or centro-parietal regions, so the patterns below emerge from training rather than design. The *w/o Ch.Att* row in Table 2 initializes the gate without the auditory warm-start (all weights set to zero); the moderate performance drop indicates that the gate's emergent non-auditory attribution is not merely an artifact of initialization. After training, auditory-prior channels are modestly down-weighted while central electrodes (C3, C4) and frontal-central regions gain importance, with occipital channels suppressed (Appendix A.11).

To further investigate which channels contribute to EEG–text alignment, we computed per-channel gradient magnitudes of the ESC alignment (Figure 7). We treat these gradient maps as model-internal diagnostics rather than evidence of neural sources, given known limitations of saliency-based attribution. The maps reveal three patterns: without a GNN, attribution concentrates over centro-parietal regions consistent with the N400 response. A GNN without a skip connection shifts emphasis to central hub electrodes, a topology artifact of dense connectivity rather than a physiological signal. The full SENSE model recovers a distributed pattern over temporal-auditory (TP7, TP8), central motor (C3, C4), and centro-parietal (CPz) regions. The prominence of C3/C4, which are not explicitly favored by design, is consistent with prior findings that speech perception engages sensorimotor cortex during passive listening [54–56]; we read this as the model capturing distributed information aligned with known properties of speech processing, not as direct evidence of neural generators. This supports our central hypothesis that EEG-to-speech generation benefits from non-auditory channels. Training dynamics are detailed in Appendix A.12.

5 Conclusions

We argued that EEG-to-speech should be approached as a joint acoustic-semantic problem, not as acoustic reconstruction alone. We introduced **SENSE**, which combines a graph-based spatial encoder over electrode geometry with EEG Semantic Conditioning (ESC). On the N400 dataset, SENSE consistently outperforms prior methods on both acoustic and semantic metrics, with particularly strong cross-subject generalization. Channel attribution further recovers patterns spanning auditory, sensorimotor, and centro-parietal regions consistent with neuroscience of speech-perception. However, reliable word-level reconstruction remains challenging, as indicated by WER values near 1, and full waveform synthesis has only been evaluated on the controlled N400 corpus despite promising representation-level adaptation to naturalistic speech. Channel attribution remains model-internal rather than neural-source localization, and evaluation on larger naturalistic and active-speech datasets is an important next step.

5.1 Broader Impact

This work uses a previously released public EEG benchmark collected under prior ethical approval, with no new participant recruitment, neural data collection, or sensing hardware, and should not be interpreted as unrestricted thought decoding. Active listening and laboratory conditions are not reliable safeguards, and future applications should protect neural data through measures such as on-device processing, explicit opt-in consent, restrictions on unauthorized access, and deployment-specific oversight and governance. These considerations follow the UNESCO Recommendation on the Ethics of Neurotechnology (2025), including its guidance on protecting neural data and information that may enable inferences about mental states.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-02217071), in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2021-II211341), and in part by the AI Seoul Tech Research Support Program of the Seoul Future Foundation.

References

[1] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[2] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976): 1037–1046, 2023.

[3] Dennis J McFarland and Jonathan R Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011.

[4] JD Simeral, Sung-Phil Kim, MJ Black, JP Donoghue, and LR Hochberg. Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array. *Journal of neural engineering*, 8(2):025027, 2011.

[5] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, pages 14453–14463, 2023.

[6] Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023.

[7] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

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

[9] Markus Ullsperger and Stefan Debener. *Simultaneous EEG and fMRI: recording, analysis, and application*. 2010.

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

[11] Reza Abiri, Soheil Borhani, Eric W Sellers, Yang Jiang, and Xiaopeng Zhao. A comprehensive review of eeg-based brain–computer interface paradigms. *Journal of neural engineering*, 16(1): 011001, 2019.

[12] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

[13] Jihwan Lee, Tiantian Feng, Aditya Kommineni, Sudarsana Reddy Kadiri, and Shrikanth Narayanan. Enhancing listened speech decoding from eeg via parallel phoneme sequence prediction. In *ICASSP*, 2025.

[14] Paul L Nunez and Ramesh Srinivasan. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press, 2006.

[15] Xun Chen, Chang Li, Aiping Liu, Martin J McKeown, Ruobing Qian, and Z Jane Wang. Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey. *IEEE Signal Processing Magazine*, 39(2):117–134, 2022.

[16] Gautam Krishna, Co Tran, Yan Han, Mason Carnahan, and Ahmed H Tewfik. Speech synthesis using eeg. In *ICASSP*, pages 1235–1238, 2020.

[17] Jihwan Lee, Aditya Kommineni, Tiantian Feng, Kleanthis Avramidis, Xuan Shi, Sudarsana Reddy Kadiri, and Shrikanth Narayanan. Toward fully-end-to-end listened speech decoding from EEG signals. In *Interspeech*, pages 2110–2114, 2024.

[18] Rini A Sharon, Shrikanth Narayanan, Mriganka Sur, and Hema A Murthy. An empirical study of speech processing in the brain by analyzing the temporal syllable structure in speech-input induced eeg. In *ICASSP*, pages 4090–4094, 2019.

[19] Kathryn K Toffolo, Edward G Freedman, and John J Foxe. Evoking the n400 event-related potential (erp) component using a publicly available novel set of sentences with semantically incongruent or congruent eggplants (endings). *Neuroscience*, 501:143–158, 2022.

[20] Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

[21] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62(1): 621–647, 2011.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[23] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

[24] Aditya Kommineni, Kleanthis Avramidis, Richard Leahy, and Shrikanth Narayanan. Knowledge-guided eeg representation learning. In *EMBC*, pages 1–6, 2024.

[25] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11 (3):532–541, 2018.

[26] Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

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

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

[29] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020.

[30] Miseul Kim, Zhenyu Piao, Jihyun Lee, and Hong-Goo Kang. Braintalker: Low-resource brain-to-speech synthesis with transfer learning using wav2vec 2.0. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE, 2023.

[31] 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.

[32] Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. Dewave: Discrete encoding of eeg waves for eeg to text translation. *NeurIPS*, 36:9907–9918, 2023.

[33] Yiqian Yang, Yiqun Duan, Qiang Zhang, Hyejeong Jo, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Neuspeech: Decode neural signal as speech. In *ICASSP*, pages 6636–6640. IEEE, 2026.

[34] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *NeurIPS*, 36: 24705–24728, 2023.

[35] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

[36] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

[37] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

[38] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[39] 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.

[40] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

[41] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

[42] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*, pages 5530–5540. PMLR, 2021.

[43] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeuIPS*, 33:17022–17033, 2020.

[44] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[46] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[47] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[48] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006.

[49] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[50] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518, 2023.

[51] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[52] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992, 2019.

[53] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5, 2023.

[54] Stephen M Wilson, Ayşe Pinar Saygin, Martin I Sereno, and Marco Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, 7(7):701–702, 2004.

[55] M Florencia Assaneo and David Poeppel. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science advances*, 4(2): eaao3842, 2018.

[56] Hyojin Park, Robin AA Ince, Philippe G Schyns, Gregor Thut, and Joachim Gross. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, 25(12):1649–1653, 2015.

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[58] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *ICASSP*, pages 4214–4217, 2010.

[59] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

A Implementation Details

This appendix collects the implementation details omitted from Section 3, including dataset preprocessing, layer-level architectures, and training configurations. The implementation is provided as supplementary material; pretrained checkpoints will be released with the de-anonymized version upon acceptance.

A.1 Dataset Preprocessing

EEG Preprocessing. We follow the preprocessing pipeline of FESDE [17]. Raw 128-channel EEG signals recorded at 512 Hz are processed as follows: (1) a 60 Hz notch filter removes power-line interference; (2) a 0.5–50 Hz Butterworth bandpass filter isolates the frequency range relevant to speech-related neural activity; (3) Independent Component Analysis (ICA) is applied to remove eye-blink artifacts, with components correlated with EOG channels removed automatically; (4) signals are resampled to 256 Hz to reduce computational cost while preserving the relevant spectral content. Trials are segmented into fixed windows aligned with sentence onsets and zero-padded to a uniform length when necessary.

Speech Preprocessing. The recorded speech audio is downsampled from its native sampling rate to 22,050 Hz to match the input rate of the HiFi-GAN vocoder [43] used in VITS. Mel-spectrogram targets are extracted with 80 mel bands using a window size of 1024 samples, hop size of 256 samples, and FFT size of 1024, following the VITS configuration [42].

Subject Selection. Following [13, 17], we exclude four subjects (sub-5, sub-10, sub-15, sub-18) due to documented issues with EEG signal quality or behavioral responses, leaving 20 subjects in the analysis. Of these, 18 are used for training and validation, with sub-23 and sub-24 held out for the unseen-subject test conditions.

A.2 Structure-Aware EEG Module

Channel Attention Gate. The gate weights $\mathbf{w} \in \mathbb{R}^{128}$ are initialized as $w_i = +2.0$ for the six auditory-related electrodes (T7, T8, FT7, FT8, TP7, TP8), giving $\sigma(2.0) \approx 0.88$ at the start of training, and $w_i = 0.0$ for all remaining channels, giving $\sigma(0) = 0.5$. All gate weights are learnable and updated by gradient descent throughout training.

Graph Adjacency Matrix. Electrode positions $p_i \in \mathbb{R}^3$ are taken from the BioSemi 128-channel standard montage available through MNE-Python [44], expressed in meters. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{128\times128}$ is constructed as:

$$A_{ij} = \begin{cases} \exp\left(-\dfrac{\|p_i - p_j\|_2^2}{2\sigma_p^2}\right) & \text{if } \|p_i - p_j\|_2 < \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{A.1}$$

with $\sigma_p = 2$ cm and threshold $\delta = 4$ cm. The diagonal is set to zero, and the matrix is row-normalized: $\mathbf{A} \leftarrow D^{-1}\mathbf{A}$, where D is the diagonal degree matrix. The resulting adjacency is registered as a fixed buffer and never updated during training.

Graph Convolution Layer. At each time step, the scalar activation of each electrode is projected to a $d_{\text{node}} = 64$ dimensional feature via a shared Linear-GELU block. Two layers of residual graph convolution are then applied, where each $f^{(\ell)}$ is a Linear $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) block with hidden dimension $d_{\text{node}} = 64$. After the two message-passing steps, node representations are averaged over all 128 nodes and linearly projected to the latent dimension $d = 192$.

Gated CNN Skip Connection. The CNN branch consists of two 1D convolutional layers with kernel size 4, stride 1, and padding 3, mapping from 128 input channels to $d = 192$ output channels. Each convolutional layer is followed by GELU activation. The branch output is normalized to match the GNN output's empirical standard deviation:

$$\tilde{\mathbf{H}}_{\text{CNN}} = \frac{\text{std}(\mathbf{H}_{\text{GNN}})}{\text{std}(\mathbf{H}_{\text{CNN}})}\,\mathbf{H}_{\text{CNN}}, \tag{A.2}$$

where $\mathrm{std}(\cdot)$ is computed over the channel and time dimensions. The scalar mixing gate $g \in \mathbb{R}$ is initialized to -2.0, giving $\sigma(-2) \approx 0.12$ at the start of training. The gate is fully learnable and adapts during optimization.

Temporal Encoding. A single Conv1d layer with kernel size 3, stride 3, and padding 1 reduces the temporal resolution by a factor of 3, followed by four S4 blocks [46] with hidden dimension $d = 192$, state dimension 64, and dropout 0.1. Each S4 block follows the standard configuration of [46] with bidirectional state-space mixing.

EEG Decoder (training only). The EEG decoder consists of a single ConvTranspose1d block with kernel size 3, stride 3, and padding 0 for temporal upsampling, followed by a two-layer ConvTranspose1d stack with kernel size 4, stride 1, and padding 3 that restores the channel dimension from $d = 192$ back to 128. The reconstruction loss is the cosine similarity loss of FESDE [17], applied with length masking that excludes zero-padded frames. The decoder is discarded at inference.

Connector. The connector is a 6-layer multi-head self-attention encoder with hidden dimension 192, 2 attention heads, feed-forward dimension 768, kernel size 3, and dropout 0.1, following the attention encoder of VITS [42]. The output is projected by a 1×1 convolution to the mean and log-variance parameters of the VITS prior. A stop-gradient is applied between $\mathbf{M}$ and the connector input, isolating the EEG Module from the speech-pathway gradients.

A.3 Phoneme Sequence Supervision

The phoneme predictor is taken from FE-Phoneme [13] without modification. It consists of one Conformer block [47] with encoder dimension 192, 8 attention heads, feed-forward expansion factor 4, convolutional kernel size 31, and dropout 0.1, followed by an LSTM attention decoder with hidden dimension 192, location-based attention, and a maximum decode length of 50 over a 51-symbol IPA vocabulary (45 IPA phonemes, 3 punctuation symbols, and pad/SOS/EOS tokens). The teacher forcing ratio is set to 0 throughout training, so the decoder is fully autoregressive on its own predictions.

CTC objective. Given a target phoneme sequence $\mathbf{z} = (z_1, \ldots, z_L)$ of length L, the LSTM attention decoder produces a sequence of L logit vectors $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_L)$ over the 51-symbol vocabulary (including a blank token). The CTC objective [48] marginalizes over alignments $\boldsymbol{\pi}$ that map to $\mathbf{z}$ under the standard CTC reduction B:

$$\mathcal{L}_{\mathrm{ctc}} = -\log \sum_{\boldsymbol{\pi} \in B^{-1}(\mathbf{z})} P(\boldsymbol{\pi} \mid \mathbf{Y}). \tag{A.3}$$

We use the PyTorch `CTCLoss` with $\mathrm{zero_infinity}=\mathrm{True}$ and $\mathrm{blank}=0$, and set both $\mathrm{input_lengths}$ and $\mathrm{target_lengths}$ to the actual phoneme sequence length (excluding the SOS token) per sample, following the FE-Phoneme implementation [13]. The predictor is discarded at inference.

A.4 EEG Semantic Conditioning

CLIP Embeddings. We use frozen CLIP ViT-B/32 [22] to compute text embeddings of dimension 512. Embeddings for all sentences in the dataset are precomputed once and cached to disk. The CLIP encoder itself is not loaded during training.

Projection layers. The decoder-side projection $W_{\mathrm{clip}} \colon \mathbb{R}^{512} \to \mathbb{R}^{192}$ is a single `nn.Linear` layer that is trained jointly with the speech decoder under the acoustic objectives (mel reconstruction, KL, adversarial, feature-matching). The EEG-side projection $W_{\mathrm{eeg}} \in \mathbb{R}^{192 \times 192}$ is a single `nn.Linear` layer trained only by $\mathcal{L}_{\mathrm{esc}}$; gradients from the conditioning bridge path are blocked at W_{eeg} so that the only gradient signal reaching W_{eeg} is the ESC alignment loss.

Length-masked temporal mean. $\mathbf{g}_{\mathrm{eeg}}$ is computed by first taking a length-masked temporal mean of $\mathbf{M}$ over its time dimension, where the mask excludes zero-padded frames according to each sample's actual EEG length. The masked mean is then projected by W_{eeg} to produce $\mathbf{g}_{\mathrm{eeg}} \in \mathbb{R}^{192}$.

Conditioning bridge schedule. At each training step, with probability $p_{\text{eeg}} = 0.33$ the decoder's conditioning is set to $\mathbf{g}_{\text{eeg}}$; gradients along this path are blocked at both $\mathbf{M}$ (so the EEG Module is not affected by the conditioning bridge) and W_{eeg} (so W_{eeg} is shaped only by $\mathcal{L}_{\text{esc}}$). With the remaining probability $1 - p_{\text{eeg}}$, standard Classifier-Free Guidance dropout [49] is applied: a per-sample Bernoulli mask with rate 0.5 selects between the CLIP text conditioning $W_{\text{clip}}(\mathbf{c}_{\text{text}})$ and a learnable null embedding (a single $\mathbb{R}^{192}$ parameter, initialized to zero). The decoder thus encounters three conditioning sources during training, with overall frequencies of approximately 33%, 33.5%, and 33.5% for $\mathbf{g}_{\text{eeg}}$, the CLIP text embedding, and the null embedding, respectively.

Inference. At inference, $\mathbf{g}_{\text{eeg}}$ is computed from $\mathbf{M}$ without any text input and replaces the null embedding as the decoder's conditioning signal. The CFG scale is set to 0, so the decoder produces speech conditioned solely on $\mathbf{g}_{\text{eeg}}$ without any text-guidance interpolation.

A.5 Training Configuration

Optimizer. Following the convention of FESDE [17] and VITS [42], we use two separate optimizers, both AdamW [57]. The first optimizer trains the EEG Module (Channel Attention Gate, GNN, CNN skip, Temporal Encoding, EEG Decoder) and the phoneme predictor. The second optimizer trains the speech pathway (Connector, W_{eeg}, W_{clip}, VITS prior, HiFi-GAN generator, and discriminators). The two optimizers are connected through the stop-gradient described in Section 3.1: $\mathbf{M}$ is computed by the first optimizer's parameters and treated as a fixed input to the second. Both optimizers use learning rate 2×10^{-4}, $\beta_1 = 0.8$, $\beta_2 = 0.99$, weight decay 0.01, and exponential learning rate decay with $\gamma = 0.999875$ per epoch.

Loss weights. The loss-weight values used in Eq. 1 are summarized in Table A.1.

Table A.1: **Loss weights used in the total training objective.**

Symbol	Value	Source
λ_{mel}	45	VITS default
λ_{kl}	1.0	VITS default
λ_{recon}	1.0	FESDE default
λ_{ctc}	0.3	FE-Phoneme default
λ_{esc}	0.5	SENSE

The adversarial loss $\mathcal{L}_{\text{adv}}$ and feature-matching loss $\mathcal{L}_{\text{fm}}$ follow the VITS implementation exactly [42] and use the standard HiFi-GAN multi-period and multi-scale discriminators [43].

Training duration and hardware. SENSE is trained for 900 epochs with 7,200 training pairs and batch size 16 per GPU, corresponding to approximately 200,000 gradient steps. Training uses full FP32 precision (no automatic mixed precision) on two NVIDIA RTX 3090 GPUs with PyTorch DistributedDataParallel and NCCL backend. We save a checkpoint every 5,000 steps and evaluate the checkpoint at step 200,000, where validation performance plateaus. Total wall-clock training time for SENSE is approximately 2 days. Baselines (FESDE, FE-Phoneme) are trained for 100,000 iterations as in their original papers; we verified that training them longer does not improve performance, as their losses plateau before this point.

Matched Training-Budget Comparison. To control for the different training schedules used in the main experiments, we additionally compare all models after 100k training iterations on the unseen-subject split. As shown in Table A.2, SENSE remains better than both baselines across all four reconstruction metrics under the same training budget.

Table A.2: **Matched training-budget comparison** on the unseen-subject split. All models are evaluated after 100k training iterations.

Model	MCD↓	Mel-Corr↑	STOI↑	WER↓
FESDE [17]	11.73	18.70	0.274	1.217
FE-Phoneme [13]	10.65	26.90	0.381	1.123
SENSE	**10.24**	**31.00**	**0.409**	**1.033**

Cross-subject scaling experiments. For Section 4.2, we use a nested subset structure $k{=}2 \subset k{=}4 \subset k{=}8 \subset k{=}18$. The $k{=}2$ subset is sub-21 and sub-22; $k{=}4$ adds sub-08 and sub-16; $k{=}8$ adds sub-04, sub-09, sub-12, and sub-13; and $k{=}18$ uses all 18 training subjects. The held-out test subjects sub-23 and sub-24 are excluded from every training subset. All scaling experiments use the same training configuration as the main experiments, with the same number of epochs and the same checkpoint selection protocol.

Reproducibility. All models are trained with five random seeds. For each seed, the random number generators of PyTorch, NumPy, and Python are initialized identically; CUDA's deterministic mode is enabled. Reported numbers are the mean across seeds, with the standard deviation reported when shown.

A.6 Evaluation Protocol

Whisper transcription. We use Whisper-large-v3 [50] with the following decoding settings: temperature 0, beam size 5, no condition on previous text. Transcriptions are normalized for WER computation by lowercasing, removing punctuation other than apostrophes, and collapsing repeated whitespace. WER is computed via the `jiwer` library against the normalized ground-truth transcript.

BERTScore. BERTScore [51] is computed using the `roberta-large` model with rescaling against a baseline. We report the F1 component.

CLIP-Sim. We use CLIP ViT-B/32 [22] to compute text embeddings of both the Whisper transcription and the ground-truth sentence, and report the cosine similarity between them.

MCD and Mel-Correlation. MCD is computed over the first 25 mel-cepstral coefficients (excluding the 0th coefficient) with dynamic time warping, and Mel-Correlation is the Pearson correlation between mel-spectrograms of generated and ground-truth speech aligned by length truncation.

STOI. We use the standard implementation of [58] from the `pystoi` library at 16 kHz, with both signals resampled from 22,050 Hz.

A.7 Symbol Reference

For convenience, Table A.3 summarizes the symbols introduced in Section 3.

Table A.3: **Notation used throughout the method.**

Symbol	Meaning
$\mathbf{X} \in \mathbb{R}^{128\times T}$	Raw EEG input (T time steps, 128 channels)
$\mathbf{X}_{\text{att}}$	Channel-gated EEG
$\mathbf{w} \in \mathbb{R}^{128}$	Channel attention gate weights (learnable)
$\mathbf{A} \in \mathbb{R}^{128\times 128}$	Fixed graph adjacency matrix
$\mathbf{H}_{\text{GNN}}, \mathbf{H}_{\text{CNN}} \in \mathbb{R}^{d\times T}$	GNN and CNN branch outputs
$\mathbf{H} \in \mathbb{R}^{d\times T}$	Combined branch output
$\mathbf{M} \in \mathbb{R}^{d\times T'}$	EEG latent ($d = 192$, T' after temporal compression)
d	EEG latent dimension ($d = 192$)
d_{node}	Per-node feature dimension in graph conv ($d_{\text{node}} = 64$)
$\mathbf{c}_{\text{text}} \in \mathbb{R}^{512}$	CLIP text embedding (frozen ViT-B/32)
W_{clip}	Decoder-side CLIP projection (learned, dim $512 \to d$)
W_{eeg}	EEG-side projection (learned, dim $d \to d$)
$\mathbf{g}_{\text{eeg}}, \mathbf{g}_{\text{clip}} \in \mathbb{R}^{d}$	EEG and CLIP conditioning vectors
$\mathcal{C}$	Subset of congruent indices in a batch
$\text{sg}[\cdot]$	Stop-gradient
p_{eeg}	Probability of using $\mathbf{g}_{\text{eeg}}$ during training ($p_{\text{eeg}} = 0.33$)

A.8 Additional Qualitative Results

We provide more visual comparison samples in Figures A.3, A.4, A.5, A.6, and A.7.

A.9 Validating the Semantic Role of ESC

To further investigate whether EEG Semantic Conditioning (ESC) captures sentence-specific information rather than merely improving acoustic reconstruction, we conduct additional analyses at both the representation and generation levels. Specifically, we examine (i) trial-specific EEG–text correspondence, (ii) sentence retrieval, (iii) the effect of matched and mismatched semantic conditioning, and (iv) the training-time and inference-time contributions of ESC.

A.9.1 Trial-Specific EEG–Text Correspondence

We first examine whether the EEG-derived representation g_{eeg} is more closely aligned with its corresponding sentence embedding than with unrelated sentences. For each test trial, we compute cosine similarity between g_{eeg} and the matched CLIP text embedding g_{clip}, and compare it with the mean similarity to the remaining 439 sentence embeddings among 440 candidates.

Table A.4: EEG–text correspondence on held-out trials. The gap denotes the difference between matched and mismatched cosine similarities.

Evaluation subset	Matched	Mismatched	Gap
All trials ($N = 800$)	0.37	0.25	+0.12
Congruent ($N = 400$)	0.40	0.25	+0.15

As shown in Table A.4, the matched sentence consistently exhibits higher similarity than mismatched sentences. The difference is statistically significant for both all trials ($p < 10^{-79}$) and congruent trials ($p < 10^{-45}$), with a larger gap for the congruent trials used for ESC supervision. These results indicate that the learned EEG representation contains trial-specific information aligned with the corresponding text embeddings.

A.9.2 Sentence Retrieval

We further evaluate whether g_{eeg} can identify the corresponding sentence among multiple candidates. Using 400 held-out congruent trials as queries, we rank 440 unique sentence embeddings by cosine similarity to the EEG-derived representation.

Table A.5: EEG-to-text retrieval among 440 candidate sentences. Random chance corresponds to uniformly random ranking.

Method	Top-1	Top-10	Top-20
EEG retrieval	2.8%	15.5%	23.0%
Random chance	0.23%	2.27%	4.55%

The EEG-derived representation achieves above-chance retrieval at all evaluated cutoffs. Although absolute retrieval accuracy remains modest, these results provide additional evidence of sentence-specific EEG–text correspondence rather than alignment solely to a shared distribution of text embeddings.

A.9.3 Matched and Mismatched Semantic Conditioning

To investigate whether the learned semantic representation contributes to speech generation, we intervene on the conditioning input at inference time while keeping all other inputs fixed. We compare the original EEG-derived conditioning g_{eeg} with the matched CLIP embedding and a mismatched CLIP embedding from another trial. Mismatched embeddings are assigned using a deterministic cyclic shift without self-pairs.

As shown in Table A.6, matched conditioning produces the highest BERTScore, followed by EEG-derived and mismatched conditioning. However, the differences are small, and the acoustic metrics

Table A.6: Effect of semantic conditioning on speech generation. Lower MCD and higher Mel-Corr and BERT are better.

Conditioning	MCD ↓	Mel-Corr ↑	BERT ↑
EEG-derived	9.942	32.14	0.096
Matched CLIP	9.778	32.81	0.097
Mismatched CLIP	9.919	32.14	0.092

do not exhibit a consistent separation between EEG-derived and mismatched conditioning. These findings suggest that conditioning identity has a modest effect on sentence-level semantic similarity. They do not establish reliable recovery of sentence meaning in the generated speech.

A.9.4 Training-Time and Inference-Time Effects

Finally, we distinguish the effect of supplying semantic conditioning at inference from the representation-shaping effect of ESC during training. For the inference-time analysis, we replace g_{eeg} with the learned null embedding in the ESC-trained model. For the training-time analysis, we compare models trained with and without ESC under identical null conditioning.

Table A.7: Contributions of ESC at inference and during training. Positive values indicate improvements from EEG-derived conditioning or ESC training, respectively.

Comparison	Δ Mel-Corr	Δ BERTScore
EEG-derived vs. null	+0.11	+0.002
ESC vs. no ESC (null)	+2.38	+0.004

Replacing EEG-derived conditioning with the null embedding results in a small decrease in semantic similarity, indicating a modest inference-time effect. Importantly, the ESC-trained model retains an advantage over the model trained without ESC even when both receive null conditioning. Since the ESC objective backpropagates through the semantic projection and EEG encoder, this result suggests that ESC also shapes the learned EEG representation during training.

Together, these experiments indicate two complementary roles of ESC: representation shaping during training and modest semantic guidance during inference. While the learned representation exhibits trial-specific EEG–text correspondence, the current reconstruction quality does not support reliable semantic speech decoding.

A.10 Generalization Beyond the N400 Paradigm

The N400 dataset uses artificially constructed sentences with controlled timing and limited prosodic variation. Although this setting facilitates EEG–stimulus alignment and comparison with prior work, it differs substantially from continuous natural speech. We therefore investigate whether ESC requires congruency-based supervision and whether the learned EEG representation can be adapted to naturalistic listened speech.

A.10.1 ESC Without Congruency-Based Selection

While the main SENSE model applies ESC only to congruent trials, the ESC objective itself requires paired EEG–text samples rather than explicit congruency labels. To evaluate the role of this restriction, we additionally train SENSE by applying ESC to all training trials, including both congruent and incongruent sentences.

Table A.8: Effect of ESC supervision selection.

ESC supervision	MCD ↓	Mel-Corr ↑	STOI ↑	WER ↓	BERT ↑
Congruent only	9.93	32.07	0.413	0.995	0.096
All trials	10.17	30.90	0.409	1.018	0.094
Without ESC	10.34	29.43	0.401	1.032	0.090

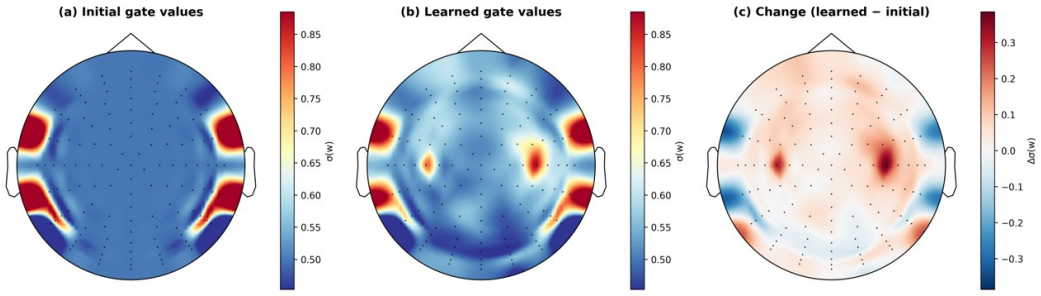


Figure A.1: **Channel attention gate before and after training.** Sigmoid-gated weights $\sigma(\mathbf{w}) \in (0, 1)$ projected onto the BioSemi 128-channel scalp layout. (a) Initialization: positive bias ($\sigma(\mathbf{w}) \approx 0.88$) on six auditory-related electrodes (T7, T8, FT7, FT8, TP7, TP8); all other channels start neutral ($\sigma(\mathbf{w}) = 0.5$). (b) Values after training. (c) Per-channel change $\Delta\sigma(\mathbf{w})$. Auditory-prior electrodes remain among the most activated but are modestly attenuated. The largest gains appear near central electrodes C3 and C4 ($\Delta\sigma \approx +0.27$ and $+0.29$), with additional gains in frontal-central regions; Occipital channels, which subserve visual processing and are not expected to contribute to auditory tasks, are suppressed below their initial value, indicating that the gate not only emphasizes auditory-related electrodes but also actively down-weights electrodes unrelated to the task.

Applying ESC to all trials improves all five metrics compared with removing ESC entirely. Restricting supervision to congruent trials provides further improvements, suggesting that congruency serves as a dataset-specific supervision filter rather than an architectural requirement.

A.10.2 Transfer to Continuous Natural Speech

To further examine generalization beyond controlled N400 stimuli, we evaluate SENSE on the Broderick naturalistic speech dataset [59], in which participants listened to continuous narrative speech. We initialize the EEG encoder and ESC projection using the N400-trained weights and fine-tune them on the naturalistic dataset. Each EEG segment is paired with its corresponding text representation, and we evaluate the cosine similarity between matched EEG–text pairs and randomly mismatched pairs.

Because the public release available to us does not provide the aligned speech waveforms required for reconstruction, we evaluate representation-level EEG–text alignment rather than speech synthesis.

Table A.9: EEG–text alignment on the Broderick naturalistic speech dataset.

Dataset	Matched	Random	Gap
Broderick	0.602	0.517	+0.085

Matched EEG–text similarity exceeds randomly mismatched similarity for all 19 participants. Furthermore, none of 5,000 random permutations produces a similarity gap as large as the observed difference ($p < 2.0 \times 10^{-4}$). These results demonstrate that the N400-pretrained encoder and ESC objective can be adapted to continuous listened-speech EEG without congruency labels or artificially segmented sentence stimuli.

Nevertheless, this experiment evaluates transfer at the representation level and does not establish naturalistic speech reconstruction performance. Full EEG-to-speech evaluation on naturalistic datasets with aligned speech waveforms remains an important direction for future work.

A.11 Channel Attention Gate Visualization

Figure A.1 shows the channel attention gate values $\sigma(\mathbf{w})$ projected onto the BioSemi 128-channel scalp layout, comparing the initialization (positive bias on six auditory-related electrodes, zero elsewhere) with the values after training and the per-channel change.

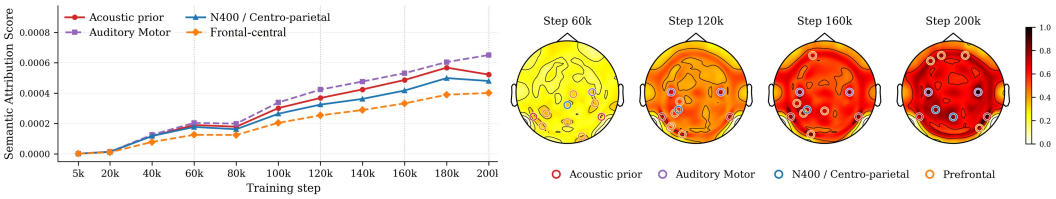


Figure A.2: **Semantic channel importance trajectory during training.** (Left) Gradient-based attribution score per EEG ROI group across training steps. Vertical dotted lines indicate steps shown in the topomaps. (Right) Spatial distribution of semantic attribution at selected checkpoints; brighter regions indicate higher channel importance for EEG–text alignment. The model initially relies on acoustic prior regions, and progressively shifts attribution toward N400/centro-parietal channels associated with semantic processing.

Table A.10: **Sensitivity analysis of loss weights λ_{ctc} and λ_{esc}.** Results are reported on the *subject* split (mean $\pm$ std over 5 runs).

λ_{ctc}	λ_{esc}	MCD ↓	Mel-Corr ↑	STOI ↑	WER ↓	CLIP-Sim ↑
0.0	0.5	10.50$\pm$0.02	27.24$\pm$0.28	0.372	1.050	0.756
0.1	0.5	10.27$\pm$0.03	27.92$\pm$0.24	0.384	1.041	0.755
0.3	**0.5**	**9.93**$\pm$0.03	**32.07**$\pm$0.15	**0.413**	**0.994**	**0.760**
1.0	0.5	10.01$\pm$0.03	30.39$\pm$0.15	0.404	1.002	0.755
0.3	0.0	10.34 $\pm$0.03	29.43 $\pm$0.27	0.402	1.032	0.760
0.3	0.1	10.33 $\pm$0.01	29.03 $\pm$0.24	0.392	1.064	0.759
0.3	**0.5**	**9.93**$\pm$0.03	**32.07**$\pm$0.15	**0.413**	**0.994**	**0.760**
0.3	2.0	10.07 $\pm$.02	29.86 $\pm$0.24	0.402	1.030	0.757

A.12 Attribution Trajectory During Training

Figure A.2 tracks the gradient-based attribution across training. Early in training, attribution is concentrated over auditory-prior regions reflecting the gate initialization. As training progresses, attribution to centro-parietal and frontal-central regions increases, while auditory contributions remain present but no longer dominant.

A.13 Sensitivity to Loss Weights

We examine sensitivity to the ESC and CTC loss weights (λ_{esc} and λ_{ctc}). Table A.10 shows the chosen values balance the reconstruction quality and semantic quality, yielding best results.

A.14 Per-Subject Results

Table A.11 reports the complete per-subject results for the cross-subject generalization experiments shown in Figure 5. For each of the three held-out pairs (sub-1, sub-2), (sub-7, sub-8), and (sub-23, sub-24), all models are retrained from scratch on the remaining 18 training subjects and evaluated on each held-out subject individually. All values are mean $\pm$ standard deviation over 5 random seeds. SENSE outperforms both baselines on every metric and every held-out subject.

Table A.11: **Complete per-subject results for cross-subject generalization.** For each held-out pair, all models are retrained from scratch on the remaining 18 subjects. Bold: best per subject.

Held-out	Model	MCD↓	Mel-Corr↑	STOI↑	WER↓	BERT↑
sub-1	FESDE	11.847 ±0.04	19.282 ±0.39	0.2793	1.1851	0.0150
	FE-Phoneme	11.095 ±0.07	24.920 ±0.51	0.3518	1.1694	0.0318
	SENSE	**10.226** ±0.01	**29.671** ±0.36	**0.4008**	**1.0289**	**0.0902**
sub-2	FESDE	11.920 ±0.08	19.407 ±0.57	0.2780	1.1939	0.0256
	FE-Phoneme	11.170 ±0.08	24.352 ±0.47	0.3442	1.1508	0.0283
	SENSE	**10.175** ±0.03	**30.311** ±0.20	**0.4065**	**1.0088**	**0.0965**
sub-7	FESDE	11.659 ±0.05	19.836 ±0.52	0.2765	1.1998	0.0371
	FE-Phoneme	11.127 ±0.05	25.029 ±0.20	0.3529	1.1731	0.0318
	SENSE	**10.156** ±0.03	**29.392** ±0.17	**0.4009**	**1.0070**	**0.0909**
sub-8	FESDE	11.609 ±0.08	19.438 ±0.43	0.2726	1.2279	0.0419
	FE-Phoneme	11.134 ±0.06	24.858 ±0.45	0.3583	1.1640	0.0194
	SENSE	**10.159** ±0.02	**30.005** ±0.30	**0.4041**	**1.0040**	**0.0826**
sub-23	FESDE	11.726 ±0.10	18.978 ±0.42	0.2774	1.2106	0.0310
	FE-Phoneme	10.645 ±0.04	27.118 ±0.16	0.3833	1.1815	0.0342
	SENSE	**10.037** ±0.03	**31.547** ±0.12	**0.4092**	**1.0132**	**0.0947**
sub-24	FESDE	11.736 ±0.10	19.103 ±0.54	0.2762	1.1977	0.0318
	FE-Phoneme	10.652 ±0.06	26.690 ±0.25	0.3777	1.1476	0.0419
	SENSE	**9.852** ±0.02	**32.557** ±0.29	**0.4168**	**0.9903**	**0.1077**

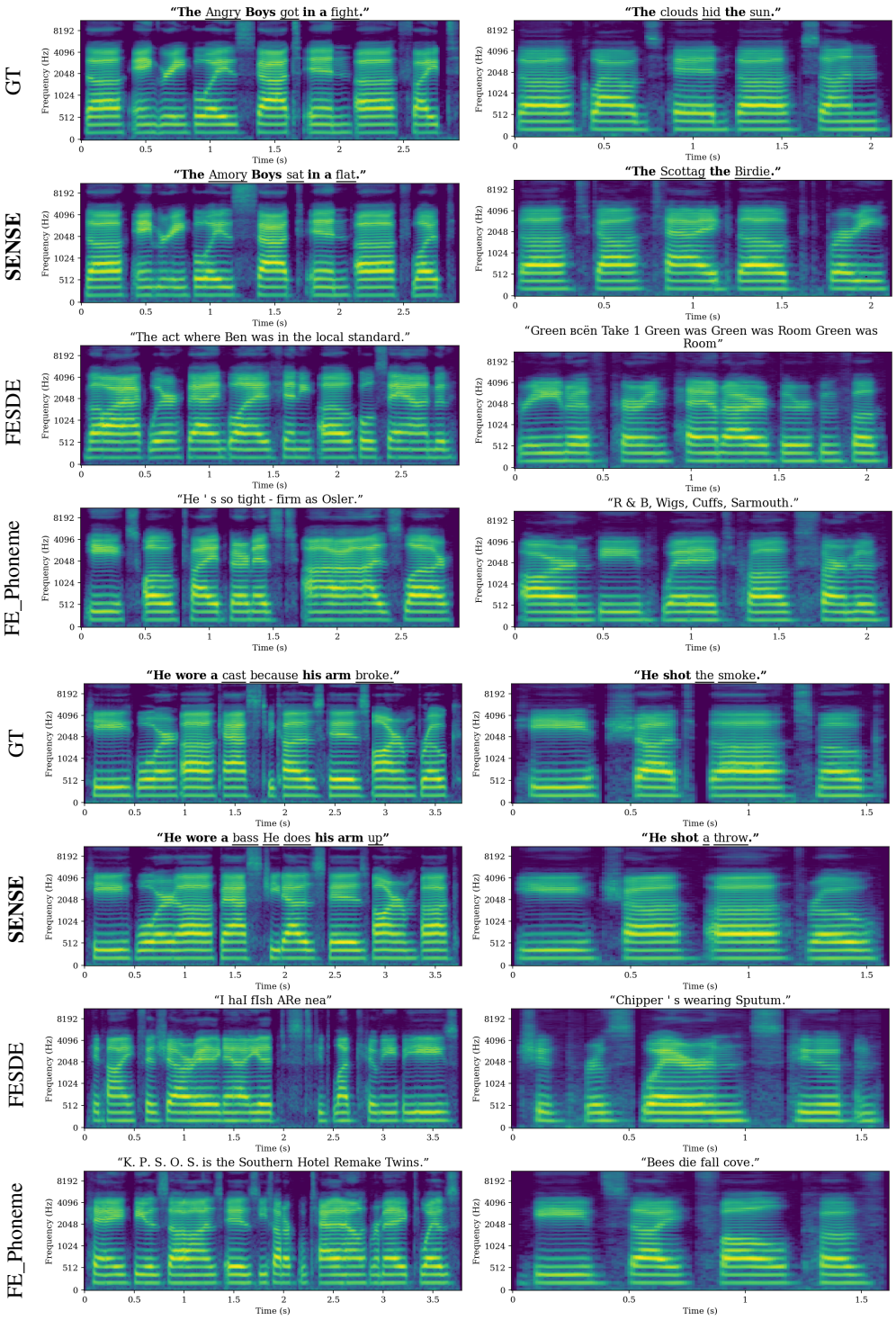


Figure A.3: Additional Melspectrogram Comparison

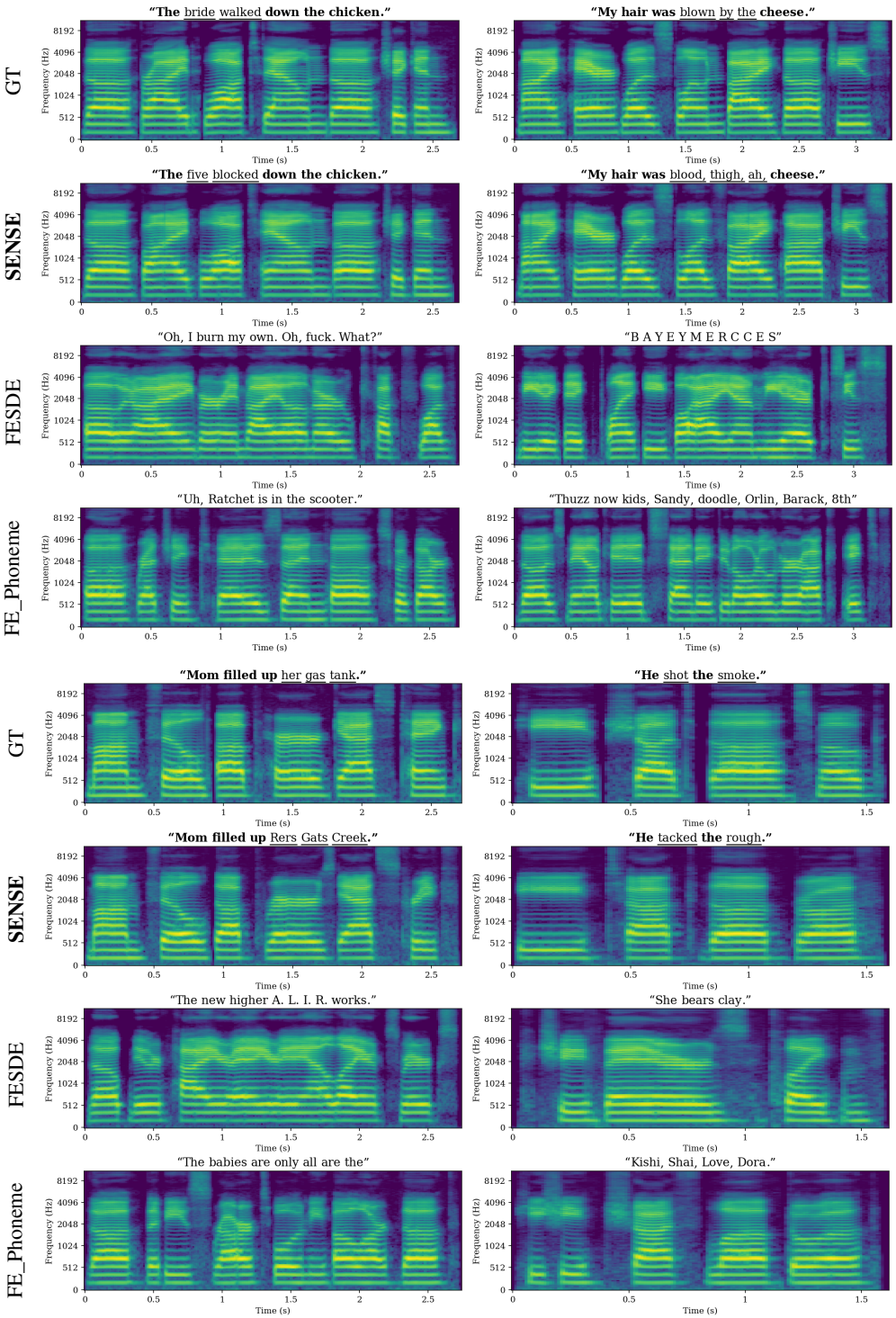


Figure A.4: Additional Melspectrogram Comparison

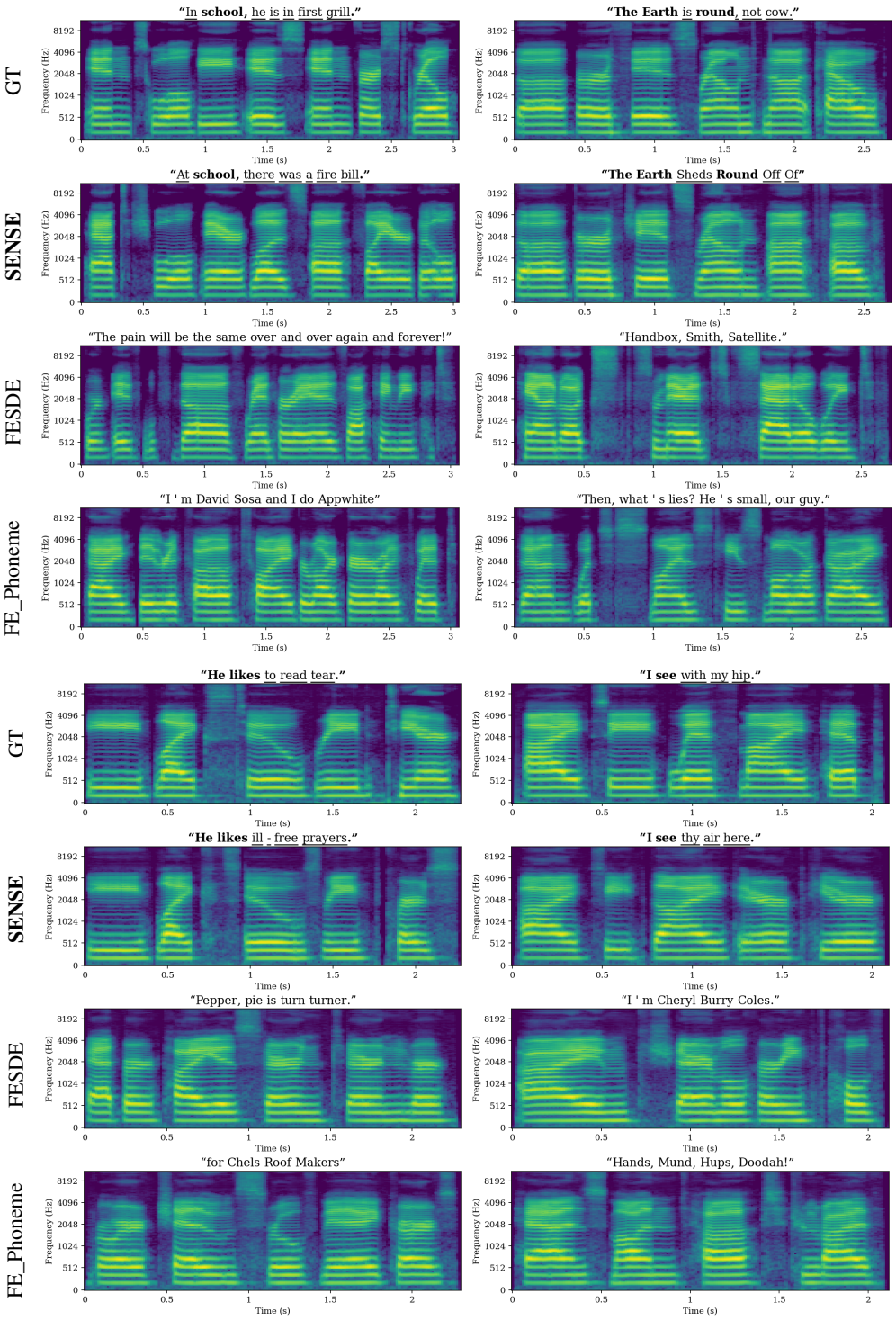


Figure A.5: Additional Melspectrogram Comparison

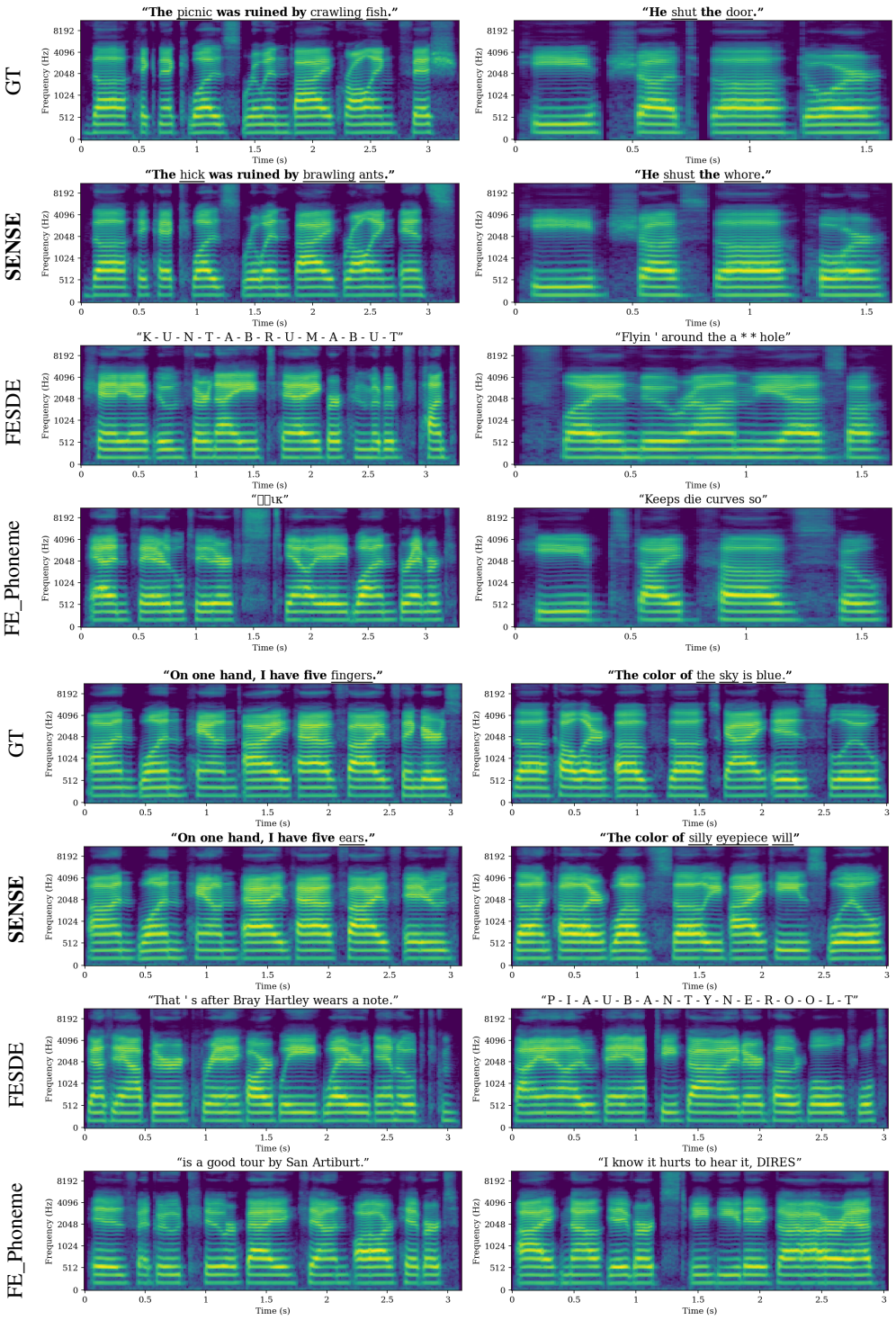


Figure A.6: Additional Melspectrogram Comparison

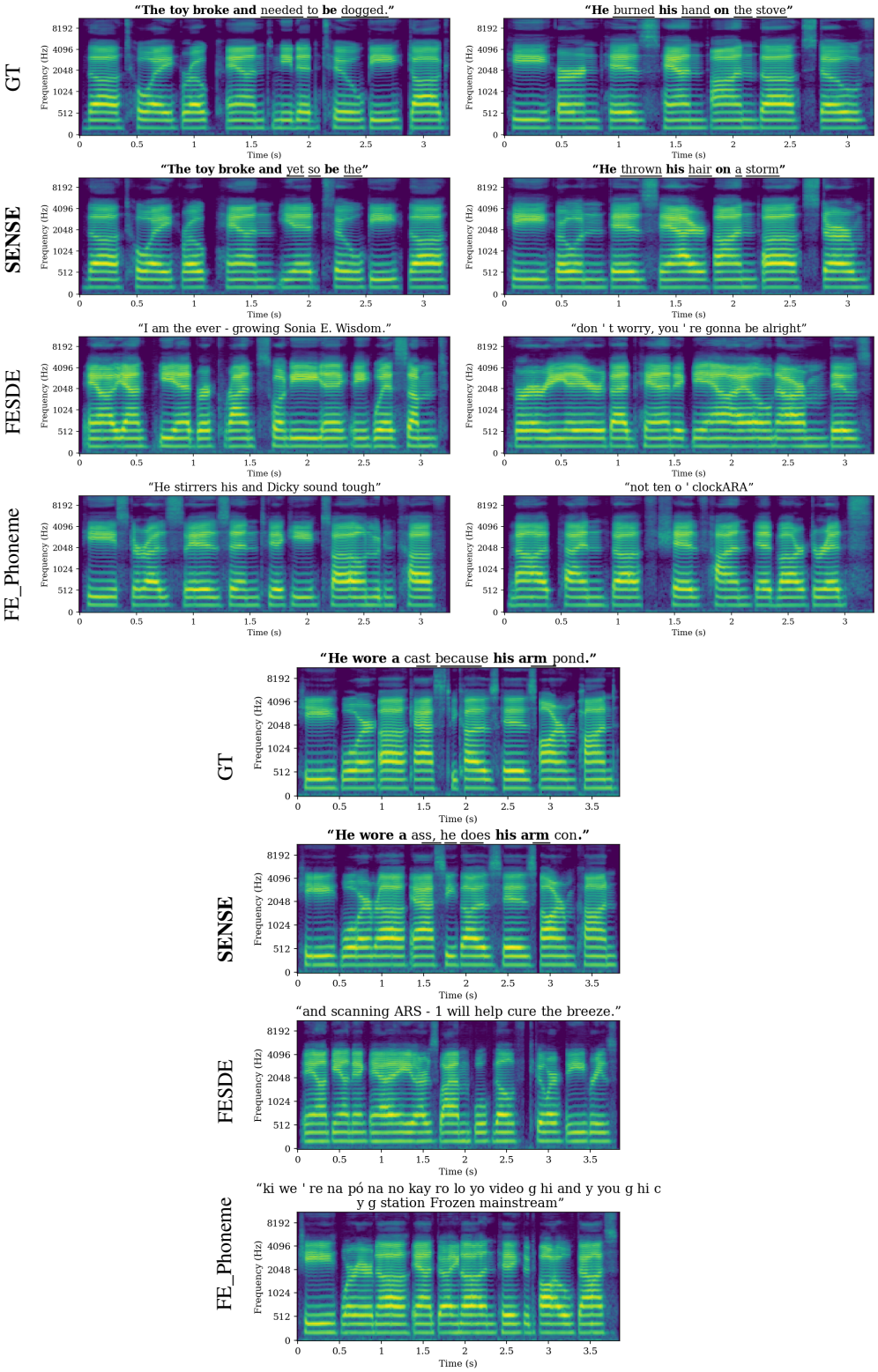


Figure A.7: Additional Melspectrogram Comparison

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 abstract and introduction clearly state the paper's contributions: (1) reformulating EEG-to-speech as a joint acoustic-semantic problem, (2) proposing SENSE with a graph-based EEG encoder and EEG Semantic Conditioning (ESC), and (3) demonstrating improved acoustic and semantic metrics including strong cross-subject generalization. These claims are supported by experimental results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5 discusses that the N400 corpus uses fixed templates with varied sentence-final words, which differs from naturalistic speech, and that absolute WER values indicate practical EEG-to-speech reconstruction remains challenging. Extension to naturalistic speech and active-speech paradigms is a natural next step.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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 does not include theoretical results requiring formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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 (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 fully describes the model architecture, and Appendix A provides comprehensive implementation details including dataset preprocessing, architecture specifications, training configuration, evaluation protocol, and loss weights. Subject splits and the nested cross-subject scaling structure are also specified. Anonymized code is provided as supplementary material for direct reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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, as described in supplemental material?

   Answer: [Yes]

   Justification: The N400 EEG corpus [19] is publicly available and cited. We provide anonymized code as supplementary material (a single zip file under 100MB). The README provides instructions for environment setup, data preparation, training, and inference. Pretrained checkpoints are not included due to file size constraints but will be released with the de-anonymized version upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies dataset splits, baselines, metrics, and training setup. Appendix A.5 details optimizer, learning rate, loss weights, training duration, and hardware. Hyperparameter selection is justified via the sensitivity analysis in Appendix A.13

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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 main results in Tables 1, 2, 4, 5, A.10, A.11 and Figures 4, 5 report mean ± standard deviation over 5 random seeds. The variability captured is across random seeds with fixed train/test splits, as noted in Section 4 and Appendix A.5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A.5 states that training a single SENSE model takes approximately 2 days on 2 NVIDIA RTX 3090 GPUs with FP32 precision and batch size 16 per GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The paper uses a publicly available EEG dataset [19], does not involve new human subject data collection, and preserves anonymity in submission.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

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 impact is discussed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release pretrained models or scraped datasets posing high misuse risk. Future released checkpoints (upon acceptance) will be trained on a publicly available EEG corpus and pose no high misuse risk, as discussed in Appendix **??**.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets are properly cited with their original publications. Specifically, the N400 EEG corpus [19] is released under CC BY-NC-ND 4.0 and is used for non-commercial research as permitted. The codebase builds upon publicly available implementations of VITS, S4, and Conformer, all of which are released under permissive open-source licenses (MIT or Apache 2.0). Our released code is provided under the MIT License, included as a LICENSE file in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release anonymized code as supplementary material. The repository includes the model implementation, configuration files, training and inference scripts, a documented README, and a LICENSE file (MIT). Pretrained checkpoints are not included due to file size constraints but will be provided in the de-anonymized release upon acceptance.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper uses a publicly available EEG dataset [19] and does not involve new crowdsourcing or human subject data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human subject research was conducted. The N400 EEG corpus [19] is publicly available and was collected by the original authors under their institutional approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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 in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a core methodological component of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.