VIBE: VISION TRANSFORMER BASED EXPERTS NETWORK FOR SSVEP DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Steady-state visual evoked potential based brain–computer interfaces (SSVEP-BCIs) have attracted wide attention for their high information transfer rate (ITR) and non-invasiveness. However, existing deep learning methods for SSVEP-BCI decoding have reached a performance bottleneck, as they struggle to fully extract the complex neural signal features required for robust performance. Motivated by advances in vision and time series modeling, here we present a **VI**sion Transformer **B**ased **E**xpert network (VIBE), a multistage deep learning framework for SSVEP classification. VIBE integrates a Vision Transformer (ViT) module to generate rich spatiotemporal representations with data and network enhancement modules in a decoder for frequency recognition. We evaluate VIBE on two large benchmark datasets, including the Benchmark and the BETA dataset spanning 105 subjects. Notably, with just 0.4 seconds of stimulation, our VIBE achieves an ITR of 263.8 bits per minute (bpm) and 202.7 bpm on the Benchmark and BETA datasets, respectively. Experimental results demonstrate that VIBE consistently outperforms state-of-the-art baselines in offline experiments, highlighting its effectiveness as a high-performance decoding strategy for SSVEP-BCIs.

1 INTRODUCTION

Brain–computer interfaces (BCIs) provide a direct communication pathway between the brain and external devices, enabling interaction without relying on neuromuscular activity. BCIs have emerged as effective tools for augmentative communication and human-machine interaction, with broad potential applications ranging from neuroprosthesis (Willett et al., 2021; 2023) to the next-generation form of human-computer interaction (Gao et al., 2025). Among noninvasive paradigms, steady-state visual evoked potential based BCIs (SSVEP-BCIs) stand out for their non-invasiveness, high ITR, robustness, and scalability. SSVEPs are frequency-tagged neural responses that can be evoked by periodic visual stimulation, including flickering squares, reversing checkerboards, and moving gratings, and they are elicited over occipital cortex at the stimulation frequency and its harmonics. These frequency-tagged responses exhibit a high signal-to-noise ratio (SNR), enabling SSVEP-BCIs to implement high-speed spellers (Chen et al., 2015b), robotic control, and smart home systems. However, achieving high decoding accuracy under short time windows remains challenging in learning effective neural representations from noisy, data-constrained EEG recordings with complex spatiotemporal and spectral dynamics.

Advancements in SSVEP-BCI decoding have been driven by both traditional linear methods and emerging deep learning models. Early approaches such as canonical correlation analysis (CCA) (Bin et al., 2009) and its filter-bank extension (FBCCA) Chen et al. (2015a) established training-free plug-and-play frequency recognition, while subsequent training-based methods like task-related component analysis (TRCA) Nakanishi et al. (2017) and task-discriminant component analysis (TDCA) (Liu et al., 2021b) designed sophisticated spatial filters using individually calibrated data to significantly boost the decoding performance. However, these linear methods remain limited in capturing the nonlinear and hierarchical patterns of EEG signals. To overcome this, convolutional neural networks (CNNs) and, more recently, Transformer-based models have been introduced to learn richer spatio-temporal representations directly from data, showing notable improvements over baselines (Li et al., 2020; Song et al., 2022). Despite these advances, challenges remain in jointly exploiting local inductive biases and global dependencies, motivating the development of new architectures tailored for high-throughput SSVEP-BCIs.

To overcome these limitations, we introduce **VIBE** (Vision Transformer Based Experts Network), a multistage framework that unifies transformer-based sequence modeling with expert-driven specialization. VIBE employs a ViT module to perform temporal generation, expanding short input sequences into richer representations that preserve multi-scale temporal dependencies. It integrates a Mixture of Experts (MoE) decoder, where experts specialize in different subband–channel–temporal dynamics, and a load-balancing loss ensures diverse expert utilization for better generalization. On top of these architectural innovations, VIBE employs a staged training scheme that progressively pretrains and fine-tunes the ViT-based temporal generation and MoE-based decoding modules, adapting from population data to subject-specific dynamics. It further integrates data augmentation strategies, including temporal stitching, channel chunk shuffling, random temporal cropping, and decorrelation, to regularize training and enhance representation learning for EEG.

We evaluate VIBE on two large benchmark datasets, the Benchmark and the BETA datasets, spanning 105 subjects. Results show that VIBE consistently outperforms both classical and deep learning baselines, achieving higher accuracy and ITR under short time windows. These findings establish VIBE as an effective decoding strategy for high-throughput SSVEP-BCIs.

In summary, our main contributions are threefold:

1. A novel hybrid framework that combines ViT-based temporal generation with MoE-based subband-channel-temporal decoding enhancement for SSVEP classification.
2. A staged training scheme that progressively adapts temporal generation and decoding modules from population to subject-specific data.
3. A suite of data augmentation methods designed for EEG, improving robust representation learning from limited and noisy neural data.

2 Related Work

Traditional methods. Early research focused on traditional frequency recognition methods, which can be broadly divided into training-free and training-based approaches. Canonical correlation analysis (CCA) (Bin et al., 2009) and its filter-bank extension (FBCCA) (Chen et al., 2015a) became widely used due to their plug-and-play traning-free capability. Prior studies using individually calibrated data introduced multiset CCA (MsetCCA) (Zhang et al., 2014), L1-regularized multiway CCA (L1MCCA) (Zhang et al., 2013), and extended CCA (eCCA) (Nakanishi et al., 2014), which improved robustness by leveraging richer reference structures and regularization strategies. More sophisticated spatial filter based methods were developed to further boost accuracy. Task-related component analysis (TRCA) (Nakanishi et al., 2017) significantly improved SSVEP decoding by maximizing trial-to-trial reproducibility. To further address redundancy in TRCA's ensemble design, task-discriminant component analysis (TDCA) (Liu et al., 2021b) eliminated the training of spatial filters class by class and leveraged spatio-temporal neural dynamics, making it a state-of-the-art method for enabling high-speed brain spellers. Despite these advances, traditional methods remain linear and limited in their ability to capture nonlinear, hierarchical representations of EEG.

Deep learning methods. Motivated by these limitations, recent research has turned toward learning complex representations from noisy signals in an end-to-end manner. Convolutional neural networks (CNNs) enabled data-driven feature extraction, analogous to filtering in EEG signal processing, and advances such as convolutional correlation analysis (ConvCA) (Li et al., 2020) and deep neural network classifiers (Guney et al., 2021) have surpassed linear baselines. Extensions incorporated fixed and dynamic template networks (Xiao et al., 2022), bidirectional Siamese correlation networks (Zhang et al., 2022), and multiscale CNNs with squeeze-and-excitation blocks (Jin et al., 2024). More recently, Transformer-based architectures leveraging self-attention to capture long-range temporal dependencies have been applied to EEG (Song et al., 2022; Wan et al., 2023), including SSVEPformer (Chen et al., 2023), DG-Conformer (Liu et al., 2024), SSVEPPoolformer (Li et al., 2025a) and MTSNet (Lan et al., 2025) for cross-subject SSVEP classification. Hybrid approaches such as TRCA-Net (Deng et al., 2023) and discriminant compacted network (Li et al., 2025b) combine spatial filters with neural networks, while ConsenNet (Zhang et al., 2024) leverage a teacher-student framework to further improve performance. Most recently, Mamba-based models such as SUMamba (Dong et al., 2026) integrated multi-scale feature fusion to facilitate classification. However, CNNs remain limited in capturing global dependencies, and Transformers often

neglect inductive biases specific to EEG, leaving the representation learning problem unresolved for high-throughput SSVEP-BCIs.

3 METHOD

We first define the notation used throughout this work. The multi-channel EEG signal is represented as $X \in \mathbb{R}^{B\times C\times T}$, where B denotes the number of filter banks, C the number of EEG electrodes (channels), and T the total number of sampled timestamps. In our experiments, we consider $B = 3$ sub-bands extracted by band-pass filtering, and $C = 9$ channels selected from classical montage for SSVEP classification (Chen et al., 2015b).

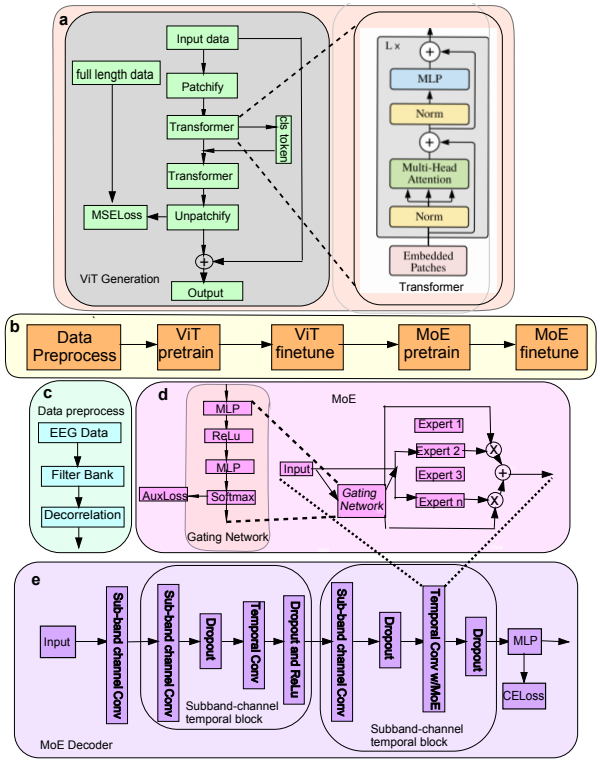

Figure 1: Framework of our proposed VIBE method. In the ViT Generation (a) Pretrain stage, the ViT module learns to generate extended temporal representations from short SSVEP EEG segments. In the ViT Fine-tune stage, subject-specific data are fed into the pretrained ViT, and both train/test data are concatenated with the regenerated temporal data. In the MoE Decoder (e) Pretrain stage, the concatenated data from all subjects, together with three forms of augmented data, are used to train the decoder. Finally, in the Decoder Fine-tune stage, only subject-specific data (without augmentation) are used for calibration. The decoder output corresponds to the classification result.

3.1 VIT-BASED TEMPORAL GENERATION.

The first component of our framework is a ViT (Dosovitskiy et al., 2021) adapted for temporal sequence generation. We represent the multi-band EEG data as an image tensor $X \in \mathbb{R}^{B\times C\times T_{\text{in}}}$, where the three dimensions correspond to sub-bands ($B = 3$), channels (C), and time samples (T_{in}). Following the ViT formulation, the signal is partitioned into non-overlapping patches of size (B, C, t), where t denotes a small temporal window (e.g., $t = 10$ samples ≈ 0.04 s). Each patch is then flattened and linearly projected into a latent embedding, forming a sequence of tokens. Like standard ViTs, we include positional embeddings added to the patchified embedding. A learnable class token is attached to the sequence, which serves to expand the representation to match the target output length T_{out}. This latent sequence is then processed by a transformer-based decoder

(also implemented as a ViT), and the resulting patches are unpatchified to reconstruct the output sequence $\hat{X} \in \mathbb{R}^{B \times C \times T_{\text{out}}}$, with $T_{\text{out}} > T_{\text{in}}$. The model was optimized using the mean squared error (MSE) loss. We clip the extended sequence of shape $\mathbb{R}^{B \times C \times (T_{\text{out}} - T_{\text{in}})}$, which represents the newly generated temporal samples. This generated segment was concatenated with the original input X, forming the final output representation.

3.2 MoE with Channel-Subband Temporal Decoder

The second part of our framework is a decoder that jointly models subband, channel, and temporal dependencies using a MoE design.

3.2.1 MoE

MoE (Shazeer et al., 2017) mechanism is designed to increase model capacity while keeping computational cost manageable through sparse activation. MoE has also been successfully applied to EEG decoding tasks (Yang et al., 2025). Instead of applying a single shared transformation to all inputs, an MoE layer maintains a set of E experts $\{f_1, f_2, \ldots, f_E\}$, each parameterized as a learnable function (e.g., convolutional filters in our case). For each input token x, a gating network produces a probability distribution over experts, and only the top-k experts are selected to process the input. Specifically, we take the mean over temporal domain as the input of the gating network, where the network consist of two layers of MLP and one ReLU as activation. The final output is then obtained as a weighted combination of the selected experts' outputs, where the gating scores act as mixture coefficients. This strategy enables different experts to specialize on distinct temporal or spectral patterns in the EEG signal, enhancing both representation power and generalization.

Auxiliary Load-Balancing Loss. To encourage balanced utilization of experts and prevent overfitting, we introduce an auxiliary load-balancing loss. For each MoE layer, the gating network computes a probability distribution over E experts for each input token. Let $G \in \mathbb{R}^{B \times E}$ denote the gate probabilities for a batch of B inputs, with G_{ij} representing the probability of assigning input i to expert j. The mean usage of each expert is then $\bar{u}_j = \frac{1}{B} \sum_{i=1}^{B} G_{ij}$. We define the auxiliary loss as the Kullback-Leibler (KL) divergence between the mean expert usage and a uniform distribution:

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{aux}} \, \text{KL}\Big(\log(\bar{u}) \,\|\, U \Big)$$

where U is a uniform vector of length E, and λ_{aux} is a weighting coefficient. This loss encourages all experts to be used approximately equally, preventing collapse onto a small subset of experts and improving generalization.

3.2.2 Channel-Subband fused Temporal Decoder

The decoder first integrates information across sub-bands and channels by treating the data as a combined subband-channel dimension of shape $(B * C, 1, T)$. A convolutional layer with kernel $(1, 1)$ is applied to extract higher-level spectral-channel features. Next, a subband-channel-temporal block is applied that consists of two convolutional layers: one over the subband-channel dimension and one over the temporal dimension. This block pools information separately from the spectral-channel and temporal dimensions. The block is repeated twice to progressively capture richer patterns across channels, sub-bands, and time. Nonlinearities and dropout are applied between layers for regularization. Finally, the extracted features are flattened and passed through a fully connected layer to produce the class logits. In our model, we replace the last temporal convolution layer with a MoE layer, where experts in the networks employ the original temporal conv.

3.3 Data Augmentation

In the following section, the dataset is denoted as $X \in \mathbb{R}^{S \times B \times N \times M \times T \times C}$, where S is the number of subjects, B the number of subbands, N the number of trials, M the number of targets, T the temporal length, and C the number of EEG channels.

Cross-Subject Temporal Stitching As a generalization of (Lotte, 2015), the EEG signals are first divided into short temporal segments, or *chunks*, across the time dimension. For each chunk, we randomly select a segment of the same duration from any subject and trial in the dataset, preserving the original subband, channel, and target labels. The selected segments are combined across time to form a new synthetic trial, maintaining the original subband, channel, and target structure. This approach allows the creation of entirely new temporal patterns by sampling from different subjects and trials, rather than modifying the original trial. Namely, $\oplus_i X_{s_i,:,n_i,:,i\tau:\tau(i+1),:} \in \mathbb{R}^{1\times B\times 1\times M\times T\times C}$is a generated piece of data of, where $\{0, \tau, 2\tau, ...\}$ is the time chunk sequence and each s_i, n_i is randomly selected among S, N,

Channel Chunk Shuffle Given a random subject, input data is first divided into consecutive chunks along the time dimension. For each chunk, with a certain probability, two channels are randomly selected and swapped, while all other dimensions—including subbands, trial, target labels remain unchanged. Explicitly, given a subject's trial and target,if time chunks $(t_i\tau, t_j\tau)$ are selected to shuffle with $(c_{i_0}, c_{i_1}), (c_{j_0}, c_{j_1}) \in S_c$ are transpositions corresponding to t_i, t_j as channel swap, $X_{s,:,n,m,t_i\tau:t_{i+1}\tau,c_{i_0}}$ replaces the channel c_{i_1} and similarly for $t_j, (c_{j_0}, c_{j_1})$.

Random Temporal Crop Inspired by (Liu et al., 2021b), a *Random Temporal Crop (RTC)* augmentation is utilized to increase temporal diversity in the training data. For any chosen subject, we preserve the original trial, target label, subband, and channel structure.With a given probability, we randomly select a short segment of time (e.g. 0.03s) from the data, keeping only the latter portion and discarding the former. The cropped segment is zero-padded at the end to restore the trial to its original temporal length.

Channel Decorrelation We adopt a covariance-based whitening procedure across channels, conditioned on each subject and subband. For each subject and subband, we first compute the mean trial across all training trials to obtain a representation. This mean trial is used to estimate the channel covariance, from which a whitening matrix is derived (He & Wu, 2019; Liu et al., 2021a). The whitening matrix is applied to both training and test data, effectively reducing subject, trial-level variability while preserving the temporal and target-related structure of the signals. The decorrelation procedure emphasizes stable patterns across different subject and subband.

3.4 Transfer Learning.

Following the transfer learning strategy of (Guney et al., 2021), we adopt a staged training procedure to strengthen representation ability. Our model comprises two main components: a ViT encoder for temporal length generation and a MoE-based decoder for subband-channel and temporal integration. Thus, the transfer learning process is added to these component.

In the first stage, the ViT encoder is trained in a generative manner, reconstructing the temporal sequence from shorter inputs. In the second stage, this encoder is fine-tuned separately for each subject, where the global model parameters are re-initialized and adapted using only subject-specific data. The decoder is trained in a similar two-step fashion: first, a global MoE decoder is optimized using the pooled training data across all subjects, and subsequently, a subject-specific fine-tuning step is applied to adapt the decoder to individual variability.

4 Experiments

4.1 Dataset

The experiments were carried out on two public 40-target SSVEP datasets: the Benchmark dataset (Wang et al., 2016) and the BETA dataset (Liu et al., 2020). Both datasets employed the joint frequency and phase modulation (JFPM) method to encode target stimuli. The data acquisition equipment for the Benchmark and BETA datasets is identical; however, the Benchmark dataset was collected in a controlled laboratory environment within an electromagnetic shielding room, whereas the BETA dataset was recorded in a more naturalistic setting, reflecting real-world conditions. All experiments, including comparisons with state-of-the-art methods, were performed on these two

datasets. This allows us to evaluate the performance of our decoding approach under both controlled and realistic acquisition conditions.

4.2 PREPROCESSING

The same preprocessing pipeline was applied to both datasets. Nine electrodes (Pz, PO5, PO3, POz, PO4, PO6, O1, Oz, and O2) were selected for analysis. The EEG signals were downsampled to 250 Hz. To account for visual response latency, we considered delays of 0.14 s for Benchmark and $0.13s$ for BETA, consistent with previous studies (Chen et al., 2015b). For each trial, data segments of length t seconds were extracted in the time windows $[0.14, 0.14 + t]$ s and $[0.13, 0.13 + t]$ s after stimulus onset for Benchmark and BETA, respectively.

We apply a filter-bank approach as a preprocessing step to enhance SSVEP signals (Chen et al., 2015a). Data passes through three band-pass filters with frequency ranges $(8N, 90)$ Hz, where $N = 1, 2, 3$, and filtered signals are concatenated along the sub-band dimension. This procedure captures multiple harmonics and improves the signal representation for subsequent decoding.

4.3 BASELINE MODELS

Deep Learning Models. **DNN** (Guney et al., 2021) is a dense convolutional neural network that processes time-series data and incorporates a fine-tuning stage to boost performance. **SSVEPformer** (Chen et al., 2023) is a transformer-based neural network that takes complex spectra as input, leveraging a transformer encoder and fully connected layer to extract phase and frequency features. **TRCANet** (Deng et al., 2023) applies TRCA-based spatial filtering to the input data, followed by a DNN for feature learning.

Traditional Models. **TDCA** (Liu et al., 2021b) addresses the redundancy of stimulus-specific spatial filters in TRCA and the underutilization of temporal information. It enhances the performance of individually calibrated SSVEP-BCIs by learning task-discriminative spatiotemporal components. **TRCA** (Nakanishi et al., 2017) derives spatial filters by maximizing SSVEP reproducibility across trials, while eTRCA extends this by ensembling filters across all frequencies. **eCCA** (Nakanishi et al., 2014) introduces a combination of spatial filters derived from canonical correlation analysis (CCA) and employs a user-specific target identification algorithm based on individual calibration data. **msTRCA** (Wong et al., 2020) extends TRCA with a multi-stimulus learning scheme that leverages data from both target and non-target stimuli.

4.4 EXPERIMENTAL SETUP

We employed k-fold cross-validation, with $k = 6$ for Benchmark and $k = 4$ for BETA. For each subject, one block of EEG data was designated as the test set, while the remaining blocks were used for training within that fold. All training follows the four-stage procedure: ViT generative pretraining, ViT subject-specific fine-tuning, MoE decoder pretraining, and MoE subject-specific fine-tuning. Further implementation details are provided in the relevant subsectionA.1 of the Appendix.

5 RESULT

To evaluate the performance of algorithms among different data lengths, we report both classification accuracy and ITR. The ITR, measured in bpm, is defined as (Wolpaw et al., 2002):

$$\mathrm{ITR}(P, T) = \left(\log_2 M + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{M - 1} \right) \frac{60}{T},$$

Here, M denotes the number of target classes, P denotes the classification accuracy, and T (in seconds) represents the total selection duration, including gaze time and a fixed gaze shift of 0.5 s.

Figures 2 and 3 present the average classification accuracy and ITR of the proposed VIBE network evaluated on Benchmark and BETA across different data lengths. At the shortest data length (0.2 s), VIBE achieved the largest accuracy advantage over all other methods, highlighting its superior capability for rapid SSVEP decoding (Benchmark: 65.5% vs. 58.8%; BETA: 53.7% vs. 46.2%).

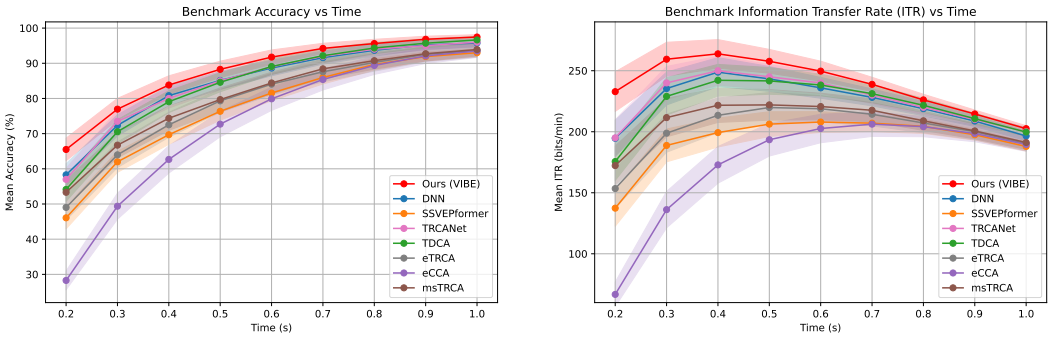

Figure 2: The left panel shows the mean classification accuracy, and the right panel shows the mean information transfer rate (ITR) across all 35 subjects in the Benchmark dataset. Shaded regions indicate the standard errors for subjects.

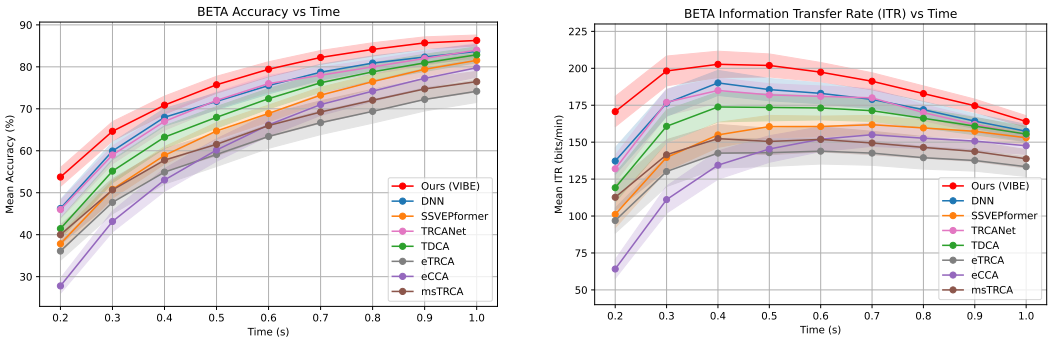

Figure 3: The left panel shows the mean classification accuracy, and the right panel shows the mean ITR across all 70 subjects in the BETA dataset. Shaded regions indicate the standard errors for subjects.

The maximum ITR for VIBE was observed at 0.4 s, reaching 263.8 ± 11.7 bpm for Benchmark and 202.7 ± 8.9 bpm for BETA, exceeding the corresponding values of the DNN baseline (248.8 ± 11.8 bpm and 190.1 ± 8.4 bpm, respectively). For a 1 s data length, VIBE maintained superior classification performance compared with DNN (Benchmark: $97.4 \pm 0.7\%$ vs. $95.7 \pm 1.1\%$; BETA: $86.3 \pm 1.3\%$ vs. $83.7 \pm 1.6\%$). Collectively, these results demonstrated that VIBE effectively decodes SSVEP responses across a range of time windows, with particularly pronounced benefits under short observation periods.

The performance of each method was evaluated in terms of decoding accuracy and ITR across data lengths. A two-way (method $\times$ data length) repeated-measures ANOVA (Greenhouse–Geisser corrected) revealed a statistically significant interaction between method and data length for both datasets (Benchmark: $F(56, 1904)$, $p < 0.001$; BETA: $F(56, 3864)$, $p < 0.001$). The eight methods included in this analysis correspond to those described in the Experiments Section. The detailed results are provided in Appendix Table 5. These findings indicate that the effect of data length on decoding performance depended on the method used, and vice versa, highlighting significant differences in performance trends across methods and data lengths.

For both Benchmark and BETA, paired t-tests revealed that our proposed VIBE method achieved significantly higher decoding accuracies than the deep learning baseline (DNN) and the traditional method (TDCA) across all evaluated data lengths (all: $p < 0.05$). The details of these results are summarized in Table 6, in Appendix. The advantage of VIBE was especially pronounced at short data lengths (e.g., 0.2 s, Benchmark: VIBE vs. DNN: $p = 1.0 \times 10^{-12}$; VIBE vs. TDCA: $p = 2.2 \times 10^{-14}$; BETA: VIBE vs. DNN: $p = 4.8 \times 10^{-18}$; VIBE vs. TDCA: $p = 3.3 \times 10^{-25}$), demonstrating that our method was more robust under very short EEG segments. As data length increases, all methods converged towards similar performance, but VIBE consistently maintained a significant edge, indicating its effectiveness in both short- and long-window SSVEP decoding.

5.1 Ablation Study

To better understand the contribution of each component in our framework, we conducted an ablation study on the Benchmark dataset with a 0.2 s data length. A brief summary of the ablation is presented in Table 8. The original model achieved an accuracy of **65.5%**.

MoE Removing the MoE module resulted in an accuracy of **64.2%**, highlighting its importance. Further analysis of MoE placement across different layers is provided in Appendix A.3.

ViT regeneration. Removing the ViT regeneration module led to a performance drop to **61.8%**, highlighting its essential role in feature representation. Further exploration of the effect of varying ViT generation time length is provided in Appendix A.3, which indicated that the optimal generation time depended on the input trial length: longer generation times benefited short trials, while shorter generation times were preferable for longer trials.

Data augmentation. Two augmentation strategies were employed: a decorrelation-based augmentation and an additional data generation module (the three methods described in Section 3.3) for the MoE decoder. When only the decoder-specific augmentation was removed, the accuracy decreased to **63.5%**; when both strategies were removed, the accuracy further dropped to **62.1%**. Further analysis of the effect of removing each augmentation is provided in Appendix A.3.

These results confirm that each module contributes positively, with all these three blocks play critical roles, and that the optimal ViT generation time length is data-length dependent.

5.2 Feature Visualization via t-SNE

To explore the reasons behind the superior performance of our model, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton, 2008) to visualize the features learned from the final fully connected layer. We examined only our model, comparing the full version with an ablated version in which the ViT, data augmentation, and MoE were removed, using a data length of 0.6 s. In Appendix Figure 4, dots of the same color in the full model (left) formed more compact and dense clusters than in the ablated model (right). The circled clusters highlighted representative examples. This increased density indicated that the three key modules contributed to generating more discriminative and tightly grouped feature representations.

5.3 Subject-wise ITR Visualization

In Appendix Figure 5, we visualized the ITR for each subject using radar plots to compare our model (VIBE) with baseline methods (DNN and TDCA) at a data length of 0.2 s. Four radar plots were presented, corresponding to the Benchmark and BETA datasets, and comparing VIBE with DNN and TDCA, respectively. Each spoke in the radar plot represented an individual subject, and the distance from the center indicated the ITR value. Across both datasets, VIBE consistently achieved higher ITR values for most subjects compared to the baseline models, illustrating its superior performance and robustness in short-duration SSVEP decoding.

6 Discussion

6.1 Neural underpinnings of the proposed modules

In our study, the effectiveness of ViT-based regeneration can be attributed to the temporal nature of the SSVEP signals. Thus, the regenerated segments effectively extended the temporal window available to the decoder, which provided richer frequency-level information (target frequencies lie within the 8-15 Hz range). The regeneration step ensured that the decoder could access more complete frequency cycles, especially when the original data length was short. Although each ViT patch embedding only encoded a tiny fraction of data length (e.g. 0.04 s), it preserved additional temporal information for the decoder to facilitate classification.

Data augmentation plays a crucial role in enhancing the robustness of the model by introducing variability and simulating real-world scenarios. Several techniques have been implemented in this study, each inspired by physiological and contextual considerations related to EEG signals. First,

the random temporal crop augmentation addresses inter-subject and task-dependent variability in SSVEP latency. While the standard latency is incorporated in the data preprocessing pipeline, the actual latency for each individual can differ, so this augmentation randomly samples temporal segments within each trial to learn latency-tolerant features rather than overfitting to a fixed window, improving generalization. Second, the channel chunk shuffle augmentation is motivated by the dipole-source origin of EEG and distortions from volume conduction and other artifacts, and it randomly shuffles chunks of channels to simulate varied electrode placements and signal quality. This promotes invariance to sensor positioning and improves generalization across hardware setups and individuals. Third, cross-subject temporal stitching encourages the decoder to focus on frequency-level information rather than subject-specific features by stitching trials across subjects, exposing it to diverse temporal patterns and yielding generalized frequency responses that reflect underlying physiology. This reduces overfitting to individual trials and strengthens subject-independent representations, improving generalization to unseen subjects.

The MoE mechanism is particularly valuable in the final temporal convolution layer of the model, where different experts can specialize in learning distinct temporal patterns relevant to specific task targets. Some experts may focus on shorter, rapid temporal responses, while others may specialize in longer, more sustained patterns, enabling the model to better capture the full range of temporal dynamics. This adaptability allows the model to allocate different experts to process different parts of the temporal signal.

6.2 Training and Testing Time Analysis

For VIBE, the two pretraining stages were performed using data from all subjects (excluding test data), while the fine-tuning stage employed data from a single subject. Table 7 in the Appendix summarizes the training times for each stage and the testing time for a single 0.4 s trial, with all experiments conducted on an NVIDIA RTX 4090 GPU. The pretraining stages accounted for the majority of the training time, whereas fine-tuning for a specific subject could be completed in approximately 17 seconds for BETA and 1 minute for Benchmark. The difference in training time between the two datasets is due to the differing number of epochs in each stage. Testing a single trial required less than 1 ms, which is negligible compared to the data duration. These findings indicate that VIBE provides a practical and efficient solution for SSVEP decoding in BCI applications.

6.3 Limitation and Future direction

One limitation of this work is that certain subjects exhibit performance that deviates markedly from the overall distribution, underscoring the need for more generalized approaches capable of handling inter-subject variability. Future investigations could therefore benefit from conducting experiments in alternative evaluation settings, such as performing cross-validation across subjects rather than across trials, or evaluate on other EEG decoding tasks (Song et al., 2024; Jiang et al., 2024; Wang et al., 2023), to provide a more rigorous assessment of generalization. Finally, an important direction for future research is the implementation of online experiments, wherein new patients are directly evaluated, to provide a realistic assessment of the model's effectiveness in practical BCI applications.

7 Conclusion

VIBE enhances SSVEP-BCI decoding by integrating a ViT-MoE architecture, decorrelation strategies, and a novel data augmentation approach that leverages knowledge from multiple subjects in model design. Evaluations on two benchmark datasets demonstrate that VIBE significantly improves both decoding accuracy and ITR. Further analyses indicate that the method effectively incorporates cross-subject information, highlighting its potential as a robust approach for SSVEP decoding. Overall, these results establish VIBE as a strong candidate for SSVEP decoding and support continued progress in BCI research.

8 Reproducibility Statement

All datasets used in this work are publicly available and open-sourced. To facilitate reproducibility, we provide the complete code for our models and experiments alongside the submission. Detailed

descriptions of model architectures, training procedures, and data preprocessing steps are included in the main text, Appendix, ensuring that independent researchers can replicate our results.

9 Ethics Statement

This work adheres to the ICLR Code of Ethics. All datasets used are publicly available and open-sourced. Specifically, the Benchmark and BETA datasets were collected under protocols approved by the respective institutions; For example, the BETA dataset protocol was approved by the Ethics Committee of Tsinghua University (No. 20190002) as reported in the original publication. No additional human subjects were involved in this study. The study focuses on computational modeling and analysis, without potential for harmful applications. All authors have read and complied with the ICLR Code of Ethics.

References

Guangyu Bin, Xiaorong Gao, Zheng Yan, Bo Hong, and Shangkai Gao. An online multi-channel ssvep-based brain–computer interface using a canonical correlation analysis method. *Journal of neural engineering*, 6(4):046002, 2009.

Jianbo Chen, Yangsong Zhang, Yudong Pan, Peng Xu, and Cuntai Guan. A transformer-based deep neural network model for ssvep classification. *Neural Networks*, 164:521–534, 2023.

Xiaogang Chen, Yijun Wang, Shangkai Gao, Tzyy-Ping Jung, and Xiaorong Gao. Filter bank canonical correlation analysis for implementing a high-speed ssvep-based brain–computer interface. *Journal of neural engineering*, 12(4):046008, 2015a.

Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyyping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain-computer interface. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44):201508080, 2015b.

Yang Deng, Qingyu Sun, Ce Wang, Yijun Wang, and S Kevin Zhou. Trca-net: using trca filters to boost the ssvep classification with convolutional neural network. *Journal of Neural Engineering*, 20(4):046005, 2023.

Liuyuan Dong, Chengzhi Xu, Xuyang Wang, Ruizhen Xie, Guangbo Lei, Yimemg Li, and Wanli Yang. Sumamba: A mamba-based deep learning model with multi-scale feature fusion for ssvep classification. *Biomedical Signal Processing and Control*, 112:108376, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Xiaorong Gao, Yijun Wang, Xiaogang Chen, Bingchuan Liu, and Shangkai Gao. Brain–computer interface—a brain-in-the-loop communication system. *Proceedings of the IEEE*, 113(5):478–511, 2025. doi: 10.1109/JPROC.2025.3600389.

Osman Berke Guney, Muhtasham Oblokulov, and Huseyin Ozkan. A deep neural network for ssvep-based brain-computer interfaces. *IEEE transactions on biomedical engineering*, 69(2):932–944, 2021.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2019.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, 2024.

Jing Jin, Xiao Wu, Ian Daly, Weijie Chen, Xinjie He, Xingyu Wang, and Andrzej Cichocki. Squeeze and excitation-based multiscale cnn for classification of steady-state visual evoked potentials. *IEEE Internet of Things Journal*, 2024.

Zhen Lan, Zixing Li, Chao Yan, Xiaojia Xiang, Dengqing Tang, Min Wu, and Zhenghua Chen. Mtsnet: Convolution-based transformer network with multi-scale temporal-spectral feature fusion for ssvep signal decoding. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Chunquan Li, Zhiyuan Liao, Yuxin Cheng, Zitao Wang, Junyun Wu, Ruijun Liu, and Peter X Liu. Ssveppoolformer: An improved poolformer model with the adaptive denoising algorithm for ssvep-eeg signal classification. *IEEE Transactions on Consumer Electronics*, 2025a.

Dian Li, Yongzhi Huang, Ruixin Luo, Lingjie Zhao, Xiaolin Xiao, Kun Wang, Weibo Yi, Minpeng Xu, and Dong Ming. Enhancing detection of ssveps using discriminant compacted network. *Journal of Neural Engineering*, 22(1):016043, 2025b.

Yao Li, Jiayi Xiang, and Thenkurussi Kesavadas. Convolutional correlation analysis for enhancing the performance of ssvep-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):2681–2690, 2020.

Bingchuan Liu, Xiaoshan Huang, Yijun Wang, Xiaogang Chen, and Xiaorong Gao. Beta: A large benchmark database toward ssvep-bci application. *Frontiers in neuroscience*, 14:627, 2020.

Bingchuan Liu, Xiaogang Chen, Xiang Li, Yijun Wang, Xiaorong Gao, and Shangkai Gao. Align and pool for eeg headset domain adaptation (alpha) to facilitate dry electrode based ssvep-bci. *IEEE Transactions on Biomedical Engineering*, 69(2):795–806, 2021a.

Bingchuan Liu, Xiaogang Chen, Nanlin Shi, Yijun Wang, Shangkai Gao, and Xiaorong Gao. Improving the performance of individually calibrated ssvep-bci by task-discriminant component analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1998–2007, 2021b.

Jiawei Liu, Ruimin Wang, Yuankui Yang, Yuan Zong, Yue Leng, Wenming Zheng, and Sheng Ge. Convolutional transformer-based cross subject model for ssvep-based bci classification. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Fabien Lotte. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6):871–890, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Masaki Nakanishi, Yijun Wang, Yu Te Wang, Yasue Mitsukura, and Tzyyping Jung. A high-speed brain speller using steady-state visual evoked potentials. *International Journal of Neural Systems*, 24(6):1450019, 2014.

Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. Enhancing detection of ssveps for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering*, 65(1):104–112, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*, 2024.

Zhijiang Wan, Manyu Li, Shichang Liu, Jiajin Huang, Hai Tan, and Wenfeng Duan. Eegformer: A transformer–based brain activity classification method using eeg signal. *Frontiers in neuroscience*, 17:1148855, 2023.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations*, 2023.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for ssvep-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2016.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, 2021.

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

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

Chi Man Wong, Feng Wan, Boyu Wang, Ze Wang, Wenya Nan, Ka Fai Lao, Peng Un Mak, Mang I Vai, and Agostinho Rosa. Learning across multi-stimulus enhances target recognition methods in ssvep-based bcis. *Journal of neural engineering*, 17(1):016026, 2020.

Xiaolin Xiao, Lichao Xu, Jin Yue, Baizhou Pan, Minpeng Xu, and Dong Ming. Fixed template network and dynamic template network: novel network designs for decoding steady-state visual evoked potentials. *Journal of Neural Engineering*, 19(5):056049, 2022.

Xiaoli Yang, Yurui Li, Jianyu Zhang, Huiyuan Tian, Shijian Li, and Gang Pan. Evomoe: Evolutionary mixture-of-experts for ssvep-eeg classification with user-independent training. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Xinyi Zhang, Shuang Qiu, Yukun Zhang, Kangning Wang, Yijun Wang, and Huiguang He. Bidirectional siamese correlation analysis method for enhancing the detection of ssveps. *Journal of Neural Engineering*, 19(4):046027, 2022.

Xinyi Zhang, Wei Wei, Shuang Qiu, Xujin Li, Yijun Wang, and Huiguang He. Enhancing ssvep-based bci performance via consensus information transfer among subjects. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Yu Zhang, Guoxu Zhou, Jing Jin, Minjue Wang, Xingyu Wang, and Andrzej Cichocki. L1-regularized multiway canonical correlation analysis for ssvep-based bci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(6):887–896, 2013.

Yu Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Frequency recognition in ssvep-based bci using multiset canonical correlation analysis. *International journal of neural systems*, 24(04):1450013, 2014.

A Appendix

A.1 Implementation details

For both ViT stages, the patch temporal length is set to 10 and the hidden dimension to 48, with a dropout rate of 0.8. For all trial durations except 0.2 s, the generative time length is 0.04 s, while for trials of 0.2 s, the generative time length equals 0.2 s. In the MoE decoder stage, the MoE is applied only to the second temporal layer, using 4 experts. The gating network consists of two MLPs, each with an intermediate dimension of 100, and only the top expert is selected. For implementation convenience, the input data is reshaped to $(B \times C, 1, T)$, where B represents the subband dimension. Consequently, all subband-channel layers use a kernel size of $(1, 1)$, and the two temporal convolution layers use a kernel size of 10. The output channels are 200 for all subband-channel layers and 120 for temporal layers. All dropout layers have a probability of 0.1,

except for the layer before the final flattening and MLP, which uses 0.95. During fine-tuning, the dropout probability of all intermediate layers is reset to 0.5.

For data augmentation, cross-subject temporal stitching is performed with a time chunk of 30. For channel chunk shuffling, the chunk size is 20 with a swap probability of 0.3. Random temporal cropping is applied with an activation probability of 0.4, selecting a short segment of 0.02 s to 0.06 s. For each augmentation method, additional data corresponding to 20% of the original dataset size is generated.

The ViT learning rate is set to 0.0001 during general pretraining and 0.00001 during subject-specific fine-tuning, while the decoder learning rate is fixed at 0.0001. The Adam optimizer is used with a weight decay of 0.0001, and an L2 regularization penalty of 0.001 is applied to the decoder. The batch size for both datasets is 32.

The number of training epochs for each stage differs between the Benchmark and BETA datasets, as summarized in Table 1.

Table 1: Stage-wise training epochs for Benchmark and BETA datasets.

Dataset	ViT Pretrain	ViT Transfer	Decoder Pretrain	Decoder Transfer
Benchmark	300	1000	1500	1000
BETA	300	500	500	700

A.2 Decorrelation Details

For each subject s and subband b, we first compute the mean across training trials:

$$\mu^{(s,b)} = \frac{1}{N} \sum_{n \in \text{train trials}} X^{(s,b)}_{n,:,:,:} \in \mathbb{R}^{M \times T \times C}.$$

The aggregated mean activity is reshaped into $\mu^{(s,b)} \in \mathbb{R}^{C \times (M \cdot T)}$, and used to compute the channel covariance matrix:

$$\text{Cov}^{(s,b)} = \frac{1}{M \cdot T} \, \mu^{(s,b)} \left(\mu^{(s,b)}\right)^{\top} \in \mathbb{R}^{C \times C}.$$

The whitening matrix is defined as

$$W^{(s,b)} = \left(\text{Cov}^{(s,b)}\right)^{-\frac{1}{2}},$$

and decorrelation is applied to both training and test data as

$$\tilde{X}^{(s,b)}_{n,:,:,:} = W^{(s,b)} X^{(s,b)}_{n,:,:,:}, \quad \forall n.$$

By using the trial-averaged activity to construct the covariance, this procedure reduces trial-level variability while preserving target and temporal structure, and ensures that whitening is guided by stable patterns rather than noisy single-trial fluctuations.

A.3 Further Ablation Study

MoE The results of applying the MoE module at different layers are summarized in Table 2. When MoE was applied to the first subband–channel layer, the accuracy was **65.0%**, and applying it to the second subband–channel layer yielded **65.2%**. In contrast, applying MoE to the first temporal layer resulted in a lower accuracy of **63.9%**, while placing it on the second temporal layer achieved an accuracy of **64.8%**. These results suggest that the second temporal layer and the subband layer were particularly important for MoE, as they contributed more significantly to improving performance compared to other layers. This highlights the importance of capturing frequency and temporal dynamics at these stages of the model.

MoE Type	Subband 1	Subband 2	Both Temporal	Temporal 1
Accuracy (%)	65.0	65.2	64.8	63.9

Table 2: MoE Ablation Study: Different MoE Configurations. The Subbands listed refer to the Sub-band channel Conv layer in Subband-channel temporal blocks. Tested on Benchmark for 0.2s.

Data Augmentation The impact of different data augmentation strategies is summarized in Table 3. Removing temporal stitching, channel shuffle, or temporal crop resulted in minor decreases in accuracy of around 1%, while omitting decorrelation caused the largest drop to **63.3%**. These results indicate that all augmentation components contributed to model performance, with decorrelation having the most significant effect. Notably, removing all three data generation methods resulted in a 2% decrease, suggesting that each method provided complementary benefits along different dimensions.

Augmentation	No Stitching	No Channel Shuffle	No Temp Crop	No Decorrelation
Accuracy (%)	64.5	64.9	64.7	63.3

Table 3: Data Augmentation Ablation Study: Different Data Augmentation Configurations. Tested on Benchmark for 0.2s.

Effect of ViT Generation Time Length Table 4 shows the impact of varying the ViT generation time length on classification accuracy. On the Benchmark dataset at 0.2 s, performance improved steadily from **64.5%** $(0.04s)$ to **65.5%** (0.2 s). On the BETA dataset, a similar trend was observed, with accuracy increasing from **52.47%** (0.04 s) to **53.74%** (0.2 s). We note that for other data lengths (0.3 s to 1.0 s), the generated data augmentation achieving the best result was fixed at 0.04 s. To illustrate the effect of longer generation times, we performed the same experiment on the 0.4 s data length. The shortest generation time of 0.04 s achieved the highest accuracy (Benchmark: **83.85%**, BETA: **70.85%**), while increasing the generation time gradually decreased performance across other settings by up to 1.5%.

Table 4: Effect of ViT generation time length on classification accuracy (%).

Dataset	0.04 s	0.08 s	0.12 s	0.16 s	0.20 s
Benchmark (0.2 s)	64.50	64.91	65.28	65.01	**65.50**
BETA (0.2s)	52.47	52.64	52.72	53.26	**53.74**
Benchmark (0.4 s)	**83.85**	83.14	82.60	82.64	82.71
BETA (0.4s)	**70.85**	70.50	69.96	69.17	69.50

A.4 Supplementary Tables and Figures

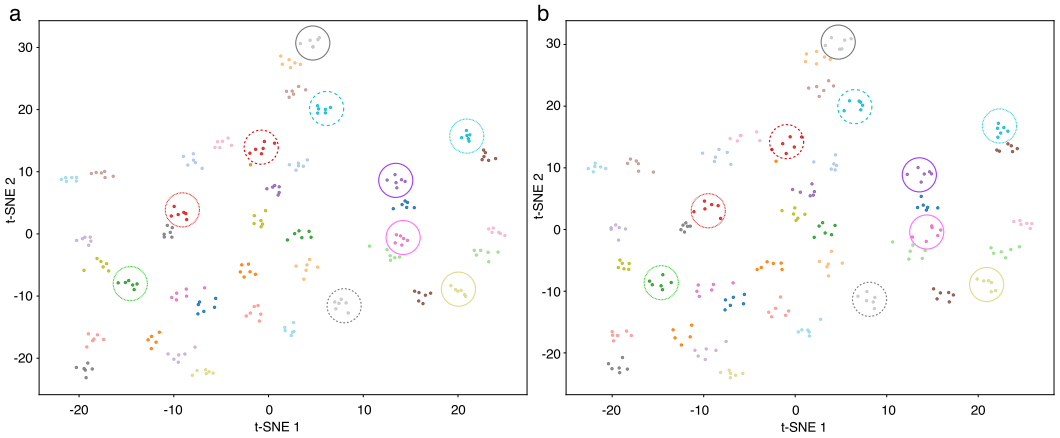

Figure 4: t-SNE visualization of learned features for a representative subject from the Benchmark dataset, using a data length of 0.6 s. Left: full model; Right: ablated model (without ViT, data augmentation, or MoE).

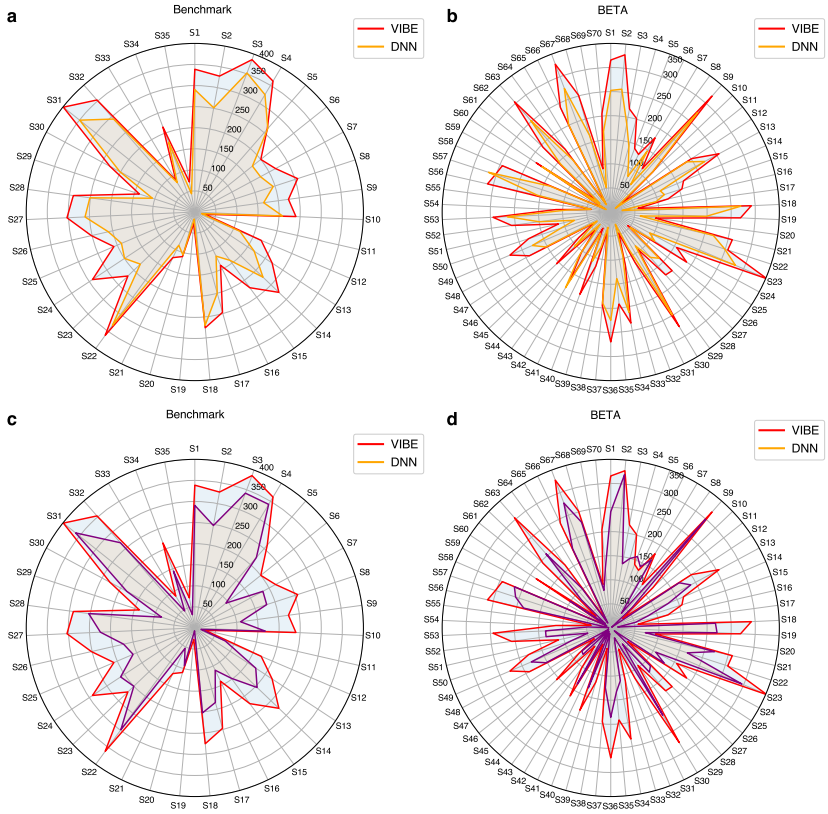

Figure 5: Subject-wise ITR plot. The methods were evaluated at a data length of 0.2s.

Table 5: Greenhouse–Geisser corrected two-way repeated-measures ANOVA results for the interaction effect between data length and method.

Effect	Benchmark (Accuracy)	Benchmark (ITR)	BETA (Accuracy)	BETA (ITR)
F	22.598	22.336	25.256	32.494
p_{GG}	3.45×10^{-7}	6.17×10^{-7}	7.05×10^{-8}	4.86×10^{-10}

Table 6: Paired t-test p-values comparing VIBE with DNN and TDCA for Benchmark and BETA across data lengths (0.2–1.0 s).

Data length (s)	Benchmark		BETA	
	VIBE vs DNN	VIBE vs TDCA	VIBE vs DNN	VIBE vs TDCA
0.2	1.044×10^{-12}	2.211×10^{-14}	4.821×10^{-18}	3.305×10^{-25}
0.3	1.713×10^{-11}	1.947×10^{-7}	7.752×10^{-13}	1.880×10^{-17}
0.4	2.147×10^{-7}	5.304×10^{-7}	1.144×10^{-7}	8.812×10^{-13}
0.5	1.269×10^{-6}	1.073×10^{-5}	2.888×10^{-9}	1.010×10^{-14}
0.6	2.425×10^{-7}	4.578×10^{-5}	1.933×10^{-10}	1.638×10^{-14}
0.7	8.017×10^{-5}	5.435×10^{-3}	1.100×10^{-8}	7.185×10^{-11}
0.8	1.284×10^{-3}	8.630×10^{-3}	1.463×10^{-9}	1.445×10^{-11}
0.9	1.398×10^{-3}	2.856×10^{-2}	6.985×10^{-10}	7.206×10^{-9}
1.0	2.875×10^{-3}	3.825×10^{-2}	1.275×10^{-6}	4.638×10^{-7}

Table 7: Training and testing times for VIBE on two datasets for time data length 0.4s. Times are in seconds, except for the test stages, which are in milliseconds.

Dataset	ViT			MoE Decoder		
	Train (s)	Finetune (s)	Test (ms)	Train (s)	Finetune (s)	Test (ms)
Benchmark	270.8	19.4	0.7	4180.3	43.2	0.09
BETA	328.9	5.7	0.7	1681.1	11.8	0.09

Model	No ViT	No MoE	No Data Augmentation
Accuracy(%)	61.8	64.2	62.1

Table 8: Ablation General Results: No ViT, No MoE, and No Data Augmentation. Tested on Benchmark for 0.2s.

A.5 LLM USAGE

Large language models (LLMs) were used solely for proofreading this manuscript to improve language clarity and readability. They were not involved in generating ideas, designing experiments, implementing methods, or analyzing results. All scientific contributions are entirely the work of the authors.