Do Speech BCIs Need Larger Models? Rethinking Neural Decoding beyond Scaling

Zehui Feng[1,2] **Weichuan Wang**[1] **Xiaohan Chen**[1] **Cuntai Guan**[2,*] **Ting Han**[1,3,*]

[1]Shanghai Jiao Tong University
[2]Nanyang Technological University
[3]Zhejiang University

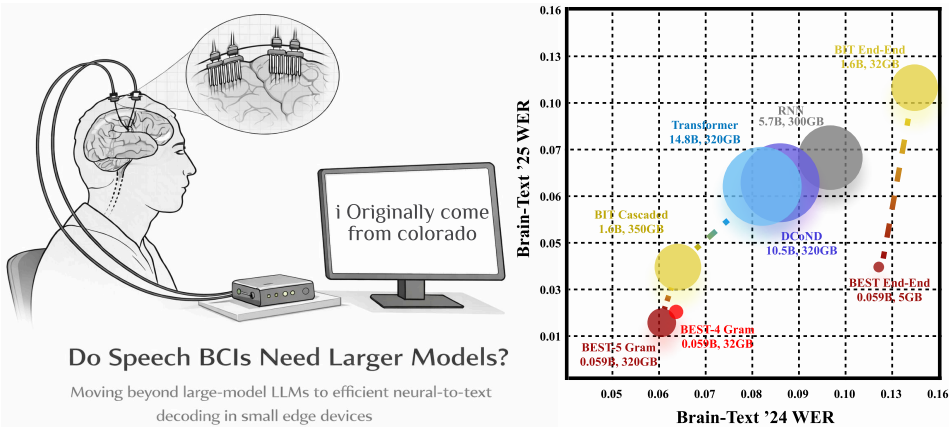


Figure 1. Method Parameter (B), RAM (GiB), and Word Error Rate (WER) on Brain-to-Text '24 and '25. Smaller circles indicate lower resource consumption.

Abstract

Recent speech brain-computer interfaces (BCIs) increasingly rely on large models and complex multi-stage pipelines for neural decoding. However, in speech BCI, where training data are scarce, noisy, and expensive to collect, the benefits of model scaling remain unclear. This work revisits a fundamental question: **are large models truly necessary for accurate neural decoding?** We propose ***BEST*** (***B***rain-***E***ncoding-***S***peech-***T***ext), a two-stage framework for efficient neural-to-text decoding. Instead of relying on large generative models or phoneme-centric pipelines, ***BEST*** introduces hierarchical intermediate representations and a compact speech-oriented decoder that directly maps neural activity to text. This design preserves essential structure while significantly reducing computational cost, enabling robust generalization in low-resource settings and practical deployment under constrained resources. On the Brain-to-Text '24 and '25 benchmarks, ***BEST*** achieves word error rates (WERs) of 6.19% and 1.62%, respectively, using only 3.85% of the parameters required by comparable methods. Beyond these results, we provide three key insights: (1) decoder scaling yields non-monotonic gains; (2) cross-modal alignment is more critical than model size; and (3) overfitting degards sequence-level performance. Representation analysis further reveals a consistent neural→audio→text alignment pathway.

1 Introduction

Speech brain-computer interfaces (BCIs) offer a promising pathway for restoring communication in individuals with severe motor impairments [47, 43]. By directly mapping neural activity to language, these systems bypass damaged neuromuscular pathways and enable brain-to-text communication [7]. Recent advances in neural decoding[52] have substantially improved transcription quality, bringing speech BCIs closer to practical deployment [58, 11, 23]. With the rapid advancement of large language models (LLMs)[54, 17, 26, 49], most existing approaches adopt a multi-stage pipeline: neural signals are first decoded into phoneme sequences via supervised classification [23], and then converted into text using LLMs, often with additional rescoring [36, 55]. While effective due to the increasingly strong linguistic priors provided by rapidly evolving LLMs, this paradigm suffers from two fundamental limitations.

First, it is poorly aligned with real-world deployment constraints. Practical BCI systems require low-latency and resource-efficient inference [32, 38], yet reliance on large pretrained models and memory-intensive rescoring introduces substantial computational overhead, limiting feasibility on edge or clinical hardware. Second, the pipeline imposes a discrete phoneme bottleneck that discards rich neural information. In data-scarce and noisy BCI settings, collapsing high-dimensional neural signals into hard phoneme labels amplifies overfitting and suppresses fine-grained linguistic structure [11]. Moreover, the strict separation between intermediate prediction and sentence generation prevents effective use of multi-level neural-linguistic representations during decoding.

These challenges point to a central requirement: an effective speech BCI must achieve high accuracy while remaining lightweight, data-efficient, and robust under limited supervision. To this end, we propose ***BEST*** (***B***rain-***E***ncoding-***S***peech-***T***ext), a two-stage framework that replaces stage-wise symbolic pipelines with hierarchical feature-level integration. In Stage 1, multi-level neural-linguistic representations are fused through a lightweight integration module, allowing phonological and lexical information to interact before discrete decoding and mitigating overfitting induced by phoneme-level supervision. In Stage 2, the fused representations are translated by a compact speech-oriented decoder, followed by beam search with a lightweight n-gram language model for efficient hypothesis scoring. This design preserves informative neural structure while introducing linguistic regularization without the computational burden of large-scale reranking.

We evaluate BEST on the Brain-to-Text '24 and '25 benchmarks, achieving word error rates (WERs) of 6.19% and 1.62%, respectively. Despite using only 3.85% of the parameters of comparable methods, BEST consistently improves decoding performance while remaining suitable for edge-class deployment. Representation analyses further reveal a coherent neural $\rightarrow$ linguistic $\rightarrow$ text alignment process underlying the model's predictions.

In summary, our contributions are threefold:

1. We propose BEST, a lightweight and practical framework for speech BCI that achieves strong performance with only 3.7% of the parameters of existing approaches.

2. We identify the discrete phoneme bottleneck and introduce a hierarchical feature-level decoding paradigm that preserves and integrates multi-level neural-linguistic representations.

3. We systematically analyze the role of decoder scale in speech BCI, showing that performance gains depend more on structured alignment than on model size under realistic constraints.

2 Related Work

Speech brain-computer interfaces. Speech BCIs decode neural activity from auditory and motor cortices into text. Early systems use RNN-based models to map neural signals to intermediate units (e.g., phonemes or characters), followed by statistical language models for text generation [5, 2, 48, 9, 28, 20]. Later work incorporates context-dependent modeling (e.g., diphones) to capture local phonetic transitions [23]. More recent approaches adopt transformer architectures for long-range dependency modeling and streaming or end-to-end decoding, often combined with LLMs [11, 28, 12]. Despite these advances, most methods rely on discrete intermediate representations and fail to preserve fine-grained temporal and spectral structure in neural signals [35].

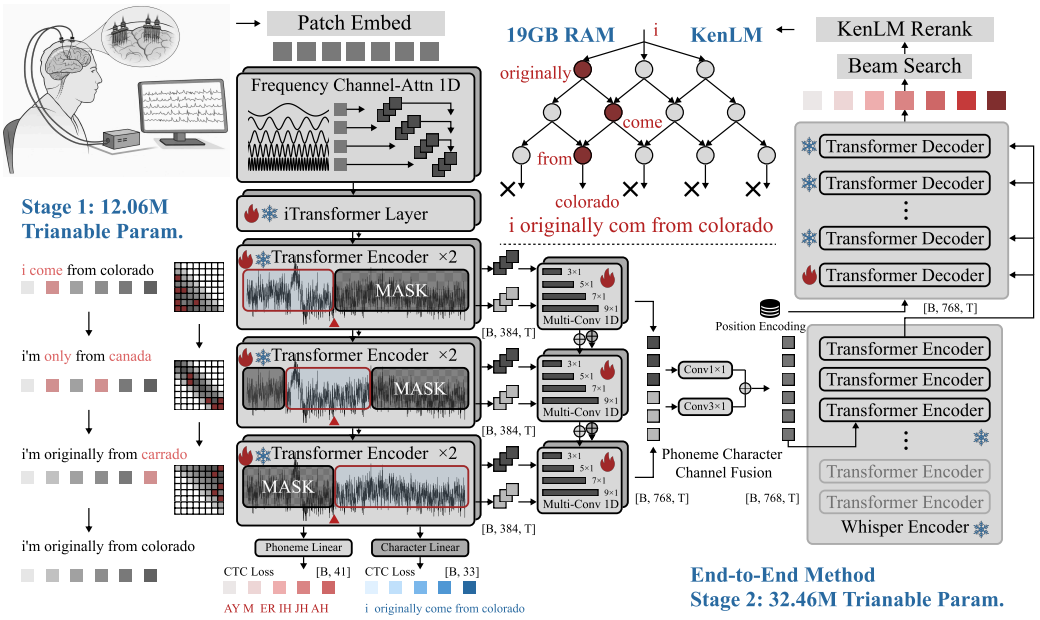


Figure 2: ***Overall illustration of BEST*. The framework consists of four core modules: (1) *Hierarchical Transformer Encoder* (Stage 1): it employs frequency-channel attention and iTransformer layers to encode neural activity into phoneme/character representations with hierarchical temporal masking; (2) *Phoneme-Character Fusion Block* (Stage 1): a lightweight convolutional fusion module aggregates hierarchical neural representation to bridge neural and linguistic spaces; (3) *Lightweight Speech Translator* (Stage 2): a lightweight speech translator generates autoregressive text sequences from fused embeddings; and (4) *KenLM Reranking* (Stage 2): a custom 4-gram language model resolves ambiguities from noisy neural signals for final selection.**

Large language models for BCI. LLMs have recently enabled neural-to-text decoding, semantic reconstruction, and cross-modal alignment[6, 57, 10]. Prior work explores LLM fine-tuning for neural decoding [34, 44, 8, 37] and autoregressive reconstruction from fMRI signals [51]. LLMs have also been used to align neural and linguistic representations [29], model semantic organization [33, 14], and support applications such as spellers and rehabilitation systems [4, 18, 16, 24, 25]. However, current approaches face key limitations: (1) end-to-end LLM decoding ignores structural mismatch between neural signals and linguistic embeddings; (2) large-scale inference incurs high computational and memory cost [58, 42]; and (3) weakly constrained decoding reduces robustness under noisy neural inputs.

Real-world BCI deployment. Practical BCI systems require efficient, robust, and accurate decoding under resource constraints [32, 38, 21, 53]. Existing lightweight models mainly target motor or intention decoding and do not generalize well to speech [30, 56]. Classical methods (e.g., random forests) are efficient but cannot model the sequential dependencies required for speech [27, 13]. Lightweight neural models face a core trade-off: limited capacity weakens representation learning, while increased complexity leads to overfitting and unstable optimization in low-resource settings [1, 15]. This highlights the need for architectures that balance efficiency with structurally aligned neural-linguistic modeling.

3 Method

In this section, we present a lightweight speech BCI framework with a hybrid decoding strategy. The overall architecture comprises a hierarchical transformer-based neural encoder, a lightweight phoneme-character fusion block, a speech translator, and a custom language model (Figure 2). The neural encoder is pretrained to map neural signals into both phoneme-level and character-level representations, providing robust linguistic features. At inference time, encoder representations are

fed into a Whisper-based decoder to generate an n-best list of hypotheses, which are subsequently rescored with a 4-gram KenLM for final selection.

3.1 Architecture

Stage 1: hierarchical transformer encoder. We introduce a hierarchical Transformer encoder [46] tailored for neural speech decoding, combining frequency-aware channel modulation with structured temporal attention. Neural signals are partitioned into temporal patches and embedded as token sequences. To incorporate spectral structure, we apply a multi-frequency channel attention mechanism (Appendix J) that projects inputs onto predefined frequency bases, aggregates channel-wise statistics, and produces adaptive reweighting coefficients, enabling frequency-selective feature refinement.

The resulting embeddings are first processed by an iTransformer [31] to capture global dependencies and stabilize latent representations, followed by a hierarchical Transformer with structured attention (Appendix J). Temporal dependencies are regulated through three complementary masking schemes: past-oriented masking for historical context, local bidirectional masking for short-range interactions, and future-oriented masking to incorporate forward cues. These masks are applied sequentially to progressively refine the temporal receptive field, enabling the model to capture localized temporal patterns while maintaining global consistency. The final representations are projected to both phoneme-level and character-level outputs via CTC heads, providing complementary supervision at different linguistic granularities.

Stage 2: phoneme–character fusion. To integrate multi-level linguistic representations, we design a lightweight multi-scale convolutional fusion module that aggregates features across encoder depths. Phoneme and character streams are processed through parallel convolutional branches with different receptive fields, capturing complementary temporal patterns. Within each stream, residual aggregation enables iterative refinement and cross-level information flow while preserving temporal structure. The fused features are then concatenated and projected into a unified representation through lightweight convolutions.

Compared to additional attention layers, this design is more computationally efficient and better suited to low-resource, high-noise BCI settings. Multi-scale convolutions capture both local and longer-range dependencies without the cost of full self-attention. The fusion module also serves as a projection interface, aligning neural representations with the acoustic embedding space and facilitating downstream linguistic decoding.

Stage 2: lightweight speech translator. To enable efficient neural-to-text mapping without large language models, we adopt a compact speech translation module based on Whisper [40]. The fused representations are projected into the acoustic embedding space and processed by the upper layers of the Whisper encoder, aligning them with pretrained speech representations and introducing implicit acoustic priors without requiring explicit audio input. The transformed embeddings are then decoded autoregressively by the Whisper decoder.

The model is trained using next-token prediction with teacher forcing. During inference, text is generated conditioned on neural-derived embeddings, optionally with beam search to improve sequence quality. This design replaces LLM-based decoding with a pretrained speech model, enabling direct neural-to-text transcription with substantially lower computational cost.

Stage 2: lightweight language model reranking. Neural-derived embeddings are inherently noisy, leading to ambiguous candidate sequences. To improve robustness, we apply lightweight n-gram rescoring using a custom 4-gram KenLM [19]. Compared to standard large-scale n-gram models (e.g., 5-gram models), this approach provides effective linguistic regularization with minimal overhead. It improves disambiguation and transcription accuracy while preserving real-time feasibility for BCI deployment.

3.2 Training objective

Pretraining stage. In contrast to masked reconstruction approaches such as BIT [58], we do not employ masked autoencoding. While masking promotes contextual inference, neural speech decoding depends more on preserving temporal–spectral structure and higher-level phonetic and linguistic information (Appendix G). To improve generalization, we adopt extensive data augmentation and hierarchical pretraining at both phoneme and word levels, enabling the encoder to learn compact

representations that capture both fine-grained acoustic structure and higher-level linguistic information from neural signals.

Fine-tuning stage. For sentence-level decoding, we freeze the neural encoder and Whisper encoder to preserve stable representations. Phoneme- and character-level features are injected into the upper layers of the Whisper encoder, while only the first layer of the Whisper decoder and the fusion module are updated. This targeted optimization restricts learning to lightweight components, reducing overfitting and training instability while enabling effective adaptation to neural-derived representations for accurate transcription.

4 Evaluation

4.1 Dataset

We evaluate our method on the Brain-to-Text '24 [48] and '25 [3] benchmarks for attempted speech decoding. The '24 dataset contains 12,100 sentences from participant T12 collected over 25 sessions using intracortical microelectrode arrays implanted in speech-related regions (Appendix A). The '25 dataset includes 10,948 sentences from participant T15 across 45 sessions with 256 electrodes in the speech motor cortex. In both datasets, participants were visually prompted with sentences and asked to attempt speech following a cue, with decoding performed on the corresponding neural activity.

The '24 dataset consists of conversational sentences with attempted vocalized speech at approximately 62 words per minute. In contrast, the '25 dataset spans more diverse corpora, and custom sets, and includes both attempted vocalized and silent speech. Neural features include threshold crossing spike counts and spike-band power, extracted at 20 ms resolution and normalized across sessions to reduce non-stationarity.

4.2 Baselines

We benchmark our method against 4 state-of-the-art baselines in neural speech BCIs:

1. ***RNN Baseline*** [48] employs a 5-layer bidirectional GRU to map neural activity to phonemes. It generates text via a 5-gram language model with OPT rescoring.
2. ***Transformer Baseline*** [11] adopts a causal transformer enhanced with time-masking augmentation to improve generalization. It operates on non-overlapping neural patches with relative positional encoding and supports real-time streaming inference.
3. ***DCoND*** [23] improves phoneme decoding by modeling diphone contexts rather than individual phonemes. It uses a GRU encoder to estimate diphone distributions and leverages an ensemble of decoders with an LLM for correction.
4. ***BIT*** [58] is an end-to-end speech decoding framework built on a cross-species, cross-task pretrained transformer. The encoder is first self-supervised pretrained via masked temporal patch reconstruction, then fine-tuned for phoneme decoding with CTC loss.

4.3 Results

Performance comparison on speech BCI decoding: We evaluate *BEST* on the Brain-to-Text '24 and '25 benchmarks (Table 1), reporting model size, memory consumption, phoneme error rate (PER), character error rate (CER), and WER on both validation and held-out test sets. Following prior work [58], we consider both end-to-end decoding and cascaded LM pipelines. *BEST* is instantiated in three configurations reflecting different accuracy-efficiency trade-offs:

1. ***Lightweight BEST*** is an end-to-end pipeline without external language models that minimizes computational cost while preserving robutsness;
2. ***Practical BEST*** incorporates a 4-gram LM to balance linguistic regularization and efficiency;
3. ***Robust BEST*** employs a 5-gram LM to maximize accuracy under higher memory budgets.

The lightweight configuration consistently outperforms the end-to-end baseline of BIT despite using substantially fewer parameters and eliminating reliance on large pretrained language models, demonstrating the effectiveness of hierarchical feature-level decoding. Incorporating lightweight n-gram LMs further reduces WER, with the practical and robust variants achieving 2.92% and 1.62% WER on the '25 benchmark, respectively. The results reveal that performance gains are primarily driven by improved neural representation learning rather than increased capacity to communicate with LMs.

Table 1: **We compare the speech-BCI framework on the Brain-to-Text '24 and '25 hold-out set. Gray background indicates the pipeline without the language model. Some results lack detailed data as the corresponding code is unavailable; see the competition leaderboard for more details. ⋆, ◇, • denote the most lightweight network, the most robust network, and the most practical network, respectively.**

Model	#Param, RAM	Brain-to-text '24 T12				Brain-to-text '25 T15			
		PER↓	CER↓	Valid WER↓	WER↓	PER↓	CER↓	Valid WER↓	WER↓
RNN(Baseline) [48]	5.70B, 300GiB	0.2122	0.2561	0.2271	0.0976	0.1310	0.2075	0.1127	0.0667
Transformer [11]	14.79B, 320GiB	0.1628	0.1732	0.1121	0.0818	0.1081	0.1573	0.0934	0.0612
DCoND [23]	10.50B, 320GiB	0.1534	-	-	0.0860	-	-	-	-
BIT [58]	1.53B, 32GiB	0.1439	-	0.1607	0.1567	0.0712	-	0.0879	0.1106
BIT [58]	1.53B, 350GiB	-	-	-	0.0635	-	-	-	0.0406
BEST⋆	0.059B, 5GiB	**0.1321**	**0.1436**	0.1504	0.1261	**0.0688**	**0.0896**	0.0681	0.0517
BEST+5-Gram LM◇	0.059B, 320GiB	**0.1321**	**0.1436**	**0.0721**	**0.0619**	**0.0688**	**0.0896**	**0.0257**	**0.0162**
BEST+4-Gram LM•	0.059B, 32GiB	**0.1321**	**0.1436**	0.0764	0.0654	**0.0688**	**0.0896**	0.0292	0.0196

Table 2: **Performance comparison across different decoders with parameter counts and fine-tuning strategies on T15. Gray background indicates the audio-based LLM decoders.**

Decoder	#Params	Dimension	Valid WER↓		
			Full fine-tuning	Partial fine-tuning	Frozen
Whisper-tiny [39]	34M	384	0.3678	0.2845	0.2996
Whisper-base [39]	74M	512	0.2409	0.2183	0.2235
DistilGPT2 [41]	82M	768	0.2091	0.1930	0.2006
Whisper-large [39]	1.55B	1280	0.1461	0.1460	0.1572
Whisper-medium [39]	769M	1024	0.1351	0.1309	0.1366
Fine-tuned DistilGPT2 [41]	82M	768	0.1323	0.1152	0.1363
Llama3.1 [45]	8B	4096	0.1256	0.1163	0.1099
Fine-tuned Llama3.1 [45]	8B	4096	0.1209	0.1131	0.1052
Qwen2 [50]	1.5B	1536	0.1133	0.1091	0.1050
Fine-tuned Qwen2 [50]	1.5B	1536	0.1103	0.0967	0.0938
Aero1-Audio [22]	1.5B	1280	0.1018	0.0989	0.0830
Whisper-small [39] (***BEST***)	244M	768	**0.0764**	**0.0681**	**0.0795**

Notably, *BEST* significantly narrows the gap between end-to-end and cascaded pipelines, indicating that mitigating the discrete phoneme bottleneck and preserving hierarchical neural-linguistic representations are critical for robust and efficient speech BCI decoding.

LLM decoder scaling and adaptation analysis: We systematically study the effect of decoder scale and adaptation strategy in end-to-end speech BCI (Table 2). Across all settings, we find that decoder scaling exhibits a distinctly non-monotonic trend. Small decoders such as Whisper-Tiny [39] and DistilGPT2 [41] are capacity-limited, while larger text-based LLMs fail to deliver consistent gains with increasing scale. Instead, mid-scale decoders such as Qwen2-1.5B [50] achieve the best performance, revealing a non-monotonic scaling trend. This suggests that, in speech BCI, increasing model capacity can amplify overfitting to noisy neural signals rather than improve generalization, making capacity-efficient decoders more effective than maximally scaled models. A second key observation is that pretraining modality is decisive. Audio-oriented models (e.g., Aero1-Audio [22], Whisper [39]) consistently outperform text-only LLMs of similar or larger size, indicating that representation alignment with neural inputs, rather than model scale, dominates decoder performance. Even with additional projection layers for adaptation, text-only LLMs [45, 50] remain mismatched to the distribution of neural features. In contrast, speech-aligned pretraining produces representations inherently compatible with neural-derived embeddings.

We further observe a scale-dependent adaptation pattern. Full fine-tuning is consistently suboptimal, reflecting instability under data-limited and high-noise conditions. For compact decoders, partial fine-tuning is sufficient to align representations, whereas larger models favor fully frozen decoding, as pretrained priors are degraded by task-specific adaptation. Decoder dimensionality (768) has a limited effect on performance. The encoder dimension is selected via hyperparameter tuning (Appendix G), as increasing dimensionality exacerbates overfitting in the neural encoding stage.

Encoder–decoder factorization. To disentangle the contributions of neural representation quality and downstream decoding, we factorize the encoder and decoder choices while keeping the remaining pipeline fixed (Table 3). The BEST encoder consistently outperforms Transformer and GRU encoders under both Whisper-small and Aero, showing that its gain is not specific to a particular downstream decoder. Conversely, Whisper-small consistently outperforms the substantially larger Aero decoder for every encoder, with WER gaps of 0.0308–0.0408. Thus, increasing decoder capacity does not compensate for weak neural representations; both the quality of the neural encoder and its compatibility with the downstream speech decoder determine sequence-level performance. Notably, the decoder gap decreases from 0.0408 with GRU to 0.0308 with BEST, suggesting that better-structured neural representations are also less sensitive to the choice of downstream decoder. These results further support our central finding that representation quality and cross-stage compatibility, rather than decoder scale alone, dominate speech BCI decoding.

Ablation analysis: We observe a monotonic improvement as each component is introduced (Table 4). Hierarchical attention improves temporal dependency modeling, while frequency attention further enhances spectral selectivity. The largest gain arises from the phoneme-character fusion module, showing that robust decoding depends on integrating multi-level linguistic representations rather than isolated phoneme prediction. *BEST* also achieves lower validation CTC loss (Appendix G and E), providing independent evidence of better-aligned neural representations and more stable optimization. The detailed architecture and hyperparameter settings are also provided (Appendix A, C, D, and G).

Overfitting analysis: The overfitting model achieves lower PER and CER but degrades WER (Table 5), exposing a misalignment between local phoneme accuracy and sequence-level decoding. We attribute to representation collapse induced by overfitting. The visualization of output probability heatmap (Appendix F) shows that the overfitting model exhibits highly concentrated and fragmented activations, reflecting overconfident and poorly calibrated predictions. These representations lack temporal coherence and fail to capture structured phoneme transitions, resulting in inconsistent sequence reconstruction. In contrast, the well-fitting model produces smoother and more distributed activations, preserving coherent neural-linguistic structure. This reveals a fundamental property of speech BCI: overfitting increases local discriminability while degrading global consistency. Reduced phoneme error does not translate to better decoding, as the resulting representations become less compatible with downstream inference. This explains why stronger phoneme supervision can harm end-to-end performance and supports the design of *BEST* in preserving structured, multi-level representations rather than relying on discrete phoneme predictions.

Table 3: **Encoder–decoder factorization on Brain-to-Text '25 validation set.**

Neural encoder	Whisper-small	Audio-LLM	Decoder gap
BEST encoder	**0.0681**	**0.0989**	**0.0308**
Transformer encoder	0.0862	0.1217	0.0355
GRU encoder	0.1054	0.1462	0.0408

Table 4: **Comparison of *BEST*'s different components on Brain-to-Text '24 and '25.**

Model	T12		T15	
	PER↓	CER↓	PER↓	CER↓
1. w. Transformer	0.1628	0.1732	0.1081	0.1573
2. w. Hierarical Attention Mask	0.1538	0.1597	0.0835	0.1103
3. w. Frequency Attention Block	0.1321	0.1436	0.0688	0.0896
4. w. Phoneme-Character Fusion	**0.1241**	**0.1394**	**0.0612**	**0.0758**

Table 5: **Performance comparison of well-fitted and over-fitted settings on validate PER, CER, and WER.**

Model	'24 T12			'25 T15		
	PER	CER	WER	PER	CER	WER
Well-fitted	0.1321	0.1436	**0.1504**	0.0688	0.0896	**0.0681**
Over-fitted	**0.1219**	**0.1324**	0.1692	**0.0621**	**0.0825**	0.0803

Table 6: **Ablation performance on training strategy, loss setting, and Gram-LM settings.**

Model Setting	T12 WER↓	T15 WER↓
BEST: End-to-End training	0.3415	0.2071
BEST: End-to-End + Multi-task Loss	0.2253	0.1067
BEST: End-to-End + Multi-task + 5-Gram	0.2060	0.0895
BEST: Two-Stage training	0.1504	0.0681
BEST: Two-Stage + 5-Gram LM	**0.0721**	**0.0257**
BEST: Two-Stage + 4-Gram KenLM	0.0764	0.0292
Transformer: Two-Stage + 5-Gram	0.0852	0.0381
GRU: Two-Stage + 5-Gram	0.1105	0.0541

Training strategy and LM-controlled analysis. Two-stage training consistently outperforms end-to-end optimization on both T12 and T15 (Table 6), while multi-task supervision only partially closes the gap. This suggests that jointly learning neural representations and sequence generation is prone

Table 7: **LM-controlled comparison on Brain-to-Text '25. All neural encoders are evaluated under identical language-model conditions. Δ denotes the absolute WER reduction relative to the corresponding baseline LM setting.**

Method	WER ↓				Δ from No LM		Δ from Mismatched 5-gram	
	No LM	Mismatched 5-gram	Current 4-gram	Current 5-gram	→ 4-gram	→ 5-gram	→ 4-gram	→ 5-gram
BEST	**0.0681**	**0.0624**	**0.0292**	**0.0257**	0.0389	0.0424	0.0332	0.0367
Transformer	0.1052	0.0875	0.0409	0.0381	**0.0643**	**0.0671**	**0.0466**	**0.0494**
GRU	0.1093	0.0926	0.0587	0.0541	0.0506	0.0552	0.0339	0.0385

to overfitting under noisy, data-limited supervision. In contrast, two-stage training first stabilizes phoneme- and character-level representations and then optimizes sentence-level decoding on top of this structured latent space.

Language modeling provides additional sequence-level refinement, but does not explain the advantage of *BEST*. Both 4-gram [19] and 5-gram [48] LMs reduce WER, with the 5-gram LM yielding the best overall performance. To isolate the effect of linguistic rescoring, we further evaluate *BEST*, Transformer, and GRU under identical LM conditions on Brain-to-Text '25 (Table 7). *BEST* achieves the lowest WER in every setting. Therefore, its advantage is already present before strong language-model priors are introduced.

Notably, weaker encoders benefit more from stronger LMs in absolute terms. Relative to no LM, the current 5-gram reduces WER by 0.0424 for *BEST*, compared with 0.0671 for Transformer and 0.0552 for GRU; the same trend holds for the 4-gram LM. Hence, *BEST*'s gain cannot be attributed to stronger LM utilization. Instead, linguistic rescoring partially compensates for encoder errors without changing the model ranking: under the same 5-gram LM, *BEST* achieves 0.0257 WER, compared with 0.0381 for Transformer and 0.0541 for GRU. Overall, these controlled results indicate that language models reduce residual sequence uncertainty, whereas the dominant performance gain arises from the quality of the neural representation and the resulting decoding space.

Representation evidence analysis. We provide representation-level evidence for the staged decoding mechanism of *BEST* (Fig. 3, Appendix B). Self-RSA reveals that character, phoneme, and fusion embeddings all exhibit clear structure, indicating that neural activity is organized into stable linguistic manifolds rather than diffuse features. Notably, the fusion embedding remains structured but less locally redundant, suggesting a more compact representation after cross-level integration. Alignment analysis shows that phoneme embeddings weakly match the audio embedding space, whereas fusion embeddings achieve substantially stronger alignment, indicating that the fusion module transforms neural features into an acoustic-compatible manifold suitable for Whisper decoding. This alignment further shifts toward the text embedding space after the Whisper encoder, forming a coherent progression from neural to audio-aligned to text-aligned representations.

Temporal activation patterns support this view: *BEST* produces smoother trajectories with shorter effective support, reflecting stronger temporal consistency and reduced redundancy. In contrast, GRU and vanilla Transformer exhibit fragmented activations, indicating weaker temporal organization. Overall, these results reveal a concrete mechanism: *BEST* performs decoding via progressive cross-modal alignment. Rather than committing early to discrete phoneme predictions, it first constructs an acoustic-compatible latent space and then maps it to text, providing evidence for a neural-to-audio-to-text decoding pathway.

Cross-day generalization analysis: We evaluate cross-day generalization by training on the first 15 or 18 days and testing on all subsequent days (Figure 4). *BEST* consistently achieves the lowest WER and degrades substantially more slowly than both the overfitted *BEST* and the Transformer baseline as

Table 8: **All end-to-end comparable frameworks are reported using inference on Jetson Orin Nano (8GiB) and training on NVIDIA RTX 4090 GPU. Epoch time is reported as mean $\pm$ standard deviation (T12, N=1200). GFLOPS stands for floating-point operations per second.**

Model	GFLOPS	Inference times (ms)	Peak Memory (GiB)	Training Epoch time (s)
BIT [58]	1736.26	1189.71±26.81	7.23±0.12	147.81±1.09
BEST (Ours)	**123.17**	**215.67±19.81**	**2.79±0.06**	**56.34±0.81**

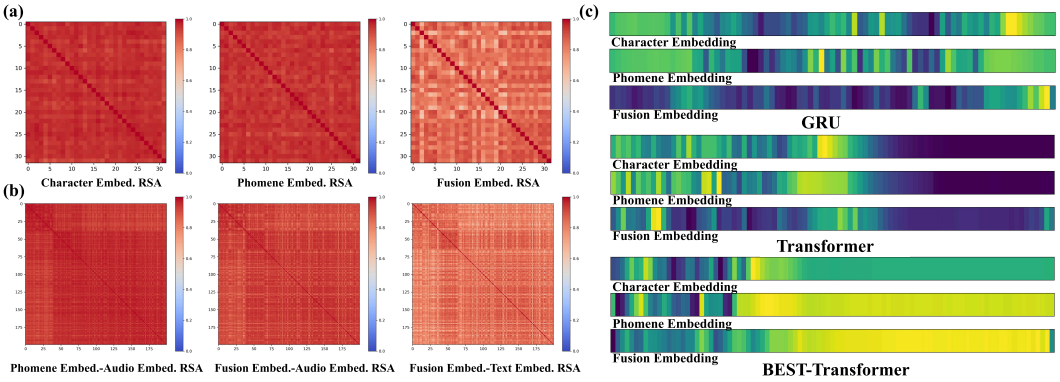


Figure 3: **(a) Self-RSA on character, phoneme, fusion embedding. (b) Cross-RSA between Phoneme Embedding-Audio Embedding, Phoneme Embedding-Audio Embedding, and Fusion Embedding-Text Embedding. (c) Temporal activation patterns of GRU, transformer, and proposed *BEST*.**

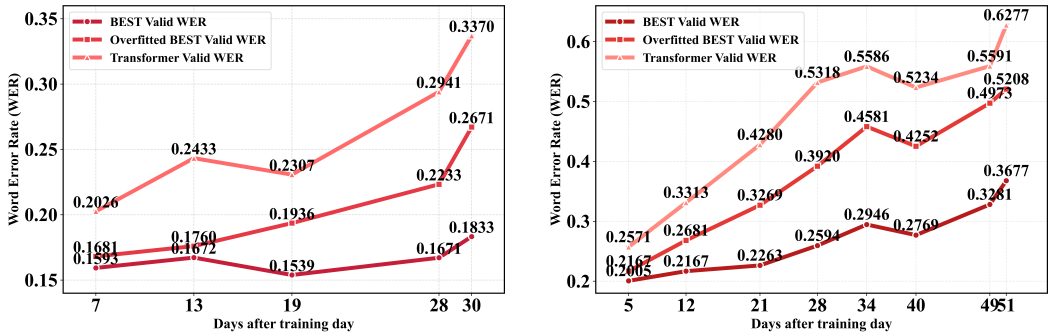


Figure 4: **Cross-day generalization under temporal shift. Models are trained on first 15 days (left) and 18 days (right), and evaluated on subsequent days.**

the temporal gap increases. This gap becomes more pronounced in the longer-horizon setting, where both baselines exhibit clear performance drift. These results indicate that the advantage of *BEST* lies in robustness to longitudinal distribution shift rather than improved in-distribution fitting. The overfitted model highlights a key failure mode: stronger fitting leads to faster degradation, revealing sensitivity to day-specific correlations. The Transformer degrades more severely, suggesting that generic sequence modeling is insufficient under cross-session non-stationarity. In contrast, the slower degradation of *BEST* implies that it learns more stable neural-linguistic representations that remain effective across sessions.

Deployment analysis of end-to-end inference: We evaluate deployment efficiency on Jetson Orin Nano (Table 8). *BEST* reduces computation from 1736.26 to 123.17 GFLOPS, lowers inference latency from 1189.71 ms to 215.67 ms, and cuts peak memory from 7.23 GiB to 2.79 GiB, with training time reduced by over a factor of two. These improvements are substantial, placing the model within the practical constraints of edge deployment. End-to-end decoding is desirable in speech BCI as it enables direct neural-to-text inference without external pipelines, reducing system complexity and latency. However, existing end-to-end frameworks incur prohibitive computational cost. *BEST* retains the advantages of end-to-end decoding while significantly improving efficiency, indicating that the overhead in prior models stems from inefficient representation processing rather than necessary computation. By constructing structured representations before decoding, *BEST* reduces optimization and inference costs, achieving a comparable accuracy-efficiency trade-off for deployment.

5 Discussion

Rethinking scaling in speech BCI: A central insight of this work is that model scaling does not directly translate to performance gains in speech BCI. Unlike conventional NLP or ASR, speech BCI

operates in a low-resource, high-noise regime where neural observations are weakly aligned with linguistic outputs. Under such conditions, increasing model capacity amplifies overfitting rather than improving generalization. Our results show a clear non-monotonic scaling behavior, where compact, modality-aligned decoders outperform larger text-based LLMs. This suggests that the dominant factor in speech BCI is not model size, but the alignment between representation and neural space.

Representation, not prediction, is the bottleneck: We identify the discrete phoneme interface as a key limitation in existing pipelines. Collapsing neural activity into hard phoneme labels discards structured information and leads to brittle decoding. *BEST* avoids this bottleneck by preserving continuous, multi-level representations and performing feature-level integration prior to decoding. Both quantitative and representation-level analyses show that this design improves temporal consistency, reduces representation collapse, and enables more stable sequence reconstruction. This highlights a broader principle: in speech BCI, decoding quality is governed by representation structure rather than intermediate prediction accuracy.

Implications for practical deployment: From a systems perspective, *BEST* demonstrates that efficient neural decoding does not require large-scale generative models. By replacing LLM-based decoding with a compact speech-aligned translator, *BEST* significantly reduces computation, latency, and memory footprint while maintaining strong performance. This is critical for clinical deployment, where real-time inference and on-device constraints are fundamental requirements rather than optimization targets.

Limitations and future directions: Despite these advances, several limitations remain. First, our evaluation is conducted on single-subject datasets, and cross-subject generalization remains an open challenge due to variability in neural representations. Second, while *BEST* improves robustness to cross-day distribution shift, long-term non-stationarity and adaptive decoding are not explicitly addressed. Third, our framework relies on pretrained speech models (e.g., Whisper), whose inductive biases may not fully capture the neural-to-linguistic mapping. Future work may explore jointly learned neural-speech representations or fully neural-native decoding architectures. Finally, while we show that large models are not necessary for current speech BCI settings, this conclusion may evolve with larger-scale neural datasets. Understanding how representation learning and model scaling interact under improved data regimes remains an important direction.

6 Conclusion

We present *BEST*, a lightweight and practical framework for speech BCI that replaces discrete phoneme-based pipelines with hierarchical feature-level decoding. *BEST* achieves strong performance with a fraction of the computational cost, while providing improved robustness to overfitting and cross-day distribution shift. Our findings challenge the prevailing assumption that larger models are inherently beneficial for speech BCI. Instead, we show that performance is governed by representation alignment, structured decoding, and controlled capacity. These results suggest a shift in focus from model scaling to representation design, offering a more efficient and robust pathway toward real-world neural speech decoding.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China under Grant No. 72471140, partly by Shanghai Jiao Tong University Key Projects of Smart Liberal Arts under Grant No. ZHWK2505, partly by Fundamental Research Funds for the Central Universities under Grant No. YG2025QNA13, partly by Shanghai Magnolia Talent Program-Pujiang Project under Grant No. 2025PJC167, partly by the Lushan Research Funding.

Conflict of Interest

The authors declare that they have no competing interests or conflicts of interest related to this work.

References

[1] Miguel Angrick, Shiyu Luo, Qinwan Rabbani, Shreya Joshi, Daniel N Candrea, Griffin W Milsap, Chad R Gordon, Kathryn Rosenblatt, Lora Clawson, Nicholas Maragakis, Francesco V Tenore, Matthew S Fifer, Nick F Ramsey, and Nathan E Crone. Real-time detection of spoken speech from unlabeled ecog signals: a pilot study with an als participant. *Journal of Neural Engineering*, 22(5):056023, 2025.

[2] Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, Mar 2013.

[3] Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. *The New England Journal of Medicine*, 391(7):609–618, 2024.

[4] Andrea Caria. Integrating large language models and brain decoding for augmented human-computer interaction: A prototype llm-p3-bci speller. In *Advances in Information and Communication*, pages 403–416, 2025.

[5] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech Language*, 13(4):359–394, 1999.

[6] Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and XINLIANG ZHOU. Uni-ntfm: A unified foundation model for eeg signal representation learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 109178–109210, 2026.

[7] Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, Nov 2025.

[8] Yi Ding, Muyun Jiang, Weibang Jiang, Shuailei Zhang, Xinliang Zhou, Chenyu Liu, Shanglin Li, Yong Li, and Cuntai Guan. Brainpro: Towards large-scale brain state-aware eeg representation learning. *arXiv preprint arXiv: 2509.22050*, 2025.

[9] Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna V Shenoy, Jaimie Henderson, and Francis Willett. Plug-and-play stability for intracortical brain-computer interfaces: A one-year demonstration of seamless brain-to-text communication. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42258–42270, 2023.

[10] Mengchen Fan, Baocheng Geng, Xi Xiao, Tianyang Wang, Siyuan Mei, Pulin Che, Xiaoqian Jiang, and Qizhen Lan. Detail consistent stage-wise distillation for efficient 3d mri segmentation. In Guang Yang, Ehsan Adeli, Marleen de Bruijne, Bartłomiej W. Papież, Stefanie Speidel, Pallavi Tiwari, Guoyan Zheng, Mohammad Yaqub, Qi Dou, and Islem Rekik, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, pages 123–133, Cham, 2027. Springer Nature Switzerland.

[11] Ebrahim Feghhi, Shreyas Kaasyap, Nima Hadidi, and Jonathan C. Kao. Time-masked transformers with lightweight test-time adaptation for neural speech decoding, 2025.

[12] Sheng Feng, Heyang Liu, Yu Wang, and Yanfeng Wang. Towards an end-to-end framework for invasive brain signal decoding with large language models. In *Interspeech 2024*, page 1495–1499, 2024.

[13] Zehui Feng, Yangge Yang, Chenqi Zhang, Junxuan Li, and Ting Han. A novel hierarchical attention-guided refinement method with eeg assistance for enhancing target speech in a multi-speaker competing environment. *Advanced Engineering Informatics*, 65:103363, 2025. ISSN 1474-0346.

[14] Zehui Feng, Chenqi Zhang, Mingru Wang, Minuo Wei, Shiwei Cheng, Cuntai Guan, and Ting Han. Unveiling deep semantic uncertainty perception for language-anchored multi-modal vision-brain alignment, 2025.

[15] Zehui Feng, Tongtong Zhou, and Ting Han. Mlst-net: Multi-task learning based spatial-temporal disentanglement scheme for video facial paralysis severity grading. *IEEE Journal of Biomedical and Health Informatics*, 29(8):5675–5686, 2025.

[16] Ismail Hossain and Mridul Banik. Hybrid eeg–driven brain–computer interface: A large language model framework for personalized language rehabilitation, 2025.

[17] Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv: 2504.19596*, 2026.

[18] Jing Jin, Yutao Zhang, Ruitian Xu, and Yixin Chen. An innovative brain-computer interface interaction system based on the large language model, 2025.

[19] KenLM. A custom 4-gram kenlm (ken language model). *https://www.kaggle.com/datasets/heyyousum/custom-4-gram-wiki-news-switchboard-updated-v3*.

[20] Erin M. Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald T. Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel R. Nason-Tomaszewski, Nicholas S. Card, Justin J. Jude, Brandon G. Jacques, Payton H. Bechefsky, Carrina Iacobacci, Leigh R. Hochberg, Daniel B. Rubin, Ziv M. Williams, David M. Brandman, Sergey D. Stavisky, Nicholas AuYong, Chethan Pandarinath, Shaul Druckmann, Jaimie M. Henderson, and Francis R. Willett. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188(17), 2025.

[21] Qizhen Lan, Mengchen Fan, Hang Zhang, Jingwei Duan, Moule Lin, Jialin Chen, Baocheng Geng, and Xiaoqian Jiang. Braintrace: Tracing longitudinal, multimodal, and volumetric evidence in brain mri clinical reasoning. *arXiv preprint: 2610.06571*, 2026.

[22] Bo Li, Chen Change Loy, Fanyi Pu, Jingkang Yang, Kaichen Zhang, Kairui Hu, Minh Thang Luu, Quang Trung Nguyen, Ba Cong Pham, Shuai Liu, Yezhen Wang, and Ziwei Liu. Aero: Audio-enhanced large language models. *https://www.lmms-lab.com/posts/aero$_a$udio/*, 2025.

[23] Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, and Eli Shlizerman. Brain-to-text decoding with context-aware neural representations and large language models, 2024.

[24] Mingyuan Li, Guangsheng Yu, Xu Wang, and Shaoxiong Ji. Frozen memory is not enough: Rethinking external memory as extraction. *arXiv preprint arXiv:2608.17050*, 2026.

[25] Mingyuan Li, Guangsheng Yu, Juyuan Zhang, Xu Wang, Zhibo Man, Haonan Zhang, and Shaoxiong Ji. Memoryathena: Adaptive routing over latent and generated memories. *arXiv preprint arXiv:2609.25853*, 2026.

[26] Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. V-cast: Video curvature-aware spatio-temporal pruning for efficient video large language models. *arXiv preprint arXiv:2603.27650*, 2026.

[27] Zhanhui Lin, Xinyu Jiang, Chenyun Dai, and Fumin Jia. Towards real time efficient and robust ecog decoding for mobile brain–computer interface. *Journal of Neural Engineering*, 22(4), 2025.

[28] Kaylo T. Littlejohn, Cheol Jun Cho, Jessie R. Liu, Alexander B. Silva, Bohan Yu, Vanessa R. Anderson, Cady M. Kurtz-Miott, Samantha Brosler, Anshul P. Kashyap, Irina P. Hallinan, Adit Shah, Adelyn Tu-Chan, Karunesh Ganguly, David A. Moses, Edward F. Chang, and Gopala K. Anumanchipalli. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature Neuroscience*, 28(4):902–912, Apr 2025.

[29] Xin Liu, Ziyue Zhang, and Jingxin Nie. Talking to the brain: Using large language models as proxies to model brain semantic representation, 2025.

[30] Yejia Liu, Shijin Duan, Xiaolin Xu, and Shaolei Ren. Scheduled knowledge acquisition on lightweight vector symbolic architectures for brain-computer interfaces, 2024.

[31] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv: 2310.06625*, 2024.

[32] Zhengwu Liu, Jie Mei, Jianshi Tang, Minpeng Xu, Bin Gao, Kun Wang, Sanchuang Ding, Qi Liu, Qi Qin, Weize Chen, Yue Xi, Yijun Li, Peng Yao, Han Zhao, Ngai Wong, He Qian, Bo Hong, Tzyy-Ping Jung, Dong Ming, and Huaqiang Wu. A memristor-based adaptive neuromorphic decoder for brain–computer interfaces. *Nature Electronics*, 8(4):362–372, Apr 2025.

[33] Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. Unimind: Unleashing the power of llms for unified multi-task brain decoding, 2025.

[34] Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2Text: Text generation from EEG signal using large language models (LLMs). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3747–3759, 2025.

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

[36] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea

Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[37] Jia Hong Puah, Sim Kuan Goh, Ziwei Zhang, Zixuan Ye, Chow Khuen Chan, Kheng Seang Lim, Si Lei Fong, Kok Sin Woon, and Cuntai Guan. Eegdm: Eeg representation learning via generative diffusion model. *arXiv preprint arXiv: 2508.14086*, 2025.

[38] Youkun Qian, Changjiang Liu, Peixi Yu, Xingchen Ran, Shurui Li, Qinrong Yang, Yang Liu, Lei Xia, Yijie Wang, Jianxuan Qi, Erda Zhou, Junfeng Lu, Yuanning Li, Tiger H. Tao, Zhitao Zhou, and Jinsong Wu. Real-time decoding of full-spectrum chinese using brain-computer interface. *Science Advances*, 11(45):eadz9968, 2025.

[39] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[40] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv: 2212.04356*, 2022.

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC Workshop*, 2019.

[42] Xiaoquan Shen, Jiajun Li, Jiaxuan Chen, and Gang Pan. Cortexvideo: A semantic-spatial dual-anchor framework for high-fidelity fmri-to-video reconstruction. In *European Conference on Computer Vision*, pages 676–693. Springer, 2026.

[43] Aditya Singh, Tessy Thomas, Jinlong Li, Greg Hickok, Xaq Pitkow, and Nitin Tandon. Transfer learning via distributed brain recordings enables reliable speech decoding. *Nature Communications*, 16(1):8749, Oct 2025.

[44] Yuting Tang, Weibang Jiang, Shanglin Li, Yong Li, Chenyu Liu, Xinliang Zhou, Yi Ding, and Cuntai Guan. Eeg-dlite: Dataset distillation for efficient large eeg model training. *arXiv preprint arXiv: 2512.12210*, 2026.

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv: 1706.03762*, 2023.

[47] Maitreyee Wairagkar, Nicholas S. Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M. Miller, Leigh R. Hochberg, David M. Brandman, and Sergey D. Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, Aug 2025.

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

[49] Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, Tianyang Wang, and Hao Xu. Staying vigilant: Mitigating visual laziness via counterfactual visual alignment in mllms. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari, editors, *Computer Vision – ECCV 2026*, pages 287–305, Cham, 2026. Springer Nature Switzerland.

[50] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[51] Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346, Mar 2025.

[52] Xuanhua Yin, Runkai Zhao, Lina Yao, and Weidong Cai. Brainvista: Modeling naturalistic brain dynamics as multimodal next-token prediction. *arXiv preprint: 2602.04512*, 2026.

[53] Lishan Yu, Derek Jiu, Qizhen Lan, and Xiaoqian Jiang. Evotreenad: Genealogy-guided evolution for llm-driven neural architecture discovery. *arXiv preprint: 2609.29016*, 2026.

[54] Hengrui Zhang, Yuhu Cheng, CL Chen, and Xuesong Wang. Diffusion subgoal planning for long-horizon offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2609.34575*, 2026.

[55] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[56] Wenqi Zhang, Yanjun Qin, and Xiaoming Tao. Decoding brain-controlled intention for uavs and ivs based on lightweight network. *IEEE Internet of Things Journal*, 11(12):21112–21124, 2024.

[57] Yanteng Zhang, Songheng Li, Zeyu Shen, Qizhen Lan, Lipei Zhang, Yang Liu, and Vince Calhoun. Lost in distortion: Uncovering the domain gap between computer vision and brain imaging - a study on pretraining for age prediction. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2026.

[58] Yizi Zhang, Linyang He, Chaofei Fan, Tingkai Liu, Han Yu, Trung Le, Jingyuan Li, Scott Linderman, Lea Duncker, Francis R Willett, Nima Mesgarani, and Liam Paninski. A cross-species neural foundation model for end-to-end speech decoding, 2026.

Contents

A. Implementation details 17

1 Data preprocessing 17

2 Training details 17

B. Audio/Text embedding extraction for RSA 17

1 Audio embedding 17

2 Text embedding 17

3 Alignment for RSA 17

C. RNN baseline hyperparameter 18

D. LLM decoder hyperparameter 18

E. Phoneme probability visualization 19

F. Validation CTC Loss and CER analysis 18

G. Hyperparameter scanning on *BEST* 19

1 Encoder dimension 19

2 Phoneme-character fusion strategy 19

3 Temporal masking 21

4 Hierarchical layer allocation 21

5 Cascade encoder Last M Layer 21

6 Frozen decoder first N Layer 21

H. Phoneme and character confusion analysis 22

I. Phoneme and character transition analysis 23

J. Method implementation 24

1 Multi-frequency attention 24

2 Hierarchical transformer 24

K. Ethical considerations 24

L. NeurIPS Paper Checklist 25

Table 9: Hyperparameter settings for T12 and T15 across Stage 1 and Stage 2.

Hyperparameter	**T12 Stage 1**	**T12 Stage 2**	**T15 Stage 1**	**T15 Stage 2**
Learning Rate	1×10^{-3}	1×10^{-5}	1×10^{-2}	1×10^{-4}
Weight Decay	1×10^{-5}	1×10^{-5}	1×10^{-3}	1×10^{-3}
Batch Size	64	32	64	32
Input Dropout	0.2	0.2	0.2	0.2
Dropout	0.35	–	0.35	–
Patch Size	(5, 256)	(5, 256)	(4, 512)	(4, 512)
Beam Size	-	50	-	50
Optimizer	AdamW			
Maximum Sequence	500			
Embedding Dimension	768			
Activation	ReLU			
MLP Dim Ratio	4			
β_1	0.9			
β_2	0.999			

A Implementation details

Data preprocessing. We follow the preprocessing protocols of T12 [48] and T15 [3]. For **T12**, 256 neural features are extracted (two features per electrode, including -4.5 RMS threshold crossings and spike band power across 128 electrodes), with White Noise Standard Deviation$= 0.8$, Gaussian smoothing width$= 2.0$, and Constant Offset Standard Deviation$= 0.2$. For **T15**, 512 neural features are used (two features per electrode across 256 electrodes), with White Noise Standard Deviation$=$ 1.0, Gaussian smoothing width$= 2.0$, and Constant Offset Standard Deviation$= 0.2$.

Training details: Stage 1 is trained from random initialization on a single NVIDIA RTX 4090 (24 GiB). Convergence is typically achieved within 80 epochs (under 1 hour for T12 and under 2 hours for T15). Beyond this point, performance gradually degrades due to overfitting, with negligible improvements after 400 epochs as indicated by PER/CER. **Stage 2** is trained on a single NVIDIA RTX 4090 (24 GiB) for 50 epochs (under 1 hour for T12 and under 2 hours for T15). Overfitting is observed beyond 50 epochs, with no significant improvement in WER. The Whisper model is initialized from pretrained weights, with only the first Transformer decoder layer and the fusion block updated during training. Peak system memory usage remains below 32 GiB. Detailed hyperparameter configurations are provided in Table 9.

B Audio/Text embedding extraction for RSA

Audio embedding: The audio embedding is obtained from the Whisper encoder. We input ground-truth speech audio into the pretrained Whisper encoder and extract the hidden states from the top encoder layer. These representations serve as the reference acoustic embedding space. For consistency with neural-derived features, we use the time-aligned encoder outputs after positional encoding, without additional pooling unless otherwise specified. In experiments involving neural inputs, the fusion embeddings are projected into the same latent space and passed through the upper layers of the Whisper encoder, ensuring that both embeddings lie in a shared acoustic representation.

Text embedding: The text embedding is obtained from the Whisper decoder. Given the text, we first tokenize the text via Whisper's tokenizer and feed it into the decoder with teacher forcing. We then extract the decoder hidden states corresponding to each token from the final decoder layer. These hidden states are used as the text embedding space. No logits or output projections are used; instead, we rely on the internal decoder representations prior to the vocabulary projection layer.

Alignment for RSA: For cross-modal RSA, neural-derived embeddings (phoneme, character, and fusion representations) are compared against audio and text embeddings after temporal alignment. We apply linear interpolation or average pooling to match sequence lengths across modalities. All embeddings are normalized before similarity computation to ensure scale invariance.

Table 10: Hyperparameters of the RNN Baseline Model

Hyperparameter	Value
GRU Layers / Units	5 / 768
GRU Dropout	0.4
Patch Size / Stride	14 / 4
Input Dropout	0.2
Optimizer	Adam (β_1=0.9, β_2=0.999)
Weight Decay	1×10^{-3}
LR Warm-up Steps	1000
Max LR	5×10^{-3}
Min LR	1×10^{-4}

Table 11: Hyperparameter settings for LoRA fine-tuning.

Hyperparameter	Value
Activation	ReLU
Learning Rate	5×10^{-5}
Weight Decay	1×10^{-5}
Batch Size	8
Gradient Accumulation Step	1 (model $<$ 7B) and 8 (model $\geq$ 7B)
LoRA Rank	8
LoRA Scaling Factor	32
LoRA Dropout Ratio	0.2

C RNN baseline hyperparameter

The hyperparameters for the RNN baseline and LoRA-based LLM decoder are summarized in Table 10. For the RNN baseline, we use a regularized training setup with moderate model capacity to ensure stable optimization under limited, noisy neural data, and adopt a warm-up learning rate schedule to enhance early-stage convergence.

D LLM decoder hyperparameter

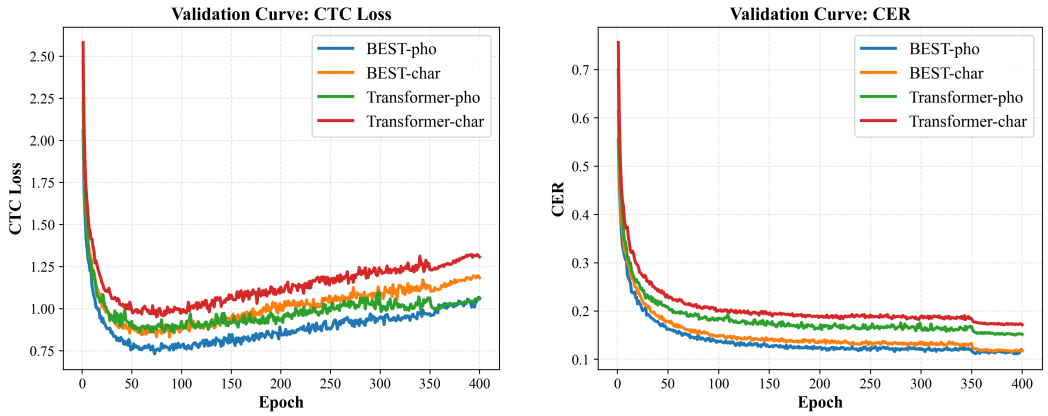
The hyperparameter of LLM-based decoders [45, 50] is summarized in Table 11. we adopt LoRA for parameter-efficient adaptation. The configuration is chosen to provide sufficient adaptation capacity while maintaining training stability and preventing overfitting. Gradient accumulation is adjusted based on model scale to accommodate hardware constraints without affecting optimization dynamics. They follow standard practices and are kept consistent across experiments to ensure fair comparison.

E Validation CTC loss and CER analysis

The validation curves (Figure. 5) show a consistent optimization advantage of *BEST* over the vanilla Transformer in both phoneme- and character-level pretraining. *BEST* reaches lower CTC loss minima and maintains lower CER throughout training, indicating more stable and more effective neural representation learning. The Transformer, by contrast, converges to a higher loss floor and consistently worse CER, reflecting earlier overfitting and weaker generalization.

A second observation is the divergence between validation CTC loss and CER, which holds across all models. After the early training stage, CTC loss rises while CER continues to decrease or remains stable. This is a property of the CTC objective rather than a model-specific effect. CTC loss depends on the full alignment distribution, whereas CER depends only on the final decoded sequence. A worse alignment distribution therefore does not necessarily imply worse decoding. This highlights a limitation of validation CTC loss as a direct proxy for sequence-level quality.

The advantage of *BEST* remains consistent under both phoneme and character supervision, indicating that the gain is not target-specific. It arises from the encoder design, which improves both local discrimination and sequence consistency. These curves therefore provide optimization-level evidence that structured representation learning, rather than increased model capacity, is the critical factor in speech BCI.



(a) CTC Loss curve

(b) CER curve

Figure 5: Validation curves of CTC loss and CER for the Transformer and proposed *BEST* on participant T12.

Table 12: Hyperparameter scanning on the dimension of the Transformer encoder. Note that overfitting becomes increasingly prominent during training as the dimension grows.

dimension	T12 PER↓	T12 CER↓	T15 PER↓	T15 CER↓
384	0.1830	0.1933	0.0971	0.1006
512	0.1628	0.1708	0.0762	0.0937
640	0.1539	0.1627	0.0730	0.0971
768	**0.1321**	**0.1436**	**0.0688**	**0.0896**
896	0.1706	0.1772	0.0893	0.0931
1024	0.2063	0.1960	0.1104	0.1052

F Phoneme probability visualization

Figure 6 compares frame-wise phoneme decoding probabilities under well-fitted and overfitted training regimes. The overfitted model produces sharper but more fragmented activations, indicating higher local confidence with weaker temporal organization. By contrast, the well-fitted model maintains smoother transitions and more coherent phoneme trajectories across time. This visualization exposes the mechanism behind the WER degradation of the overfitted model. Overfitting does not primarily destroy local phoneme separability; it disrupts the temporal structure required for consistent sequence decoding. The result is a representation that appears more confident locally but is less compatible with downstream inference. This provides qualitative support for the central claim of the paper: in speech BCI, sequence-level performance depends more on temporal coherence than on isolated phoneme accuracy.

G Hyperparameter scanning on *BEST* configuration

Hyperparameter scanning on encoder dimension: We observe a clear non-monotonic dependence on encoder dimension (Table 12). Performance improves as the dimension increases from 384 to 768, but degrades beyond this point on both T12 and T15, indicating that larger hidden sizes quickly push the model into an overfitting regime under noisy and data-limited supervision. The optimum at 768 is not incidental: besides providing the best balance between capacity and generalization, it also matches the hidden dimension of the Whisper-based decoder. This dimensional consistency likely reduces projection mismatch between neural representations and the downstream speech embedding space, leading to more effective cross-stage alignment.

Phoneme-Character fusion strategy: Table 13 shows that fusion quality is determined by structural bias rather than parameter count alone. Linear and MLP fusion remain limited, while cross-attention performs worst despite its larger capacity, suggesting that heavy global interaction is both inefficient and unstable in this low-resource setting. By contrast, convolutional fusion is consistently more effective. In particular, the multi-scale design achieves the best performance with moderate parameter

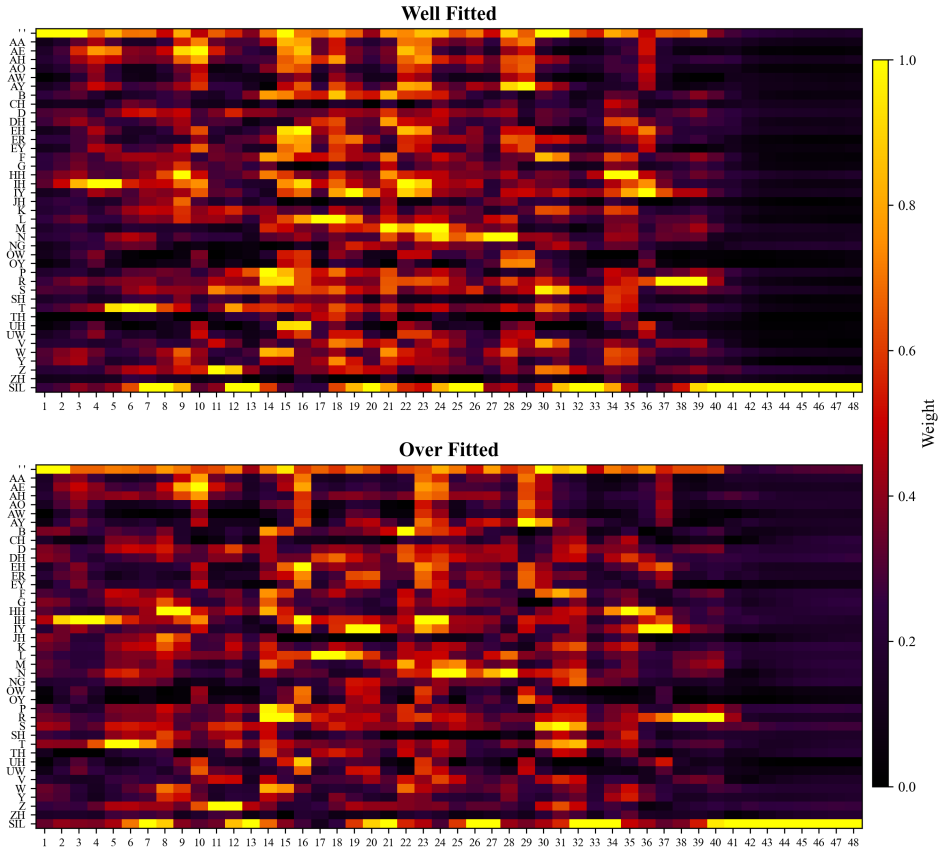


Figure 6: Phoneme decoding probability heatmaps under well-fitted (top) and overfitted (bottom) conditions. Overfitting produces fragmented and overconfident activations, indicating representation collapse despite lower phoneme-level error.

Table 13: Ablation on different fusion strategies for Phoneme-Character Fusion.

Phon.-Char. Fusion	**#Param.**	**T12 WER↓**	**T15 WER↓**
Linear	0.59M	0.3215	0.1329
MLP	2.36M	0.2371	0.1135
Cross Attention	3.54M	0.4962	0.4219
Conv1×1	0.59M	0.1653	0.0765
Conv3×3	1.77M	0.1560	0.0782
Conv1×1, 3×3	2.36M	**0.1504**	**0.0681**

cost, indicating that it provides the right balance between efficiency and expressiveness. The 1×1 branch enables lightweight channel mixing, while the 3×3 branch captures short-range temporal context; their combination yields structured feature integration without the excessive flexibility that would encourage overfitting.

Masked autoencoder ablation: Table 14 reports the impact of introducing a masked autoencoder (MAE) objective during pretraining. Contrary to prior work (e.g., BIT [58]), incorporating masking leads to a consistent degradation across both benchmarks, with PER increasing from 0.1321 to 0.1420 (+7.5%) on T12 and from 0.0688 to 0.0727 (+5.7%) on T15. A similar trend is observed for CER. This result suggests a fundamental mismatch between MAE-style reconstruction and the requirements of speech BCI decoding. Specifically, MAE encourages pointwise signal recovery under heavy corruption, which biases the model toward low-level local structure reconstruction. However, neural speech signals are inherently noisy, temporally diffuse, and only indirectly aligned with acoustic realizations. As a result, emphasizing reconstruction can divert model capacity away from learning invariant, linguistically meaningful representations. In particular, phonetic identity and word-level

Table 14: Ablation on masked autoencoder-based reconstruction for the Brain-to-Text '24 and '25.

Ablation	Brain-to-text '24 T12		Brain-to-text '25 T15	
	PER	CER	PER	CER
BEST	0.1321	0.1436	0.0688	0.0896
w. Masked Auto	0.1420	0.1483	0.0727	0.0915

Table 15: Hyperparameter scanning on the number of time masking and the number of hierarchical layer configuration in phoneme decoding.

Number of time masks	**CTCLoss↓**	**Number of layers**	**CTCLoss↓**
5	0.7238±0.0081	1:2:3	0.7562±0.0048
10	0.6870±0.0064	2:2:2	**0.7293±0.0039**
15	**0.7293±0.0039**	3:2:1	0.7503±0.0040
20	0.7519±0.0071	4:1:1	0.7761±0.0052
25	0.7636±0.0049	2:1:2	0.7438±0.0063

semantics are encoded in distributed temporal-spectral patterns rather than in precisely reconstructable signal segments. Furthermore, masking disrupts the already sparse and information-limited neural inputs, exacerbating ambiguity in the latent space. This is especially detrimental in low-SNR regimes typical of intracranial or non-invasive recordings, where preserving global temporal coherence is more critical than recovering masked fragments. The relatively larger degradation on T12 also indicates that MAE objectives are less robust under more challenging decoding conditions.

Temporal masking: Performance improves when masking is increased from a sparse setting, but degrades once masking becomes too aggressive (Table 15). This pattern suggests that temporal masking is beneficial only when it regularizes local redundancy without destroying the continuity of speech-relevant dynamics. Too little masking fails to suppress over-reliance on local correlations, whereas too much masking removes the temporal evidence required for stable phoneme alignment. The optimum therefore lies in a moderate regime, where masking improves invariance while preserving sequence structure.

Hierarchical layer allocation: As shown in Table 15, the symmetric $2:2:2$ configuration consistently achieves the lowest CTC loss, while increasingly imbalanced allocations lead to noticeable degradation. This suggests that phoneme decoding from neural signals critically depends on a balanced distribution of representational capacity across hierarchical stages. Allocating excessive depth to early layers encourages overfitting to local temporal fluctuations and low-level neural dynamics, limiting the emergence of stable phonetic abstractions. In contrast, concentrating depth in later layers weakens early-stage temporal grounding and delays the progressive aggregation of contextual information. The symmetric design mitigates both issues by enforcing a more uniform transformation pipeline, enabling a gradual transition from fine-grained temporal encoding to higher-level phonetic and contextual representations. This balanced hierarchy is better aligned with the inherently multi-scale and distributed nature of neural speech signals, leading to improved generalization.

Cascade encoder last M layer: Performance is maximized in an intermediate regime (Table 16). Cascading too few encoder layers leads to weak adaptation, as the neural representation remains insufficiently aligned with the downstream decoder space. Cascading too many layers, however, begins to erode the stability of the pretrained structure and introduces unnecessary flexibility, which is harmful under limited supervision. The optimum at $M = 4$ indicates that effective neural decoding requires substantial but controlled representational transformation before sentence-level decoding.

Froze Decoder first N layer: The decoder freezing results are consistent with this interpretation (Table 16). Freezing only a small number of initial decoder layers yields the best performance, while freezing more layers degrades CTC loss. This suggests that the decoder must retain enough fixed pretrained structure to stabilize optimization, but still leave sufficient trainable capacity near the interface for adaptation to neural-derived inputs. Excessive freezing restricts this interface and prevents effective alignment; insufficient freezing would instead expose too much of the decoder to noisy supervision. Together, these results support a broader conclusion of the paper: in speech BCI,

Table 16: Hyperparameter scanning on cascaded last layers in the encoder and frozen last layers in the decoder for neural phoneme decoding.

Encoder Last Layer M	CTCLoss↓	Frozen FirstLayer N	CTCLoss↓
1	0.8614±0.0420	2	**0.7293±0.0039**
2	0.8426±0.0068	4	0.7415±0.0033
3	0.7682±0.0052	8	0.7436±0.0067
4	**0.7293±0.0039**	9	0.7823±0.0060
5	0.7554±0.0106	10	0.8069±0.0092

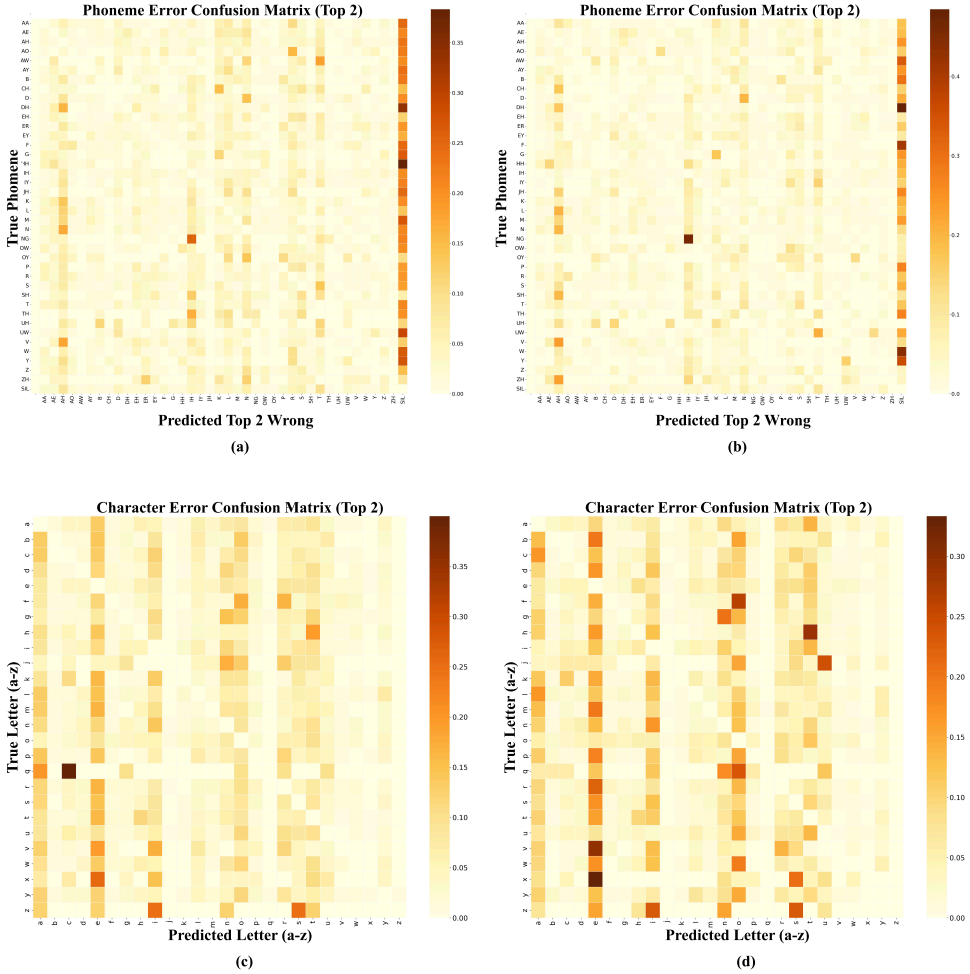


Figure 7: Top-2 phoneme and character confusion matrices on Brain-to-Text '24 and '25 benchmarks.

the best operating point is not obtained by maximizing flexibility, but by carefully controlling where adaptation is allowed and where pretrained structure is preserved.

H Phoneme and character confusion analysis

We examine the top-2 phoneme and character confusion matrices on the Brain-to-Text '24 and '25 benchmarks (Figure 7). The patterns reveal strong structural regularity: phoneme confusions cluster within articulatorily and acoustically similar groups, while character confusions, though broader due to phoneme-to-grapheme mapping, remain highly organized. These results indicate that neural decoding preserves linguistic structure rather than producing arbitrary errors.

The consistency of confusion patterns across both datasets demonstrates that these error structures are intrinsic to neural representations rather than dataset-specific artifacts. Notably, the dominant phoneme confusions align with natural phonotactic constraints, and character confusions reflect

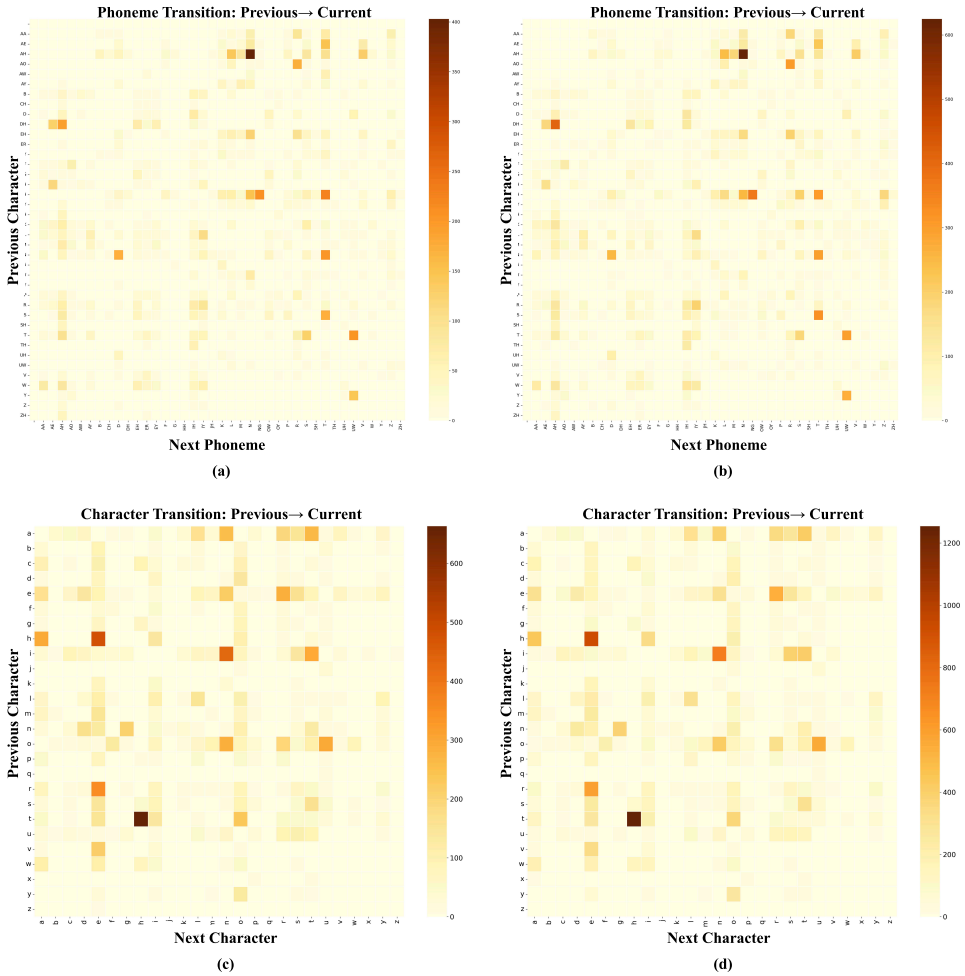


Figure 8: Phoneme and character transition matrices (previous → current) on Brain-to-Text '24 and '25 benchmarks. The matrices capture empirical transition patterns in neural decoding at both phonological and orthographic levels.

coherent orthographic patterns, highlighting a systematic propagation of uncertainty from phoneme to text space. Critically, these observations expose a fundamental limitation of discrete phoneme-based decoding. While local predictions are largely plausible, the model accumulates errors across sequences because early commitment to discrete units discards continuous evidence embedded in neural representations. This provides mechanistic support for *BEST*'s design: delaying discrete decisions and integrating multi-level representations preserves structured neural-linguistic information, enabling more reliable sequence-level decoding.

I Phoneme and character transition analysis

We analyze phoneme and character transition statistics (previous → current) on Brain-to-Text '24 and '25 (Figure 8 and observe highly structured and sparse patterns. Transitions are concentrated on a small subset of dominant pairs, indicating that neural decoding preserves strong local sequential dependencies rather than producing independent predictions.

At the phoneme level, transitions are largely confined within phonotactically plausible neighborhoods, reflecting preserved articulatory continuity. At the character level, the patterns are more diffuse due to phoneme-to-grapheme ambiguity, yet remain strongly structured. Notably, the transition distributions are highly consistent across both datasets, despite differences in data composition and recording conditions, suggesting that these patterns reflect intrinsic properties of neural speech representations rather than dataset-specific effects.

These results reveal a fundamental limitation: neural decoding captures local continuity but lacks sufficient capacity to resolve ambiguity over longer contexts. Consequently, while local transitions are largely correct, errors accumulate at the sequence level. This further supports the design of *BEST*, which mitigates this limitation by delaying discrete decisions and leveraging richer contextual representations for decoding.

J Method implementation

Multi-Frequency Attention: We introduce a lightweight multi-frequency channel attention module to capture informative temporal-spectral patterns in neural signals. Given input $x \in \mathbb{R}^{B\times C\times T}$, we first align the temporal dimension to a fixed length via adaptive average pooling. The signal is then projected onto a set of fixed discrete cosine transform (DCT) bases to extract frequency-specific responses. Specifically, we employ **16 fixed low-frequency DCT branches** corresponding to indices **1, 2, 4, 8, 16, 32**, which are selected by the top-K strategy. For each frequency branch, we compute three complementary channel-wise statistics: mean, max, and min. We aggregate these statistics and feed them into a shared bottleneck MLP to produce attention weights.The fixed DCT bases are stored as non-trainable buffers, and the input is reweighted by the sigmoid-activated attention to form final output.This design imposes negligible overhead while boosting robustness to noise and temporal distribution shifts in neural signals.

Pseudo-code:

Input: $x \in \mathbb{R}^{B\times C\times T}$
$x_{\text{pool}} = \text{AdaptiveAvgPool}(x, L)$
$\text{feat} = 0$
for each DCT basis f_i :
  $s = x_{\text{pool}} * f_i$
  $\text{feat} += \text{Mean}(s) + \text{Max}(s) + \text{Min}(s)$
$\text{feat} = \text{feat}/K$
$\text{attn} = \text{Sigmoid}(\text{MLP}(\text{feat}))$
Output: $x * \text{attn.unsqueeze}(-1)$

Hierarchical Transformer: To capture temporal dependencies at multiple scales, we employ a hierarchical Transformer with progressive masking. A shallow initial Transformer layer models global context, followed by stacked Transformer blocks with distinct temporal constraints, including causal (past), local-window, and forward (future) attention. This staged design separates long- and short-range temporal dependencies, improving stability and robustness under noisy neural inputs.

Pseudo-code:

Input: x
$h = \text{iTransformer}(x)$
$h = \text{Transformer1}(h, \text{mask=past_mask})$
$h = \text{Transformer2}(h, \text{mask=local_mask})$
$h = \text{Transformer3}(h, \text{mask=future_mask})$
Output: $\text{Projection}(h)$

K Ethical considerations

This work presents a lightweight speech brain-computer interface (BCI) for neural-to-text decoding, with careful consideration of privacy, ethics, and responsible deployment. All datasets (Brain-to-Text '24/'25) are publicly available and collected under IRB approval with informed consent. Data are fully anonymized, and no new human experiments are conducted. Speech BCIs raise risks of neural privacy leakage and unintended cognitive inference. The proposed BEST framework mitigates these risks via a compact architecture that avoids reliance on large third-party models, reducing external data transmission. It is explicitly designed to decode only speech-related neural features, adhering to minimal and task-specific information use.

Clinically, this work targets communication restoration for patients with severe motor impairments. Its lightweight design enables edge deployment, promoting accessibility. However, further clinical validation, safety standardization, and long-term evaluation are required before real-world use. We further advocate efficient model design over unnecessary scaling, reducing computational cost and

environmental impact. Limitations, including single-subject evaluation and limited cross-subject generalization, are clearly acknowledged to avoid overstatement. Future work will follow medical ethics standards and incorporate secure data governance (e.g., encryption and access control), alongside interdisciplinary collaboration to ensure BCI deployment aligns with privacy, and human dignity.

NeurIPS Paper Checklist

1. **Claims Answer: Yes.**
   The claims in the abstract and introduction accurately reflect the contributions and experimental results presented in the paper. All stated assumptions and limitations are consistent with the scope of the work.
2. **Limitations Answer: Yes.**
   The paper includes a discussion of limitations, including assumptions, dataset scope, and potential failure cases, and analyzes their impact on generalization.
3. **Theory, Assumptions and Proofs Answer: N/A.**
   The paper does not include theoretical results or formal proofs.
4. **Experimental Result Reproducibility Answer: Yes.**
   We provide sufficient details of the model architecture, training procedure, and implementation to enable reproducibility of the results.
5. **Open Access to Data and Code Answer: Yes.**
   We provide code, data processing scripts, and instructions necessary to reproduce the main experimental results.
6. **Experimental Setting / Details Answer: Yes.**
   The paper specifies all key experimental details, including dataset splits, hyperparameters, and training configurations.
7. **Experiment Statistical Significance Answer: Yes.**
   We report results averaged over multiple runs with appropriate measures of variability (e.g., standard deviation).
8. **Experiments Compute Resource Answer: Yes.**
   The paper reports the computational resources used, including hardware specifications and training time.
9. **Code of Ethics Answer: Yes.**
   We confirm that the research complies with the NeurIPS Code of Ethics.
10. **Broader Impacts Answer: Yes.**
    The paper discusses potential broader impacts, including both beneficial applications and possible negative societal implications.
11. **Safeguards Answer: N/A.**
    The proposed method does not involve high-risk models requiring specific safeguards for deployment.
12. **Licenses Answer: Yes.**
    We properly cite all datasets, codebases, and pretrained models used in this work and respect their respective licenses.
13. **Assets Answer: Yes.**
    We release new assets associated with this work and provide documentation describing their usage, training details, and limitations.
14. **Crowdsourcing and Research with Human Subjects Answer: Yes.**
    We include details of instructions provided to participants and relevant experimental setup information. Compensation details are also described where applicable.
15. **IRB Approvals Answer: Yes.**
    The study involving human subjects was reviewed and approved by the appropriate institutional review board (IRB) or equivalent ethics committee.
16. **Declaration of LLM Usage Answer: Yes.**
    Large Language Models were used only for writing assistance and did not influence the core methodology or experimental results.