NEURAL INFORMATION PROCESSING SYSTEMS

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]

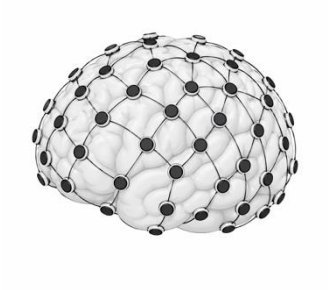
[1]Department of Artificial Intelligence, Chung-Ang University, Korea
[2]Department of Computer Science and Engineering, Chung-Ang University, Korea
[3]School of Psychology, Centre for Human Brain Health (CHBH), University of Birmingham, UK

40th Conference on Neural Information Processing Systems

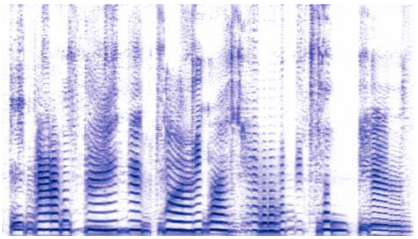
Task: What is EEG-to-speech generation?



LISTEN
Speech is heard



RECORD EEG
128 scalp channels



SYNTHESIZE
Generate waveform

- A listener hears a sentence while 128-channel EEG is recorded.
- A model learns to generate the corresponding speech waveform from EEG.
- At inference, only EEG is given. This is perceived speech, not thought reading.

Motivation: EEG contains more than acoustic information

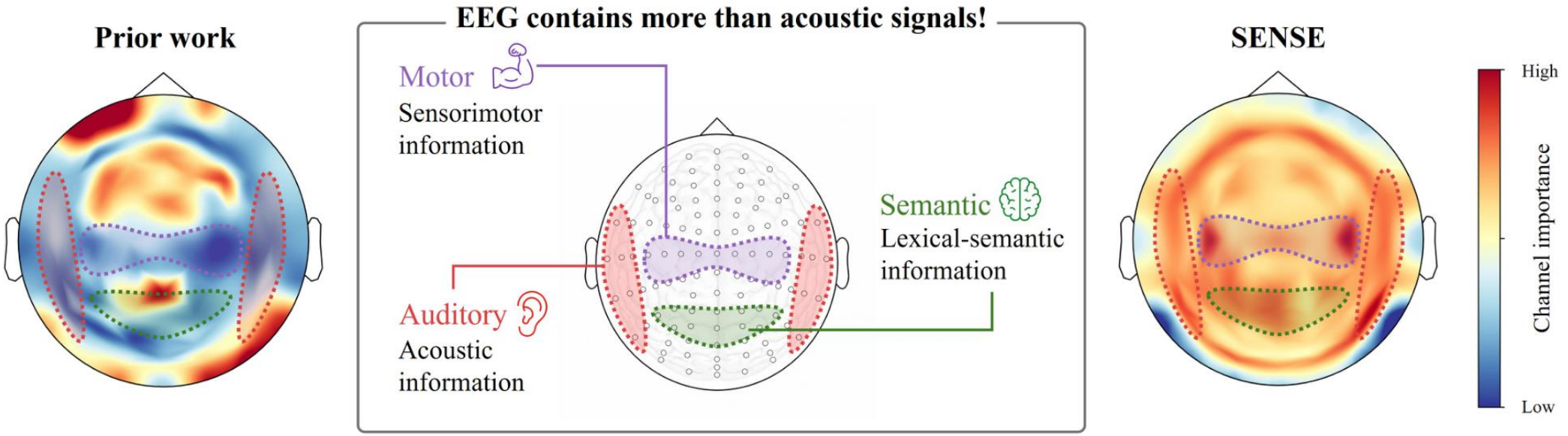


Figure 1. Acoustic, sensorimotor and semantic information across the scalp.

- Existing EEG-to-speech models mainly optimize acoustic reconstruction.
- Electrode geometry is underused.
- Our goal: combine spatial modeling with semantic supervision.

SENSE framework: structure-aware EEG plus semantics

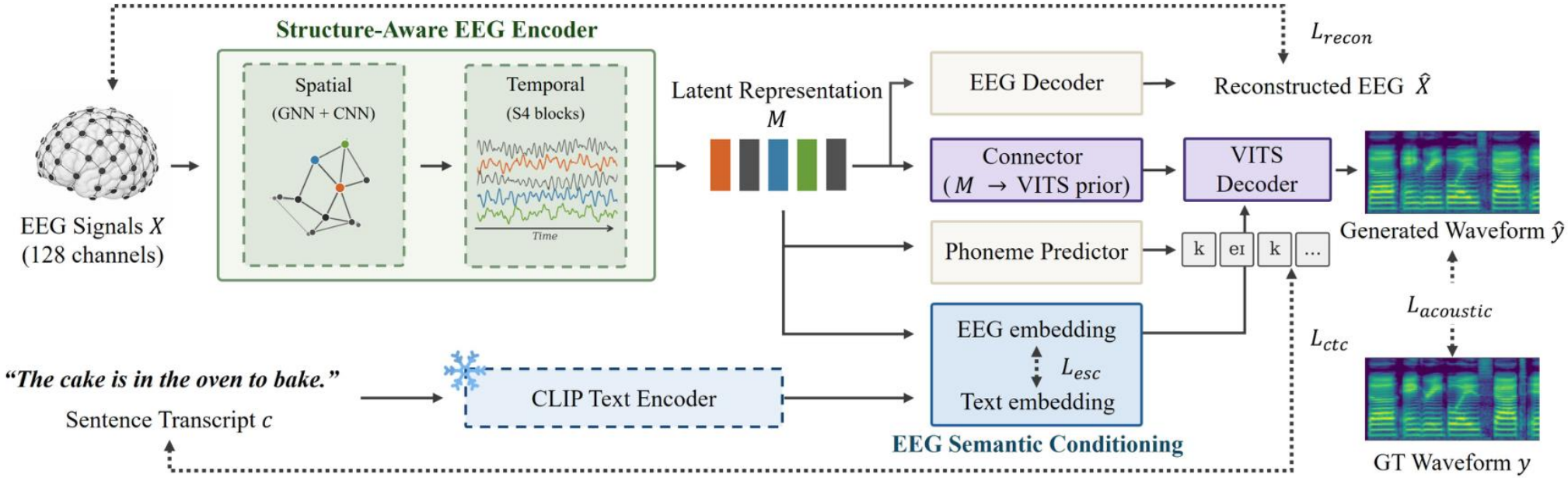


Figure 2. SENSE training pipeline (original paper figure).

- Spatial EEG encoder: channel gate, geometry-based GNN, CNN skip, then S4.
- Training adds EEG reconstruction and phoneme supervision.
- EEG Semantic Conditioning aligns EEG to frozen CLIP text embeddings.
- At inference, a VITS decoder generates speech using EEG only.

EEG Semantic Conditioning: use N400 congruency

ESC alignment: ON

CONGRUENT TRIAL

The ending fits the context

"The boys got in a fight."

ESC alignment: OFF

INCONGRUENT TRIAL

The ending violates the context

"The boys got in a mouse."

Both trial types still supervise acoustic waveform reconstruction and phoneme prediction.

- N400 trials contain congruent and incongruent sentence endings.
- ESC aligns EEG with CLIP targets on congruent trials only.
- All trials still train acoustic and phoneme objectives.
- A conditioning bridge prepares the decoder for EEG-only inference.

K. K. Toffolo et al., "Evoking the N400 Event-Related Potential (ERP) Component," Neuroscience, 2022.

Main results: acoustic and semantic reconstruction

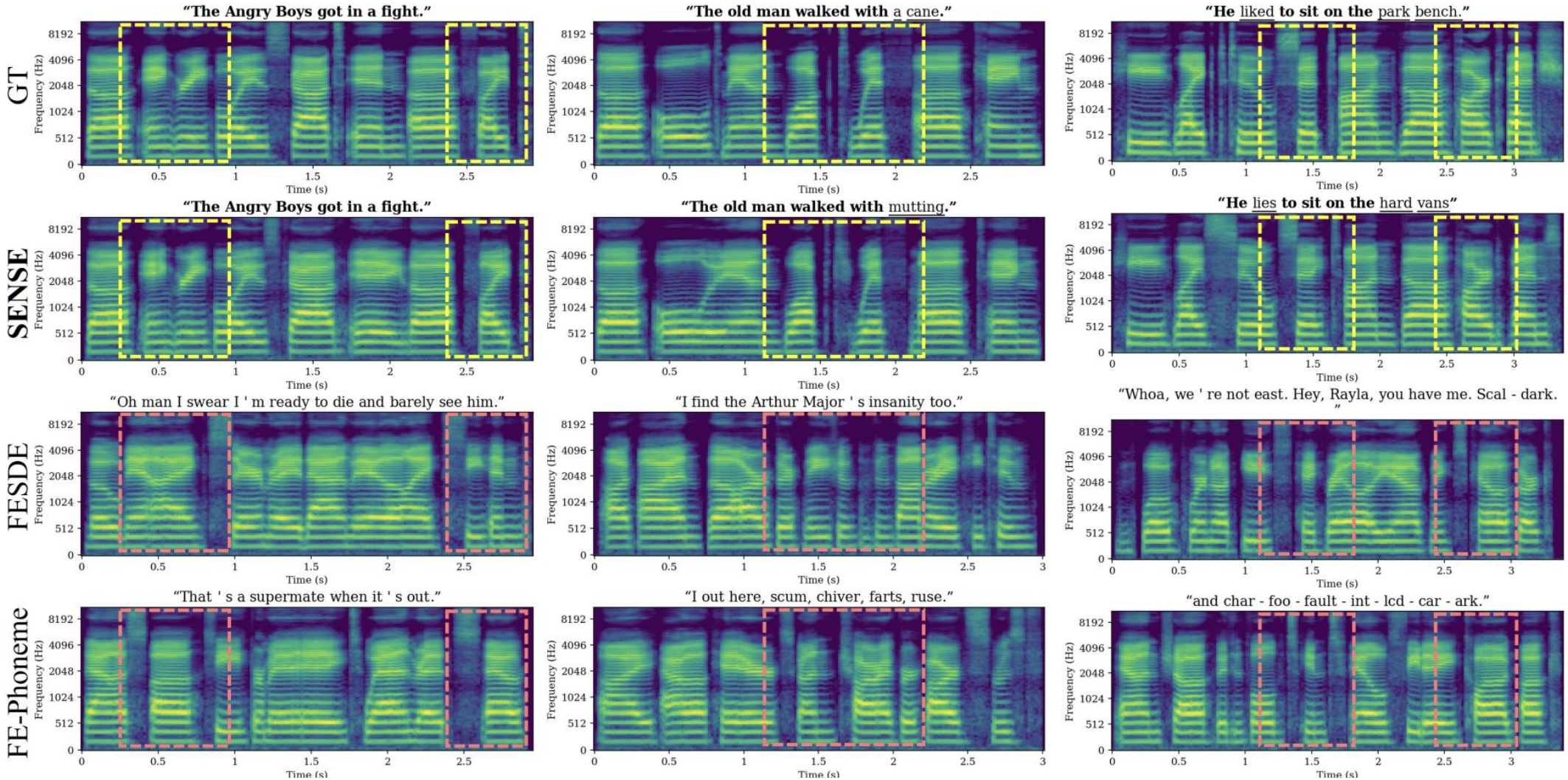
Table 1: **Comparison of end-to-end EEG-to-speech models** across three evaluation conditions.

	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**

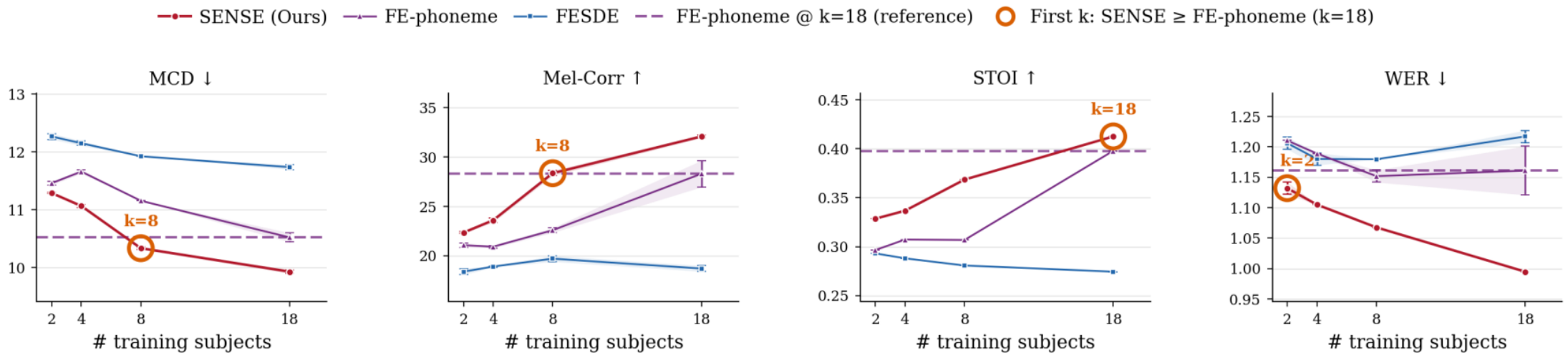
- N400 benchmark: unseen both, unseen audio and unseen subjects.
- Unseen subjects: Mel-Corr improves 26.90 to 32.07; MCD drops 10.65 to 9.93.
- WER improves 1.123 to 0.995, but word-level recovery remains challenging.

[17] J. Lee et al., “Toward Fully-End-to-End Listened Speech Decoding from EEG Signals,” Interspeech, 2024.
[13] J. Lee et al., “Enhancing Listened Speech Decoding from EEG via Parallel Phoneme Sequence Prediction,” ICASSP, 2025.

Main results: acoustic and semantic reconstruction



Generalization: fewer training subjects, stronger transfer



- We train with 2, 4, 8 or 18 subjects and test on held-out people.
- With only 2 subjects, SENSE beats the 18-subject FE-Phoneme WER.
- By 8 subjects, SENSE matches or exceeds baseline acoustic metrics.

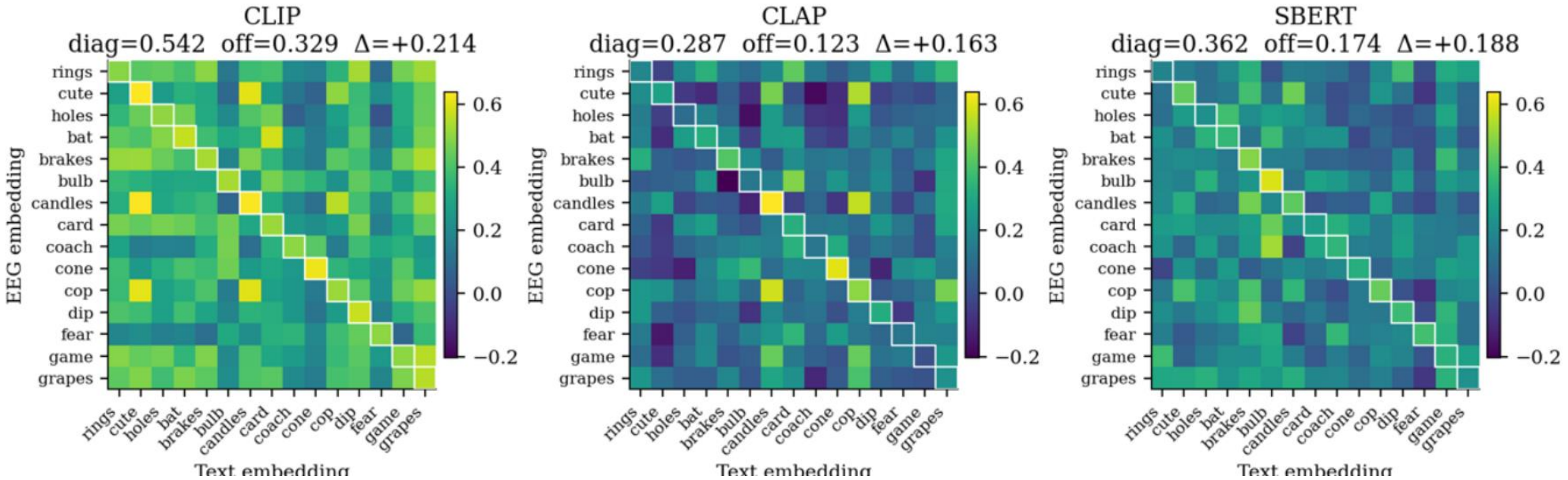
Ablation: both structure and semantic targets matter

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

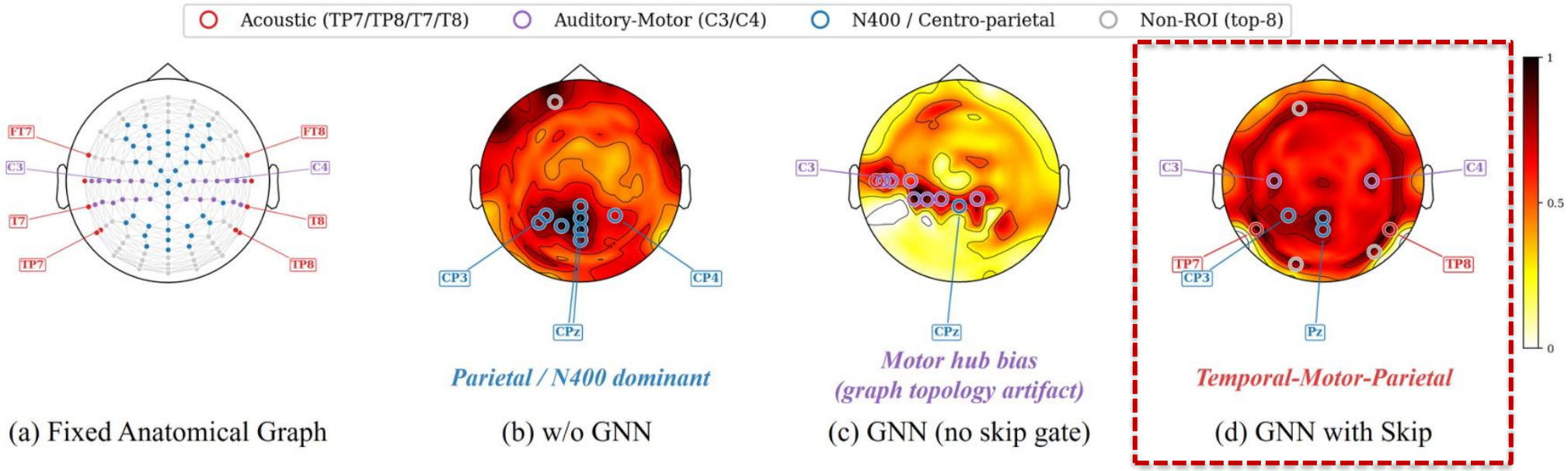
- Removing ESC reduces both acoustic and semantic performance.
- Shuffling CLIP targets nearly removes the ESC benefit.
- CTC, GNN, CNN skip and channel attention also help.

Analysis: does EEG align with the intended meaning?



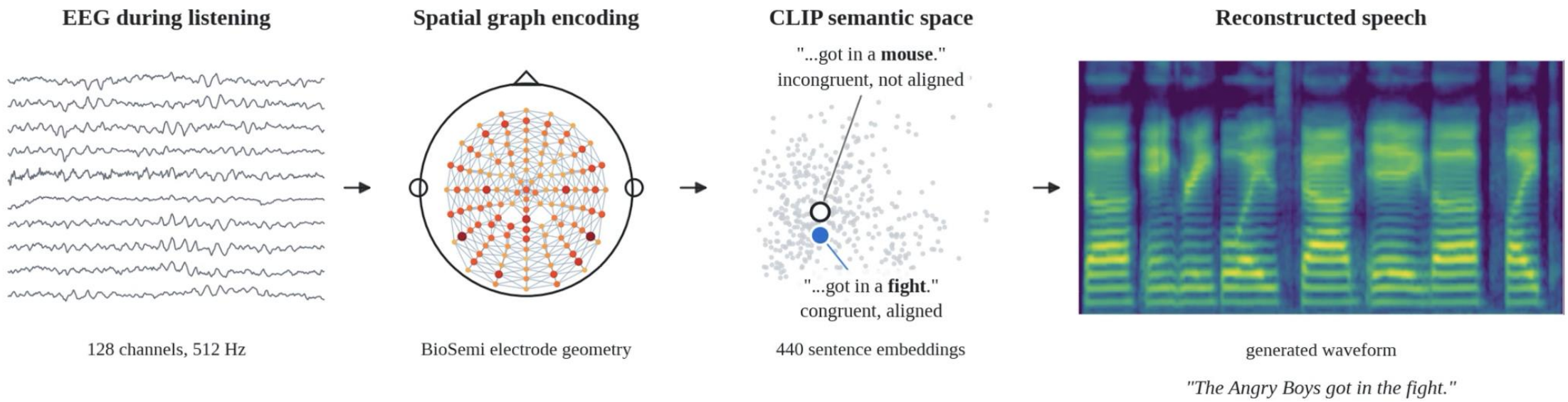
- A bright diagonal means the correct EEG-text pairs are more similar.
- CLIP shows the clearest diagonal and the largest gap: 0.214.

Analysis: which EEG regions support semantic alignment?



- Without a GNN, attribution centers on centro-parietal electrodes.
- Full SENSE highlights auditory, motor and centro-parietal regions.

Takeaways



- SENSE combines electrode geometry and congruency-aware semantics.
- It improves reconstruction, especially on unseen subjects.
- Word-level reconstruction remains difficult.

Code + Audio Samples