ASPEN: SPECTRAL-TEMPORAL FUSION FOR CROSS-SUBJECT BRAIN DECODING*

Megan Lee[1] **Seung Ha Hwang**[1,2] **Inhyeok Choi**[1,3] **Shreyas Darade**[1]

Mengchun Zhang[4] **Kateryna Shapovalenko**[1]

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Kyung Hee University, Yongin, South Korea
[3]Korea Advanced Institute of Science and Technology, Deajeon, South Korea
[4]University of Pittsburgh, Pittsburgh, PA, USA

ABSTRACT

Cross-subject generalization in EEG-based brain-computer interfaces (BCIs) remains challenging due to individual variability in neural signals. We investigate whether spectral representations offer more stable features for cross-subject transfer than temporal waveforms. Through correlation analyses across three EEG paradigms (SSVEP, P300, and Motor Imagery), we find that spectral features exhibit consistently higher cross-subject similarity than temporal signals. Motivated by this observation, we introduce ASPEN, a hybrid architecture that combines spectral and temporal feature streams via multiplicative fusion, requiring cross-modal agreement for features to propagate. Experiments across six benchmark datasets reveal that ASPEN is able to dynamically achieve the optimal spectral-temporal balance depending on the paradigm. ASPEN achieves the best unseen-subject accuracy on three of six datasets and competitive performance on others, demonstrating that multiplicative multimodal fusion enables effective cross-subject generalization.

1 INTRODUCTION

Cross-subject generalization remains a fundamental bottleneck in EEG-based brain-computer interfaces (BCIs). Models trained on multi-subject data often degrade substantially when deployed to new users, requiring lengthy subject-specific calibration that undermines the goal of plug-and-play systems (Wan et al., 2021; Liang et al., 2024b). This is due to inherent differences between individuals such as skull thickness, cortical folding, and electrode placement that can produce substantial variation in signal amplitude, timing, and spatial distribution (Lu et al., 2024; Roy et al., 2019).

A growing body of work has addressed this limitation through increasingly expressive temporal modeling, progressing from compact CNN-based decoders (Lawhern et al., 2018) to Transformer architectures that capture global dependencies (Song et al., 2022). However, temporal waveforms are highly sensitive to phase shifts, latency jitter, and amplitude scaling across subjects. The hypothesis we investigate here is that spectral representations provide a more stable basis for cross-subject transfer. Frequency-domain features abstract away precise timing information while preserving the oscillatory signatures, such as μ (8–12 Hz) and β (13–30 Hz) rhythms, that serve as primary biomarkers for BCI paradigms (Ang et al., 2008; Mane et al., 2020).

To test this hypothesis, we first conduct a systematic correlation analyses comparing temporal and spectral representations across SSVEP, P300, and Motor Imagery paradigms. Our analysis reveals that spectral features exhibit substantially higher cross-subject similarity than temporal signals, suggesting that frequency-domain representations offer a more robust foundation for generalization. Motivated by this finding, we introduce ASPEN (**A**daptive **Sp**ectral **E**ncoder **N**etwork, Figure 1), a hybrid framework that processes EEG signals through parallel temporal and spectral streams and

*Code available at `https://github.com/meganmlee/ASPEN`

Preprint.

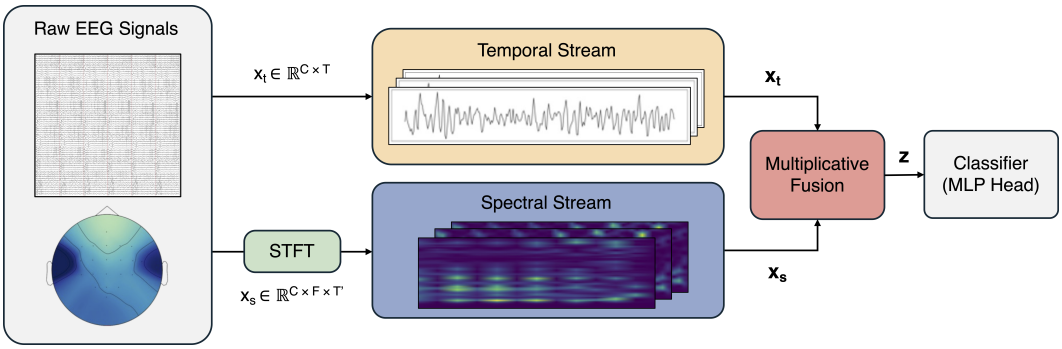

Figure 1: ASPEN architecture. Raw EEG is processed through parallel temporal and spectral streams, combined via multiplicative fusion before classification.

combines them via multiplicative fusion. Unlike prior approaches that concatenate or average multimodal features (Li et al., 2021; 2025), multiplicative fusion computes element-wise products of projected stream representations, requiring both streams to agree for a feature to propagate. This cross-modal gating naturally suppresses artifacts and noise that appear prominently in only one view, while amplifying genuine neural patterns that manifest consistently across both temporal and spectral domains.

We evaluate ASPEN across six benchmark datasets spanning three paradigms. Our experiments reveal that the optimal spectral-temporal balance varies by task: P300 decoding benefits strongly from spectral emphasis, while Motor Imagery requires greater temporal contribution. ASPEN achieves the best unseen-subject accuracy on three datasets (Lee2019 SSVEP, BNCI2014 P300, and Lee2019 MI), outperforming both specialized temporal models and recent multimodal transformers. These results demonstrate that our model is able to perform cross-subject generalization across different BCI tasks while maintaining robustness across diverse neural signatures.

1.1 Related Work

Temporal modeling: Deep learning for EEG signals has evolved from high-capacity architectures like DeepConvNet (Schirrmeister et al., 2017) toward compact, neurophysiologically-informed models. EEGNet (Lawhern et al., 2018) introduced depthwise and separable convolutions that mirror traditional spatial filtering, achieving strong performance with minimal parameters. Transformer-based models such as EEG Conformer (Song et al., 2022) and hybrid CNN-Transformer architectures like CTNet (Zhao et al., 2024) capture long-range temporal dependencies. Temporal convolutional networks (TCNs) offer improved sequential modeling with training stability advantages over recurrent approaches (Ingolfsson et al., 2020; Musallam et al., 2021).

Spectral and filter-bank approaches: Filter-bank methods decompose EEG into frequency subbands before learning spatial filters. The foundational FBCSP algorithm (Ang et al., 2008) demonstrated that isolating discriminative frequency bands improves motor imagery classification. Deep learning extensions apply this principle with learnable filters (Mane et al., 2020; Liu et al., 2022), while IFNet (Wang et al., 2023) models cross-frequency interactions. Time-frequency representations via wavelets have also shown promise for capturing non-stationary dynamics (Morales & Bowers, 2022).

Multimodal fusion: Recent work has begun combining temporal and spectral features. Li et al. (Li et al., 2021) proposed a temporal-spectral squeeze-and-excitation network for motor imagery. TSformer-SA (Li et al., 2025) integrates temporal signals with wavelet spectrograms through cross-view attention for RSVP decoding. Dual-branch architectures have also been explored for emotion recognition (Luo et al., 2023). However, these approaches typically employ additive fusion strategies like concatenation, averaging, or learned weighted sums that allow each stream to contribute independently. Our multiplicative approach instead enforces cross-modal agreement, acting as a feature-wise AND gate that filters unreliable activations.

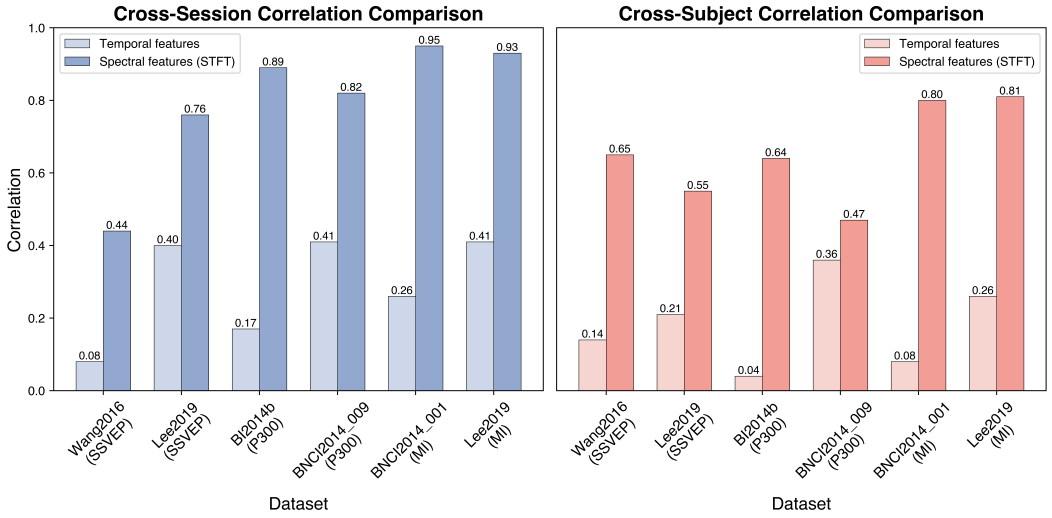

Figure 2: Cross-session (left) and cross-subject (right) correlation comparison between temporal and spectral representations. Spectral features exhibit consistently higher cross-subject similarity across all datasets.

Cross-subject generalization: Domain adaptation techniques including adversarial alignment (Ganin et al., 2016), distribution matching, and adaptive transfer learning (Zhang et al., 2021) have been studied to reduce calibration requirements. MultiDiffNet (Zhang et al., 2025) takes a different route, learning a shared latent space jointly optimized via diffusion, contrastive, and reconstruction objectives to improve cross-subject generalization without explicit distribution alignment.

2 METHODOLOGY

2.1 DATASETS

We evaluated our framework across 3 EEG paradigms (SSVEP, P300, and Motor Imagery) using 6 benchmark datasets: Wang2016 (Wang et al., 2016) and Lee2019 (Lee et al., 2019) for SSVEP; BI2014b (Korczowski et al., 2019) and BNCI2014_009 (Aricò et al., 2014) for P300; and BNCI2014_001 (Tangermann et al., 2012) and Lee2019 (Lee et al., 2019) for MI. Data were partitioned into training, validation, and two distinct test sets: a seen-subject (cross-session) split and an unseen-subject (cross-subject) split to evaluate generalization. Dataset specifications are summarized in Appendix A.1.

2.2 DOMAIN CHARACTERIZATION

To inform our architectural choices, we carried out systematic correlation analyses in both the temporal and spectral domains across all EEG datasets. For each dataset, we constructed class-conditional representative patterns by averaging all trials sharing the same label, producing a time-domain representation of shape (C, T). For the spectral analysis, we applied an STFT to each trial (using task-specific parameters from the corresponding dataset configuration), took the power representation $|Z(f, t)|^2$, and averaged within each class to obtain a spectral representative pattern of shape (C, F, T'). To quantify the similarity between any two representative patterns, we flattened and z-normalized each one, then computed their Pearson correlation coefficient.

We examined two complementary notions of signal consistency: cross-session and cross-subject correlation. Cross-session correlation measures the agreement between representative patterns derived from different recording sessions of the same subject, aggregated across all session pairs and labels. Cross-subject correlation measures the agreement between representative patterns from different subjects, again aggregated across all subject pairs and labels. As shown in Figure 2, spectral representative patterns consistently exhibited higher cross-subject similarity than their raw tempo-

ral counterparts across datasets. This suggests that time-frequency representations capture a more subject-invariant signature of task-relevant neural dynamics, and directly motivates our modeling framework's emphasis on spectral representations as a robust foundation for cross-subject generalization.

2.3 Signal Processing

We process the raw EEG signals into two modalities to accommodate the dual-stream architecture of ASPEN.

Temporal Modality: To ensure signal stability and cross-subject consistency, raw EEG data are preprocessed using task-specific bandpass filtering and trial-wise Z-score normalization to achieve zero mean and unit variance. Frequency ranges are tailored to each paradigm: 6-90 Hz for SSVEP to capture high-frequency harmonics (Wang et al., 2016), 1-24 Hz for P300 to isolate the evoked response, and 4-40 Hz for Motor Imagery (MI) to focus on μ and β rhythms (Lee et al., 2019). Signals are downsampled to optimize computational efficiency (see Appendix A.1 for details).

Spectral Modality: To obtain frequency-domain representations, we apply the Short-Time Fourier Transform (STFT) to each preprocessed trial on a per-channel basis. For each paradigm, we fix a task-specific set of STFT parameters $(f_s, n_{\text{perseg}}, n_{\text{overlap}}, n_{\text{fft}})$ drawn from the dataset configuration. These parameters are validated through a 27-way ablation study with varying window length, overlap ratio, and Fast Fourier Transform (FFT) size, ensuring that all subjects share a consistent $\text{frequency} \times \text{time}$ grid at the input of the spectral stream. Concretely, we compute spectrograms with a Hann window and convert them to power spectra by taking the squared magnitude $|Z_c(f,t)|^2$ for each channel c, yielding a 3D tensor of shape (C, F, T) per trial. For the purposes of our comparative analysis, we refer to the standalone Spectral Encoder Network as SPEN, while ASPEN represents the full hybrid framework utilizing multiplicative fusion. These power spectrograms are fed into the SPEN convolutional blocks, where the chosen window and hop sizes control the trade-off between temporal resolution $\Delta t = (n_{\text{perseg}} - n_{\text{overlap}})/f_s$ and frequency resolution $\Delta f = f_s/n_{\text{fft}}$, allowing ASPEN to capture task-relevant harmonic structure for SSVEP, evoked components for P300, and μ/β band dynamics for MI in a unified spectral representation.

2.4 Model Architecture

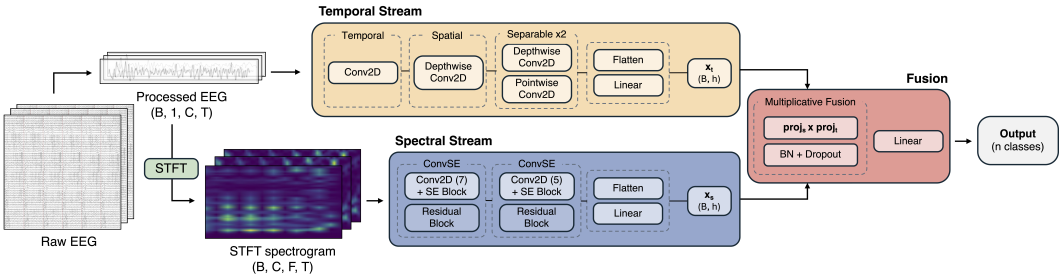

Figure 3: Detailed view of temporal stream, spectral stream, and multiplicative fusion components.

The high-level architecture is shown in Figure 1. ASPEN consists of two primary components: a Temporal Stream and a Spectral Stream. Given a raw EEG trial $\mathbf{X} \in \mathbb{R}^{C \times T}$, the two complementary inputs are the raw signal $\mathbf{X}_{\text{time}}$ for the temporal stream and the per-channel STFT magnitude spectrograms $\mathbf{X}_{\text{spec}} \in \mathbb{R}^{C \times F \times T'}$ for the spectral stream.

A detailed view of the two streams and the fusion mechanism are illustrated in Figure3. The spectral stream extracts frequency-time patterns through a two-stage CNN with squeeze-and-excitation (SE) attention (Hu et al., 2018) and residual blocks. SE modules adaptively recalibrate channel responses by learning to emphasize informative spectral patterns, while residual connections improve gradient flow. After two stages of convolution, SE attention, and pooling, features are projected and averaged across EEG channels to yield $\mathbf{x}_s \in \mathbb{R}^d$. The temporal stream follows an EEGNet-inspired design (Lawhern et al., 2018) where the temporal convolution learns frequency-specific filters analogous to bandpass filtering, the depthwise spatial convolution learns channel combinations analogous

to CSP (Ang et al., 2008), and the separable convolutions efficiently extract higher-order features. The output is projected to $\mathbf{x}_t \in \mathbb{R}^d$.

We combine stream representations via element-wise multiplication after learned linear projections

$$\mathbf{z} = (\mathbf{W}_s \mathbf{x}_s) \odot (\mathbf{W}_t \mathbf{x}_t) \tag{1}$$

where $\mathbf{W}_s, \mathbf{W}_t \in \mathbb{R}^{d \times d}$ are learnable matrices and $\odot$ denotes the Hadamard product. This multiplicative interaction acts as cross-modal gating. Dimension z_i is large only when both streams produce strong activations, effectively requiring agreement between spectral and temporal evidence. Features that appear prominently in only one view, which often indicative of artifacts or noise, are suppressed. The fused representation passes through batch normalization and a linear classifier.

The model is optimized using task-specific loss functions. Binary cross-entropy with logits ($\mathcal{L}_{BCE}$) are used for two-class paradigms like P300, which includes automated positive-weight scaling to mitigate class imbalance. For multi-class tasks such as SSVEP and Motor Imagery, standard cross-entropy loss ($\mathcal{L}_{CE}$) is employed. The learned weights w_s and w_t provide an interpretable measure of each modality's contribution, enabling an analysis of which representation the model prioritizes for different paradigms and individual trials.

3 EXPERIMENTS

3.1 BASELINES

To evaluate the performance of our proposed method, we benchmarked against five baselines that emphasize cross-subject generalization and novel data representations.

EEGNet (Lee et al., 2019) serves as a compact convolutional baseline, leveraging depthwise and separable convolutions to efficiently extract spatial and frequency-specific features with minimal parameters. To model global dependencies, EEGConformer (Song et al., 2022) adopts a hybrid design that combines CNNs for local feature extraction with Transformer modules for long-range temporal modeling. CTNet (Zhao et al., 2024) is included for its emphasis on cross-task and cross-subject robustness, utilizing domain-invariant representations to mitigate EEG non-stationarity. TSformer-SA (Li et al., 2025) integrates temporal and spectral features through cross-view self-attention, enabling joint modeling of time-domain signals and wavelet-based time-frequency representations for improved cross-subject decoding. Finally, MultiDiffNet (Zhang et al., 2025) incorporates multi-scale differential transformations of the input signal to better capture complex distributions and enhance training stability in noisy data environments.

All models were evaluated using identical training schedules and hyperparameters unless otherwise specified by architectural constraints. The performance of these baselines are shown in Table 2.

3.2 ABLATIONS

Table 1: Ablation study summary. Best STFT parameters and fusion strategy per dataset, selected by unseen-subject accuracy. Best Acc = best fusion accuracy on held-out subjects (%), Mult Acc = multiplicative fusion accuracy (%), Δ = absolute difference. Global Attn = Global Attention, Bilinear = Low-rank Bilinear.

Dataset	nperseg	noverlap	nfft	Best Fusion	Best Acc	Mult Acc	Δ
Wang 2016 SSVEP	256	128	256	Global Attn	72.76	69.47	-3.3
Lee2019 SSVEP	256	128	1024	Bilinear	86.68	85.71	-1.0
BI2014b P300	32	16	512	Bilinear	73.52	66.12	-7.4
BNCI2014-009 P300	128	120	256	Multiplicative	89.82	89.82	–
BNCI2014-001 MI	512	256	512	Multiplicative	30.73	30.73	–
Lee2019 MI	32	30	32	Multiplicative	75.70	75.70	–

STFT Settings: Since SPEN operates on time-frequency representations, we first optimized the STFT frontend through a controlled ablation study. For each task, we evaluated 27 configurations

by sweeping three values per parameter: window length (`nperseg`), overlap ratio (0.50, 0.75, 0.9375), and FFT size (`nfft`). Window lengths were defined in a task-aware manner (half, default, and maximum resolution), constrained to never exceed trial length; `noverlap` was derived from the overlap ratio subject to `noverlap` $<$ `nperseg`; and `nfft` was drawn from a pool containing the smallest power of two $\geq$ `nperseg` and the task default, with `nfft` $\geq$ `nperseg` enforced.

For each configuration, we trained the same SPEN backbone and training protocol to isolate the effect of STFT settings, then evaluated on validation and test splits using accuracy and F1/recall (plus ROC-AUC and PR-AUC for imbalanced binary P300 tasks). For binary tasks, we re-optimized the decision threshold on the validation set (F1-maximizing sweep) and applied it to test evaluation; for P300 tasks we additionally used a WeightedRandomSampler to balance training batches. The best STFT setting was selected per task (F1 for P300, accuracy otherwise), and the top 3 STFT configurations were carried forward into subsequent fusion ablations to avoid confounding fusion comparisons with suboptimal preprocessing. Full details are provided in Appendix A.2.1.

Fusion Strategies: We evaluated seven fusion strategies for combining the temporal and spectral streams, drawing from foundational methods in multimodal learning (Liang et al., 2024a): (1) static equal weighting, (2) global attention with learned trial-level weights, (3) spatial attention with per-channel weighting, (4) gated linear units (GLU) for noise suppression, (5) element-wise multiplicative fusion, (6) low-rank bilinear pooling, and (7) multi-head cross-attention between streams. Each strategy was evaluated across all benchmark tasks using the top 3 best performing STFT parameters from our spectrogram ablation.

While optimal fusion varied by dataset, multiplicative fusion achieved the highest unseen-subject accuracy on three of six tasks (BNCI2014 P300, BNCI2014-001 MI, Lee2019 MI) and remained competitive on two others (within 1% on Wang2016 SSVEP and 3.3% on Lee2019 SSVEP). Bilinear fusion outperforms on BI2014b P300 by 7.4%, but we prioritize cross-paradigm consistency over peak single-task performance. Given multiplicative fusion's stable cross-paradigm performance and its alignment with our cross-modal gating hypothesis (Equation 1), we adopt it as the unified fusion strategy for ASPEN. This selection also determined which STFT configuration to use for final evaluation. Best fusion method and STFT parameters are summarized in Table 1. Full mathematical details and per-task results are provided in Appendix A.3.1.

3.3 RESULTS

Table 2: Final results across tasks and models. Cross-subject generalization accuracy (%). Bold indicates best cross-subject performance per dataset. Mean ± STD across three seeds.

			Method						
Task	**Dataset**	**Cross-**	EEGNet	EEGConf.	MultiDiff.	TSformer-SA	CTNet	SPEN (Ours)	ASPEN (Ours)
SSVEP	Wang2016	Session	81.96±6.02	56.32±4.15	91.74±1.62	47.96±4.16	88.37±4.20	85.71±1.57	73.98±1.46
		Subject	74.25±5.27	49.95±2.81	**87.95**±2.56	39.93±5.25	83.60±0.82	78.82±6.41	67.20±4.83
	Lee2019	Session	95.04±0.35	92.00±0.94	93.67±0.67	79.98±6.60	95.36±0.61	70.58±0.80	95.50±0.33
		Subject	86.51±0.09	81.99±1.05	85.04±0.76	63.38±6.43	87.25±0.69	58.51±0.98	**87.53**±0.29
P300	BI2014b	Session	64.74±2.71	78.84±3.46	81.46±4.56	82.25±3.04	76.06±6.45	84.15±0.93	77.95±5.52
		Subject	62.64±1.01	77.55±4.59	80.95±3.91	**83.13**±0.35	74.55±8.41	82.96±0.64	77.01±7.16
	BNCI2014	Session	84.65±0.62	84.25±1.09	85.42±0.89	86.75±0.68	86.26±0.38	78.31±4.40	89.65±0.48
		Subject	84.05±1.84	83.20±1.29	84.91±1.68	86.92±1.06	85.28±0.90	77.97±4.66	**88.57**±0.76
MI	BNCI2014	Session	61.21±3.27	56.35±3.05	58.53±4.46	35.91±6.48	57.34±7.68	33.63±3.32	51.59±8.40
		Subject	**37.29**±4.45	33.91±7.33	29.05±4.00	28.99±4.82	36.29±9.82	26.91±3.69	32.00±0.53
	Lee2019	Session	78.89±0.38	76.28±0.91	76.54±2.47	71.90±0.74	77.15±0.77	55.28±0.45	77.93±0.52
		Subject	75.88±1.98	74.55±1.39	74.67±1.95	71.40±2.61	74.98±2.13	53.50±1.69	**76.27**±0.73

The performance of SPEN, ASPEN, and five baselines across six benchmark datasets is summarized in Table 2. Our results demonstrate that ASPEN achieves superior cross-subject generalization in three of the six evaluated datasets: Lee2019 SSVEP (87.53%), BNCI2014 P300 (88.57%), and Lee2019 MI (76.27%) datasets. Notably, on the BNCI2014 P300 task, ASPEN outperforms TSformer-SA by nearly 2%, despite the latter being specifically designed for evoked potential decoding. This suggests that our multiplicative gating mechanism is more effective at filtering the inter-subject noise prevalent in large-scale P300 datasets.

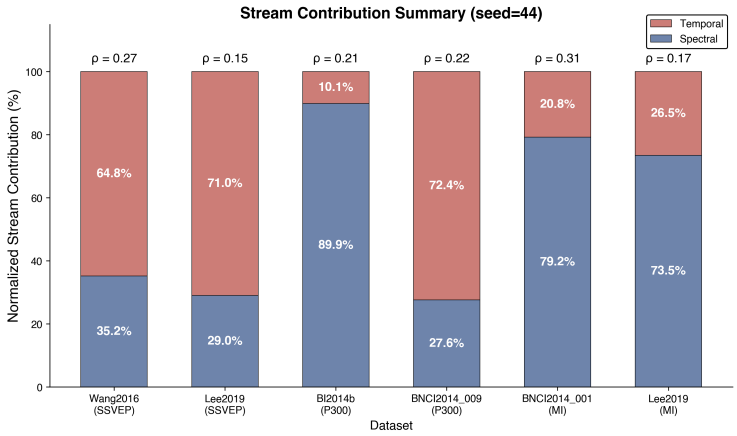

Figure 4: Stream contributions and feature correlation (ρ) across datasets. Low correlation values confirm that streams capture distinct information.

We also observe that while TSformer-SA performs competitively on P300-like tasks (83.13% on BI2014b), its performance degrades significantly on SSVEP and MI tasks (39.93% on Wang2016 SSVEP). In contrast, ASPEN maintains competitive performance across all three tasks. This indicates that multiplicative fusion acts as a universal architectural prior that adapts to the specific spectral-temporal demands of the underlying neural signal.

The standalone spectral encoder (SPEN) struggles on Motor Imagery tasks, achieving only 26.91% on BNCI2014-001 MI (barely above the 25% chance level) and 53.50% on Lee2019 MI. These results reveal that cross-subject stability and discriminative power are distinct properties. While spectral representations are more consistent across individuals, Motor Imagery classification relies on precise temporal dynamics of sensorimotor rhythms that are lost in the STFT magnitude representation. The performance recovery from SPEN to ASPEN on Lee2019 MI (53.50% to 76.27%) demonstrates that neither modality alone suffices and that cross-modal fusion is essential for robust generalization.

4 Analysis

Our experimental results demonstrate that ASPEN achieves superior or competitive cross-subject generalization compared to state-of-the-art baselines. To understand the drivers of this performance, we analyze the impact of multiplicative fusion, the paradigm-specific reliance on spectral versus temporal features, and the role of spectral stability.

4.1 Mechanism of Multiplicative Spectral-Temporal Gating

To investigate how ASPEN leverages dual-stream information, we analyze the features during inference. The fused representation is defined as:

$$\mathbf{z}_{\text{fused}} = \text{proj}_S(\mathbf{x}_S) \odot \text{proj}_T(\mathbf{x}_T) \tag{2}$$

where $\mathbf{x}_S$ and $\mathbf{x}_T$ denote the spectral and temporal features respectively, and $\odot$ represents element-wise multiplication. We quantify the relative spectral magnitude w_S via the normalized L2 norm of the projected features:

$$w_S = \frac{\|\text{proj}_S(\mathbf{x}_S)\|_2}{\|\text{proj}_S(\mathbf{x}_S)\|_2 + \|\text{proj}_T(\mathbf{x}_T)\|_2} \tag{3}$$

with $w_T = 1 - w_S$ representing the relative temporal magnitude. Stream complementarity is measured through the feature correlation ρ, defined as the cosine similarity between the projected features:

$$\rho = \frac{\langle \text{proj}_S(\mathbf{x}_S), \text{proj}_T(\mathbf{x}_T) \rangle}{\|\text{proj}_S(\mathbf{x}_S)\|_2 \cdot \|\text{proj}_T(\mathbf{x}_T)\|_2} \tag{4}$$

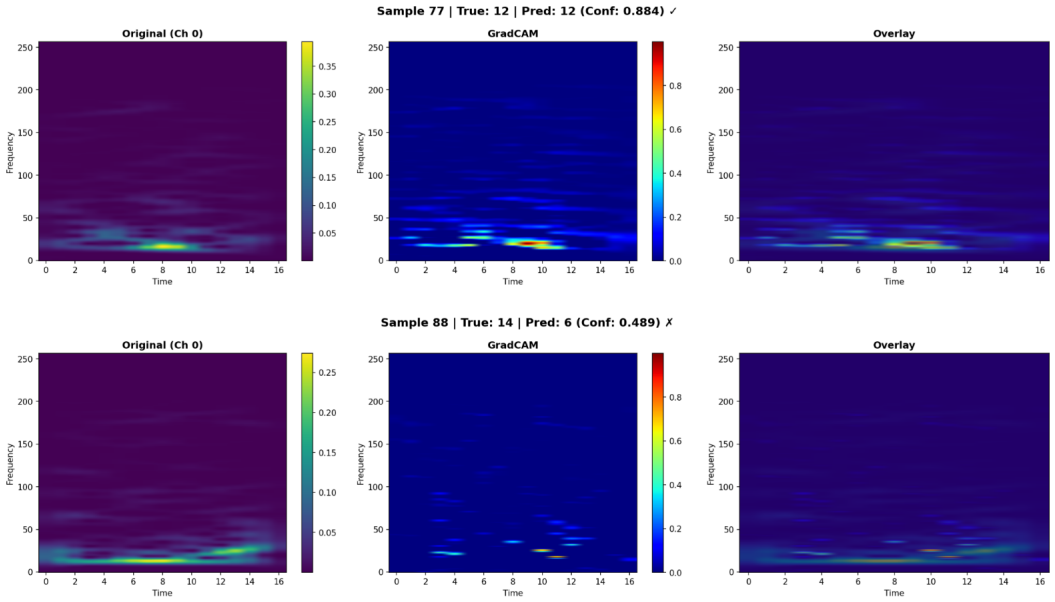

Figure 5: Grad-CAM visualization of feature importance for P300 classification. Correct prediction (top) shows focused attention on physiologically relevant low-frequency bands. Misclassification (bottom) reveals scattered attention towards high-frequency noise artifacts.

As illustrated in Figure 4, ASPEN adaptively shifts its reliance on spectral vs. temporal features based on the task at hand. The P300 task (BI2014b) exhibits strong spectral dominance ($w_S = 89.9\%$), while SSVEP tasks (Wang2016, Lee2019) shift toward the temporal stream, with w_T reaching 64.8% and 71.0%, respectively. Motor imagery datasets show a spectrally-biased distribution ($w_S \approx 73\%$–79%). Across all datasets, the low correlation values ($0.15 \leq \rho \leq 0.31$) confirm that the streams capture distinct, non-redundant information.

This architecture functions as a strict cross-modal gating mechanism. Unlike additive fusion, where high-magnitude artifacts in one modality can bias the decision boundary, our multiplicative approach acts as a logical *AND* gate. A feature is only activated in the fused representation if it receives concurrent support from both streams. Consequently, transient artifacts, such as muscle noise that appears in the temporal domain but lacks spectral consistency, are naturally suppressed. As evidenced in Table 1, forcing this cross-view agreement encourages the model to prioritize features robust to the phase shifts and amplitude variations inherent in cross-subject transfer.

4.2 Visualizing Decision Boundaries

To validate the interpretability of our framework and justify the necessity of cross-modal fusion, we visualized the learned features using Grad-CAM Selvaraju et al. (2017). We analyzed the spectral regions contributing most to the model's decisions in both successful and failed prediction scenarios on the P300 dataset.

Fig.5 illustrates the Grad-CAM activation maps for two representative samples. As shown in the top row of Fig. 5, when the model *correctly* identifies the target class with high confidence, the activation hotspot is tightly concentrated in the low-frequency band and specific temporal windows. This aligns perfectly with neurophysiological knowledge, as P300 components are primarily characterized by low-frequency energy deflections. The model successfully ignores high-frequency background activity, confirming that it has learned robust, physiologically valid features.

Conversely, the bottom row of Fig. 5 shows a misclassified sample with low confidence. Here, the model's attention is fragmented and scattered across high-frequency bands, likely driven by muscle artifacts or instrument noise rather than neural signals. This distraction by high-frequency noise

highlights the vulnerability of single-stream interactions where artifactual high-amplitude spikes can propagate to the decision layer.

5 Conclusion and Outlook

In this work, we introduced ASPEN, a multimodal framework designed to overcome the challenges of cross-subject generalization in EEG-based BCIs. By leveraging the inherent stability of spectral representations, ASPEN utilizes a multiplicative fusion mechanism to enforce cross-modal agreement. Our experiments across six benchmark datasets show that this approach effectively suppresses non-neural artifacts and prioritizes robust features, achieving the best unseen-subject performance on three datasets spanning SSVEP, P300, and Motor Imagery paradigms.

While ASPEN significantly reduces the performance gap for new users, several avenues for future research remain. Future work will investigate automated configuration optimization, perhaps through learnable time-frequency transforms, to move toward a truly "one-size-fits-all" zero-shot model. Additionally, we aim to explore the integration of self-supervised pre-training on large multi-subject EEG corpora to further enhance the richness of the shared latent space.

Acknowledgments

We would like to thank Professor Bhiksha Raj of Carnegie Mellon University for his guidance and support throughout this project. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

References

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 2390–2397. IEEE, 2008.

Pietro Aricò, F Aloise, Francesca Schettini, Serenella Salinari, D Mattia, and Febo Cincotti. Influence of p300 latency jitter on event related potential-based brain–computer interface performance. *Journal of neural engineering*, 11(3):035008, 2014.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 2958–2965. IEEE, 2020.

Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain invaders solo versus collaboration: Multi-user P300-based brain-computer interface dataset (bi2014b)*. PhD thesis, GIPSA-lab, 2019.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019.

Xujin Li, Wei Wei, Shuang Qiu, and Huiguang He. A temporal–spectral fusion transformer with subject-specific adapter for enhancing rsvp-bci decoding. *Neural Networks*, 181:106844, 2025.

Yang Li, Lianghui Guo, Yu Liu, Jingyu Liu, and Fangang Meng. A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery eeg decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1534–1545, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024a.

Shuang Liang, Linzhe Li, Wei Zu, Wei Feng, and Wenlong Hang. Adaptive deep feature representation learning for cross-subject eeg decoding. *BMC bioinformatics*, 25(1):393, 2024b.

Ke Liu, Mingzhao Yang, Zhuliang Yu, Guoyin Wang, and Wei Wu. Fbmsnet: A filter-bank multi-scale convolutional neural network for eeg-based motor imagery decoding. *IEEE Transactions on Biomedical Engineering*, 70(2):436–445, 2022.

Wei Lu, Xiaobo Zhang, Lingnan Xia, Hua Ma, and Tien-Ping Tan. Domain adaptation spatial feature perception neural network for cross-subject eeg emotion recognition. *Frontiers in Human Neuroscience*, 18:1471634, 2024.

Jie Luo, Weigang Cui, Song Xu, Lina Wang, Xiao Li, Xiaofeng Liao, and Yang Li. A dual-branch spatio-temporal-spectral transformer feature fusion network for eeg-based visual recognition. *IEEE Transactions on Industrial Informatics*, 20(2):1721–1731, 2023.

Ravikiran Mane, Neethu Robinson, A Prasad Vinod, Seong-Whan Lee, and Cuntai Guan. A multi-view cnn with novel variance layer for motor imagery brain computer interface. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pp. 2950–2953. Ieee, 2020.

Santiago Morales and Maureen E. Bowers. Time-frequency analysis methods and their application in developmental eeg data. *Developmental Cognitive Neuroscience*, 54:101067, 2022. doi: 10.1016/j.dcn.2022.101067.

Yazeed K Musallam, Nasser I AlFassam, Ghulam Muhammad, Syed Umar Amin, Mansour Alsulaiman, Wadood Abdul, Hamdi Altaheri, Mohamed A Bencherif, and Mohammed Algabri. Electroencephalography-based motor imagery classification using temporal convolutional network fusion. *Biomedical Signal Processing and Control*, 69:102826, 2021.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 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.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Zitong Wan, Rui Yang, Mengjie Huang, Nianyin Zeng, and Xiaohui Liu. A review on transfer learning in eeg signal analysis. *Neurocomputing*, 421:1–14, 2021.

Jiaheng Wang, Lin Yao, and Yueming Wang. Ifnet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1900–1911, 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.

Kaishuo Zhang, Neethu Robinson, Seong-Whan Lee, and Cuntai Guan. Adaptive transfer learning for eeg motor imagery classification with deep convolutional neural network. *Neural Networks*, 136:1–10, 2021.

Mengchun Zhang, Kateryna Shapovalenko, Yucheng Shao, Eddie Guo, and Parusha Pradhan. Multi-DiffNet: A multi-objective diffusion framework for generalizable brain decoding. 2025. Preprint: arXiv:2511.18294.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. Ctnet: a convolutional transformer network for eeg-based motor imagery classification. *Scientific reports*, 14(1):20237, 2024.

A Appendix

A.1 Dataset Specifications and Splitting

To evaluate model robustness across datasets, we employed a multi-tier splitting strategy with a 60/20/20 ratio for subjects included in training. For each task, a subset of subjects was designated as "seen" and their data was partitioned into three splits. The Training split (60%) was used for model optimization, the Validation split (20%) for hyperparameter tuning, and Test 1 (20%) for cross-session evaluation. Test 1 specifically evaluates the model's ability to generalize across different recording sessions from the same individuals, thereby assessing robustness to within-subject temporal variations.

To assess zero-shot generalization capabilities, a separate cohort of subjects was entirely withheld from the training process. Test 2 (Cross-Subject) evaluates model performance on individuals with previously unencountered physiological profiles, providing a rigorous test of subject-independent performance.

Table 3: EEG Dataset Specifications and Preprocessing Parameters. Z-score normalization was applied to all datasets to standardize signal amplitudes across different subjects and sessions, ensuring that high-voltage artifacts or individual physiological variations do not disproportionately influence model training.

Task	Dataset	Subj.	Chan.	Cls.	Bandpass	Epoch	SR
SSVEP: Frequency-tagged visual decoding	Wang2016	35	64	26	6–90 Hz	1.0s	250 Hz
	Lee2019	54	62	4	6–90 Hz	1.0s	250 Hz
P300: Binary target vs. non-target ERP	BI2014b	38	32	2	0.1–30 Hz	1.0s	256 Hz
	BNCI2014_009	10	16	2	1–24 Hz	1.0s	256 Hz
MI: Motor imagery decoding	BNCI2014_001	9	22	4	4–40 Hz	4.0s	250 Hz
	Lee2019	54	22	2	4–40 Hz	4.0s	250 Hz

A.2 STFT Ablation

A.2.1 STFT Methods

STFT-based Spectral Representation. For each EEG trial, we construct a spectral (time–frequency) representation by applying the Short-Time Fourier Transform (STFT) independently to

each channel. Let $x_c[n]$ denote the preprocessed discrete-time signal of channel $c \in \{1, \ldots, C\}$ with sampling rate f_s (Hz), and let $w[m]$ be a Hann window of length n_{perseg}. We compute the complex STFT as

$$Z_c(f_k, t_\ell) = \sum_{m=0}^{n_{\text{perseg}}-1} x_c[\ell H + m]\, w[m]\, e^{-j2\pi km/n_{\text{fft}}}, \tag{5}$$

where n_{fft} is the FFT size, $H = n_{\text{perseg}} - n_{\text{overlap}}$ is the hop size (in samples), and (f_k, t_ℓ) index frequency and time frames, respectively. In practice, we use the one-sided spectrum for real-valued signals, yielding $F = \lfloor n_{\text{fft}}/2 \rfloor + 1$ frequency bins. We then convert complex spectrograms to *power* spectrograms by taking the squared magnitude:

$$S_c(f_k, t_\ell) = |Z_c(f_k, t_\ell)|^2. \tag{6}$$

Each trial is thus represented as a tensor $\mathbf{S} \in \mathbb{R}^{C \times F \times T}$, where T denotes the number of time frames determined by the trial length and hop size. These power spectrograms are provided directly to the spectral stream of the model, ensuring a consistent $(F \times T)$ grid across subjects within each paradigm. The STFT parameters control the time–frequency trade-off: the frequency resolution is $\Delta f = f_s/n_{\text{fft}}$ (Hz per bin) and the frame step is $\Delta t = H/f_s = (n_{\text{perseg}} - n_{\text{overlap}})/f_s$ seconds. Larger n_{perseg} and n_{fft} improve frequency resolution, while higher overlap (smaller H) improves temporal granularity.

Task-aware STFT parameter search space. Because optimal STFT settings can vary by paradigm and trial length, we perform a systematic ablation over STFT parameters for each task. The ablation evaluates a fixed set of 27 configurations formed by selecting three values for each of: (i) window length n_{perseg}, (ii) overlap ratio $r \in \{0.50,\ 0.75,\ 0.9375\}$ with $n_{\text{overlap}} = \lfloor r\, n_{\text{perseg}} \rfloor$ and the hard constraint $n_{\text{overlap}} < n_{\text{perseg}}$, and (iii) FFT size n_{fft}. To ensure feasibility, we enforce $n_{\text{perseg}} \leq L_{\text{trial}}$ where L_{trial} is the number of samples per trial for the dataset. Window-length candidates are generated from the task default $n_{\text{perseg}}^{(0)}$ as

$$\mathcal{N}_{\text{perseg}} = \Big\{ \max(32, \lfloor \tfrac{1}{2} n_{\text{perseg}}^{(0)} \rfloor),\ n_{\text{perseg}}^{(0)},\ \min(L_{\text{trial}},\ 2n_{\text{perseg}}^{(0)}) \Big\}, \tag{7}$$

after removing duplicates. For each n_{perseg}, we build an FFT pool

$$\mathcal{P} = \text{unique}\Big(\{ 2^{\lceil \log_2(n_{\text{perseg}}) \rceil},\ 256,\ 512,\ 1024,\ 2048,\ n_{\text{fft}}^{(0)} \} \Big), \tag{8}$$

retain only valid $n_{\text{fft}} \geq n_{\text{perseg}}$, and choose three FFT candidates centered around $n_{\text{fft}}^{(0)}$ when available (otherwise the first three valid values). If fewer than three candidates remain, we pad by doubling the maximum value until three candidates are obtained. Each configuration is assigned an identifier

$$\texttt{nperseg}\{n_{\text{perseg}}\}\texttt{\_ov}\{100r\}\texttt{\_nfft}\{n_{\text{fft}}\}. \tag{9}$$

Ablation protocol and model selection. For each task and STFT configuration, we train the same model architecture with identical optimization hyperparameters to isolate the effect of the STFT frontend. We report accuracy and complementary metrics (macro-F1/macro-recall for multi-class; precision/recall, ROC-AUC, and PR-AUC for binary tasks). For binary tasks, we additionally optimize the decision threshold on the validation set by sweeping $t \in \{0.01, 0.02, \ldots, 0.99\}$ and maximizing validation F1, then reuse the selected threshold for all test evaluations to avoid test-time tuning. For P300-style binary datasets, we mitigate class imbalance using weighted sampling, assigning example weights inversely proportional to class frequency. The best STFT configuration is selected per task using validation accuracy for multi-class tasks and validation PR-AUC (fallback to ROC-AUC if needed) for binary P300 tasks; we log all configurations and metrics (CSV/JSON) to enable reproducibility.

A.2.2 STFT Ablation Results

Table 4: STFT Ablation SSVEP

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen loss
128	**64**	**256**	**83.05**	**80.06**	**0.77**	**71.71**	**71.90**	**71.71**	**96.08**	**1.15**
128	64	512	82.34	81.20	0.69	69.23	69.25	69.23	96.16	1.17
128	64	1024	81.34	80.06	0.76	65.46	65.71	65.46	95.30	1.26
128	96	256	83.76	82.76	0.70	70.83	70.91	70.83	96.21	1.13
128	96	512	82.48	82.91	0.66	71.31	71.41	71.31	96.57	1.10
128	96	1024	82.19	83.33	0.66	70.11	70.18	70.11	96.06	1.13
128	120	256	82.19	83.05	0.69	70.27	70.34	70.27	96.09	1.14
128	120	512	81.77	81.77	0.70	70.35	70.57	70.35	96.02	1.13
256	**128**	**256**	**84.47**	**82.91**	**0.69**	**72.60**	**72.69**	**72.60**	**96.75**	**1.08**
256	**128**	**512**	**83.76**	**84.19**	**0.64**	**71.47**	**71.59**	**71.47**	**96.82**	**1.07**
256	128	1024	84.19	83.33	0.64	68.99	69.30	68.99	95.83	1.19
256	192	256	81.91	82.19	0.72	71.39	71.54	71.39	96.05	1.13
256	192	512	82.48	82.62	0.67	71.31	71.51	71.31	95.98	1.15
256	192	1024	83.19	82.34	0.67	70.59	70.80	70.59	96.06	1.16
256	240	256	83.62	84.47	0.64	70.35	70.48	70.35	96.23	1.14
256	240	512	83.33	83.48	0.65	69.55	69.62	69.55	95.91	1.15

Table 5: STFT Ablation Lee2019 SSVEP

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen loss
64	32	256	61.06	61.09	0.95	48.14	48.22	48.14	72.63	1.28
64	32	512	67.09	67.00	0.85	53.70	53.74	53.70	77.73	1.18
64	32	1024	66.19	65.75	0.87	53.12	53.20	53.12	77.35	1.14
64	48	256	69.34	68.22	0.85	54.77	54.77	54.77	78.36	1.21
64	48	512	69.28	68.13	0.81	55.52	55.56	55.52	78.56	1.14
64	48	1024	66.63	66.09	0.87	53.50	53.53	53.50	77.09	1.22
64	60	256	58.66	60.00	0.98	46.52	46.39	46.52	71.48	1.24
64	60	512	67.31	66.44	0.85	52.84	52.91	52.84	76.35	1.22
128	64	256	67.56	67.44	0.86	54.70	54.88	54.70	78.39	1.21
128	**64**	**512**	**71.53**	**71.38**	**0.83**	**58.75**	**58.81**	**58.75**	**81.73**	**1.21**
128	64	1024	70.75	69.63	0.78	57.59	57.78	57.59	80.61	1.07
128	96	256	70.84	70.69	0.83	57.27	57.27	57.27	79.93	1.25
128	96	512	72.66	70.81	0.77	57.41	57.53	57.41	80.44	1.12
128	96	1024	71.53	71.19	0.76	57.39	57.49	57.39	81.00	1.10
128	120	256	71.22	69.88	0.80	56.21	56.25	56.21	78.98	1.15
128	120	512	69.50	68.53	0.82	56.29	56.21	56.29	78.56	1.12
128	120	1024	71.38	70.03	0.80	57.34	57.52	57.34	80.22	1.17
256	128	256	68.94	68.25	0.86	55.38	55.44	55.38	78.80	1.28
256	128	512	71.56	71.56	0.78	58.30	58.44	58.30	81.17	1.15
256	**128**	**1024**	**72.38**	**72.16**	**0.74**	**57.95**	**58.00**	**57.95**	**80.98**	**1.11**
256	192	256	70.94	69.59	0.93	56.05	56.31	56.05	79.08	1.37
256	192	512	72.19	70.56	0.78	58.14	58.25	58.14	80.60	1.12
256	**192**	**1024**	**72.53**	**72.59**	**0.74**	**58.98**	**59.03**	**58.98**	**81.57**	**1.06**
256	240	256	69.81	67.94	0.82	54.84	54.92	54.84	77.82	1.16
256	240	512	70.16	70.41	0.80	56.34	56.39	56.34	79.41	1.19
256	240	1024	71.25	70.69	0.76	58.32	58.40	58.32	80.83	1.12

Table 6: STFT Ablation BI2014b P300

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen pr_auc	unseen loss
32	16	32	82.14	15.40	1.13	16.67	28.50	99.67	54.10	18.75	1.18
32	16	256	82.14	19.10	1.12	19.35	28.35	95.72	52.85	18.02	1.17
32	**16**	**512**	**82.14**	**19.17**	**1.13**	**20.29**	**28.44**	**95.07**	**51.54**	**17.54**	**1.18**
32	24	32	82.00	24.08	1.11	22.92	28.05	90.13	53.30	19.84	1.16
32	24	256	82.14	32.48	1.12	26.64	28.37	87.17	53.06	18.34	1.16
32	24	512	82.21	44.32	1.12	38.71	27.50	69.74	51.80	18.82	1.16
32	**30**	**32**	**82.14**	**16.27**	**1.12**	**17.38**	**28.48**	**98.68**	**54.68**	**20.63**	**1.17**
32	30	256	82.21	34.77	1.11	38.93	27.09	68.09	52.77	18.40	1.16
64	32	64	82.14	15.47	1.13	16.89	28.49	99.34	52.02	17.31	1.18
64	32	256	82.14	17.28	1.12	20.67	28.19	93.42	50.15	17.21	1.17
64	32	512	82.07	27.98	1.12	22.26	28.09	91.12	52.85	17.96	1.17
64	48	64	82.21	16.48	1.12	16.89	28.49	99.34	53.81	18.14	1.16
64	48	256	82.14	28.18	1.12	32.57	28.65	81.25	53.09	19.12	1.16
64	48	512	82.14	22.06	1.12	19.79	28.39	95.39	50.37	16.45	1.16
64	60	64	82.14	15.74	1.11	18.91	28.38	96.38	54.22	19.93	1.16
64	60	256	82.21	25.62	1.12	22.42	28.06	90.79	53.34	18.50	1.15
64	60	512	82.14	19.30	1.12	16.67	28.57	100.00	50.34	16.71	1.15
128	64	128	82.14	34.97	1.12	43.86	27.48	63.82	52.49	17.80	1.17
128	64	256	81.87	24.21	1.12	26.54	29.32	91.45	54.09	18.71	1.16
128	64	512	82.00	15.87	1.12	16.83	28.61	100.00	52.11	17.23	1.16
128	96	128	80.30	15.60	1.12	17.00	28.52	99.34	52.84	17.59	1.15
128	96	256	82.14	18.77	1.12	20.56	28.37	94.41	51.90	17.64	1.17
128	96	512	82.14	23.87	1.13	37.12	27.54	71.71	51.31	17.38	1.18
128	**120**	**128**	**82.14**	**15.40**	**1.13**	**16.67**	**28.57**	**100.00**	**54.67**	**19.83**	**1.18**
128	120	256	82.14	29.12	1.13	24.73	28.30	89.14	50.85	17.30	1.18
128	120	512	82.14	20.91	1.12	18.20	28.61	98.36	48.39	16.50	1.15

Table 7: STFT Ablation BNCI2014 P300

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen pr_auc	unseen loss
128	64	256	74.44	83.32	0.23	83.33	90.91	100.00	60.09	87.14	0.23
128	64	512	72.83	83.61	0.23	83.64	91.05	99.86	59.84	86.38	0.23
128	64	1024	72.87	83.32	0.23	83.33	90.91	100.00	58.68	86.10	0.23
128	96	256	78.32	83.57	0.23	84.34	91.40	99.86	62.86	87.18	0.23
128	96	512	78.16	83.40	0.23	83.60	91.03	99.88	64.72	88.37	0.23
128	**96**	**1024**	**71.92**	**82.99**	**0.23**	**84.41**	**91.42**	**99.61**	**65.19**	**88.60**	**0.22**
128	**120**	**256**	**74.03**	**83.69**	**0.22**	**83.95**	**91.21**	**99.88**	**65.05**	**89.14**	**0.22**
128	120	512	72.54	83.53	0.22	83.78	91.13	100.00	62.40	87.68	0.22
128	120	1024	83.44	83.61	0.24	84.05	91.26	99.93	62.56	87.81	0.24
256	**192**	**256**	**71.80**	**83.48**	**0.23**	**84.14**	**91.29**	**99.70**	**64.28**	**88.64**	**0.22**
256	192	512	77.95	83.36	0.23	83.31	90.90	99.95	60.43	86.58	0.23
256	192	1024	75.19	83.32	0.23	83.33	90.91	100.00	55.78	85.53	0.23
256	240	256	82.41	83.28	0.23	83.33	90.91	100.00	63.01	88.43	0.23
256	240	512	81.92	83.28	0.23	83.35	90.91	99.95	63.81	88.15	0.23
256	240	1024	82.91	83.32	0.23	83.33	90.91	100.00	61.28	87.20	0.23

Table 8: STFT Ablation MI

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen loss
256	128	512	35.42	29.46	1.34	24.13	19.51	24.13	50.03	1.39
256	128	1024	39.58	35.71	1.32	27.08	26.13	27.08	51.73	1.49
256	128	2048	35.12	35.71	1.34	23.44	20.72	23.44	51.72	1.41
256	192	512	35.42	33.63	1.37	26.91	21.55	26.91	51.57	1.40
256	**192**	**1024**	**38.39**	**35.71**	**1.34**	**28.47**	**23.27**	**28.47**	**53.26**	**1.41**
256	192	2048	31.85	31.25	1.38	25.35	18.44	25.35	51.86	1.39
256	240	512	37.50	31.55	1.36	25.17	23.32	25.17	51.86	1.42
512	**256**	**512**	**40.77**	**37.50**	**1.31**	**27.43**	**26.84**	**27.43**	**53.99**	**1.45**
512	256	1024	35.71	32.74	1.37	28.30	19.11	28.30	51.12	1.42
512	256	2048	34.23	30.95	1.37	24.83	16.85	24.83	50.66	1.43
512	384	512	39.29	33.04	1.32	27.08	22.40	27.08	53.65	1.42
512	**384**	**1024**	**33.93**	**33.04**	**1.37**	**28.82**	**20.34**	**28.82**	**52.73**	**1.39**
512	384	2048	34.82	30.36	1.37	25.17	17.59	25.17	52.45	1.40
512	480	512	35.71	31.85	1.37	27.08	23.93	27.08	53.41	1.39
512	480	1024	36.31	32.74	1.35	27.26	23.36	27.26	53.15	1.39
1000	500	1024	35.71	29.76	1.34	26.39	22.38	26.39	51.52	1.45
1000	500	2048	33.63	27.08	1.39	23.44	18.06	23.44	51.11	1.41
1000	500	4096	34.23	28.57	1.37	25.69	20.95	25.69	52.13	1.39
1000	750	1024	36.90	33.63	1.35	23.44	21.25	23.44	51.61	1.41
1000	750	2048	36.90	32.44	1.35	26.04	24.53	26.04	49.74	1.49
1000	750	4096	37.20	32.14	1.32	23.78	22.23	23.78	49.91	1.50
1000	937	1024	34.82	33.93	1.37	24.83	23.70	24.83	50.73	1.39
1000	937	2048	33.63	29.17	1.35	23.78	21.87	23.78	49.78	1.41

Table 9: STFT Ablation Lee2019 MI

nperseg	noverlap	nfft	val acc	seen acc	seen loss	unseen acc	unseen f1	unseen recall	unseen auc	unseen loss
32	16	32	57.34	51.06	0.67	51.88	66.81	96.89	56.01	0.69
32	16	256	55.75	51.31	0.68	51.46	66.23	95.18	55.16	0.69
32	16	512	55.00	49.31	0.68	50.84	66.78	98.82	54.10	0.69
32	**24**	**32**	**57.66**	**53.81**	**0.67**	**52.52**	**66.80**	**95.54**	**56.37**	**0.69**
32	24	256	55.47	50.41	0.68	51.20	66.42	96.54	54.53	0.69
32	**30**	**32**	**56.28**	**53.50**	**0.67**	**52.32**	**66.30**	**93.79**	**56.47**	**0.69**
64	32	64	55.16	49.25	0.68	50.70	66.76	99.04	55.21	0.70
64	32	256	55.44	49.28	0.69	51.14	66.78	98.21	54.55	0.70
64	32	512	55.63	49.19	0.68	50.77	66.81	99.11	54.65	0.69
64	48	64	56.53	51.13	0.68	51.54	66.48	96.11	55.95	0.69
64	48	256	55.47	52.91	0.68	52.13	66.40	94.61	53.91	0.70
64	48	512	55.44	49.06	0.68	50.64	66.75	99.07	53.78	0.69
64	**60**	**64**	**55.75**	**53.09**	**0.68**	**52.48**	**66.67**	**95.04**	**54.93**	**0.69**
128	64	128	54.78	48.00	0.69	49.95	66.58	99.71	53.55	0.70
128	64	256	55.69	48.56	0.69	50.34	66.63	99.18	53.87	0.70
128	64	512	54.63	48.72	0.68	50.66	66.67	98.71	53.40	0.69
128	96	128	56.47	50.31	0.67	51.04	66.63	97.75	55.14	0.69
128	96	256	56.13	49.53	0.68	51.38	67.06	99.00	55.48	0.69
128	96	512	55.78	49.06	0.68	50.55	66.67	98.93	55.06	0.69
128	120	128	55.66	52.72	0.68	51.63	66.12	94.39	54.25	0.69
128	120	256	55.28	49.41	0.69	51.16	66.83	98.39	53.22	0.69

A.3 Fusion Ablation

A.3.1 Fusion Methods

Here is a breakdown of each fusion method that we tried in the fusion ablation study.

Static Fusion: The simplest baseline approach using fixed equal weighting of both streams. Provides a baseline assuming both modalities contribute equally, with no learnable fusion parameters.

$$\mathbf{f} = \frac{1}{2}\mathbf{x}_s + \frac{1}{2}\mathbf{x}_t \tag{10}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}(\mathbf{f}) \tag{11}$$

Global Attention Fusion: Trial-level dynamic weighting that learns to adaptively balance the two streams based on the input. Allows the model to dynamically weight each stream per trial, enabling task-adaptive fusion. The residual connection improves gradient flow.

$$\mathbf{c} = [\mathbf{x}_s \| \mathbf{x}_t] \in \mathbb{R}^{B \times 2D} \tag{12}$$

$$\mathbf{a} = \text{softmax}\left(\frac{\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1\mathbf{c} + \mathbf{b}_1) + \mathbf{b}_2}{\tau}\right) \in \mathbb{R}^{B \times 2} \tag{13}$$

$$\mathbf{f} = a_0 \cdot \mathbf{x}_s + a_1 \cdot \mathbf{x}_t \tag{14}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}\left(\mathbf{f} + \sigma(\alpha) \cdot \frac{\mathbf{x}_s + \mathbf{x}_t}{2}\right) \tag{15}$$

where τ is a temperature parameter, α is a learnable residual scaling factor, $\mathbf{W}_1 \in \mathbb{R}^{D \times 2D}$, $\mathbf{W}_2 \in \mathbb{R}^{2 \times D}$, and $\|$ denotes concatenation.

Spatial Attention Fusion: Channel-level attention that computes different fusion weights for each EEG channel before global pooling. Different brain regions may benefit from different spectral/temporal balances. This allows the model to learn region-specific fusion strategies.

Let $\mathbf{X}_s \in \mathbb{R}^{B \times C \times D}$ be the per-channel spectral features (before averaging), and $\mathbf{x}_t \in \mathbb{R}^{B \times D}$ be the temporal features broadcast to $\mathbf{X}_t \in \mathbb{R}^{B \times C \times D}$.

$$\mathbf{C}_i = [\mathbf{X}_{s,i} \| \mathbf{X}_{t,i}] \in \mathbb{R}^{B \times 2D}, \quad \forall i \in \{1, \ldots, C\} \tag{16}$$

$$\mathbf{a}_i = \text{softmax}\left(\frac{\text{MLP}(\mathbf{C}_i)}{\tau}\right) \in \mathbb{R}^{B \times 2} \tag{17}$$

$$\mathbf{F}_i = a_{i,0} \cdot \mathbf{X}_{s,i} + a_{i,1} \cdot \mathbf{X}_{t,i} \tag{18}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}\left(\frac{1}{C}\sum_{i=1}^{C} \mathbf{F}_i\right) \tag{19}$$

Gated Linear Unit (GLU) Fusion: A gating mechanism that learns to suppress noise and control information flow. The gating mechanism can learn to suppress noisy or irrelevant feature dimensions while preserving informative ones, acting as a learned noise filter.

$$\mathbf{c} = \mathbf{x}_s + \mathbf{x}_t \tag{20}$$

$$\mathbf{g} = \sigma\left(\mathbf{W}_g[\mathbf{x}_s \| \mathbf{x}_t] + \mathbf{b}_g\right) \in \mathbb{R}^{B \times D} \tag{21}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}(\mathbf{c} \odot \mathbf{g}) \tag{22}$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, and $\mathbf{W}_g \in \mathbb{R}^{D \times 2D}$.

Multiplicative Fusion: Element-wise product of projected features to capture second-order interactions. Multiplicative interactions can capture correlations between the two streams that additive fusion cannot, enabling the model to learn feature conjunctions.

$$\mathbf{f} = (\mathbf{W}_s \mathbf{x}_s + \mathbf{b}_s) \odot (\mathbf{W}_t \mathbf{x}_t + \mathbf{b}_t) \tag{23}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}(\text{Dropout}(\mathbf{f})) \tag{24}$$

where $\mathbf{W}_s, \mathbf{W}_t \in \mathbb{R}^{D \times D}$.

Low-Rank Bilinear Fusion: Full bilinear interaction approximated with low-rank factorization for efficiency. Captures rich pairwise interactions between all feature dimensions while maintaining computational efficiency through the low-rank bottleneck.

A full bilinear model would compute $\mathbf{x}_s^\top \mathbf{W} \mathbf{x}_t$ with $\mathbf{W} \in \mathbb{R}^{D \times D \times D}$, which is computationally prohibitive. Instead, we use a low-rank approximation:

$$\mathbf{z}_s = \mathbf{U}_s \mathbf{x}_s \in \mathbb{R}^{B \times R} \tag{25}$$

$$\mathbf{z}_t = \mathbf{U}_t \mathbf{x}_t \in \mathbb{R}^{B \times R} \tag{26}$$

$$\mathbf{z} = \mathbf{z}_s \odot \mathbf{z}_t \in \mathbb{R}^{B \times R} \tag{27}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}(\text{Dropout}(\mathbf{W}_{\text{out}} \mathbf{z} + \mathbf{b}_{\text{out}})) \tag{28}$$

where $\mathbf{U}_s, \mathbf{U}_t \in \mathbb{R}^{R \times D}$ are low-rank projection matrices, $R \ll D$ is the rank (typically 8-32), and $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D \times R}$.

Cross-Attention Fusion: Multi-head cross-attention allowing one stream to attend to the other, followed by global weighting. Cross-attention allows the spectral stream to selectively incorporate information from the temporal stream based on learned relevance, providing a more flexible fusion mechanism than simple weighting.

Step 1: Multi-Head Cross-Attention

For H attention heads with head dimension $d_h = D/H$:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{x}_s, \quad \mathbf{K} = \mathbf{W}_K \mathbf{x}_t, \quad \mathbf{V} = \mathbf{W}_V \mathbf{x}_t \tag{29}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\tau\sqrt{d_h}}\right)\mathbf{V} \tag{30}$$

$$\mathbf{x}_{\text{attn}} = \mathbf{W}_O \cdot \text{Concat}(\text{head}_1, \ldots, \text{head}_H) \tag{31}$$

Step 2: Global Weighting

$$\mathbf{w} = \text{softmax}\left(\frac{\text{MLP}([\mathbf{x}_s \| \mathbf{x}_{\text{attn}}])}{\tau}\right) \in \mathbb{R}^{B \times 2} \tag{32}$$

$$\mathbf{f} = w_0 \cdot \mathbf{x}_s + w_1 \cdot \mathbf{x}_{\text{attn}} \tag{33}$$

$$\mathbf{f}_{\text{out}} = \text{BatchNorm}\left(\mathbf{f} + \sigma(\alpha) \cdot \frac{\mathbf{x}_s + \mathbf{x}_t}{2}\right) \tag{34}$$

A.3.2 FUSION ABLATION RESULTS

Table 10: Fusion Strategy Ablation on Wang 2016 SSVEP

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
nperseg=128 noverlap=64 nfft=256	Static	71.510	68.109	68.419	68.109	96.504
	Global Attn	73.932	69.311	69.490	69.311	97.114
	Spatial Attn	69.943	64.663	64.718	64.663	95.877
	GLU	76.211	68.990	69.355	68.990	96.537
	Multiplicative	71.795	67.147	67.440	67.147	95.512
	Bilinear	69.658	64.663	64.687	64.663	95.171
	Cross Attn	73.362	68.349	68.863	68.349	96.155
nperseg=256 noverlap=128 nfft=256	Static	72.222	68.990	69.393	68.990	96.580
	Global Attn	**76.211**	**72.756**	**73.000**	**72.756**	**97.543**
	Spatial Attn	72.080	67.788	67.929	67.788	96.727
	GLU	76.638	72.035	72.333	72.035	96.937
	Multiplicative	71.083	69.471	69.721	69.471	95.530
	Bilinear	70.513	68.750	69.210	68.750	95.647
	Cross Attn	75.214	72.035	72.180	72.035	97.073
nperseg=256 noverlap=128 nfft=512	Static	70.085	63.942	64.216	63.942	94.849
	Global Attn	72.650	67.388	67.596	67.388	94.758
	Spatial Attn	71.225	64.022	64.132	64.022	94.683
	GLU	75.783	71.474	71.592	71.474	96.872
	Multiplicative	72.365	66.026	66.262	66.026	95.219
	Bilinear	67.379	62.981	63.480	62.981	93.359
	Cross Attn	75.783	72.516	72.692	72.516	96.668

Table 11: Fusion Strategy Ablation on Lee2019 SSVEP

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
nperseg=128 noverlap=64 nfft=512	Static	94.906	85.964	85.955	85.964	97.716
	Global Attn	94.906	83.250	83.245	83.250	96.912
	Spatial Attn	94.906	84.821	84.800	84.821	97.363
	GLU	95.219	85.214	85.244	85.214	97.534
	Multiplicative	95.250	85.518	85.498	85.518	97.460
	Bilinear	94.562	85.411	85.394	85.411	97.338
	Cross Attn	94.938	85.250	85.243	85.250	97.227
nperseg=256 noverlap=128 nfft=1024	Static	95.125	86.464	86.504	86.464	97.844
	Global Attn	95.156	85.750	85.711	85.750	97.558
	Spatial Attn	94.758	86.000	86.024	86.000	97.597
	GLU	94.781	84.786	84.768	84.786	97.197
	Multiplicative	94.844	85.714	85.687	85.714	97.573
	Bilinear	**95.000**	**86.679**	**86.637**	**86.679**	**97.783**
	Cross Attn	95.031	85.286	85.254	85.286	97.376
nperseg=256 noverlap=192 nfft=1024	Static	94.281	85.714	85.696	85.714	97.467
	Global Attn	94.938	85.893	85.893	85.893	97.500
	Spatial Attn	94.156	85.375	85.353	85.375	96.641
	GLU	94.969	85.518	85.493	85.518	97.329
	Multiplicative	94.438	84.679	84.672	84.679	97.363
	Bilinear	94.719	84.429	84.432	84.429	96.868
	Cross Attn	94.531	82.929	82.925	82.929	96.564

Table 12: Fusion Strategy Ablation on BI2014b P300

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
	Static	69.121	68.202	32.084	45.066	60.730
	Global Attn	65.372	67.599	33.220	48.355	62.038
nperseg=32	Spatial Attn	65.372	69.737	33.654	46.053	63.271
noverlap=16	GLU	62.645	63.322	29.801	46.711	55.929
nfft=512	Multiplicative	65.167	66.118	25.721	35.197	57.451
	Bilinear	**74.506**	**73.520**	**30.101**	**34.211**	**59.848**
	Cross Attn	64.963	66.338	27.078	37.500	59.078
	Static	67.144	66.228	28.205	39.803	58.639
	Global Attn	69.325	71.382	29.839	36.513	60.829
nperseg=32	Spatial Attn	60.123	59.978	28.988	49.013	58.069
noverlap=30	GLU	63.258	62.555	29.515	47.039	57.759
nfft=32	Multiplicative	62.509	62.171	28.571	45.395	58.654
	Bilinear	61.963	61.897	28.424	45.395	57.899
	Cross Attn	64.894	61.952	28.454	45.395	58.056
	Static	72.529	72.423	30.042	35.526	60.018
	Global Attn	65.235	64.748	27.508	40.132	57.483
nperseg=128	Spatial Attn	62.849	62.664	28.691	45.066	59.455
noverlap=120	GLU	62.781	66.118	34.255	52.961	62.957
nfft=128	Multiplicative	61.486	61.732	30.478	50.329	61.736
	Bilinear	62.236	63.103	27.712	42.434	58.111
	Cross Attn	63.122	66.118	28.966	41.447	60.058

Table 13: Fusion Strategy Ablation on BNCI2014-009 P300

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
	Static	85.301	87.751	92.520	90.903	89.936
	Global Attn	86.210	88.580	93.089	92.292	90.390
nperseg=128	Spatial Attn	85.756	88.619	93.070	91.713	90.766
noverlap=96	GLU	83.485	87.191	92.042	88.889	91.374
nfft=1024	Multiplicative	86.581	89.410	93.624	93.310	90.953
	Bilinear	86.664	89.275	93.529	93.009	90.647
	Cross Attn	85.260	87.519	92.353	90.440	90.394
	Static	84.682	88.465	92.940	91.111	91.036
	Global Attn	87.903	88.214	92.955	93.310	88.599
nperseg=128	Spatial Attn	85.343	88.484	92.932	90.856	90.969
noverlap=120	GLU	84.765	87.596	92.333	89.630	91.103
nfft=256	**Multiplicative**	**87.903**	**89.815**	**93.859**	**93.403**	**90.400**
	Bilinear	88.233	89.699	93.814	93.727	91.694
	Cross Attn	86.168	89.178	93.427	92.292	91.685
	Static	86.788	89.352	93.587	93.241	91.220
	Global Attn	84.393	86.748	91.791	88.912	90.294
nperseg=256	Spatial Attn	83.237	86.632	91.715	88.796	89.918
noverlap=192	GLU	86.210	89.063	93.372	92.454	91.126
nfft=256	Multiplicative	85.838	88.677	93.113	91.852	91.135
	Bilinear	84.228	87.905	92.508	89.606	92.356
	Cross Attn	85.136	87.654	92.321	89.051	91.523

Table 14: Fusion Strategy Ablation on BNCI2014-001 MI

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
nperseg=512 noverlap=384 nfft=1024	Static	47.917	28.472	26.085	28.472	54.502
	Global Attn	43.750	27.083	26.188	27.083	53.981
	Spatial Attn	36.607	24.479	22.860	24.479	50.203
	GLU	36.012	24.479	24.411	24.479	50.760
	Multiplicative	47.321	28.299	27.440	28.299	55.923
	Bilinear	50.000	28.993	28.586	28.993	56.170
	Cross Attn	55.060	28.125	26.986	28.125	57.382
nperseg=512 noverlap=256 nfft=512	Static	54.762	28.125	27.264	28.125	55.060
	Global Attn	50.298	28.472	26.788	28.472	55.892
	Spatial Attn	34.226	25.521	23.843	25.521	52.102
	GLU	44.048	26.563	24.716	26.563	54.395
	Multiplicative	**54.464**	**30.729**	**30.029**	**30.729**	**56.532**
	Bilinear	30.655	26.042	15.248	26.042	48.723
	Cross Attn	50.000	27.778	23.785	27.778	58.496
nperseg=256 noverlap=192 nfft=1024	Static	44.643	25.000	24.727	25.000	54.229
	Global Attn	36.905	26.563	22.866	26.563	52.112
	Spatial Attn	46.131	27.083	25.778	27.083	55.271
	GLU	36.905	24.826	21.697	24.826	50.911
	Multiplicative	52.083	29.688	28.651	29.688	57.098
	Bilinear	50.893	29.514	27.590	29.514	59.252
	Cross Attn	42.262	27.431	25.301	27.431	52.867

Table 15: Fusion Strategy Ablation on Lee2019 MI

STFT Config	Strategy	Val Acc	Unseen Acc	Unseen F1	Unseen Rec	Unseen AUC
nperseg=32 noverlap=24 nfft=32	Static	76.969	74.714	76.266	81.250	83.905
	Global Attn	77.031	74.589	75.520	78.393	83.287
	Spatial Attn	78.000	74.607	75.287	77.357	83.576
	GLU	77.844	74.125	74.566	75.857	83.285
	Multiplicative	78.344	75.500	77.149	82.714	84.523
	Bilinear	77.625	74.214	72.620	68.393	84.000
	Cross Attn	77.406	75.125	75.271	75.714	83.925
nperseg=32 noverlap=30 nfft=32	Static	76.563	74.714	76.129	80.643	83.005
	Global Attn	74.813	74.679	75.288	77.143	83.193
	Spatial Attn	77.375	75.250	76.437	80.286	84.481
	GLU	76.719	75.357	76.067	78.321	84.637
	Multiplicative	**78.406**	**75.696**	**76.723**	**80.107**	**83.754**
	Bilinear	77.281	75.232	76.910	82.500	84.301
	Cross Attn	75.375	74.714	76.870	84.036	83.780
nperseg=64 noverlap=60 nfft=64	Static	76.313	74.429	75.293	77.929	83.844
	Global Attn	77.281	73.286	75.777	83.571	83.165
	Spatial Attn	77.781	75.071	75.806	78.107	84.327
	GLU	77.656	75.357	75.958	77.857	84.013
	Multiplicative	78.563	74.875	76.285	80.821	83.780
	Bilinear	77.438	75.482	75.746	76.571	84.003
	Cross Attn	76.906	74.946	75.356	76.607	84.132

A.4 Training Details

To ensure reproducibility, we adopt a standardized training protocol across all evaluated tasks. The model is trained for a maximum of 100 epochs using the Adam optimizer with a batch size of 64. We utilize an initial learning rate of 3×10^{-4} with a weight decay of 5×10^{-4}. The learning rate is managed by a `ReduceLROnPlateau` scheduler, which reduces the rate by a factor of 0.5 following 5 epochs of stagnant validation loss. Regularization is enforced through a global dropout rate of 0.3, a CNN-specific dropout of 0.25, and gradient norm clipping at a threshold of

1.0. Training terminates early if validation accuracy does not improve for 15 consecutive epochs. All experiments are conducted with a fixed random seed of 44, 36, and 10 to ensure deterministic behavior.