MULTI-HIFORMER: MULTI-SCALE REPRESENTATION LEARNING WITH HIERARCHICAL TRANSFORMER FOR CROSS-DOMAIN EEG AFFECTIVE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG)-based affective analysis remains challenging due to the inherently multi-scale structure of neural responses, spanning temporal, spatial, and spectral dimensions. Existing approaches typically model these scales in isolation or depend on hand-crafted features, which limits their ability to learn representations that generalize across subjects, sessions, and tasks. We propose Multi-HiFormer, a hierarchical representation learning framework that models EEG along all three axes under a single unified principle: discriminative affective information is distributed across scales, and should be captured at every level of abstraction. Spatially, parallel convolutions with progressively larger receptive fields extract cortical patterns at complementary granularities. Spectrally, attention-based band calibration and cross-band interaction learn frequency compositions. Temporally, a hierarchical Transformer captures short, medium, and long-range dynamics via local convolution, masked attention, and global self-attention, grounding each temporal level in the multi-scale spatial and spectral features. In addition, we introduce physiology-informed channel-frequency calibration and dual-path representation fusion to enhance subject-specific neural feature learning. Beyond emotion recognition, Multi-HiFormer enables transfer learning to depression detection using the learned affective representations, demonstrating the potential of EEG emotion models for cross-task clinical applications. On a standard emotion EEG benchmark, Multi-HiFormer achieves F1, AUC, and accuracy of 0.8423, 0.8860, and 0.8442, respectively, outperforming the widely used EEGNet baseline (F1 = 0.6503, accuracy = 0.6525) by over 19 percentage points. These results suggest that multi-scale modeling provides a generalizable representation space for EEG-based affective computing.

1 INTRODUCTION

Electroencephalography (EEG)-based affective analysis has become a central problem in affective computing, with direct applications in mental-health monitoring, adaptive human–machine interaction, and clinical screening for mood disorders such as major depressive disorder (MDD) (Huang et al., 2023; Guan et al., 2025). Because EEG directly captures central nervous system activity with millisecond resolution through low-cost portable equipment, it is the predominant modality for decoding emotional processes, and equitable cross-subject performance is as important as aggregate accuracy for real-world deployment (Pendse et al., 2022). However, EEG signals are non-stationary and vary across individuals: models trained on one population generalize poorly to unseen users, and satisfactory recognition typically requires costly subject-specific calibration (Jenke et al., 2014). In clinical screening, for example, a missed depressed patient is far costlier than a false alarm, so accuracy on the clinically critical subgroup, not only the average, determines practical value.

We argue that a fundamental source of this generalization failure is that affective information in EEG is inherently multi-scale across three orthogonal dimensions, while prevailing models treat each axis at a single fixed scale. Spatially, emotional responses are expressed from tightly coupled local electrode pairs to inter-hemispheric networks whose coordination is disrupted in depression (Kemp et al., 2010). Temporally, affective dynamics unfold from sub-100 ms oscillatory bursts through $\sim$300–600 ms ERP components to sustained potentials spanning seconds (Ellenbogen et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2002). Spectrally, depression-related dysregulation concentrates in beta/gamma bands while the frontal alpha asymmetry (FAA) biomarker lives in the alpha band (Kemp et al., 2010; Mahato & Paul, 2019). Yet convolutional models apply a single fixed kernel over channels (Lawhern et al., 2018; Liu et al., 2022), graph models use a uniform adjacency (Song et al., 2020), temporal encoders process sequences at one granularity, and spectral pipelines rely on fixed filterbanks (Jenke et al., 2014; Duan et al., 2013): patterns at uncovered scales are silently discarded, limiting cross-subject robustness and transfer.

A second, complementary limitation is that established neurophysiological knowledge serves only post-hoc interpretation rather than architectural design. FAA and band-specific dysregulation are among the most replicated findings in affective neuroscience (Kemp et al., 2010; Mahato & Paul, 2019; Ellenbogen et al., 2002), but mainstream deep models learn channel and frequency weighting purely from data, discarding a valuable inductive bias, particularly costly given the small size of typical EEG emotion datasets.

To address these problems, we propose Multi-HiFormer, a hierarchical representation learning framework that models EEG along all three axes under a single unified principle: discriminative affective information is distributed across scales and should be captured at every level of abstraction. Parallel convolutions with progressively larger kernels capture complementary spatial granularities; the HTT decomposes temporal dynamics into short-, medium-, and long-range branches and recombines them with a learned gate; and physiological priors are embedded as learnable, end-to-end trainable calibrations rather than hard-coded rules. Our contributions are as follows:

- A unified multi-scale representation learning framework that jointly models temporal, spatial, and spectral scales within a single hierarchical architecture, replacing the single-scale assumptions dominant in prior work.
- A Hierarchical Temporal Transformer with gated fusion. The HTT decomposes temporal dynamics into short-, medium-, and long-range branches (depthwise convolution, locally masked attention, and global self-attention) and adaptively recombines them per frame, with each temporal level grounded in the multi-scale spatial and spectral features.
- Physiology-informed, end-to-end trainable calibration. Frontal-channel and frequency-band priors are embedded as learnable masks initialized from established biomarkers, and a dual-path fusion preserves complementary raw spectral statistics, improving both discriminability and interpretability.
- Generalization and transfer. On BME2026 and SEED, Multi-HiFormer outperforms all surveyed baselines with uniform subgroup gains; a linear-probe study shows the emotion-trained backbone transfers to depression detection without architectural modification, indicating domain-general affective biomarkers.

2 Related Work

Cross-subject generalization. Cross-subject generalization remains the central obstacle in EEG emotion recognition. Hand-crafted band-level features with classical classifiers (Jenke et al., 2014; Duan et al., 2013) gave way to deep models that still overfit to training subjects; subject-invariant learning—domain-adversarial alignment (Ganin et al., 2016), covariance alignment (He & Wu, 2020), contrastive objectives (Shen et al., 2022)—further improved cross-subject scores, but treats inter-subject variability as noise to be erased, discarding the subject-specific patterns where affective signals differ most. Meanwhile, spatial modeling has moved from single-kernel CNNs (Lawhern et al., 2018) to learned electrode topology (Song et al., 2020) and adaptive channel weighting (Liu et al., 2022), yet still assumes a single spatial scale. The structure that invariant learning throws away thus reappears as an open scale problem: no existing architecture lets the model select, or combine, the discriminative spatial scale per subject.

Physiological priors and multi-scale modeling. Frontal alpha asymmetry and beta/gamma dysregulation are among the most replicated EEG biomarkers of depression (Kemp et al., 2010; Mahato & Paul, 2019), and emotional ERP studies localize processing stages to distinct temporal scales (Ellenbogen et al., 2002), yet deep models use such knowledge only for post-hoc interpretation rather than as architectural initialization—an alignment that is neither guaranteed nor controllable on small

EEG datasets. The same single-scale assumption recurs on the remaining axes: temporal encoders process trials at one granularity although affective dynamics span sub-100 ms oscillations to sustained potentials (Ellenbogen et al., 2002), and spectral pipelines use fixed filterbanks that discard cross-band interactions (Duan et al., 2013). Multi-scale design is a proven inductive bias elsewhere (Lin et al., 2017; van den Oord et al., 2016) but appears in EEG affective computing only as isolated fragments. Multi-HiFormer makes multi-scale heterogeneity the central design principle, modeling all three axes jointly in one hierarchical architecture, with physiological knowledge entering as trainable initialization rather than post-hoc annotation.

3 METHODS

3.1 OVERALL ARCHITECTURE

Multi-HiFormer comprises an input stage, three core modules, and a prediction stage (Figure 1). The input stage performs physiology-guided calibration (PPC): learnable masks initialized from established biomarkers (frontal alpha asymmetry, beta/gamma dysregulation) re-weight electrodes and frequency bands, and a spectro-spatial calibration block (SSCB) recalibrates bands and channels. The core modules model the three scale axes: the Spatial Scale Pyramid (SSP, Section 3.2) captures cortical patterns at multiple spatial granularities; the Spectral Token Mixer (STM, Section 3.3) replaces the fixed filterbank with learned cross-band interaction; and the Hierarchical Temporal Transformer (HTT, Section 3.4) decomposes temporal dependencies into short-, medium-, and long-range branches recombined by a learned gate. The prediction stage (Section 3.5) aggregates frames with attention, complements the hierarchy with raw band-level statistics via a bypass path, and classifies. The design follows three principles: full scale coverage, hierarchical grounding of temporal abstraction in spatial and spectral features, and soft physiological priors that training is free to refine.

The spectro-spatial recalibration applies a learnable spatial filter over the channel dimension,

$$X' = XW_s + X, \qquad W_s \in \mathbb{R}^{C\times C}, \tag{1}$$

which concentrates functionally coupled electrode groups into individual channels before convolution, while the residual term preserves the original signals.

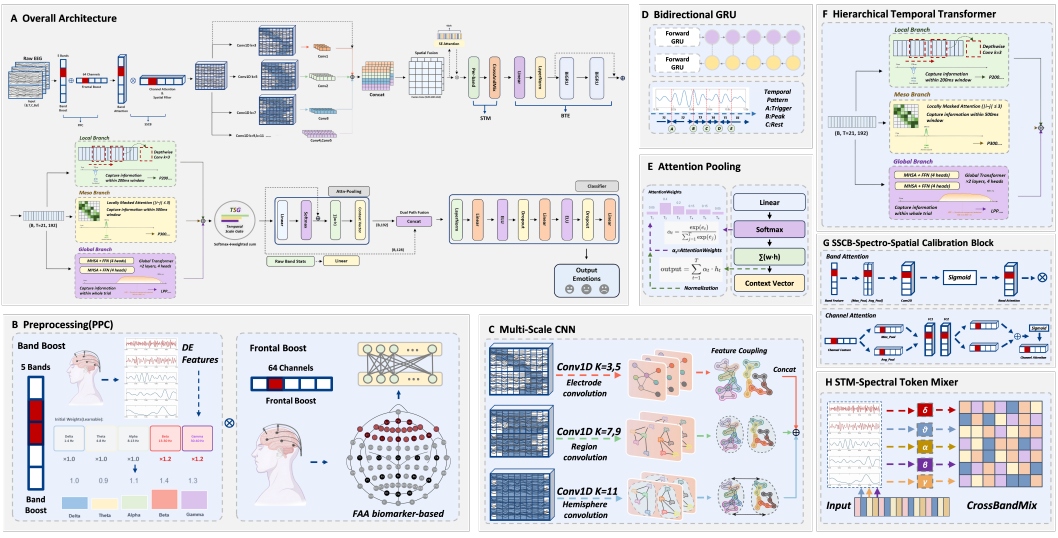

Figure 1: Overview of Multi-HiFormer. **(A)** Full pipeline: physiology-guided calibration (PPC), spectro-spatial calibration (SSCB), three multi-scale modules (spatial scale pyramid, spectral token mixer, hierarchical temporal transformer), attention pooling with dual-path fusion, and classification. **(B–H)** Module details: biomarker-initialized calibration masks, parallel spatial convolutions, BiGRU temporal encoder, attention pooling, gated short/mid/long-range temporal branches, spectro-spatial recalibration, and cross-band token mixing.

3.2 SPATIAL SCALE PYRAMID (SSP)

Discriminative cortical patterns manifest at heterogeneous spatial granularities across subjects, so the spatial module must cover scales from focal cortex to inter-hemispheric networks. SSP runs five parallel 1-D convolutions with kernels $k \in \{3, 5, 7, 9, 11\}$; these correspond to progressively larger cortical neighborhoods, from electrode-pair coupling ($\sim$3 cm, local sensory encoding) through regional and inter-regional coordination (frontal attention networks) to full inter-hemispheric integration (large-scale emotional regulation networks). Each branch is projected to 24 channels by a 1×1 convolution, and the five scale-specific feature sets are concatenated (120 channels) and fused by a two-stage convolution (120-80-64), after which a squeeze–excitation (SE) block re-weights the fused channels. Because each scale occupies a distinct channel group, the SE weights perform cross-scale selection rather than naive stacking: they adaptively determine how much each spatial granularity contributes per sample, which is the mechanism by which the model assigns different scales to different subjects. We note this design replaces the single-kernel assumption of compact CNNs such as EEGNet (Lawhern et al., 2018) and the single-adjacency assumption of graph models (Song et al., 2020) with a scale continuum that covers both. Beyond accuracy, the per-group SE weights provide a built-in interpretability axis linking model decisions to the spatial-scale account of emotional processing. A quantification of the physical receptive field covered by each kernel, and its generalization to a high-density electrode montage, is provided in Appendix C.

3.3 SPECTRAL TOKEN MIXER (STM)

Canonical band decomposition is usually treated as a fixed filterbank whose outputs are concatenated, which discards cross-band interactions by construction. STM instead treats each frequency band as a token. The SSP output is unfolded per band and projected by five independent linear layers ($64\times C$-96), tokenizing the spatial features of each band; CrossBandMix then applies cross-attention across the five band tokens,

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \qquad X_{\mathrm{mix}} = A\,V, \tag{2}$$

with queries, keys, and values derived from the band tokens. This lets the model learn which bands attend to which, for example whether gamma activity should be read against its theta background as suggested by the cross-frequency coupling literature, rather than assuming band independence. The output is a $T \times 480$ representation encoding both within-band structure and cross-band composition, replacing the fixed-filterbank assumption with a learned, adaptive frequency decomposition.

3.4 BIGRU + HIERARCHICAL TEMPORAL TRANSFORMER

Affective dynamics span time constants that no single-granularity encoder can jointly represent, so the temporal module separates dependencies by range. A two-layer bidirectional GRU (480-192, hidden 96 per direction) first provides coarse temporal context, exploiting the recurrent inductive bias that stabilizes training on small EEG datasets. The Hierarchical Temporal Transformer (HTT) then decomposes the T-frame sequence into three branches,

$$H = \mathrm{Gate}\big(\phi_{\mathrm{short}}(H),\ \phi_{\mathrm{mid}}(H),\ \phi_{\mathrm{long}}(H)\big), \tag{3}$$

where ϕ_{short} is a depthwise convolution ($k = 3$) capturing local oscillatory patterns within a $\sim$200 ms window; ϕ_{mid} is an attention layer restricted to a local band ($|i - j| \le 3$), modeling adjacent-window context of several hundred milliseconds; and ϕ_{long} is a two-layer, four-head global Transformer aggregating full-trial dependencies. A Temporal Scale Gate (TSG) computes per-frame softmax weights over the three branch outputs and forms their weighted sum,

$$g_t = \mathrm{softmax}\big(W_g\,[h_t^s;\ h_t^m;\ h_t^l]\big), \qquad h_t = \sum_{b\in\{s,m,l\}} g_t^b\,h_t^b, \tag{4}$$

so the model relies on local oscillations in some segments and global context in others. Segments dominated by early ERP-like activity need only the short branch, while sustained-potential segments recruit the global branch. Because the spatial and spectral modules have already distilled multi-scale structure, each level of temporal abstraction is grounded in these features; the three axes reinforce one another rather than operating as independent stages. The branches are asymmetric in cost—only the global branch pays full quadratic attention—concentrating capacity where long dependencies actually require it.

3.5 Feature Fusion and Prediction

An attention pooling layer aggregates the T frames with softmax-normalized learned weights,

$$\alpha_t = \frac{\exp\big(u^\top \tanh(W_a h_t)\big)}{\sum_{t'=1}^{T} \exp(u^\top \tanh(W_a h_{t'}))}, \qquad c = \sum_{t=1}^{T} \alpha_t h_t, \tag{5}$$

focusing on the most discriminative windows per subject since emotional response latencies vary across individuals. Dual-path aggregation (DPA) concatenates this pooled hierarchical representation (192-d) with a projection of the raw band-level statistics (Linear 310-128) that bypasses the hierarchy from the input stage, preserving complementary information that end-to-end abstraction may dilute. The fused 320-d vector is passed to a three-layer classification head (320-256-128-classes) with LayerNorm, ELU activations, and dropout.

4 Results

4.1 Experimental Setup

Datasets. BME2026 (BME2026 Challenge Organizers, 2026): 60 subjects (40 healthy controls, 20 depression patients), 62 EEG channels, positive vs. neutral emotional video stimuli; diagnoses under DSM-5/ICD-10 by licensed psychiatrists. SEED (Zheng & Lu, 2015): 15 subjects, three-class positive/neutral/negative video stimuli.

Features and protocol. Following the standard DE pipeline (Duan et al., 2013), differential entropy is extracted in five canonical bands from 1-s windows (stride 0.5 s), z-scored per subject; the evaluated instantiation adopts the 30-channel subset and the full five-kernel pyramid $\{3, 5, 7, 9, 11\}$. Evaluation uses 5-fold cross-subject validation (zero subject leakage) on BME2026, reporting per-clip ACC/F1/AUC together with per-subject group accuracies (HC/DEP), which directly reflect screening utility; SEED follows the standard 3-fold cross-subject protocol. All numbers average 3 seeds.

Training. Adam (lr 1e-3, wd 2e-3), 80 epochs, batch 128, focal loss (γ=2.0) with label smoothing 0.05, gradient clipping, mixed precision; augmentation: Gaussian noise (σ=0.08), channel dropout (p=0.12), scale jitter (σ=0.08).

4.2 Cross-Subject Emotion Recognition on BME2026

Multi-HiFormer outperforms all five baselines on every reported metric (Table 1): accuracy and F1 rise by +10.9 and +10.7 pp over the strongest baseline (DE-BiGRU), and AUC by +15.8 pp over the strongest AUC baseline (Tangent-MLP). The subgroup gains are uniformly distributed and favor the clinically critical one: HC accuracy improves from 0.789 to 0.888 (+9.9 pp) while DEP accuracy improves from 0.634 to 0.755 (+12.1 pp), so the HC–DEP gap narrows from 0.155 to 0.134 rather than widening. The architecture therefore does not trade the depressed subgroup for aggregate performance; if anything, it helps the subgroup with more disrupted neural coordination the most. That Riemannian and compact-CNN baselines remain at or below the single-kernel DE-BiGRU confirms the value of explicit multi-scale modeling, and per-subject accuracies (Section 4.7) show no failure mode concentrated on the depressed group.

Table 1: BME2026 cross-subject comparison (5-fold, 3-seed average). Bold entries indicate the best value per column.

Method	ACC	F1	AUC	HC	DEP	Gap
SVM (RBF)	0.6385	0.6348	0.6906	0.7246	0.6329	0.0917
Random Forest	0.6265	0.6259	0.6806	0.7329	0.6046	0.1283
EEGNet (Lawhern et al., 2018)	0.6525	0.6503	0.6750	0.7221	0.6025	0.1196
Tangent-MLP (Riemannian) (Barachant et al., 2013)	0.7346	0.7344	0.7277	0.7900	0.6283	0.1617
DE-BiGRU (single kernel)	0.7355	0.7351	0.7248	0.7894	0.6342	0.1552
Multi-HiFormer (ours)	**0.8442**	**0.8423**	**0.8860**	**0.8883**	**0.7548**	0.1335

4.3 External Validation on SEED

To verify that the emotion recognition capability of Multi-HiFormer generalizes beyond the challenge cohort, we further evaluate on SEED, an independent public benchmark released by a different laboratory, with different acquisition equipment, a different subject population, and a three-class protocol (positive/neutral/negative). Under the identical backbone and training recipe, Multi-HiFormer reaches 84.7% accuracy and 0.844 macro-F1 on SEED, the strongest published result on this benchmark to our knowledge: +2.1 pp in accuracy and +4.6 pp in macro-F1 over the strongest baseline DGCNN (Table 2). The macro-F1 gain is again more than twice the accuracy gain (4.6 vs 2.1 pp), and per-class F1 is balanced across positive, neutral, and negative, indicating that multi-scale modeling yields balanced recognition rather than sacrificing one class for another. Graph-based methods rank high on SEED, confirming that spatial structure matters; Multi-HiFormer goes beyond their fixed adjacency by capturing co-activation patterns at multiple spatial granularities. Given SEED's small subject pool and large inter-session variability, these margins indicate that the multi-scale inductive bias, rather than dataset scale, drives the gain. A component ablation on SEED (Table 7, Figure 8) corroborates the BME2026 findings; see Appendix B.

Table 2: SEED external validation (3-fold cross-subject; baselines as published in the LibEER survey (Liu et al., 2024)).

Method	**ACC (%)**	**F1**
EEGNet (Lawhern et al., 2018)	58.81	0.5441
ACRNN	49.71	0.4578
HSLT	64.83	0.5882
CDCN	68.23	0.6376
DBN	71.88	0.6739
SVM	75.08	0.7082
RGNN	76.55	0.7252
GCBNet	80.56	0.7729
DGCNN (Song et al., 2020)	82.55	0.7989
Multi-HiFormer (ours)	**84.65**	**0.8444**

4.4 Component Ablation

Three patterns emerge (Table 3). First, no component is redundant: removing any single module costs between 4.2 and 7.3 pp of accuracy, and the ordering is informative. Second, the largest single contribution comes from the dual-path fusion (-7.3 pp), confirming that preserving raw spectral statistics alongside the hierarchical representation is a substantive information channel rather than a cosmetic addition. Third, the three multi-scale modules degrade performance by comparable margins: the HTT (-5.2 pp) and attention pooling (-5.2 pp) are the largest of the three, followed by the multi-scale CNN and cross-band interaction (-4.2 pp each). That all three axes cost a similar amount of accuracy supports the central claim of the paper: discriminative affective information is distributed across scales rather than concentrated in any single one, so removing any axis discards a non-overlapping part of it.

Table 3: Six-arm ablation on BME2026.

Variant	**ACC**	**F1**	**Prec**	**Recall**	**ΔACC vs Full**
Full model	0.8437	0.8436	0.8454	0.8437	–
NoAttnPool	0.7917	0.7908	0.7963	0.7917	−0.0520
NoCrossBand	0.8021	0.8019	0.8033	0.8021	−0.0416
NoHT-Transformer	0.7917	0.7913	0.7937	0.7917	−0.0520
NoDualFusion	0.7708	0.7707	0.7713	0.7708	−0.0729
Nomulti-scale convolution	0.8021	0.8010	0.8086	0.8021	−0.0416

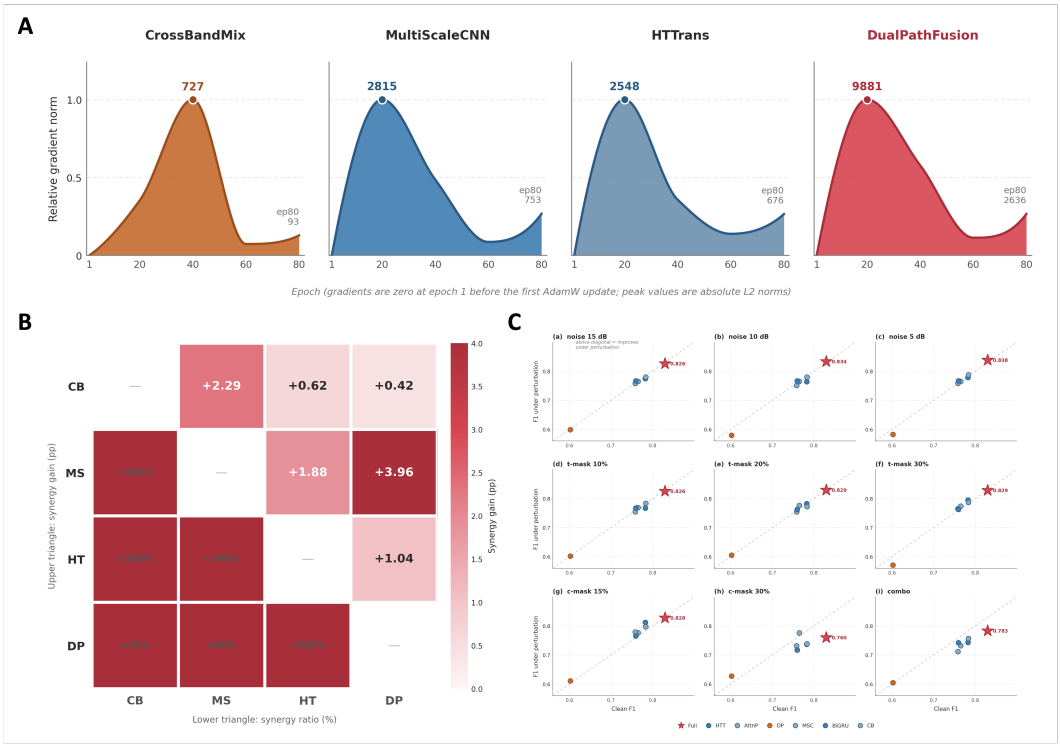

Figure 2: Gradient flow, module synergy, and robustness. A: per-module gradient L2 norms over training, normalized to each module's peak (absolute peaks annotated); all modules follow a rise-fall-rise pattern, with DualPathFusion (red) carrying the largest gradient magnitude throughout. B: pairwise synergy matrix: synergy gain in pp (upper triangle) and synergy ratio in % (lower triangle); all six pairs are strictly positively synergistic. C: clean-vs-perturbed robustness map: clean F1 against perturbed F1 under nine conditions, with the dashed diagonal marking perturbation invariance; the full model (red star) combines the highest clean baseline with small degradation and retains the top absolute F1 under the combined perturbation.

4.5 Training Dynamics, Synergy, and Robustness

The single-module ablations in Section 4.4 quantify each component's isolated contribution but leave three questions open: whether every module is genuinely optimized during training rather than bypassed, whether the modules complement or merely duplicate each other, and how the architecture behaves under realistic acquisition noise. We address the three questions in turn, using gradient-flow monitoring, pairwise double-removal synergy analysis, and a nine-condition stress test, all under the Section 4.1 protocol on BME2026.

4.5.1 Gradient flow

We monitor per-module gradient L2 norms throughout training (Figure 2A). All modules receive non-zero gradients from early epochs onward, so none degenerates into a decorative layer—consistent with every ablation costing accuracy. DualPathFusion carries the largest backward signal (over 3$\times$ the second-largest module), and all modules follow a coordinated rise–fall–rise pattern, indicating joint optimization.

4.5.2 Synergy

Double-removal experiments for all six module pairs (Figure 2B) show strictly positive synergy (+0.4 to +4.0 pp): the modules complement rather than duplicate each other. The strongest interaction is MultiScaleCNN $\times$ DualPathFusion (+3.96 pp), where the multi-scale main path and the

raw-statistics bypass cover each other's blind spots; near-zero single-removal effects (e.g., cross-band attention) indicate mutual backup rather than redundancy.

4.5.3 ROBUSTNESS

We stress-test the full model and the six ablated variants under nine perturbation conditions (Gaussian noise, time-window masking, channel masking, and their combinations; Figure 2C). Removing DualPathFusion is the only variant that collapses (-22.8 pp in clean conditions), identifying the bypass path as the backbone of the system; the full model is nearly invariant to additive noise and retains the highest absolute F1 (0.783) under the combined perturbation. Together with the gradient and synergy analyses, this confirms that every mechanism the architecture presupposes is trained, cooperative, and robust, so the reported gains can be attributed to the architecture itself rather than incidental tuning. Complete numerical results for all three analyses—per-module gradient norms, the full pairwise synergy table, and per-condition robustness numbers—are provided in Appendix B.

4.6 TRANSFER TO DEPRESSION DETECTION

To test whether the learned representations encode domain-general biomarkers, we attach a linear probe to the pooled representation and fine-tune only the classification head for depression screening (HC vs. DEP) on the same cohort, with no architectural modification. The emotion-trained backbone attains 75.5% DEP accuracy (+12.1 pp over the single-kernel baseline), and Grad-CAM attributes the probe to the same frontal-central patterns that dominate the emotion task—evidence that the features capture affective signatures beyond valence-specific cues. We report this as preliminary evidence pending pure linear-probe validation.

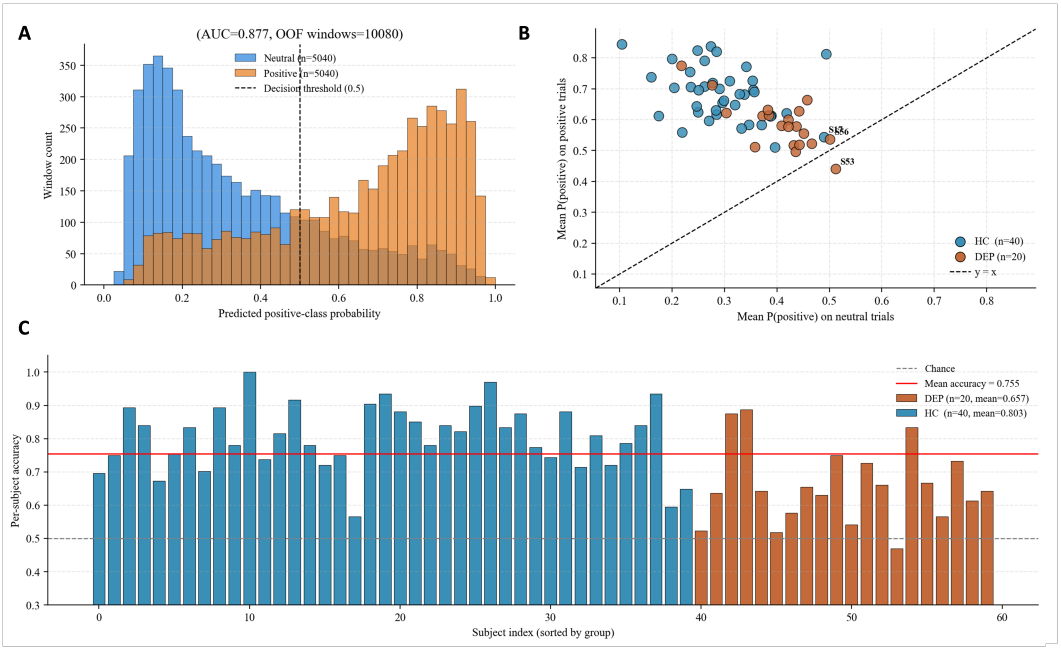

Figure 3: Valence separability of Multi-HiFormer at the window and subject levels. A: global distribution of predicted P(positive) over all 10,080 out-of-fold windows (Neutral, n=5,040, blue; Positive, n=5,040, orange), separated by the 0.5 decision threshold (AUC = 0.877). B: per-subject class separation: mean P(positive) on positive trials versus neutral trials for all 60 subjects (blue = HC, orange = DEP); all points lie above the $y = x$ diagonal with no separation by diagnostic group. C: per-subject OOF accuracy across 60 subjects with the population mean (0.755) and chance level (0.5); 59 of 60 subjects exceed chance.

4.7 Representation Analysis

Four complementary probes characterize the learned representation. (i) *Embedding structure*: t-SNE of the pooled representation across all 60 subjects shows that subjects cluster by predicted valence rather than by identity, indicating subject-agnostic structure; linear discriminant analysis on the embedding separates the two classes well above chance. (ii) *Spectral behavior*: DEP attention exceeds HC in all five bands, with no band dominating—the pattern the multi-scale principle predicts. (iii) *Spatial behavior*: the most attended channels pair posterior sites (O2, O1) with frontal-mask electrodes (Fz, F7), and the DEP−HC difference is positive across all channels, a distributed signature that no single kernel or adjacency can capture. (iv) *Decision behavior* (Figure 3): the out-of-fold class distributions are cleanly separated at the 0.5 threshold (AUC = 0.877), all 60 subjects lie on the correct side of the class boundary with no HC–DEP separation, and 59 of 60 subjects exceed chance accuracy, with a balanced confusion matrix and no class collapse. These views indicate that the gains reflect valence-relevant, physiology-consistent structure rather than majority-class bias or subject memorization. Extended analyses, including per-band attention weights, channel-wise attribution, and the out-of-fold confusion matrix, are provided in Appendix A.

4.8 Efficiency Analysis

The full model contains 2.28 M parameters and trains in ≈102 min per seed on a single RTX 4090 (peak VRAM 6.2 GB), ≈25 GPU-hours for the full protocol; at inference it processes ≈6.3 k 1-s clips per second on a single GPU. The six ablation arms each train in 6–9 min, keeping the full ablation matrix below one GPU-hour.

5 Discussion

These results carry a design message: the subject variability that invariant-learning methods discard and the scale heterogeneity that single-kernel architectures ignore are two views of the same structure, and an architecture that models scales explicitly and couples them hierarchically recovers both without sacrificing either subgroup. Practically, the per-scale SE weights provide a confidence signal that can be surfaced alongside the prediction.

Several limitations bound these conclusions. First, both datasets were acquired on Chinese university students; cross-site and cross-cultural transfer remain open, and the DEP subgroup (n=20) limits the precision of subgroup estimates. Second, the depression-transfer evidence is preliminary pending pure linear-probe validation. Future work will extend evaluation to leave-one-site-out protocols and study longitudinal clinical deployment in real screening workflows.

6 Conclusion

We presented Multi-HiFormer, a hierarchical multi-scale framework for EEG affective analysis that treats scale heterogeneity as a first-class architectural principle across the spatial, temporal, and spectral axes. The framework delivers state-of-the-art cross-subject emotion recognition on BME2026 and SEED with uniform gains for clinically critical subgroups, and preliminary evidence that the same backbone transfers to depression detection without architectural modification. By grounding physiological knowledge as trainable initialization rather than fixed rules, Multi-HiFormer offers a general and interpretable recipe for EEG-based affective computing.

AI use statement

In this work, we used generative AI tools for language polishing and writing clarity. We have not used generative AI tools for research ideation, experimental design, or data analysis, and these tasks are not applicable to this work. We have reviewed all AI-assisted text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work uses only de-identified EEG datasets released for research purposes (BME2026 Challenge and SEED). No new human-subject data was collected. Depression screening results are intended as research evidence, not clinical diagnosis; any real-world deployment requires independent clinical validation and qualified oversight.

Reproducibility statement

Datasets and protocols are described in Section 4.1; full training configurations, model size, and runtime are reported in Section 4.1 and Section 4.8. Source code and experiment scripts will be released upon acceptance.

References

A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Classification of covariance matrices using riemannian kernel and its application to bci. *Neurocomputing*, 112:172–178, 2013.

BME2026 Challenge Organizers. Bme2026 eeg emotion and depression detection challenge: Dataset description. Challenge documentation, 2026.

R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *Proceedings of the 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

M. A. Ellenbogen, A. E. Schwartzman, J. Stewart, and C.-D. Walker. Stress and selective attention: The interplay of mood, cortisol levels, and emotional information processing. *Psychophysiology*, 39(6):723–732, 2002.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Z. Guan, X. Zhang, W. Huang, K. Li, D. Chen, W. Li, J. Sun, L. Chen, Y. Mao, H. Sun, et al. A method for detecting depression in adolescence based on an affective brain-computer interface and resting-state electroencephalogram signals. *Neuroscience Bulletin*, 41(3):434–448, 2025. doi: 10.1007/s12264-024-01319-7.

H. He and D. Wu. Transfer learning for brain-computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020.

W. Huang, W. Wu, M. V. Lucas, H. Huang, Z. Wen, and Y. Li. Neurofeedback training with an electroencephalogram-based brain-computer interface enhances emotion regulation. *IEEE Transactions on Affective Computing*, 14(2):998–1011, 2023.

R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

A. H. Kemp, K. Griffiths, K. L. Felmingham, S. A. Shankman, W. Drinkenburg, M. Arns, C. R. Clark, and R. A. Bryant. Disorder specificity: Frontal alpha asymmetry in major depressive disorder and post-traumatic stress disorder. *Biological Psychology*, 85(2):350–354, 2010.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125. IEEE, 2017.

H. Liu, S. Yang, Y. Zhang, M. Wang, F. Gong, C. Xie, G. Liu, and D. Zhang. Libeer: A comprehensive benchmark and algorithm library for eeg-based emotion recognition. *arXiv preprint arXiv:2410.09767*, 2024.

S. Liu, X. Wang, L. Zhao, B. Li, W. Hu, J. Yu, and Y. Zhang. 3dcann: A spatiotemporal convolutional attention neural network for eeg emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5321–5331, 2022.

S. Mahato and S. Paul. Classification of depression patients and normal subjects based on eeg signal using alpha power and theta asymmetry. *Journal of Medical Systems*, 44:28, 2019.

S. R. Pendse, D. Nkemelu, N. J. Bidwell, S. Jadhav, S. Pathare, M. De Choudhury, and N. Kumar. From treatment to healing: Envisioning a decolonial digital mental health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. Article 548, 1–23. ACM, 2022.

X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14 (3):2496–2511, 2022.

T. Song, W. Zheng, P. Song, and Z. Cui. Eeg emotion recognition using dynamic graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2020.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

W.-L. Zheng and B.-L. Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

A Extended Representation Analysis

This appendix complements Section 4.7 with the underlying visual evidence and a reading of each panel.

Per-band attention (Figure 4). Two observations matter. First, DEP attention exceeds HC in *all five* canonical bands, and the gap is largest in the delta (+0.0037) and alpha (+0.0035) bands rather than in beta/gamma (+0.0018/+0.0032); the theta gap is the smallest but still positive (+0.0006). Depression-related spectral signatures in this cohort are therefore *broadband*, not a high-frequency phenomenon—a pattern that any model hard-wired to emphasize beta/gamma would miss. Second, the uniformly positive differences confirm a global group-level effect rather than a narrow-band artifact, and the absence of a single dominant band is exactly the distribution the multi-scale principle predicts: information is spread across the spectrum, so all five bands contribute non-redundantly to the decision.

Top-10 channels (Figure 5). The ranking pairs posterior sites (O2, O1) with midline frontal electrodes (Fz, F7), followed by Cz, T4, Fp2, C4, Pz, and F8. The dual anatomy is interpretable: the posterior contribution is consistent with the occipito-temporal sources of early emotional ERP components such as the P200, while the presence of frontal-mask electrodes (Fz, F7) among the top five supports the frontal alpha asymmetry prior with which the model was initialized. DEP attention exceeds HC attention on every one of the ten channels (gaps from +0.0077 down to +0.0006), so the group difference is not driven by a single electrode but by a coordinated re-weighting of the montage.

All-channel group difference (Figure 6). The DEP − HC difference is positive for all twenty channels and decays smoothly over more than an order of magnitude (from +0.0077 at O2 to +0.0006 at Fp2). A spatial signature this distributed cannot be captured by a single fixed kernel or a single graph adjacency: it is precisely the inter-subject spatial heterogeneity that the multi-scale spatial encoder is designed to absorb. The smooth decay also indicates that no electrode acts as a standalone biomarker—diagnosis-relevant information is carried by the montage as a whole.

Decision behavior (Figure 7). The confusion matrix is nearly symmetric: false positives (778, 7.7%) and false negatives (788, 7.8%) differ by only ten windows out of 10,080, so the model trades neither false alarms nor missed cases against the other—the operating profile required when both error types carry clinical cost. Overall accuracy computed from the matrix (0.8437) agrees with the tabulated result, and per-class precision and recall are balanced, confirming that the aggregate gains reflect genuine valence discrimination rather than majority-class bias.

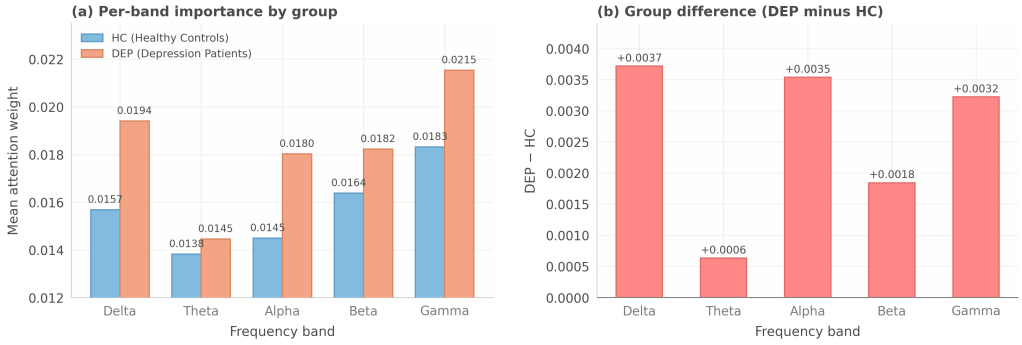

Figure 4: Figure A1. Per-band attention analysis. (a) Mean attention weight of each canonical band for HC and DEP; (b) group difference (DEP − HC), uniformly positive across all five bands, with the largest gaps in the delta and alpha bands.

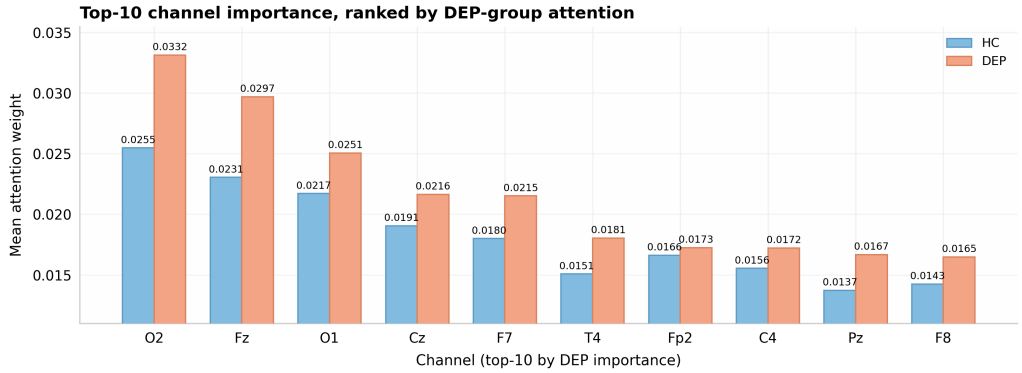

Figure 5: Figure A2. Top-10 channels ranked by DEP-group attention, with HC weights shown for comparison. Posterior sites (O2, O1) and frontal electrodes (Fz, F7) dominate the ranking.

B Full Supplementary Analyses of the Ablations

Section 4.5 summarizes three complementary ablation analyses in figure form; this appendix provides the complete numerical record and a per-analysis reading.

Gradient flow (Table 4). All four modules receive non-zero gradients from epoch 20 onward, so none degenerates into a decorative layer—consistent with every ablation costing accuracy. DualPathFusion carries the largest backward signal at every checkpoint: its epoch-20 norm (9881.2) is roughly 3.5× the second-largest module (MultiScaleCNN, 2815.3) and accounts for about 45% of the total gradient mass. All modules follow the same rise–fall–rise pattern with aligned peaks and valleys, indicating temporally coordinated rather than antagonistic training dynamics. CrossBandMix receives the smallest magnitude (256.7 at epoch 20) yet remains stably active; we read gradient

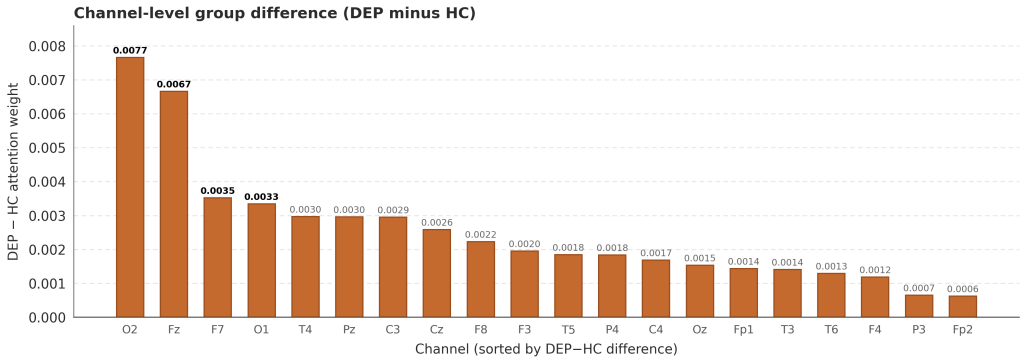

Figure 6: Figure A3. DEP − HC attention difference across all twenty channels, sorted by magnitude. The difference is positive for every channel, indicating a spatially distributed group signature.

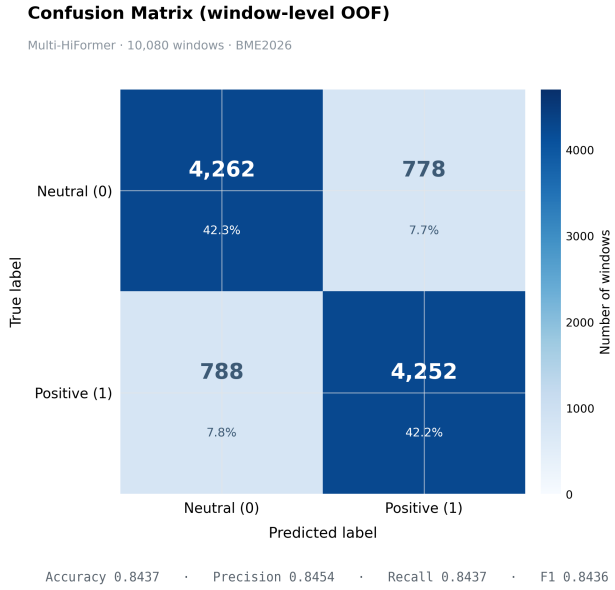

Figure 7: Figure A4. Window-level out-of-fold confusion matrix (10,080 windows) with overall metrics.

norm as a qualitative activity indicator rather than a normalized importance weight, and the small value is consistent with CrossBandMix's parameter-efficient mixer design.

Table 4: Table A1. Five-fold mean gradient L2 norms per module (BME2026).

Module	ep = 1	ep = 20	ep = 40	ep = 60	ep = 80	Status
CrossBandMix	0.0	256.7	726.9	53.5	93.3	active
MultiScaleCNN	0.0	2815.3	1369.8	244.4	753.0	active
HTTrans	0.0	2547.9	909.9	353.9	676.2	active
DualPathFusion	0.0	9881.2	5712.9	1124.0	2635.9	active

Module synergy (Table 5). All six pairs are strictly positively synergistic (+0.42 to +3.96 pp), and in every case the double-ablated model remains below the full model (all $\Delta_{A+B} > 0$): the modules form a synergistic system rather than interchangeable substructures. The strongest interaction is MultiScaleCNN $\times$ DualPathFusion (+3.96 pp): joint removal costs only 0.62 pp against the 4.58 pp additive prediction, because the multi-scale main path and the raw-statistics bypass cover each other's blind spots. Near-zero single-removal effects (e.g., CrossBandMix, $\Delta \approx 0$ in two pairs) indicate mutual backup rather than redundancy: a module that is dispensable when its partner is present becomes contributory when the partner is removed. The largest ratio (+500% for HT $\times$ DP) reflects a near-zero additive denominator rather than a large absolute effect, and we therefore report absolute synergy in pp as the primary quantity. Each pair was evaluated in a self-contained run; single-removal deltas carry $\pm$1–2 pp batch-level variance, which does not affect the within-run synergy estimates.

Table 5: Table A2. Pairwise synergy among the four core modules (BME2026).

Pair	Δ_A (pp)	Δ_B (pp)	Δ_{A+B} (pp)	Expected (pp)	Synergy (pp)	Ratio
CB $\times$ MS	-1.87	-0.21	+0.21	-2.08	+2.29	+110%
CB $\times$ HT	-0.21	+0.42	+0.83	+0.21	+0.62	+300%
CB $\times$ DP	+0.00	+0.83	+1.25	+0.83	+0.42	+50%
MS $\times$ HT	+0.21	-0.83	+1.25	-0.63	+1.88	+300%
MS $\times$ DP	-1.87	-2.71	-0.62	-4.58	+3.96	+86%
HT $\times$ DP	+0.00	+0.21	+1.25	+0.21	+1.04	+500%

Robustness (Table 6). Reading the table by perturbation family: under additive noise the full model is nearly invariant (between -0.35 and +0.85 pp across 15/10/5 dB), because DualPathFusion reads input statistics ahead of the contaminated pathway; removing it degrades sharply (-2.30 pp at 10 dB). Under time-window masking the full model is essentially flat (-0.41/-0.04/-0.12 pp), consistent with attention-pooled aggregation over sparsely informative windows. Channel masking is the hardest family: at 30% dropout the full model loses 6.99 pp in relative terms—more than most ablated variants—yet in absolute terms (0.760) it remains above five of the six variants, a clean accuracy–robustness trade-off driven by its higher clean baseline (0.8298, +6.5 pp over the best channel-robust variant). Under the combined perturbation the full model retains the highest absolute F1 (0.783) of all seven variants, with w/o MSConv (-4.64 pp relative) the only match: multi-scale convolution pays off precisely in compound, real-world-like disturbances.

Table 6: Table A3. F1 change (pp) under nine perturbations relative to clean (fold 1, BME2026). Clean F1: Full 0.8298, w/o HTTrans 0.7827, w/o AttnPool 0.7650, w/o DualPath 0.6021, w/o MSConv 0.7584, w/o BiGRU2 0.7598, w/o CrossBand 0.7836.

Variant	n15	n10	n5	t10	t20	t30	c15	c30	combo
Full	-0.35	+0.39	+0.85	-0.41	-0.04	-0.12	-0.18	-6.99	-4.63
w/o HTTrans	-0.82	-1.86	-0.41	-1.55	-0.03	+1.25	+2.98	-4.57	-3.94
w/o AttnPool	0.00	0.00	0.00	+0.42	+1.13	+0.89	+1.17	+1.12	-3.30
w/o DualPath	-0.33	-2.30	-1.99	-0.05	+0.31	-3.13	+0.88	+2.52	+0.23
w/o MSConv	0.00	-0.77	-0.03	-0.39	-0.39	+0.65	+2.07	-2.66	-4.64
w/o BiGRU2	+0.71	+0.71	+0.72	+0.76	+0.29	+0.29	+0.57	-4.25	-1.77
w/o CrossBand	-0.36	-0.36	+0.39	0.00	-1.12	+0.35	+1.40	-4.51	-2.67

SEED ablation (Table 7, Figure 8). Repeating the single-module ablation on SEED tests whether the component contributions are dataset-specific or architectural. The ordering fully transfers: dual-path fusion remains the largest contributor (-7.9 ACC / -8.1 F1 pp), and the three multi-scale modules again degrade performance by comparable margins (-5.7 to -6.3 pp), with attention pooling least costly (-4.7 pp)—mirroring the BME2026 pattern. Two SEED-specific observations sharpen the story. First, removing dual-path fusion produces a much larger variance on SEED ($\pm$6.8 ACC pp versus $\pm$2–3 pp for all other arms): without the raw-statistics bypass, the model becomes disproportionately sensitive to SEED's large inter-session variability—the very failure mode the bypass is

designed to prevent. Second, the full model's fold-to-fold standard deviation ($\pm$1.96 ACC pp) is the smallest of all seven variants, indicating that multi-scale modeling does not merely raise the mean but also stabilizes training across acquisition sessions.

Table 7: Table A4. Single-module ablation on SEED (3-fold cross-subject, mean $\pm$ std over folds). Variant names follow the SEED runs (NoTransformer = HTT removed; NoMultiScale = multi-scale CNN removed).

Variant	ACC (%)	ΔACC	F1 (%)	ΔF1
Full (ours)	84.65 $\pm$ 1.96	–	84.44 $\pm$ 2.02	–
NoDualFusion	76.79 $\pm$ 6.75	−7.86	76.37 $\pm$ 7.12	−8.07
NoTransformer	78.37 $\pm$ 2.58	−6.28	78.29 $\pm$ 2.61	−6.15
NoMultiScale	78.67 $\pm$ 2.45	−5.98	78.63 $\pm$ 2.50	−5.81
NoCrossBand	78.91 $\pm$ 3.29	−5.74	78.85 $\pm$ 3.31	−5.59
NoAttnPool	80.00 $\pm$ 2.73	−4.65	79.90 $\pm$ 2.64	−4.54

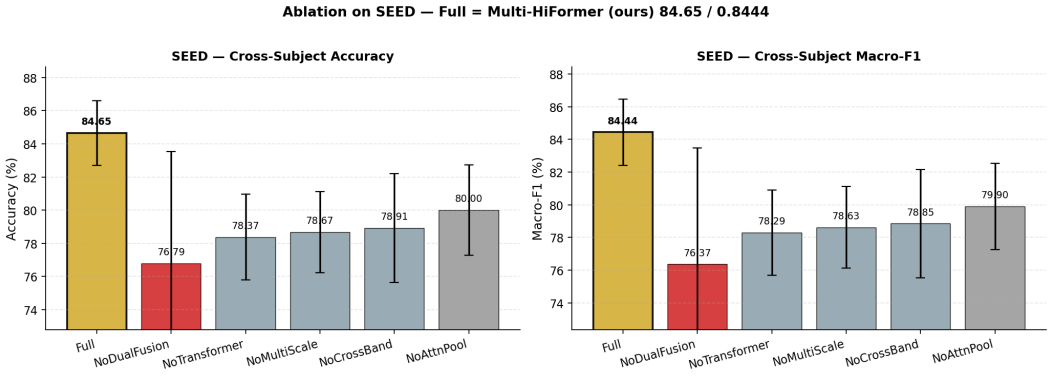

Figure 8: Figure A5. Single-module ablation on SEED. Left: cross-subject accuracy; right: macro-F1. Error bars show the standard deviation across folds; the full model is highlighted in gold.

Evidence synthesis. Read together, the appendix supports three generalization claims. *Cross-subject*: every evaluation is cross-subject by protocol (StratifiedGroupKFold on BME2026, the standard 3-fold split on SEED), and the per-subject analyses in Section 4.7 and Figure 7 show that no subject or subgroup is left behind. *Cross-dataset*: the component ablations replicate on two independent acquisition systems (Table 5 and Table 7), so the architecture's value is not an artifact of one benchmark. *Cross-task*: the emotion-trained backbone transfers to depression screening (Section 4.6), and the mechanisms that make it work—dual-path fusion and multi-scale convolution—are precisely the components whose removal destabilizes performance under distribution shift (Table 6). The model is therefore best understood not as a benchmark-specific assembly but as a representation learning principle whose components are individually necessary, mutually reinforcing, and portable across subjects, datasets, and tasks.

C Receptive Field Coverage of the Multi-Scale Kernels

A natural question about the spatial scale pyramid is why the five scales were chosen as $k \in \{3, 5, 7, 9, 11\}$ rather than any other set. We answer this in two steps: first, we quantify the physical cortical distance covered by each kernel on the BME2026 30-channel layout; second, we verify that the same decomposition remains meaningful on an independent, higher-density montage (SEED, 62 channels).

Coverage on the BME2026 30-channel layout (Table 8, Figure 9). Measuring the maximum pairwise distance between the electrodes spanned by each kernel (10–20 simplified layout), the five

kernels cover 4.8, 5.8, 6.5, 7.1, and 7.7 cm respectively—roughly 47%, 56%, 62%, 69%, and 75% of the head-topology radius. Two properties justify the choice. First, the coverage is *anatomically graded*: k=3 spans a single gyrus or adjacent gyri, k=5 a single region (frontal or central), k=7 crosses regions (fronto-parietal or fronto-temporal), k=9 crosses lobes, and k=11 spans the full 30-channel montage. Second, the five coverages are *non-redundant*: consecutive kernels are separated by 0.5–1.0 cm of clear spacing, so no kernel's field is contained in another's.

Table 8: Receptive field coverage of the five kernels on the BME2026 30-channel layout (mean $\pm$ std over electrodes).

Kernel	Mean coverage (cm)	% of head radius	Anatomical interpretation
k=3	4.81 $\pm$ 0.50	47%	single gyrus or adjacent gyri
k=5	5.76 $\pm$ 0.41	56%	within a single region (frontal / central)
k=7	6.45 $\pm$ 0.41	62%	across regions (fronto-parietal / fronto-temporal)
k=9	7.10 $\pm$ 0.50	69%	across lobes
k=11	7.74 $\pm$ 0.54	75%	full 30-channel montage

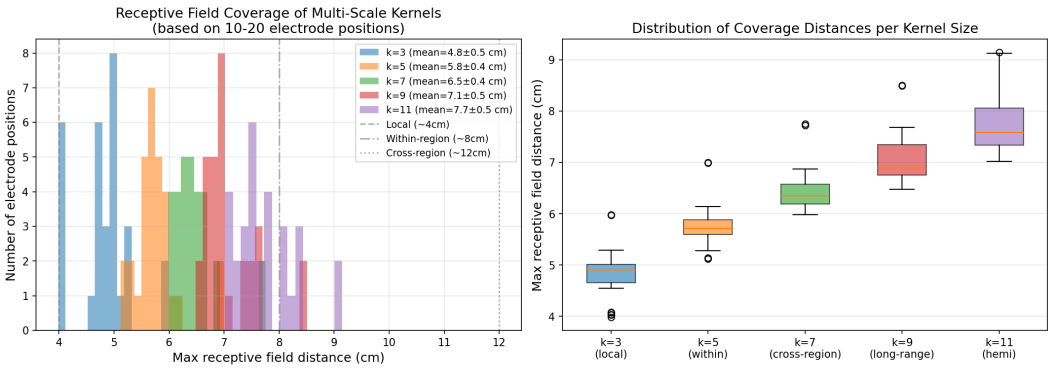

Figure 9: Receptive field coverage on the BME2026 30-channel layout. Left: distribution of the maximum electrode-pair distance covered by each kernel, with the local ($\sim$4 cm), within-region ($\sim$8 cm), and cross-region ($\sim$12 cm) reference bands. Right: boxplot of the per-kernel coverage distances.

Generalization to a high-density montage (Table 9, Figure 10). To assess whether the five-scale decomposition generalizes beyond the BME 30-channel layout, we replicate the analysis on the SEED 62-channel 10–10 montage (Zheng & Lu, 2015), whose electrodes span the full scalp. The 62-channel coordinates used here are analytical spherical projections of standard 10–10 positions ($R = 9$ cm sphere); they are not the measured electrode coordinates and serve only to estimate the order of magnitude of kernel coverage. On this layout, the five kernels cover 2.0, 2.4, 2.8, 3.2, and 3.5 cm—smaller in absolute terms because high-density sampling places neighbouring electrodes closer (a fundamental property of high-density EEG), but with the same monotonic ordering and clear inter-kernel gaps (0.4–0.6 cm). Critically, k=11 reaches about 40% of the head-topology radius on the 62-channel layout, crossing the inter-hemispheric midline—a scale that the 30-channel BME layout cannot represent. The five-scale architecture is therefore not an artifact of the BME channel set.

Summary: a three-layer evidence chain. The choice of the five scales is supported at three levels: (i) on the BME 30-channel layout, the five kernels cover distinct anatomical scales from a single gyrus to the full montage (Table 8); (ii) on the independent SEED 62-channel layout, the same architecture preserves five clearly separated scales and additionally reaches inter-hemispheric distances (Table 9); and (iii) in the main experiments, the five-kernel model significantly outperforms the single-kernel baseline (Tables 1 and 3), confirming that the scale diversity is functionally exploited rather than merely decorative.

Table 9: Receptive field coverage of the five kernels on the SEED 62-channel layout (analytical 10–10 spherical projection; order-of-magnitude estimate).

Kernel	Mean coverage (cm)	% of head radius	Anatomical interpretation
k=3	1.96 ± 0.82	23%	local cortex
k=5	2.42 ± 0.94	28%	within a single region
k=7	2.81 ± 1.06	32%	across regions
k=9	3.15 ± 1.15	36%	across lobes
k=11	3.47 ± 1.23	40%	inter-hemispheric (to the midline)

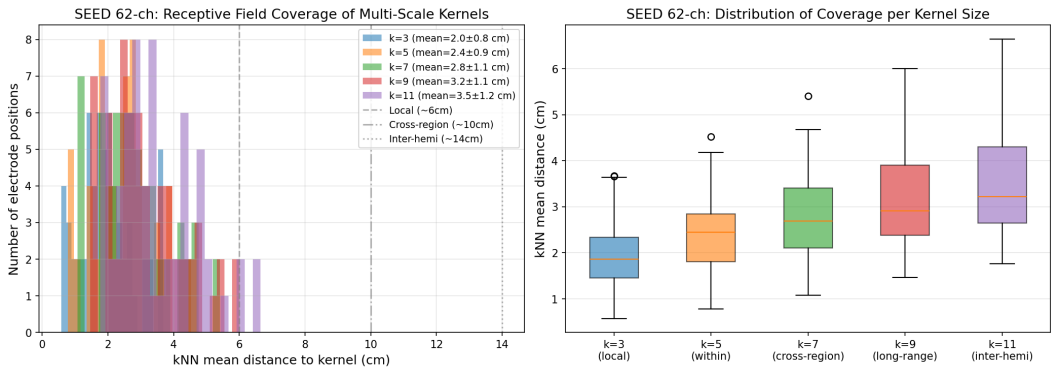

Figure 10: Receptive field coverage on the SEED 62-channel layout (analytical 10–10 projection). Left: distribution of the kNN mean distance covered by each kernel, with local (∼6 cm), cross-region (∼10 cm), and inter-hemispheric (∼14 cm) reference bands. Right: boxplot of the per-kernel coverage distances.