Neuro-Composer: Bridging Intention and Articulation via Dual-Stream Predictive Coding and Metacognitive Feedback

Liying Zhang[1,2], **Qi Zhu(zhuqinuaa@163.com)**[1,2,*], **Peiliang Gong**[1,2], **Chuhang Zheng**[3], **Kun Wang**[1,2] **&**
Daoqiang Zhang(dqzhang@nuaa.edu.cn)[1,2,*]

[1]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing, China
[3]The International Joint Institute of Tianjin University, Tianjin University, Fuzhou, China

Abstract

Direct speech synthesis from neural activity offers a transformative lifeline for individuals with severe paralysis, yet decoding high-dimensional, noisy intracranial signals remains a formidable challenge. Current approaches often rely on passive regression, overlooking the hierarchical and generative nature of human speech production. To bridge this gap, we propose ***Neuro-Composer***, a biologically grounded framework that reframes decoding as the dynamic integration of top-down semantic planning and bottom-up acoustic execution. Grounded in the Dual-Stream Theory, our architecture decomposes decoding into a ventral stream, which extracts abstract intent anchored by multimodal foundation models (BERT/Wav2Vec 2.0), and a dorsal stream that captures fine-grained articulatory dynamics. These streams are synthesized via a latency-aware Predictive Refiner. Crucially, we introduce a bio-mimetic auditory feedback loop that enforces semantic consistency, endowing the system with self-monitoring capabilities. Experiments demonstrate that Neuro-Composer significantly outperforms baselines in reconstruction fidelity. Furthermore, qualitative analyses reveal emergent semantic manifolds and brain-like gating patterns, providing computational validation for neurocognitive theories of language production.

Keywords: Speech neuroprosthesis; Deep learning; sEEG;

Introduction

Spoken language is fundamental to human social interaction. Neurological disorders such as stroke or neurodegenerative diseases can severely impair speech production, imposing profound psychological and functional burdens on affected individuals and their caregivers (Rousseau et al., 2015). Brain–computer interfaces (BCIs) that decode speech directly from neural activity offer a promising avenue for restoring communication (Silva et al., 2024; Zhang et al., 2024). In particular, direct speech reconstruction from stereoelectroencephalography (sEEG) signals has gained increasing attention, as it bypasses damaged peripheral motor pathways and translates cortical activity directly into linguistic representations (Anumanchipalli et al., 2019; Herff et al., 2019).

Early studies in neural speech decoding demonstrated the feasibility of speech reconstruction using linear mappings between high-gamma cortical activity and acoustic representations (Pasley et al., 2012). As the field progressed, increasingly expressive deep learning models, including convolutional and recurrent neural networks, were introduced to capture the nonlinear temporal and spectral dynamics of speech

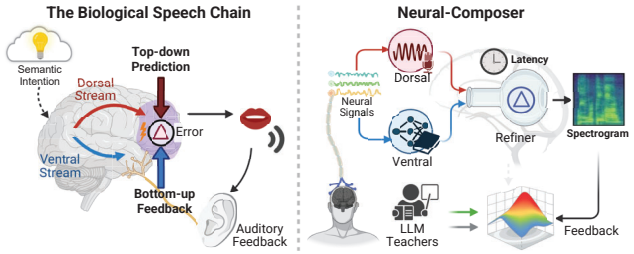


Figure 1: Conceptual overview. **(Left)** Biological speech production modeled as active inference, integrating top-down semantic (ventral) and bottom-up articulatory (dorsal) streams via predictive coding.**(Right)** The bio-inspired Neuro-Composer, which utilizes this dual-stream organization for neural speech reconstruction via predictive refinement and closed-loop feedback.

production (Berezutskaya et al., 2023; Herff et al., 2015; Petrosyan et al., 2022). More recently, large-scale self-supervised foundation models have further advanced decoding performance by aligning neural embeddings with discrete latent representations learned from massive speech corpora, such as Wav2Vec 2.0 and HuBERT (Défossez et al., 2023; Metzger et al., 2023). By framing decoding as a translation from neural activity to pretrained acoustic codebooks, these approaches exploit rich phonotactic and acoustic priors, achieving unprecedented reconstruction fidelity (Stavisky, 2025).

However, this engineering-driven paradigm contrasts with the biological reality of speech production. Cognitive neuroscience characterizes speech not as a passive signal transmission, but as an active, generative process governed by a hierarchical dual-stream architecture (Hickok & Poeppel, 2007; Saur et al., 2008). As illustrated in the left panel of Figure 1, this process emerges from the dynamic interaction between a ventral stream, which constructs high-level semantic and linguistic plans (the "what"), and a dorsal stream, which supports fine-grained articulatory execution (the "how") (Pickering & Garrod, 2013). Within this framework, predictive coding plays a central role: the brain continuously generates top-down predictions about intended speech content to minimize mismatches against noisy neural evidence (Friston, 2010), while auditory feedback further stabilizes this process through self-monitoring and error correction (Indefrey & Levelt, 2004). Despite these insights, most existing decoding frameworks

*Corresponding authors.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

treat speech reconstruction as a predominantly passive engineering challenge. Current models typically adopt a feedforward regression paradigm, viewing the brain as a static signal source and the decoder as a function approximator that maps neural activity directly to acoustic outputs. By neglecting the hierarchical separation between planning and execution, as well as the stabilizing role of feedback-driven error correction, these "black-box" approaches are prone to conflating signal and noise. Consequently, while they may achieve high quantitative scores, they offer limited mechanistic insight into how intelligible speech emerges from noisy cortical dynamics.

To bridge this gap between prevailing computational decoding paradigms and neurocognitive theories of speech production, we propose *Neuro-Composer*, a biologically grounded neural speech decoding framework. As shown in the right panel of Figure 1, Neuro-Composer reframes speech reconstruction as a hierarchical generative process. The architecture consists of two complementary pathways: a ventral semantic stream that constructs high-level linguistic priors and a dorsal acoustic stream that preserves fine-grained temporal and articulatory dynamics. These streams are integrated through a Predictive Refiner that enables iterative, precision-weighted alignment between top-down expectations and bottom-up neural evidence. In addition, Neuro-Composer incorporates an explicit auditory self-monitoring loop, emulating intrinsic feedback mechanisms to constrain reconstructed speech toward semantic consistency with decoded intent.

In summary, this work makes the following contributions:

- We develop Neuro-Composer, a bio-inspired framework that reframes neural speech reconstruction as a hierarchical generative process by integ rating top-down semantic planning with bottom-up articulatory execution.
- We propose a predictive refiner module incorporating a precision-weighted gating mechanism to model physiological latencies and synchronize neural evidence with linguistic priors across representational levels.
- Extensive experiments on public sEEG dataset demonstrate the effectiveness of Neuro-Composer. Moreover, multi-level probing of internal representations substantiates the model's mechanistic interpretability across hierarchical generative and semantic dimensions.

Methodology

This section introduces the proposed *Neuro-Composer*, a biologically grounded neural speech decoding framework. The overall architecture is illustrated in Fig. 2.

Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote the intracranial neural recordings, where C is the number of electrode channels and T is the temporal length. The objective of neural speech decoding is to reconstruct the corresponding speech representation $\mathbf{Y} \in \mathbb{R}^{F\times T'}$, where F denotes the number of Mel-frequency bands and T' is the temporal resolution of the acoustic representation. In practice, T and T' may differ due to temporal downsampling and representation mismatch between neural and acoustic domains.

Rather than directly modeling the conditional distribution $P(\mathbf{Y} \mid \mathbf{X})$ via feedforward regression, we formulate speech decoding as a hierarchical generative inference problem. We assume that speech production arises from the interaction between high-level semantic planning and low-level articulatory execution, mediated by predictive error correction and auditory self-monitoring. Under this formulation, decoding corresponds to inferring latent semantic intentions and refining them using temporally precise neural evidence.

Dual-Stream Neural Encoder

Neuro-Composer employs a dual-stream EEG encoder to disentangle intention-level representations from execution-level neural dynamics. Unlike conventional parallel feature extractors, the two streams are assigned asymmetric inferential roles within a hierarchical generative process. The ventral stream constrains *what* can be said by defining a semantic hypothesis space, whereas the dorsal stream informs *how* speech unfolds over time by providing local neural evidence.

Ventral Semantic Stream The ventral stream extracts slow-varying, intention-level representations associated with semantic planning. To capture these dynamics, the ventral encoder $\Phi_{\mathcal{V}}$ is implemented as a series of temporal-spatial convolution layers followed by a Transformer-based bottleneck, designed to aggregate information over extended windows.

Formally, $\Phi_{\mathcal{V}}$ processes the neural input $\mathbf{X}$ into a high-level feature manifold. This manifold is then bifurcated into two distinct representations:

$$\mathbf{z}_{sem}, \mathbf{M}_{coarse} = \Phi_{\mathcal{V}}(\mathbf{X}) \tag{1}$$

Specifically, the global semantic variable $\mathbf{z}_{sem} \in \mathbb{R}^D$ is derived via global average pooling across the temporal dimension, representing articulatory-invariant intent. Simultaneously, the coarse spectro-temporal prior $\mathbf{M}_{coarse} \in \mathbb{R}^{F\times T'}$ is synthesized through a linear projection layer (or a transposed convolution) that maps the latent features back to a reduced-resolution acoustic space.

This coarse prior $\mathbf{M}_{coarse}$ serves as a top-down structural template of the target spectrogram, capturing fundamental spectral peaks while abstracting away fine-grained temporal fluctuations. To facilitate multimodal alignment, $\mathbf{z}_{sem}$ is mapped to a joint embedding space $\mathbf{z}_{clip}$ via a lightweight MLP projection head.

Dorsal Acoustic Stream The dorsal stream captures fast-varying neural dynamics associated with phonetic execution. To preserve fine-grained temporal resolution, the dorsal encoder $\Phi_{\mathcal{D}}$ utilizes a series of dilated temporal convolutions that expand the receptive field without sacrificing the sampling rate of articulatory transients. Formally, $\Phi_{\mathcal{D}}$ maps the sEEG input $\mathbf{X}$ into a high-resolution feature sequence:

$$\mathbf{Z}_{aco} = \mathrm{PosEnc}\big(\Phi_{\mathcal{D}}(\mathbf{X})\big) \in \mathbb{R}^{T\times D} \tag{2}$$

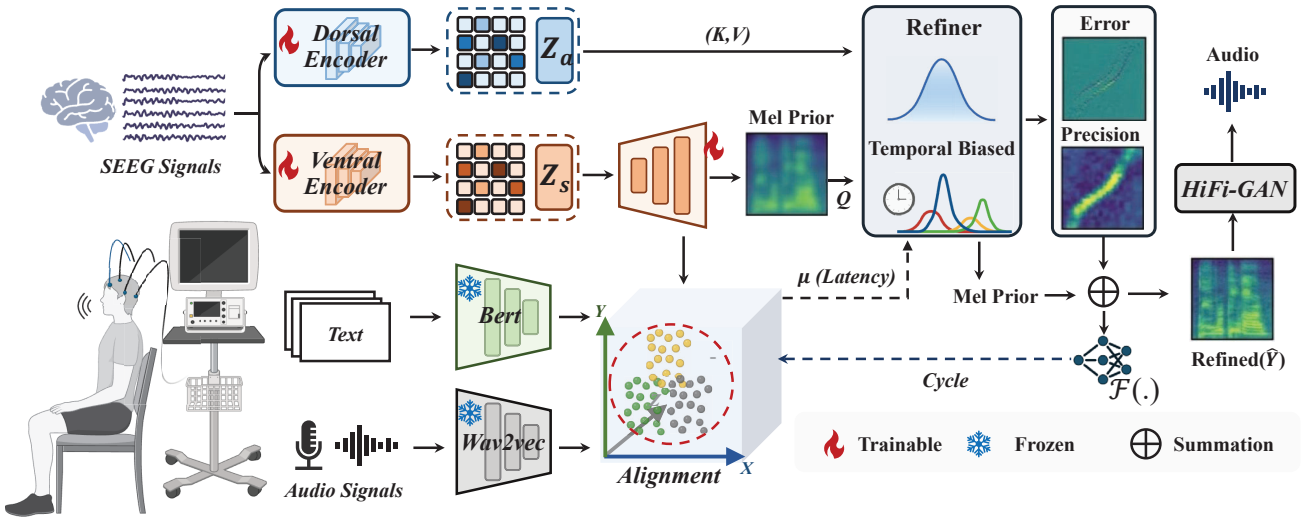


Figure 2: Overview of the proposed Neuro-Composer architecture. Neural signals are encoded through a dual-stream EEG encoder consisting of a ventral semantic stream and a dorsal acoustic stream. The ventral stream extracts a global semantic latent representation and generates a coarse spectro-temporal prior, while the dorsal stream encodes temporally precise acoustic evidence. A latency-aware predictive refiner integrates top-down semantic predictions with bottom-up neural evidence via precision-weighted error correction. The refined spectrogram is further constrained by cross-modal knowledge distillation and an auditory self-monitoring loop enforcing cycle consistency between intended and perceived speech.

where PosEnc$(\cdot)$ injects sin-cos positional information to maintain temporal order. Unlike the ventral stream, $\Phi_{\mathcal{D}}$ avoids global pooling, ensuring $\mathbf{Z}_{aco}$ maintains a frame-rate consistent with the acoustic target.

Within our framework, $\mathbf{Z}_{aco}$ is not a direct generator of speech but serves as bottom-up sensory evidence. These embeddings encapsulate the residual articulatory details required to refine the coarse semantic prior. Specifically, $\mathbf{Z}_{aco}$ provides the keys and values for the subsequent Predictive Refiner, enabling the model to update top-down expectations with actual neural observations.

Latency-Aware Predictive Refinement

To synchronize top-down semantic expectations with bottom-up acoustic evidence under physiological constraints, *Neuro-Composer* introduces a predictive refinement module that implements an iterative error-correction mechanism aligned with the principles of predictive coding.

We implement the refinement process as a cross-attention operation, where the semantic prior $\mathbf{M}_{coarse}$ provides predictive queries ($\mathbf{Q}$), and the dorsal stream embeddings $\mathbf{Z}_{aco}$ serve as keys ($\mathbf{K}$) and values ($\mathbf{V}$). To account for the intrinsic neural-to-articulatory latency, the attention weights $\mathbf{A}_{i,j}$ are modulated by a Gaussian temporal bias that penalizes distant alignment:

$$\mathbf{A}_{i,j} = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sqrt{D}} - \frac{(j - c_i)^2}{2\sigma^2}\right) \tag{3}$$

where i and j denote the acoustic and neural time indices, and D is the embedding dimension. The attention center c_i dynamically tracks the expected alignment point:

$$c_i = i \cdot \frac{T}{T'} - \big(\mu + \psi(\mathbf{z}_{sem})\big) \tag{4}$$

where μ is the baseline physiological delay, and $\psi(\mathbf{z}_{sem})$ is a context-dependent offset that accounts for variable neural processing demands based on semantic complexity.

Following the hierarchy of predictive coding, the decoder updates the top-down prior by integrating sensory innovation from the dorsal stream. The dorsal embeddings $\hat{\mathbf{E}}_{dorsal} \in \mathbb{R}^{F \times T'}$ are first projected into the acoustic manifold to generate a residual evidence map:

$$\hat{\mathbf{E}}_{dorsal} = \text{Proj}_{\mathcal{D}}(\mathbf{Z}_{aco}) \tag{5}$$

To manage neural uncertainty, the model estimates a spatio-temporal precision map $\mathbf{\Lambda} \in \mathbb{R}^{F \times T'}$, reflecting the reliability of bottom-up signals. This precision-gating mechanism is formulated as:

$$\mathbf{\Lambda} = \sigma\big(\text{MLP}(\mathbf{Z}_{aco} \oplus \mathbf{z}_{sem})\big) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid activation. The final reconstructed spectrogram $\hat{\mathbf{Y}}$ is synthesized by augmenting the prior with the precision-weighted residual:

$$\hat{\mathbf{Y}} = \mathbf{M}_{coarse} + \mathbf{\Lambda} \odot \hat{\mathbf{E}}_{dorsal} \tag{7}$$

This formulation ensures that the model maintains long-term structural coherence through $\mathbf{M}_{coarse}$ while dynamically incorporating fine-grained articulatory details through $\hat{\mathbf{E}}_{dorsal}$ only when the neural evidence is deemed reliable.

Latent Alignment and Self-Monitoring

Neural speech decoding is inherently under-constrained due to sparse and noisy intracranial recordings. In biological speech production, this uncertainty is mitigated by grounding internal representations in long-term linguistic priors and continuously monitoring generated speech through auditory feedback. To emulate these mechanisms, we introduce a neuro-mimetic strategy that combines latent manifold alignment with closed-loop self-monitoring.

Manifold Grounding via Foundation Models. To impose structured semantic and acoustic priors, we leverage pre-trained foundation models as fixed reference manifolds. Wav2Vec 2.0 ($\mathcal{T}_a$) defines the acoustic space, while BERT ($\mathcal{T}_s$) structures the semantic space. Given speech $\mathbf{Y}$ and transcript $\mathbf{L}$, teacher embeddings are extracted as:

$$\mathbf{z}^a = \text{Pool}(\mathcal{T}_a(\mathbf{Y})), \quad \mathbf{z}^s = \text{Pool}(\mathcal{T}_s(\mathbf{L})). \tag{8}$$

The sEEG-derived intention embedding $\mathbf{z}_{clip}$ is aligned with both manifolds via a contrastive objective. This grounding constrains neural representations to linguistically plausible regions, stabilizing decoding under high neural uncertainty.

Closed-Loop Auditory Self-Monitoring. Speech production involves continuous self-monitoring through internal feedback mechanisms. Inspired by this process, we introduce an auditory feedback module $\mathcal{F}(\cdot)$ that re-encodes the generated spectrogram into the semantic space:

$$\mathbf{z}_{cycle} = \mathcal{F}(\hat{\mathbf{Y}}) \tag{9}$$

A cycle consistency constraint is imposed between $\mathbf{z}_{clip}$ and $\mathbf{z}_{cycle}$, penalizing reconstructions that deviate from the decoded semantic intent.

Objective Function

To facilitate the hierarchical generative process, *NeuroComposer* is optimized via a multi-task objective function designed to harmonize spectral accuracy with neuro-mimetic consistency. This composite loss function ensures a synergy across three critical dimensions: **signal fidelity** for acoustic reconstruction, **semantic grounding** via cross-modal alignment, and **biological plausibility** through predictive-coding-based regularization.

First, to ensure high-fidelity speech synthesis, we minimize the acoustic reconstruction loss $\mathcal{L}_{rec}$, which combines L_1 distance for spectral accuracy and Pearson Correlation Coefficient (PCC) for temporal envelope coherence.

$$\mathcal{L}_{rec} = \|\hat{\mathbf{Y}} - \mathbf{Y}\|_1 + \alpha(1 - \text{PCC}(\hat{\mathbf{Y}}, \mathbf{Y})) \tag{10}$$

Simultaneously, to overcome data scarcity, we apply a multimodal alignment loss $\mathcal{L}_{align}$. This term anchors the normalized neural embedding $\bar{\mathbf{z}}_{eeg}$ to the linguistic priors provided by the teacher models ($\bar{\mathbf{z}}^a, \bar{\mathbf{z}}^s$), maximizing their mutual information via symmetric cross-entropy:

$$\mathcal{L}_{align} = \text{CE}(\tau\bar{\mathbf{z}}_{eeg}, \bar{\mathbf{z}}^a) + \text{CE}(\tau\bar{\mathbf{z}}_{eeg}, \bar{\mathbf{z}}^s) \tag{11}$$

To mimic the brain's active inference mechanisms, we introduce two regularization terms. $\mathcal{L}_{cycle}$ simulates the auditory feedback loop by enforcing cosine similarity between the initial intention and the re-encoded feedback. Furthermore, drawing from the Free Energy Principle, we employ an uncertainty-aware loss $\mathcal{L}_{pc}$, where the model estimates a precision matrix $\mathbf{\Lambda}$ to down-weight reconstruction errors in noisy regions dynamically:

$$\mathcal{L}_{cycle} = 1 - \text{CosSim}(\mathbf{z}_{eeg}, \mathbf{z}_{cycle}) \tag{12}$$

$$\mathcal{L}_{pc} = \mathbb{E}\left[\|\mathbf{\Lambda} \odot (\hat{\mathbf{Y}} - \mathbf{Y})\|_1 - \beta \log(\mathbf{\Lambda})\right] \tag{13}$$

The final loss is a weighted sum: $\mathcal{L}_{total} = \lambda_r\mathcal{L}_{rec} + \lambda_a\mathcal{L}_{align} + \lambda_c\mathcal{L}_{cycle} + \lambda_p\mathcal{L}_{pc}$.

Experiments

Dataset We evaluate our model on the Dutch-iBIDS dataset (Verwoert et al., 2022), comprising recordings from 10 participants with pharmacoresistant epilepsy (mean age: 32 years; 5 male and 5 female). Neural signals were recorded using platinum–iridium sEEG electrodes while participants read aloud 100 randomly selected Dutch words from the IFA corpus. Each stimulus was displayed on a screen for 2 s, followed by a 1-s interval, resulting in 300 s of recording per subject.

Preprocessing Raw sEEG recordings are first resampled to 1024 Hz. To isolate task-relevant cortical activity, we apply a fourth-order Butterworth bandpass filter (70–170 Hz) and additional notch filters at 98–102 Hz and 148–152 Hz to suppress line noise harmonics. All filtering is executed via a zero-phase forward-and-reverse procedure to eliminate phase distortion. High-gamma activity is extracted by computing the analytic signal's magnitude via the Hilbert transform. Simultaneously, synchronized audio is downsampled to 16 kHz and anonymized through randomized pitch modulation. Following established protocols (He et al., 2025; Kohler et al., 2021), data are segmented using a sliding-window scheme, pairing 1-s sEEG windows with center-aligned 464-ms speech segments at a 25-ms stride.

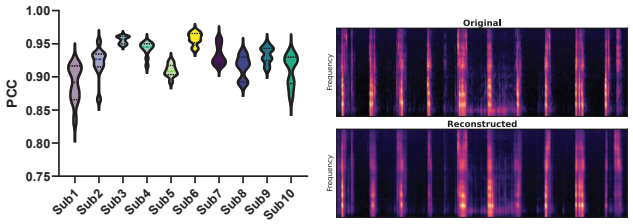
Implementation Details We compare our model with following related methods: Regression (Verwoert et al., 2022), CNN (Angrick et al., 2019), 3D-CNN (Arthur & Csapó, 2024), GRUs (Krishna et al., 2020), RNNSeq2Seq (Kohler et al., 2021), Seq2Seq (Duraivel et al., 2023), NeuroIncept (Khanday et al., 2025), MiSTR (Al-Radhi et al., 2025).

All experiments are implemented in PyTorch and conducted on an NVIDIA GeForce RTX 4090D GPU. To ensure fair comparison, all methods shared the same data preprocessing pipeline and experimental protocol. A subject-dependent evaluation strategy is adopted, where models are trained and evaluated separately for each subject using 10-fold cross-validation. The model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A cosine annealing learning rate scheduler is applied, with the minimum learning rate set to 1×10^{-6}. Pre-trained wav2vec 2.0 and BERT

Table 1: Performance comparison across different models.

Model	PCC (↑)	RMSE (↓)	SSIM (↑)
Regression	$0.6820 \pm 0.1207^{\dagger}$	$2.4938 \pm 0.9184^{\dagger}$	$0.5995 \pm 0.0441^{\dagger}$
GRUs	$0.6793 \pm 0.0811^{\dagger}$	$2.3415 \pm 0.9265^{\dagger}$	$0.6113 \pm 0.0462^{\dagger}$
RNNSeq2Seq	$0.7270 \pm 0.0926^{\dagger}$	$2.0992 \pm 0.8648^{\dagger}$	$0.6476 \pm 0.0434^{\dagger}$
CNN	$0.8036 \pm 0.0732^{\dagger}$	$1.9847 \pm 0.8065^{\dagger}$	$0.6468 \pm 0.0379^{\dagger}$
3D-CNN	$0.8224 \pm 0.0681^{\dagger}$	$1.8729 \pm 0.7583^{\dagger}$	$0.6597 \pm 0.0416^{\dagger}$
Seq2Seq	$0.8491 \pm 0.0576^{\dagger}$	$1.1248 \pm 0.7089^{\dagger}$	$0.6704 \pm 0.0387^{\dagger}$
NeuroIncept	$0.9067 \pm 0.0314^{*}$	0.5326 ± 0.1183	$0.7018 \pm 0.0310^{\dagger}$
MiSTR	$0.9093 \pm 0.0346^{*}$	$0.5524 \pm 0.1391^{*}$	$0.7047 \pm 0.0298^{*}$
Proposed	$\mathbf{0.9264 \pm 0.0221}$	$\mathbf{0.5078 \pm 0.0967}$	$\mathbf{0.7389 \pm 0.0254}$

Note. $^{*}p < 0.05$ compared to Proposed; $^{\dagger}p < 0.001$ compared to Proposed.



(a) 10-fold PCC distribution. (b) Spectrogram comparison

Figure 3: **Reconstruction stability and spectral fidelity.**

models are employed as frozen teacher networks to extract acoustic and semantic representations, which were precomputed and cached prior to training. All experiments are conducted with a batch size of 64, and gradient clipping with a maximum norm of 1.0 was applied to stabilize training. Loss weights are $\{\lambda_r, \lambda_a, \lambda_c, \lambda_p\} = \{1.0, 0.5, 0.1, 0.1\}$, with $\alpha = 0.5, \tau = 0.07$, and $\beta = 0.01$. Finally, the predicted Mel spectrograms are converted into raw speech waveforms using a pre-trained HiFi-GAN vocoder (Kong et al., 2020).

Reconstruction quality is evaluated using three complementary metrics: PCC for temporal envelope correlation, Root Mean Square Error (RMSE) for spectral magnitude accuracy, and Structural Similarity Index Measure (SSIM) for global spectro-temporal perceptual similarity.

Results Table 1 presents a comparative analysis of our proposed model with advanced baseline models on the Dutch-iBIDS dataset. To conduct statistical analysis, the *Wilcoxon signed-rank test* with *Holm-Bonferroni correction* is utilized. The results demonstrate that the proposed model achieves consistently best reconstruction performance across all metrics, and most improvements are statistically significant. Specifically, compared to the suboptimal results (achieved by MiSTR), the PCC and SSIM of the proposed method are improved by margins of 0.017 ($p < 0.05$) and 0.034 ($p < 0.05$), respectively, while the RMSE is reduced by 0.045 ($p = 0.0739$). In addition, it can be found that deep learning models generally achieve superior performance compared to the traditional Linear Regression due to the capability of modeling non-linear neural dynamics. GRUs and RNNSeq2Seq achieve relatively inferior results among deep models, likely because they rely solely on temporal recurrence without efficient spatial feature extraction from the electrode array. However, CNN and 3D-CNN achieve more considerable results by explicitly capturing local spectro-temporal patterns. Additionally, Seq2Seq and NeuroIncept obtain better reconstruction results by leveraging encoder-decoder structures and multi-scale feature integration, respectively. MiSTR extracts highly robust features by introducing masked representation learning, achieving performance closest to ours. Despite this, by utilizing the dual-stream predictive coding strategy and closed-loop feedback, our proposed Neuro-Composer can more adequately capture the hierarchical interaction between semantic intent and acoustic details, thereby achieving state-of-the-art decoding performance.

Table 2: Ablation study on the Dutch-iBIDS dataset.

Method Variant	PCC (↑)	RMSE (↓)	SSIM (↑)
Full Architecture	**0.926**	**0.508**	**0.739**
Structural Ablation (Dual-Stream)			
w/o Ventral Stream	0.845 (↓ 0.081)	0.682 (↑ 0.174)	0.657 (↓ 0.082)
w/o Dorsal Stream	0.772 (↓ 0.154)	0.895 (↑ 0.387)	0.614 (↓ 0.125)
Mechanism Ablation			
w/o Refiner	0.884 (↓ 0.042)	0.633 (↑ 0.125)	0.692 (↓ 0.047)
w/o Feedback	0.901 (↓ 0.025)	0.542 (↑ 0.034)	0.718 (↓ 0.021)

To further validate the stability of our model, Figure 3(a) visualizes the PCC distribution specifically for Neuro-Composer across all test trials, revealing a consistent high-fidelity performance. Furthermore, the qualitative spectral comparison for a representative subject in Figure 3(b) confirms that our model effectively recovers fine-grained acoustic structures, such as harmonic stacks and formant transitions, while maintaining precise temporal alignment with the ground truth.

Ablation Study To quantify the contribution of each component, we evaluate four ablated variants (Table 2). The results reveal distinct mechanistic roles: **1) Dual-Stream Complementarity:** Removing the *dorsal acoustic stream* caused the most severe drop in signal fidelity (PCC −16.6%), identifying it as the carrier of articulatory dynamics. In contrast, removing the *ventral semantic stream* disproportionately degraded SSIM (−11.1%), a dissociation confirming that semantic priors are indispensable for structural coherence. **2) Latency Modeling:** Substituting the *Latency-Aware Refiner* with standard attention significantly increased reconstruction error (RMSE +0.125). This underscores the necessity of the temporal inductive bias to align semantic intent with delayed neural evidence. **3) Feedback Regulation:** Disabling *closed-loop feedback* led to consistent performance attenuation, validating the self-monitoring mechanism as a critical regularizer against decoding drift. Overall, the full model's superiority shows these components synergistically bridge the gap between neural activity and intelligible speech.

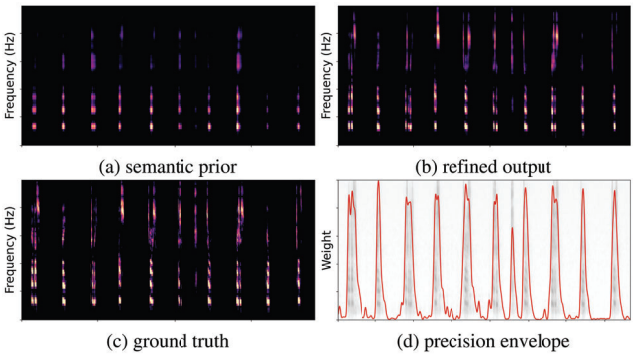


Figure 4: **Visualizing the Predictive Coding Mechanism.** (a-c) The model first generates a blurry semantic prior (Intention) and then refines it with acoustic details, mirroring a coarse-to-fine cognitive process. (d) The precision gate (red line) dynamically modulates the integration of neural signals, exhibiting high confidence during vocalization and suppressing noise during silence.

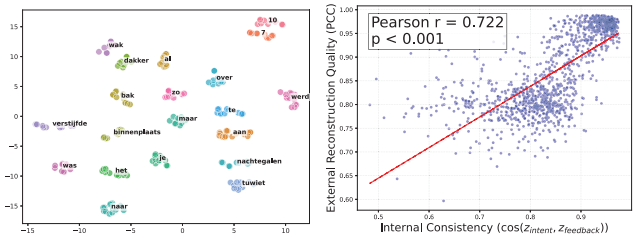
Neurocognitive Interpretability

To substantiate *Neuro-Composer* as a mechanistically interpretable speech production model, we conduct a multi-level analysis probing internal representations across two neurocomputational dimensions: the hierarchical generative process and the latent manifold's semantic structure.

Hierarchical Generative Mechanism via Predictive Coding. Our architecture posits that the brain operates as a generative model, synthesizing speech by integrating top-down predictions (priors) with bottom-up residuals (prediction errors) derived from neural evidence.

As illustrated in Figure 4, the decoding process follows a predictive coding schema. The *Semantic Prior* (Figure 4(a)), generated by the ventral stream, captures the global "intention" of the utterance, which is visible as the coarse-grained spectral envelope. This prior serves as a low-frequency scaffold deficient in textural details. The *Refined Output* (Figure 4(b)) is then synthesized by integrating high-frequency residuals (representing articulatory details) from the dorsal stream to correct this scaffold. This hierarchical reconstruction closely mirrors the ground truth (Figure 4(c)), demonstrating that the model successfully functionally disentangles semantic planning from articulatory execution.

Further analyzing the generative mechanism, Figure 4(d) visualizes the precision-weighted error integration by overlaying the predicted *Precision Gate* (red curve) onto the refined spectrogram. The gate acts as a dynamic filter modulating the integration of dorsal stream residuals: it approaches 1.0 (high confidence) during active speech segments, allowing dorsal details to update the ventral prior, and drops to near 0.0 during silence or transition periods. This mimics the brain's biological mechanism of suppressing neural noise when evidence is ambiguous (low precision), thereby effectively preventing the generation of hallucinatory artifacts during silence.



(a) Latent Semantic Clusters (b) Metacognitive Calibration

Figure 5: **Visualization of representation quality and self-monitoring.** (a) t-SNE projection of decoded latent intention vectors ($\mathbf{z}_{clip}$) colored by semantic category. (b) Scatter plot showing the correlation between internal cycle consistency (model confidence) and external reconstruction quality (PCC).

Latent Semantic Structure and Metacognitive Calibration. To validate the quality of the learned representations, we analyze the organization of the latent space and the reliability of the self-monitoring mechanism.

We visualize the internal organization by projecting latent vectors ($\mathbf{z}_{clip}$) into 2D using t-SNE (Maaten & Hinton, 2008) (Maaten & Hinton, 2008). As shown in Figure 5(a), the representations exhibit clear clustering patterns corresponding to distinct semantic categories (color-coded). Despite the high variability inherent in single-trial sEEG signals, samples belonging to the same category exhibit high intra-class cohesion. This compactness confirms that the cross-modal distillation has successfully regularized the neural embeddings into a structured semantic manifold, effectively aligning stochastic brain signals with abstract linguistic concepts.

We further examine whether the model possesses *metacognitive calibration*—defined here as the ability to estimate its own decoding reliability. We hypothesize that the internal cycle consistency (computed as the cosine similarity between the forward intention $\mathbf{z}_{clip}$ and the feedback reconstruction $\mathbf{z}_{cycle}$) serves as a proxy for the model's confidence. Figure 5(b) reveals a robust positive correlation (Pearson's $r > 0.5$) between this internal metric and the external reconstruction quality (PCC). This significant relationship indicates that the model can autonomously gauge the quality of its output in the absence of ground truth, which is critical for reliable error detection in practical BCI applications.

Conclusion

In this paper, we propose *Neuro-Composer*, a bio-inspired framework that advances speech decoding from passive regression to a hierarchical generative process. Guided by the Dual-Stream Theory, our model integrates top-down semantic priors and bottom-up acoustic details via a latency-aware predictive refiner and a closed-loop feedback mechanism. Results on Dutch-iBIDS dataset demonstrate that this model achieves state-of-the-art reconstruction fidelity. These bio-mimetic constraints bolster decoding robustness and interpretability, providing a principled solution for neural dynamics.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62136004, Grant 62276130, and Grant 62371234, in part by the National Key R&D Program of China under Grant 2023YFF1204803, in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231438, and in part by the Basic Research Program of the Bureau of Science and Technology (ILF24001).

References

Al-Radhi, M. S., Németh, G., & Gerazov, B. (2025). Mistr: Multi-modal ieeg-to-speech synthesis with transformer-based prosody prediction and neural phase reconstruction. *arXiv preprint arXiv:2508.03166*.

Angrick, M., Herff, C., Johnson, G., Shih, J., Krusienski, D., & Schultz, T. (2019). Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing*, *342*, 145–151.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493–498.

Arthur, F. V., & Csapó, T. G. (2024). Speech synthesis from intracranial stereotactic electroencephalography using a neural vocoder. *Infocommunications Journal*, *16*(1).

August, B. C., Benally, C. D., & Cadena, D. E. (2007). An example conference paper. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*, 100–105. https://escholarship.org/uc/item/k0e1y2

Berezutskaya, J., Freudenburg, Z. V., Vansteensel, M. J., Aarnoutse, E. J., Ramsey, N. F., & van Gerven, M. A. (2023). Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. *Journal of neural engineering*, *20*(5), 056010.

Daphne, E. F., & Echo, F. G. (2022). An example journal article. *Journal of Example Studies*, *10*(3), 200–215. https://doi.org/10.1234/eg.article

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, *5*(10), 1097–1107.

Duraivel, S., Rahimpour, S., Chiang, C.-H., Trumpis, M., Wang, C., Barth, K., Harward, S. C., Lad, S. P., Friedman, A. H., Southwell, D. G., et al. (2023). High-resolution neural recordings improve the accuracy of speech decoding. *Nature communications*, *14*(1), 6938.

Fitzgerald, G. H., & Galli, H. I. (1985). *An example technical report* (Report No. EX-TR-85-1). Some University, Department of Hypothetical Studies.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Hakuole, I. J. (2001). *An example thesis title* [Doctoral dissertation, Some University]. SU Campus Repository. https://hdl.handle.net/20.1000/5555

He, V. Y., Liu, A., Duan, S., Gao, Y., Qian, R., & Chen, X. (2025). Enhancing seeg-based speech decoding via convolutional encoder-decoder and scale-recursive reconstructor. *IEEE Sensors Journal*.

Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M. C., Goldrick, M. A., Krusienski, D. J., Slutzky, M. W., & Schultz, T. (2019). Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Frontiers in neuroscience*, *13*, 1267.

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, *8*, 141498.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393–402.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144.

Issa, J. K. (1963). An example book chapter in an edited volume. In K. L. Juliet & L. M. Kapoor (Eds.), *Example edited volume title* (pp. 300–315). Some Academic Press.

Khanday, O. M., Pérez-Córdoba, J. L., Mir, M. Y., Najar, A. A., & Gonzalez-Lopez, J. A. (2025). Neuroincept decoder for high-fidelity speech reconstruction from neural activity. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Kohler, J., Ottenhoff, M. C., Goulis, S., Angrick, M., Colon, A. J., Wagner, L., Tousseyn, S., Kubben, P. L., & Herff, C. (2021). Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*.

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, *33*, 17022–17033.

Krishna, G., Tran, C., Han, Y., Carnahan, M., & Tewfik, A. H. (2020). Speech synthesis using eeg. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1235–1238.

Lobsang, M. N. (Ed.). (2023). *Example edited volume title* (2nd ed.). Some University Press. https://doi.org/10.4321/eg.vol2

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, *620*(7976), 1037–1046.

Mitanni, N. O., & November, O. P. (1972). *Example book title*. Some Publisher.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang,

E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, *10*(1), e1001251.

Petrosyan, A., Voskoboinikov, A., Sukhinin, D., Makarova, A., Skalnaya, A., Arkhipova, N., Sinkin, M., & Ossadtchi, A. (2022). Speech decoding from a small set of spatially segregated minimally invasive intracranial eeg electrodes with a compact and interpretable neural network. *Journal of Neural Engineering*, *19*(6), 066016.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329–347.

Rousseau, M.-C., Baumstarck, K., Alessandrini, M., Blandin, V., Billette de Villemeur, T., & Auquier, P. (2015). Quality of life in patients with locked-in syndrome: Evolution over a 6-year period. *Orphanet journal of rare diseases*, *10*(1), 88.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences*, *105*(46), 18035–18040.

Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A., & Chang, E. F. (2024). The speech neuroprosthesis. *Nature Reviews Neuroscience*, *25*(7), 473–492.

Stavisky, S. D. (2025). Restoring speech using brain–computer interfaces. *Annual Review of Biomedical Engineering*, *27*.

Verwoert, M., Ottenhoff, M. C., Goulis, S., Colon, A. J., Wagner, L., Tousseyn, S., Van Dijk, J. P., Kubben, P. L., & Herff, C. (2022). Dataset of speech production in intracranial electroencephalography. *Scientific data*, *9*(1), 434.

Zhang, L., Zhou, Y., Gong, P., & Zhang, D. (2024). Speech imagery decoding using eeg signals and deep learning: A survey. *IEEE Transactions on Cognitive and Developmental Systems*.