HFFN-ID: A Hierarchical Feature Fusion Network with Bi-Phase Subject ID Modulation for EEG Mel-Spectrogram Reconstruction

Chi Huang* , **Zhaohu Liu*** , **Yong Peng†** and **Wanzeng Kong**
School of Computer Science, Hangzhou Dianzi University, Hangzhou, 310018, China
{huangc, zhaohuliu, yongpeng, kongwanzeng}@hdu.edu.cn

Abstract

High-fidelity reconstruction of mel-spectrograms from EEG signals remains a formidable challenge, primarily due to the inherent inter-subject variability of neural patterns and the semantic gap between heterogeneous feature representations of these two modalities. To alleviate both issues, this paper proposes a Hierarchical Feature Fusion Network with bi-phase subject IDentifier modulation (HFFN-ID) for reconstructing mel-spectrograms from EEG signals. The primary improvements of the HFFN-ID framework are from two aspects, a Binary-Phase Subject-ID Modulation (BiPSM) mechanism for explicit subject conditioning and a Condition-guided Hierarchical Fusion (CHF) component for dynamic multi-layer feature synthesis, which respectively aim to address the problems of inter-subject variability in neural patterns and heterogeneous cross-modality feature fusion. Moreover, a speech envelope feature-based pre-training strategy is incorporated to initialize the parameter space, inspired by the shared low-level representations across different speech features. On the SparrKULee dataset, HFFN-ID establishes a new benchmark with a pearson correlation coefficient of 0.0723, representing a substantial 44% relative improvement over existing baseline. Importantly, HFFN-ID is highly efficient, achieving a 38.7% parameter reduction and a 2.5× training speedup. These results highlight the effectiveness of subject-conditioned and hierarchically fused architectures for advancing high-fidelity neural speech decoding.

1 Introduction

Decoding speech intentions from brain activity has long been a goal in medicine and neuroscience. Speech brain-computer interfaces (BCIs) offer a promising solution for patients who cannot speak due to brain injuries, strokes, or neurodegenerative diseases to communicate with the surrounding environment [Anumanchipalli *et al.*, 2019; Metzger *et al.*, 2022; Kunz *et al.*, 2025]. Previous research has successfully decoded various aspects of speech from intracranial neural recordings. Studies have achieved the decoding of acoustic features such as spectral content and speech envelopes [Akbari *et al.*, 2019; Accou *et al.*, 2023; Xu *et al.*, 2024], articulatory movements [Moses *et al.*, 2021; Sun *et al.*, 2024], individual words [Chan *et al.*, 2011], phrases or sentences [Wang and Ji, 2022]. Some models even performed end-to-end speech decoding from EEG, directly reconstructing listened speech waveforms [Lee *et al.*, 2024; Guo *et al.*, 2023; Ma *et al.*, 2025]. While invasive BCIs achieve high precision in decoding fundamental linguistic features, this capability is severely offset by significant clinical barriers, precluding their widespread use for most patients.

The pursuit of speech decoding from non-invasive EEG signals is gaining considerable importance [Sato *et al.*, 2024]. While the non-surgical nature of EEG-based BCIs offers a safer and more universally applicable solution, the approach is constrained by two major obstacles, poor signal quality and artifact vulnerability [Sharma and Meena, 2024], coupled with the difficulty of interpreting signals from distributed language networks [Fedorenko *et al.*, 2024]. Existing studies demonstrated the feasibility through low-level semantic classification; for example, the 3DCNN-BiLSTM model achieved 75.2% accuracy in phrase classification [Alharbi and Alotaibi, 2024] and over 85% accuracy was obtained by 1D-CNN/LSTM in the character/digit classification tasks [Mahapatra and Bhuyan, 2023]. However, the applicability of these classification methods was limited to closed vocabulary sets and simple semantic hierarchies, severely restricting their utility in processing complex, real-world language sequences. This limitation spurred a critical shift toward speech feature reconstruction from EEG. Although some generative frameworks such as NeuroTalk which implements word reconstruction using GANs [Lee *et al.*, 2023] and sEMG-based models which estimate articulatory features for speech synthesis [Lee *et al.*, 2025] have emerged, they typically suffer from a shared over-reliance on small-scale datasets and confinement to word-level speech. This creates a substantial gap from the long-term continuous speech decoding from EEG signals required for practical applications.

The high inter-subject variability of EEG patterns, stemming from variations in brain structure and the subjects' current psychological states, and distinct neurophysiological

†Corresponding author

processing strategies [Huang *et al.*, 2023], poses a significant challenge for building models that generalize well across subjects. This variability, combined with the complex and high-dimensional nature of EEG signals, complicates the precise extraction of robust, speech-related neural features; therefore, the long-sequence neural speech decoding research faces a significant challenge. Existing decoding frameworks often face a dilemma between insufficient accuracy and computational extravagance. The inherent limitation of conventional feature fusion methods lies in their static and non-adaptive nature, which compromises both the discriminative power and robustness of existing models. This shortcoming stems from their inability to effectively bridge the hierarchical semantic gap between heterogeneous EEG and speech representations.

To overcome existing limitations, we introduce a novel HFFN-ID framework for accurate and efficient EEG-based mel-spectrogram reconstruction. This framework integrates three core components, i.e., a Binary-Phase Subject-ID Modulation (BiPSM) component to explicitly model and capture the inter-subject variability, a Condition-guided Hierarchical Fusion (CHF) component to address the hierarchical semantic gap and effectively integrate multi-layer features, and a selective state space model, i.e., Bidirectional Mamba (BiMamba)-based Global Feature Extraction (GFE) component [Gu and Dao, 2024] for efficiently modeling long-term dependencies. The main contributions of this work are summarized below.

- A novel BiPSM component is introduced to mitigate the inter-subject variability by mapping subject identifiers into affine transformation parameters, which enables our model to adaptively calibrate internal features via learnable scaling and shifting, ensuring robust performance across diverse individual EEG characteristics.
- An hierarchical feature fusion component is proposed to resolve semantic misalignment via multi-head attention and adaptive gating. This mechanism ensures cross-layer consistency and semantic complementarity, preserving both fine-grained spectral details and global structural integrity.
- Experiments show that HFFN-ID framework significantly outperforms the baseline in mel-spectrogram reconstruction, improving the PCC metric by 44% to 0.0723 on SparrKULee. Ablation studies confirm the contribution of each component and the two-stage training strategy. Besides, HFFN-ID framework exhibits reduced parameter size and running time.

2 Related Work

The primary challenge in achieving long-term speech decoding from non-invasive EEG signals involves overcoming the inherently low signal-to-noise ratio, mitigating substantial inter-subject variability in neural recordings, and effectively balancing the modeling of long-term temporal dependencies with the capture of short-term acoustic features. Early pioneering work established the feasibility of decoding coarse speech information from EEG signals, with a major focus on reconstructing the speech envelope—a low-dimensional, slow-varying representation of speech energy. In the early years, the decoding task was conventionally formulated as a linear regression problem [O'sullivan *et al.*, 2015]. Recently, some representative works have been proposed. In [Accou *et al.*, 2023], the VLAAI network was proposed to enhance the temporal dependency modeling by incorporating nonlinear modules and output context modules. The HappyQuakka model utilizes a pre-layer normalized transformer architecture augmented with a global conditioning mechanism for within-subject EEG-to-speech envelope decoding [Piao *et al.*, 2023]. The Sea-Wave architecture [Van Dyck *et al.*, 2023] effectively adapts the WaveNet's dilated causal convolutions and residual connections to make it particularly suitable for EEG signal processing, given its proficiency in capturing both localized temporal details and extended contextual patterns. Generally, EEG speech decoding has evolved from simple linear models to deep learning models that better handle the cross-modality feature alignment and individual variability.

Mel-spectrograms remain the paramount representation for high-fidelity speech feature reconstruction, owing to their information-rich time-frequency characteristics. ConvconcatNet [Xu *et al.*, 2024] utilizes a deep CNN architecture with extensive feature concatenation, further incorporating an envelope-based auxiliary training strategy to refine spectral details. In [Li *et al.*, 2024], CAT-WaveNet was proposed to address the cross-modal gap (e.g., EEG to audio) by integrating a cross-attention mechanism within a WaveNet backbone, and employing a coarse-to-fine generation strategy. For better temporal dependency modeling and long-sequence decoding, DMF2Mel network was proposed to enhance local transient details and global semantic context [Fan *et al.*, 2025a]. SSM2Mel [Fan *et al.*, 2025b] employs a hybrid architecture, by combining a Mamba module for long-sequence dependency modeling, an S4-UNet for local feature extraction, and an Embedding Strength Modulator (ESM) for subject adaptation. However, the ESM module relies solely on one-hot encoding and linear projection for subject identifiers, which lacks expressivity and fundamentally limits the model capacity in adapting to individual neural signatures.

Generally, current EEG-to-speech decoding models still struggle with two key challenges, significant inter-subject variability in EEG signals and a persistent hierarchical semantic gap in feature representations. Therefore, it is necessary to develop advanced decoding models with improved performance and efficiency.

3 Methodology

3.1 Overview of HFFN-ID

As illustrated in Figure 1, the proposed HFFN-ID framework comprises three core components that operate in a hierarchical manner. The first component is the BiPSM mechanism, which takes raw EEG signals and a sub-id as inputs. The sub-id acts as a unique embedding for each participant, enabling this component to model individual differences through a two-stage modulation mechanism. This process yields physiologically and semantically enriched feature representations, which are subsequently passed to the S4-Unet [Ronneberger *et al.*, 2015; Gu *et al.*, 2022; Fan *et al.*, 2025b] for local spatio-temporal feature extraction.

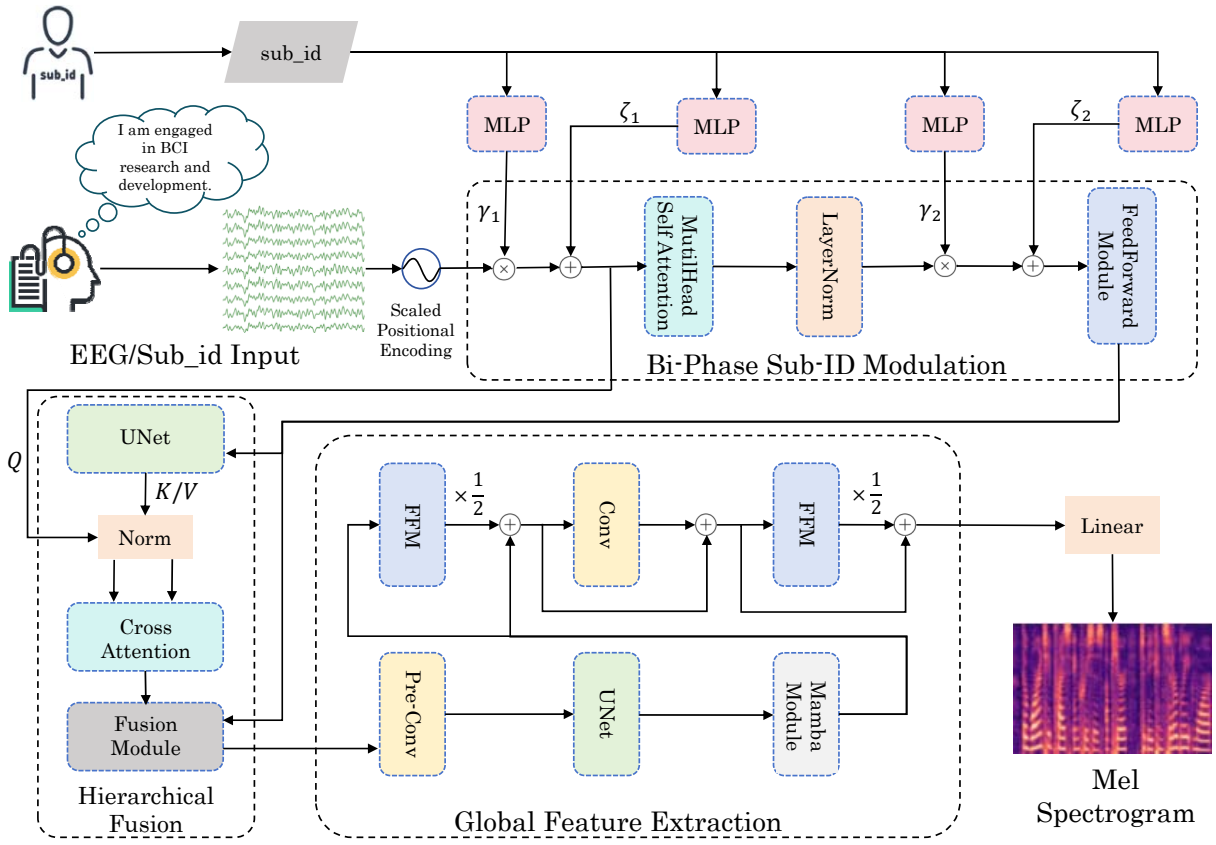

Figure 1: The HFFN-ID framework overview, which has three components, i.e., the Bi-Phase Subject-ID Modulation (BiPSM) component, the Condition-guided Hierarchical Fusion (CHF) component, and the BiMamba-based Global Feature Extraction (GFE) component.

The second component is the CHF network, which integrates the outputs from both modulation stages of the BiPSM and the S4-Unet via a structured attention-and-gating mechanism [Shazeer, 2020]. Specifically, in this component, an attention operation [Vaswani *et al.*, 2017] is first applied between the first modulation result of the BiPSM and the S4-Unet outputs. The resulting representation is then adaptively fused with the second modulation result from the BiPSM through a gating mechanism, thereby ensuring conditional consistency and semantic complementarity across different feature layers. The third component, the fused features are fed into a BiMamba-based sequence model to capture global dependencies, and the output is projected into a mel-spectrogram through a linear layer for subsequent speech reconstruction.

3.2 Binary-Phase Subject-ID Modulation (BiPSM)

EEG signals are characterized by significant inter-subject variability, which is evident in both spontaneous brain activities and evoked responses to external stimuli. To alleviate this variability, a unique sub-id information is incorporated into the model input. As a result, the model more effectively captures individual differences in neurophysiological signals, which in turn enhances the accuracy and generalizability of non-invasive speech decoding [Piao *et al.*, 2023].

The BiPSM component begins with processing the subject identify information (i.e., index number) through sinusoidal positional encoding, followed by a linear layer. A multi-layer perceptron (MLP) with non-linear activation functions then transforms this representation into a subject-embedded condition vector, which is subsequently split into four distinct affine parameters, $scale_{att}$, $shift_{att}$, $scale_{ffn}$, and $shift_{ffn}$. These parameters are precisely injected into two core computational modules—the attention and feed-forward module—to implement Binary-phase modulation. The architectural details of the BiPSM component are illustrated in the upper section of Figure 1.

We generate subject-specific embeddings by mapping each subject identifier (sub-id) through a time multi-layer perceptron (Time MLP). The design of this module is conceptually inspired by diffusion models [Ho *et al.*, 2020; Peebles and Xie, 2023; Hu *et al.*, 2026], where a Time MLP employs sinusoidal positional encoding to transform discrete temporal inputs into continuous conditioning vectors. Just as each diffusion timestep is associated with a unique noise profile, each subject may exhibit a distinct neural signature that should be encoded into a dedicated embedding. This approach provides a theoretically coherent and experimentally effective framework for personalized model adaptation, ensuring that individual neural characteristics are preserved and utilized during mel-spectrogram reconstruction.

Formally, the subject identifier is processed through a Time MLP to obtain an initial embedding vector. This represen-

tation is then non-linearly transformed by a follow-on multi-layer perceptron (MLP). The output of this cascaded structure is divided into four separate channels, producing the four sets of affine parameters. The detailed procedure is expressed by

$$h_1 = W_1 \cdot PE(\text{sub-id}) + b_1, \tag{1}$$

$$h_2 = W_2 \cdot SimpleGate(h_1) + b_2, \tag{2}$$

$$SimpleGate(h_1) = h_1^a \odot h_1^b, \tag{3}$$

where $PE(\cdot)$ denotes the sinusoidal positional encoding, W_1, W_2, b_1, b_2 respectively represent the weight and bias of the linear transformation. The $SimpleGate(\cdot)$ operation splits the input h_1 into two segments h_1^a and h_1^b, followed by an element-wise multiplication between them [Dauphin *et al.*, 2017]. The sub-id is processed by the Time MLP to generate the sub-embedding h_2, which is subsequently transformed by a MLP into h_2'. The output is partitioned into four consecutive blocks of equal size, corresponding to the affine parameters $scale_{att}$, $shift_{att}$, $scale_{ffn}$, and $shift_{ffn}$, i.e.,

$$h_2' = MLP(h_2), \tag{4}$$

$$[scale_{att}, shift_{att}, scale_{ffn}, shift_{ffn}] = split(h_2'). \tag{5}$$

To deeply integrate subject identity into EEG feature representations, a bi-phase modulation mechanism is implemented, which applies learnable affine transformations at two strategic stages. This approach ensures that the neural patterns are effectively personalized, enhancing the model ability in distinguishing among different individuals.

Phase I: Attention Layer Input Modulation. Prior to the multi-head attention operation, the input EEG features along with the scaled positional encoding (denoted as X) undergo an affine transformation to integrate initial subject identity cues. The resulting modulated features X_{att} are obtained by

$$X_{att} = X \odot (\text{scale}_{att} + 1) + shift_{att}, \tag{6}$$

where $\odot$ signifies the element-wise multiplication. X_{att} is then fed into the self-attention mechanism, followed by layer normalization, yielding the intermediate feature representation X'_{att}, i.e.,

$$X'_{att} = LN(MH(X_{att}, X_{att}, X_{att})). \tag{7}$$

Here, $MH(\cdot)$ denotes the multi-head self-attention function and $LN(\cdot)$ denotes layer normalization.

Phase II: Feed-Forward Layer Input Modulation. This second modulation refines the features X'_{att} immediately preceding the feed-forward network (FFN) by

$$X_{ffn} = X'_{att} \odot (scale_{ffn} + 1) + shift_{ffn}. \tag{8}$$

The fully modulated feature X_{ffn}, now carrying complete subject identity information, is passed through the FFN to generate the latent feature representation X'_{ffn}, i.e.,

$$X'_{ffn} = FFN(X_{ffn}). \tag{9}$$

The bi-phase integration of subject identity ensures that the network exerts rigorous, fine-grained control over adaptive feature transformation, enhancing the model's personalized decoding capabilities.

3.3 Condition-guided Hierarchical Fusion (CHF)

To ensure effective integration of multi-source representations spanning from subject-specific conditions to global contextual dependencies, we propose a dedicated Condition-guided Hierarchical Fusion (CHF) component. Considering that the fusion methods, such as naive concatenation or element-wise summation, are prone to semantic misalignment, this CHF component is designed to dynamically recalibrate and integrate these representations. Functioning as a gated filtering mechanism, it adaptively weights and consolidates semantic information in accordance with subject identity, thereby ensuring coherent cross-layer feature alignment.

Input Representation and Semantic Roles. This component operates on three distinct feature streams listed below, each serving a specific semantic role in the fusion hierarchy.

- X_{guide}: The feature representation following the initial binary-phase modulation, which acts as a Query Guidance to encapsulate the immediate subject-specific biases to direct the attention mechanism.
- X_{base}: The backbone feature processed through the complete modulation block, which serves as the Identity-Invariant Base to provide a stable and robust semantic foundation.
- X_{ctx}: The multi-scale contextual feature enhanced by the structured state-space model (S4-Unet), which acts as a Contextual Reservoir, offering rich and fine-grained details that may be selectively retrieved.

Cross-Layer Semantic Alignment. The first stage aims to align the contextual details (X_{ctx}) with the guidance signal (X_{guide}). To mitigate internal covariate shifts and ensure distribution consistency between heterogeneous streams, we first perform layer normalization on them. We formulate the attention mechanism by defining the Query (Q), Key (K), and Value (V) projections as

$$Q = LN(X_{guide}), K = LN(X_{ctx}), V = LN(X_{ctx}). \tag{10}$$

Then a scaled dot-product attention mechanism is employed to compute the relevance map, i.e.,

$$H_{att} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{11}$$

where H_{att} represents the contextual features that have been explicitly filtered and re-weighted by the guidance signal, and d_k is the scaling factor corresponding to the feature dimension. This operation ensures that only the contextual information most relevant to the current subject's state is propagated.

Adaptive Gated Aggregation. In the second stage, we introduce a non-linear gating mechanism to dynamically balance the impacts between the refined context (H_{att}) and the stable base (X_{base}). This is achieved via a parameterized gating network that learns a spatial-wise fusion coefficient. The gating coefficient $g \in [0, 1]$ is adaptively learned through a lightweight encoder-decoder structure. That is

$$g = \sigma\left(W_4 \cdot SiLU\left(W_3 \cdot [H_{att} \oplus X_{base}]\right)\right), \tag{12}$$

where $[\cdot \oplus \cdot]$ denotes the concatenation operation, W_3 and W_4 are learnable linear projection matrices, and the Sigmoid Linear Unit $SiLU$ is utilized as the hidden activation function to

facilitate smooth gradient flow. The final sigmoid activation $\sigma(\cdot)$ constraints the gating coefficient to the range $(0, 1)$. Finally, the integrated feature representation F_{out} is derived via a gating mechanism, i.e.,

$$F_{out} = g \odot H_{att} + (1 - g) \odot X_{base}. \tag{13}$$

This adaptive combination allows the network to effectively preserve the stable backbone information while selectively injecting context-aware details, thereby achieving a robust and personalized feature representation.

3.4 Global Feature Extraction (GFE)

Accurate mel-spectrogram reconstruction from EEG requires effective modeling of long-term temporal dependencies. To this end, we incorporate a selective state space module for long-sequence modeling. Specifically, we adopt the Bidirectional Mamba (BiMamba) model [Liang *et al.*, 2024], which robustly captures global sequence dependencies while mitigating the information fading and computational bottlenecks common in traditional recurrent architectures. BiMamba model incorporates convolutional layers and the S4-UNet architecture to strategically enhance the convolutional module. This design synergistically strengthens the model's intrinsic convolutional capabilities, rendering it significantly more perceptive to local, fine-grained structural nuances. Following this, a feed-forward module (FFM) applies nonlinear transformations through fully connected layers, projecting the features into a higher-dimensional semantic space. This step enriches the representation and optimizes the features for the final reconstruction stage.

3.5 Loss Function

To ensure training stability, we employ a combined loss function consisting of the mean absolute loss (i.e., $\mathcal{L}_1$) and the negative pearson correlation coefficient (i.e., R). The former quantifies the absolute deviation between the predicted and ground-truth mel-spectrograms, while the latter measures their linear relationship. The overall loss $\mathcal{L}$ is defined as

$$\mathcal{L} = -R + \alpha \cdot \mathcal{L}_1, \tag{14}$$

where

$$R = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \mu_{\hat{Y}})(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \mu_{\hat{Y}})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \mu_Y)^2}}, \tag{15}$$

$$\mathcal{L}_1 = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right|. \tag{16}$$

where $\hat{Y}_i$ and Y_i respectively denote the predicted and actual mel-spectrogram values, μ denotes the mean, and α is a hyperparameter to balance the contribution of both terms.

4 Experiments

4.1 Data Preparation and Experimental Setup

The ICASSP 2024 Auditory EEG Challenge dataset (SparrKULee) [Accou *et al.*, 2024] was utilized in this study, which contains EEG recordings from 85 normal-hearing Dutch-speaking participants listening to continuous Flemish narratives. EEG data was pre-processed by multichannel Wiener filtering for artifact removal, common average re-referencing, and final downsampling to 64 Hz. Raw audio was processed into 64 Hz auditory features by extracting Gammatone envelopes and mel-spectrograms, followed by polyphase resampling for synchronization with neural data. The dataset is subsequently partitioned into three parts with ratio 80%, 10% and 10%, respectively for model training, validation and testing.

We implemented the HFFN-ID framework in PyTorch and performed training for 1,000 epochs using the Adam optimizer, supplemented by a StepLR scheduler (decay factor of 0.9 every 50 epochs). The platform consists of an Intel Core i9-13900KF CPU and an NVIDIA GeForce RTX 4080 GPU (16GB VRAM). During model training, five-second signal segments were randomly cropped from each EEG and speech mel-spectrogram segment to promote training stability. For inference, the input EEG signal was divided into multiple non-overlapping five-second segments. Each segment was processed independently and the outputs were concatenated to reconstruct the complete mel-spectrogram.

We evaluate the effectiveness of our method against the following established EEG-to-Speech baselines, including

- **VLAAI** [Accou *et al.*, 2023], which enhances speech envelope decoding from EEG using innovative nonlinear modules and output context modules;
- **HappyQuokka** [Piao *et al.*, 2023], which employs a pre-layer normalized Transformer architecture with a global conditioner for within-subject speech envelope decoding from EEG signals;
- **ConvConcatNet** [Xu *et al.*, 2024], which reconstructs the mel-spectrograms from EEG data by combining iterative concatenation across blocks and leveraging speech envelope features for auxiliary training;
- **DMF2Mel** [Fan *et al.*, 2025a], which designs a dynamic multiscale fusion network to balance the efficiency of temporal dependency modeling and information retention in long-sequence decoding;
- **SSM2Mel** [Fan *et al.*, 2025b], which uses a hybrid architecture that integrates Mamba (long-sequence modeling), S4-UNet (local feature extraction), and an ESM (subject adaptation) to reconstruct the mel-spectrogram.

4.2 The Training Strategy

Inspired by the principle of knowledge transfer and the established fact that neural responses track the speech envelope [Accou *et al.*, 2023], we propose a two-stage training strategy. In the initial stage, the model undergoes pre-training on the condensed speech envelope-based feature representations to capture fundamental temporal and spectral dynamics. Subsequently, the pre-trained weights are fine-tuned on the richer mel-spectrogram inputs. This approach facilitates highly effective knowledge transfer from the general temporal domain to the specific frequency domain, with the expectation of significantly enhancing final model performance.

To empirically validate our pre-training strategy, we computed the Pearson correlation coefficients (PCC) between the speech envelope and mel-spectrogram representations for all matched samples in the SparrKULee dataset. A mean PCC of

0.33 was observed across the mel-frequency bands, indicating a consistent positive correlation. This supports the use of the utilizing the envelope as an effective initial feature space for pre-training. The full distribution of the correlation coefficients is visualized in Figure 2.

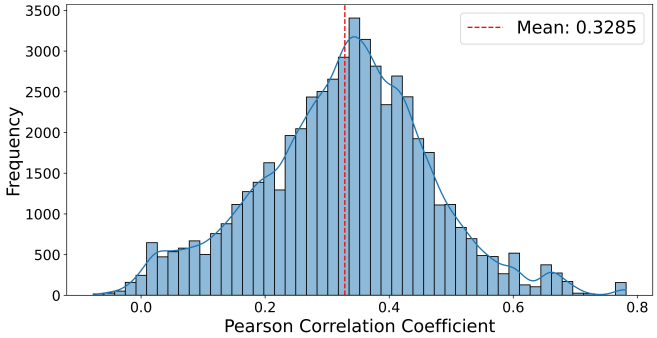

Figure 2: The PCC distribution between speech features of mel-frequency bands and envelope.

4.3 Results and Analysis

We evaluated the representational fidelity using the PCC for acoustic feature recovery. As summarized in Table 1, our proposed HFFN-ID framework demonstrates superior performance in mel-spectrogram reconstruction, achieving a compelling PCC of 0.072. This score notably exceeds the 0.069 correlation achieved by the strongest baseline, SSM2Mel, and represents a 44% relative enhancement over the VLAAI baseline (0.050). This significant performance gain strongly attests to the combined effect of our proposed bi-phase subject-ID modulation, conditioned hierarchical feature fusion, and the two-stage pre-training strategies in HFFN-ID.

Model	PCC(↑)	Improvement
VLAAI	0.050	–
HappyQuokka	0.052	4.0%
ConvConcatNet	0.058	16.0%
ConvConcatNet ensembled	0.063	26.0%
DMF2Mel	0.063	26.0%
SSM2Mel	0.069	38.0%
HFFN-ID (ours)	**0.072**	**44.0%**

Table 1: Mel-spectrogram recovery performance in terms of PCC metric achieved by different models.

To evaluate our proposed HFFN-ID framework, we conducted a rigorous comparison with SSM2Mel, which acts as a leading method for mel-spectrogram reconstruction. As shown in Table 2, our model achieves a superior trade-off between performance and complexity, outperforming SSM2Mel across all the evaluation metrics. Here, higher Structural Similarity Index Measure (SSIM) and lower Mel-Cepstral Distortion (MCD) indicate better reconstruction quality. Specifically, HFFN-ID attains a lower MCD (15.3845 vs. 15.3948), a higher SSIM (0.8258 vs. 0.8256), and an improved PCC (0.0723 vs. 0.0692), confirming its enhanced performance across multiple metrics.

It is worth mentioning that these comprehensive performance gains are achieved with significantly reduced computational overhead, as illustrated by the running time and parameter-size metrics in Table 2; specifically, 1) *on the running time*, there is a significant reduction from 44,149.62 (SSM2Mel) to 17,390.66 seconds (HFFN-ID), representing an approximate 2.5× speedup; and 2) *on the size of parameter space*, our model requires only 4.22M parameters, which has a 38.7% reduction compared to 6.88M in SSM2Mel. These results underscore the success of our lightweight design, where the optimal reconstruction fidelity is maintained while the architecture effectively prunes feature redundancy. This leads to the achievement of better representational power at a substantially lower computational cost.

Model	PCC(↑)	MCD(↓)	SSIM(↑)	Runtime(↓)	Params(↓)
SSM2Mel	0.0692	15.3948	0.8256	44149.62s	6.88M
HFFN-ID	**0.0723**	**15.3845**	**0.8258**	**17390.66**s	**4.22**M

Table 2: Performance comparison in mel-spectrogram reconstruction between SSM2Mel and our model, where 's' and 'M' respectively denote 'second' and 'million'.

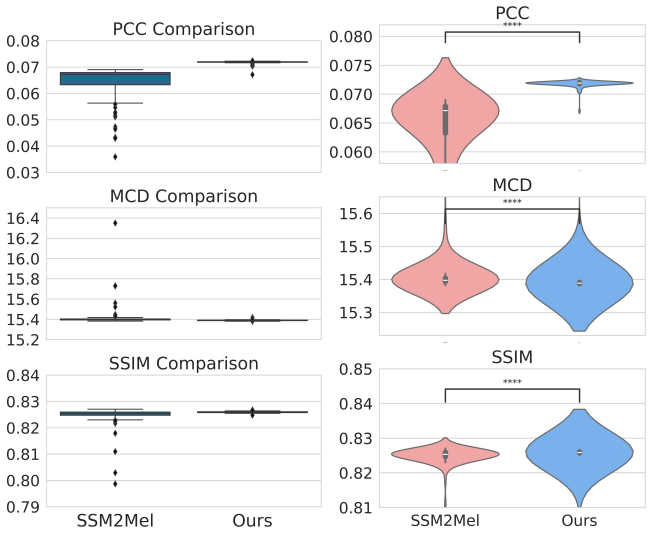

Figure 3: Box plot of mel-spectrogram test results (left) and distribution comparison of the metrics (right).

Beyond aggregate performance metrics, the reconstruction reliability of our model was evaluated by tracking PCC, MCD, and SSIM over 1000 test epochs. The distributions are summarized in Figure 3. Our model achieves markedly greater stability, evidenced by a narrower Inter-Quartile Range (IQR) and a higher median PCC. The baseline model shows a pronounced long tail of poor-performing outliers, suggesting inconsistent performance and periodic failure. In contrast, our model maintains a compact, outlier-sparse distribution that MCD values are tightly grouped near the lower bound, and SSIM remains consistently high with

minimal dispersion. These results confirm enhanced robustness and generalization. Overall, the BiPSM component and envelope-based pre-training not only improve accuracy but also deliver more stable and predictable performance.

Figure 4 presents a per-sample comparison on 100 randomly selected test cases, demonstrating our model's advantage over SSM2Mel in reconstruction fidelity. Our method demonstrates a systematic improvement, with its error distribution (red) sharply concentrated at lower MSE values (peak ≈ 0.010) and its SSIM distribution shifted toward the superior range of [0.85, 0.95]. In contrast, SSM2Mel yields broader, more dispersed distributions for both metrics. This consistent advantage is corroborated by scatter plots, where a majority of samples cluster on the favorable side of the $y = x$ identity line, confirming that our model delivers superior fidelity for most individual cases. This sample-level improvement supports robust performance in downstream applications such as neural decoding and speech-related feature generation.

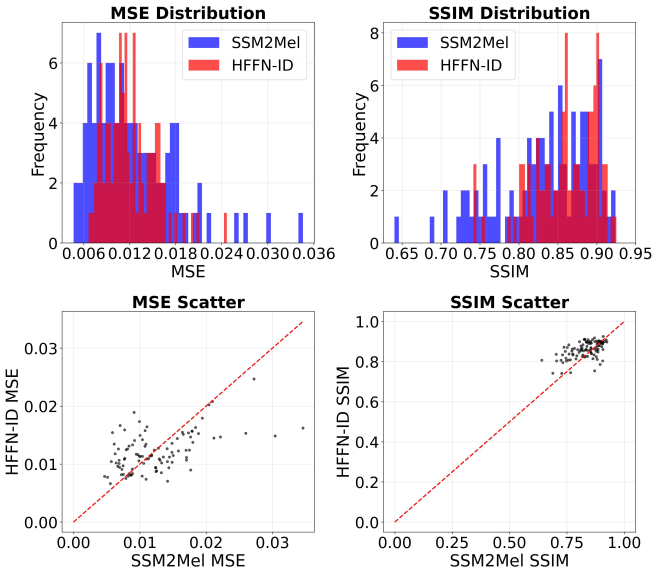

Figure 4: Sample-wise comparison between SSM2Mel (blue) and HFFN-ID (red) in mel-spectrogram reconstruction in terms of MSE (left) and SSIM (right) metrics.

4.4 Ablation and Extended Studies

To evaluate the contribution of each component in HFFN-ID, we performed a comprehensive ablation study. Each component was removed during pre-training, and the model was subsequently fine-tuned to isolate its impact. The absence of the CHF component caused the largest performance decline (8.2%), highlighting its importance in bridging heterogeneous feature representations. Removing the BiPSM component led to a 7.5% reduction, confirming the necessity of explicit subject identity for personalized decoding. Omitting the envelope-based pre-training resulted in a 7.6% drop, underscoring its role in ensuring temporal stability and faster convergence. The best performance (PCC: 0.0723) is achieved only when all three components are integrated, demonstrating their synergistic effect.

Model	PCC(↑)	Improvement
w/o BiPSM	0.0669	-7.5%
w/o CHF	0.0664	-8.2%
w/o Pretrain	0.0668	-7.6%
HFFN-ID(ours)	**0.0723**	–

Table 3: Ablation studies on the impact of each network component in the proposed HFFN-ID framework.

The advantage of our proposed HFFN-ID framework also extends to the speech envelope recovery task, as summarized in Table 4. Benefiting from the mel-spectrogram-based pre-training, our model achieves a peak PCC of 0.211. Even without pre-training, it attains a PCC of 0.209, maintaining a consistent lead over SSM2Mel (0.208). This result underscores the robustness of our framework in tracking temporal amplitude dynamics.

Model	PCC(↑)	Improvement
VLAAI	0.161	–
HappyQuokka	0.185	14.9%
DMF2Mel	0.204	26.7%
SSM2Mel	0.208	29.2%
HFFN-ID(ours) w/o Pretrain	**0.209**	**29.8%**
HFFN-ID(ours)	**0.211**	**31.1%**

Table 4: Envelope recovery performance in terms of PCC metric achieved by different models. ConvConcatNet results are not available for envelope recovery and are therefore not reported.

5 Conclusion

In this paper, we introduced a HFFN-ID framework for EEG-based high-fidelity mel-spectrogram reconstruction, effectively addressing the primary challenges of subject variability and the gap between heterogeneous feature representations. Our main contributions are in two aspects, a Binary-Phase Subject-ID Modulation for explicitly adapting feature learning on subject identity, and a condition-guided Hierarchical Fusion for enforcing the consistency between representations at varying levels. As a secondary contribution, an envelope-based pre-training strategy was introduced for stabilizing temporal alignment and accelerating convergence. Experimental results depicted that HFFN-ID achieved a compelling PCC of 0.0723, which constitutes a substantial 44% improvement over the VLAAI baseline. Importantly, this superiority was attained with improved efficiency with 38.7% parameter size reduction and 2.5× running time acceleration compared to SOTA SSM2Mel model. Furthermore, rigorous statistical analysis validated the exceptional robustness and reliability of our model. These results demonstrate that effectively modeling subject identity and leveraging efficient hierarchical feature fusion are vital for advancing neural speech decoding performance, promoting the application of speech BCIs in assisting individuals with speech impairments.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported in part by the 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang Province (2025C04001), the National Natural Science Foundation of China (62571171), and the MoE Humanities and Social Sciences Project (24YJCZH225).

Contribution Statement

Chi Huang and Zhaohu Liu contributed equally to this work.

References

[Accou *et al.*, 2023] Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Scientific Reports*, 13(1):812, 2023.

[Accou *et al.*, 2024] Bernd Accou, Lies Bollens, Marlies Gillis, Wendy Verheijen, Hugo Van Hamme, and Tom Francart. SparrKULee: A speech-evoked auditory response repository from KU Leuven, containing the EEG of 85 participants. *Data*, 9(8):94, 2024.

[Akbari *et al.*, 2019] Hassan Akbari, Bahar Khalighinejad, Jose L Herrero, Ashesh D Mehta, and Nima Mesgarani. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1):874, 2019.

[Alharbi and Alotaibi, 2024] Yasser F Alharbi and Yousef A Alotaibi. Decoding imagined speech from EEG data: A hybrid deep learning approach to capturing spatial and temporal features. *Life*, 14(11):1501, 2024.

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

[Chan *et al.*, 2011] Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage*, 54(4):3028–3039, 2011.

[Dauphin *et al.*, 2017] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, pages 933–941, 2017.

[Fan *et al.*, 2025a] Cunhang Fan, Sheng Zhang, Jingjing Zhang, Enrui Liu, Xinhui Li, Gangming Zhao, and Zhao Lv. DMF2Mel: A dynamic multiscale fusion network for EEG-driven mel spectrogram reconstruction. In *Proceedings of ACM International Conference on Multimedia*, pages 6977–6985, 2025.

[Fan *et al.*, 2025b] Cunhang Fan, Sheng Zhang, Jingjing Zhang, Zexu Pan, and Zhao Lv. SSM2Mel: State space model to reconstruct mel spectrogram from the EEG. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[Fedorenko *et al.*, 2024] Evelina Fedorenko, Anna A Ivanova, and Tamar I Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, 2024.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, pages 1–16, 2024.

[Gu *et al.*, 2022] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Inernational Conference on Learning Representations*, pages 1–12, 2022.

[Guo *et al.*, 2023] Yina Guo, Ting Liu, Xiaofei Zhang, Anhong Wang, and Wenwu Wang. End-to-end translation of human neural activity to speech with a dual–dual generative adversarial network. *Knowledge-Based Systems*, 277:110837, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[Hu *et al.*, 2026] Renzhi Hu, Ting Luo, Gangyi Jiang, Leiming Liu, Yeyao Chen, Haiyong Xu, and Zhouyan He. LatentDark: Reflectance guided latent diffusion model for low-light image enhancement. *Signal Processing*, 238:110125, 2026.

[Huang *et al.*, 2023] Gan Huang, Zhiheng Zhao, Shaorong Zhang, Zhenxing Hu, Jiaming Fan, Meisong Fu, Jiale Chen, Yaqiong Xiao, Jun Wang, and Guo Dan. Discrepancy between inter-and intra-subject variability in EEG-based motor imagery brain-computer interface: Evidence from multiple perspectives. *Frontiers in Neuroscience*, 17:1122661, 2023.

[Kunz *et al.*, 2025] 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, et al. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188(17):4658–4673, 2025.

[Lee *et al.*, 2023] Young-Eun Lee, Seo-Hyun Lee, Sang-Ho Kim, and Seong-Whan Lee. Towards voice reconstruction from EEG during imagined speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6030–6038, 2023.

[Lee *et al.*, 2024] Jihwan Lee, Aditya Kommineni, Tiantian Feng, Kleanthis Avramidis, Xuan Shi, Sudarsana Kadiri, and Shrikanth Narayanan. Toward fully-end-to-end listened speech decoding from EEG signals. In *Interspeech*, pages 1500–1504, 2024.

[Lee *et al.*, 2025] Jihwan Lee, Kevin Huang, Kleanthis Avramidis, Simon Pistrosch, Monica Gonzalez-Machorro, Yoonjeong Lee, Björn Schuller, Louis Goldstein, and Shrikanth Narayanan. Articulatory feature prediction from surface EMG during speech production. In *Interspeech*, pages 320–324, 2025.

[Li *et al.*, 2024] Hao Li, Yuan Fang, Xueliang Zhang, Fei Chen, and Guanglai Gao. Cross-attention-guided WaveNet for EEG-to-MEL spectrogram reconstruction. In *Proc. Interspeech 2024*, pages 2620–2624, 2024.

[Liang *et al.*, 2024] Aobo Liang, Xingguo Jiang, Yan Sun, Xiaohou Shi, and Ke Li. Bi-mamba+: Bidirectional mamba for time series forecasting. *arXiv preprint arXiv:2404.15772*, 2024.

[Ma *et al.*, 2025] Chen Ma, Yue Zhang, Yina Guo, Xin Liu, Hong Shangguan, Juan Wang, and Luqing Zhao. Fully end-to-end EEG to speech translation using multi-scale optimized dual generative adversarial network with cycle-consistency loss. *Neurocomputing*, 616:128916, 2025.

[Mahapatra and Bhuyan, 2023] Nrushingh Charan Mahapatra and Prachet Bhuyan. EEG-based classification of imagined digits using a recurrent neural network. *Journal of Neural Engineering*, 20(2):026040, 2023.

[Metzger *et al.*, 2022] Sean L Metzger, Jessie R Liu, David A Moses, Maximilian E Dougherty, Margaret P Seaton, Kaylo T Littlejohn, Josh Chartier, Gopala K Anumanchipalli, Adelyn Tu-Chan, Karunesh Ganguly, et al. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(1):6510, 2022.

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

[O'sullivan *et al.*, 2015] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, 25(7):1697–1706, 2015.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Piao *et al.*, 2023] Zhenyu Piao, Miseul Kim, Hyungchan Yoon, and Hong-Goo Kang. Happyquokka system for ICASSP 2023 auditory EEG challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–2, 2023.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Sato *et al.*, 2024] Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of EEG data. *arXiv preprint arXiv:2407.07595*, 2024.

[Sharma and Meena, 2024] Ramnivas Sharma and Hemant Kumar Meena. Emerging trends in EEG signal processing: A systematic review. *SN Computer Science*, 5(4):415, 2024.

[Shazeer, 2020] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[Sun *et al.*, 2024] Sinan Sun, Longxiang Zhang, Bo Wang, Xihong Wu, and Jing Chen. Representation of articulatory features in EEG during speech production tasks. In *IEEE International Symposium on Chinese Spoken Language Processing*, pages 219–223, 2024.

[Van Dyck *et al.*, 2023] Bob Van Dyck, Liuyin Yang, and Marc M Van Hulle. Decoding auditory EEG responses using an adapted wavenet. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–2, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11, 2017.

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

[Xu *et al.*, 2024] Xiran Xu, Bo Wang, Yujie Yan, Haolin Zhu, Zechen Zhang, Xihong Wu, and Jing Chen. ConvConcatNet: a deep convolutional neural network to reconstruct mel spectrogram from the EEG. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, pages 113–114, 2024.