STFSFORMER: INTEGRATING NEURONAL DYNAMICS AND FULL-SPIKE ATTENTION FOR ENERGY-EFFICIENT EEG MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient electroencephalography (EEG) modeling requires representations that retain rapid signal variations while capturing temporal dependencies under limited energy budgets. Spiking neural networks provide an energy-efficient computing substrate, but conventional leaky integrate-and-fire dynamics can attenuate rapid changes through temporal smoothing, while global attention incurs quadratic computational cost in sequence length. We introduce STFSFormer, a spiking Transformer that integrates transient-sensitive neuronal dynamics with local full-spike attention for energy-efficient EEG modeling. A high-frequency-enhanced leaky integrate-and-fire neuron combines a differential synaptic pathway with refractory feedback to emphasize input changes and regulate repeated firing. Parallel encoding branches with distinct decay profiles capture complementary temporal responses. Spatial-temporal full-spike attention integrates these representations using binary queries, keys, and values within local temporal blocks. Its softmax-free formulation enables a dimension-aware adaptive multiplication strategy that selects the lower-cost of two equivalent orders from the block length and head dimension, preserving the output while retaining linear scaling in sequence length for fixed block and feature dimensions. Experiments on SEED, SEED-V, and DREAMER demonstrate competitive emotion recognition performance. Under the evaluated input configuration, STFSFormer uses 1.80 million parameters and 0.46 billion multiply–accumulate operations per segment. A consistent synthesis-based FPGA evaluation estimates whole-model inference power at 1.2 W and energy consumption at 1.5 mJ per segment, corresponding to 50.0% and 44.4% lower energy than the implemented SpikingGCN and SCNN baselines, respectively. These results demonstrate a favorable balance between recognition performance, computational cost, and energy-efficient inference.

Keywords: Electroencephalography; Emotion recognition; Spiking neural networks; Transformer; Energy-efficient inference

1 INTRODUCTION

Automatic emotion recognition aims to infer affective states from observable signals and provides a computational basis for affective sensing and human–computer interaction (Li et al., 2022a). Electroencephalography (EEG) is particularly relevant to this task because it provides a non-invasive view of neural activity with millisecond temporal resolution. This temporal precision makes EEG useful for studying the evolving neural responses associated with emotion. Reliable recognition nevertheless remains challenging because affective activity is complex (Li et al., 2022b).

Wearable EEG systems extend the motivation for emotion recognition to repeated and potentially continuous sensing in daily life (Tian et al., 2023; 2025). In such settings, a model must process successive signal segments under limited computational and energy budgets. Recognition performance therefore needs to be considered together with the cost of temporal processing. Reducing model

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.

size can help, but efficient inference also depends on how the model represents signal changes and computes interactions over time.

Deep EEG models use convolutional, graph, and attention-based representations to learn discriminative features and dependencies (Chen et al., 2024; Dong et al., 2024; Liu et al., 2023; Gao et al., 2024; Fan et al., 2024). These approaches provide a range of mechanisms for integrating temporal and multichannel information. Lightweight architectures and frequency-aware designs further motivate efficient representations adapted to EEG (Su et al., 2022; Yao et al., 2025; Jin et al., 2025). For continuous sensing, an additional design question is how to retain informative temporal variations while limiting the computation devoted to sequence interactions.

This design question is closely connected to the structure of EEG signals. Emotion-related information spans oscillatory bands, short-lived fluctuations, and coordinated activity across channels (Zheng & Lu, 2015; Li et al., 2022b; Gao et al., 2024). Temporal integration provides context for these patterns, but excessive smoothing may attenuate rapid changes. At the same time, global attention computes interactions across the entire token sequence, increasing the cost of processing longer segments. These considerations motivate a joint design of the temporal response and the interaction operator: the representation should remain sensitive to localized changes, while the interaction mechanism should capture their dependencies at a manageable cost.

Spiking neural networks (SNNs) represent activity with discrete events and offer a substrate for energy-efficient execution on suitable hardware (Maass, 1997; Roy et al., 2019; Davies et al., 2018). Learnable spike encoders connect continuous inputs to spike-based computation (Gong et al., 2023), and spiking Transformers extend this representation to attention (Zhou et al., 2023). Applying these ideas to EEG requires addressing two coupled choices. First, the neuron dynamics determine which signal variations elicit spikes. Conventional leaky integrate-and-fire (LIF) neurons integrate inputs with exponential decay, providing temporal memory but potentially attenuating fast variations. Second, the attention operator determines which interactions among the encoded features are computed. Combining transient-sensitive spike generation with local temporal interactions provides a way to address both choices within a single model.

We propose STFSFormer, an end-to-end spiking Transformer for EEG emotion recognition that combines these two design choices. Its High-Frequency Enhanced Leaky Integrate-and-Fire (HELIF) neuron introduces a recurrent differential pathway to emphasize rapid changes and a refractory trace to regulate repeated firing. Parallel encoding branches with distinct decay profiles capture complementary temporal responses, which are fused into a compact token sequence. Spatial-Temporal Full-Spike Attention (STFSAtten) then models feature dependencies within local temporal blocks using binary query, key, and value representations. STFSAtten further introduces a dimension-aware adaptive multiplication strategy: because no softmax or other nonlinearity separates the two matrix products, it selects token-first or feature-first evaluation according to the block length and head dimension, minimizing the core multiplication cost and intermediate size without changing the output. Our contributions are:

- We introduce the HELIF neuron, which combines differential-driven synaptic dynamics with refractory feedback, and construct a parallel multi-frequency spike encoder with complementary temporal response profiles.
- We develop STFSAtten, a local spike attention module with a dimension-aware adaptive multiplication strategy that selects between equivalent token-first and feature-first orders to reduce core computation and intermediate storage.
- We evaluate the integrated STFSFormer on SEED, SEED-V, and DREAMER under subject-dependent protocols. Recognition comparisons, component ablations, computational analysis, and FPGA-based energy estimates characterize its performance and inference cost. A leave-one-subject-out evaluation on SEED and DREAMER reports performance on unseen subjects.

2 Related Work

EEG-Based Emotion Recognition. Deep EEG emotion-recognition models commonly combine temporal feature extraction with interchannel dependency modeling. Convolutional and recurrent architectures integrate local patterns across successive feature frames, whereas graph-based methods

encode relationships among recording channels through learned connectivity structures (Li et al., 2022a; Gao et al., 2024). Lightweight and frequency-aware designs further seek to reduce processing cost while retaining task-relevant temporal–spectral information (Su et al., 2022; Jin et al., 2025). Together, these studies show that effective EEG modeling depends on both temporal response and cross-feature interaction. For continuous sensing, these functions must also operate under tight computation and energy budgets, motivating spike-based representations and restricted temporal interactions.

Spiking Transformers. Spiking Transformers combine attention-based feature interaction with the discrete representations of SNNs. A representative mechanism is spiking self-attention, which uses spike-form queries, keys, and values without softmax normalization (Zhou et al., 2023). Subsequent work has explored more scalable architectures and training strategies, including spike firing approximation with integer-valued training and spike-driven inference (Yao et al., 2025). These developments establish a basis for combining expressive feature interactions with efficient spike-based processing. Adapting this family of models to EEG also requires specifying the temporal response of the spike generator and the extent of interactions along the signal sequence. STFSFormer combines differential-driven neurons with attention restricted to local temporal blocks. Within each block, softmax-free associativity permits either a token-first or feature-first multiplication order, and a dimension-aware adaptive rule selects the lower-cost order and smaller intermediate representation. This design connects the temporal sensitivity of the encoder to the locality of the attention operator, targeting the joint requirements of EEG representation and inference efficiency.

3 METHOD

Given an EEG segment $\mathbf{X} \in \mathbb{R}^{C\times T}$ with C recording channels and T time samples, a learnable multi-scale temporal encoder extracts features and downsamples the temporal axis, producing $\mathbf{H} \in \mathbb{R}^{T'\times D}$. Here, T' is the sequence length after downsampling, and D is the feature dimension. Parallel HELIF branches map the shared representation $\mathbf{H}$ to binary spike streams with complementary temporal responses. The streams are concatenated along the feature dimension and projected back to D dimensions; batch normalization then yields real-valued fused tokens $\mathbf{Z} \in \mathbb{R}^{T'\times D}$ without changing the downsampled sequence length. The token sequence is processed by stacked blocks, each combining STFSAtten with a spike-based MLP, and a classification head produces the emotion prediction (Figure 1). Within STFSAtten, $\mathbf{Z}$ is partitioned into non-overlapping temporal blocks of L consecutive tokens.

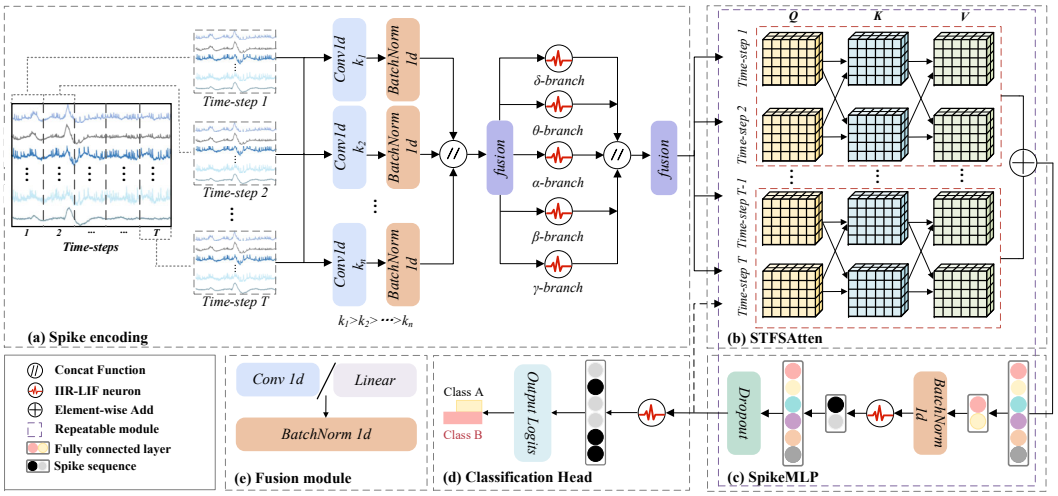

Figure 1: The overall architecture of the proposed STFSFormer for emotion recognition.

3.1 EEG REPRESENTATION

We filter EEG with an offline zero-phase band-pass filter of 0.5–$50\,\mathrm{Hz}$ and apply channel-wise z-score normalization within each trial. Trials are assigned to data partitions before their fixed-length windows are used as samples. Each sample has C channels and T time points, with dataset-dependent dimensions. Training augmentations and preprocessing details appear in Appendix A.

3.2 HIGH-FREQUENCY ENHANCED SPIKING NEURON

The conventional LIF neuron integrates input with exponential decay (Roy et al., 2019). This recurrence provides temporal memory, but its smoothing response can attenuate fast input changes. For EEG encoding, the temporal response should retain recent context while remaining sensitive to localized variations. HELIF addresses these requirements through two coupled components. First, a differential-driven synaptic path updates the transient state as

$$f[t] = \gamma f[t-1] + \big(x[t] - x[t-1]\big), \tag{1}$$

where $x[t]$ is the input current, we set $x[-1] = 0$ at the start of each EEG window, and $\gamma \in (0, 1)$ controls the persistence of the differential response. Second, the membrane potential, spike output, and refractory trace are updated in causal order:

$$\begin{aligned} v[t] &= \eta v[t-1] + f[t] - V_{\mathrm{th}} r[t-1], \\ s[t] &= H_{\mathrm{sg}}(v[t] - V_{\mathrm{th}}), \\ r[t] &= \lambda r[t-1] + s[t]. \end{aligned} \tag{2}$$

Here $v[t]$ is the membrane potential, $s[t]$ the binary output, $r[t]$ the refractory trace, and V_{th} the firing threshold. The coefficients $\eta, \lambda \in (0, 1)$ control membrane and refractory decay. H_{sg} is a step function in the forward pass, with surrogate gradients for training (Tavanaei et al., 2019).

In Equation 1, the difference $x[t] - x[t-1]$ supplies a signed change signal: a constant input contributes no new differential drive, whereas changing inputs update $f[t]$. The recurrent term $\gamma f[t-1]$ retains a decaying history of these changes, and Equation 2 integrates this state before thresholding. Thus, transient emphasis is introduced at the synaptic input while recurrent memory remains available. Here, high-frequency enhancement refers to the response to changes in the input current sequence; the effective response also depends on the decay coefficients and spike threshold.

Refractory feedback regulates the resulting activity. Each emitted spike increases $r[t]$, strengthening the inhibitory term $-V_{\mathrm{th}} r[t-1]$ in subsequent membrane updates. In the absence of further spikes, the trace decays at a rate controlled by λ. This couples sensitivity to input changes with feedback from recent firing, rather than controlling both through membrane decay alone. HELIF adds two auxiliary states and uses elementary recurrent operations. States reset at each EEG window, so temporal memory remains local to the segment. Figure 2 summarizes the neuron; Appendix B gives the conventional LIF comparison and expanded equations.

3.3 MULTI-FREQUENCY SPIKE ENCODING

A learnable multi-scale temporal encoder maps the input into $\mathbf{H} = \Phi_{\mathrm{ms}}(\mathbf{X}) \in \mathbb{R}^{T' \times D}$, with temporal downsampling to reduce subsequent sequence length. Unlike fixed rate or delay coding, the representation is trained jointly with the classifier. We apply K parallel HELIF branches with distinct decay profiles to the same feature stream:

$$\mathbf{S}^{(k)} = \Psi_k(\mathbf{H}) \in \{0, 1\}^{T' \times D}, \quad \mathbf{Z} = \mathrm{BN}\Big(\mathrm{Concat}_{k=1}^{K}(\mathbf{S}^{(k)})\mathbf{W}_{\mathrm{fuse}}\Big). \tag{3}$$

The projection $\mathbf{W}_{\mathrm{fuse}} \in \mathbb{R}^{KD \times D}$ merges complementary temporal responses. Each branch outputs binary spikes; the projected, normalized tokens $\mathbf{Z} \in \mathbb{R}^{T' \times D}$ are real-valued inputs to subsequent spiking projections. The branches provide different temporal response profiles. We use $K = 5$ in the reported configuration, with the branch-count analysis in Appendix D.1.

3.4 SPATIAL-TEMPORAL FULL-SPIKE ATTENTION

STFSAtten operates on the fused tokens $\mathbf{Z} \in \mathbb{R}^{T' \times D}$. HELIF projections produce $\mathbf{Q} = \phi(\mathbf{Z}\mathbf{W}_Q)$, $\mathbf{K} = \phi(\mathbf{Z}\mathbf{W}_K)$, and $\mathbf{V} = \phi(\mathbf{Z}\mathbf{W}_V)$, where ϕ denotes the HELIF dynamics applied along the

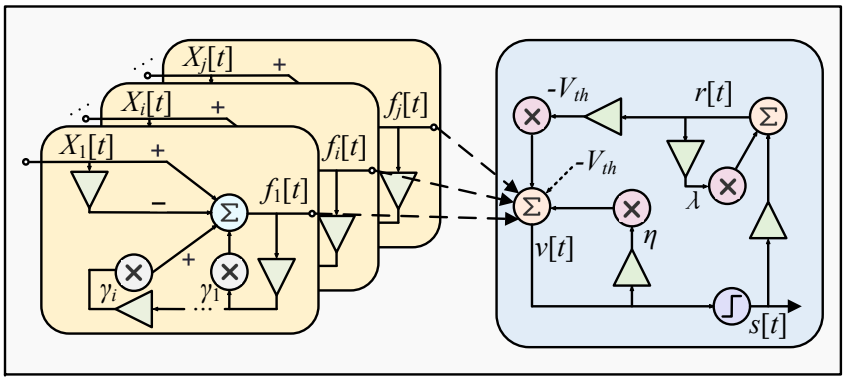

Figure 2: Overview of the proposed HELIF neuron. By introducing a differential-driven synaptic path and a refractory trace, the neuron enhances sensitivity to rapid EEG variations while stabilizing spiking activity for efficient event-driven processing.

downsampled sequence. Before block extraction, the binary tensors have shape $N_h \times T' \times d_h$, with $d_h = D/N_h$. We partition each sequence into non-overlapping blocks of L consecutive tokens. The final block is zero-padded when necessary, and padded outputs are discarded after attention. The parameter L therefore sets the temporal neighborhood for attention on the downsampled sequence. Thus $\mathbf{Q}^{(i)}, \mathbf{K}^{(i)}, \mathbf{V}^{(i)} \in \{0,1\}^{N_h \times L \times d_h}$.

Since no softmax or other nonlinear operation is applied between the two matrix multiplications, associativity gives $(\mathbf{Q}^{(i)}(\mathbf{K}^{(i)})^\top)\mathbf{V}^{(i)} = \mathbf{Q}^{(i)}((\mathbf{K}^{(i)})^\top\mathbf{V}^{(i)})$. We therefore introduce a dimension-aware adaptive multiplication strategy that selects the evaluation order according to L and d_h:

$$\mathbf{O}^{(i)} = \alpha \begin{cases} \left(\mathbf{Q}^{(i)}(\mathbf{K}^{(i)})^\top\right)\mathbf{V}^{(i)}, & L \le d_h, \\ \mathbf{Q}^{(i)}\left((\mathbf{K}^{(i)})^\top\mathbf{V}^{(i)}\right), & L > d_h, \end{cases} \qquad \alpha = d_h^{-1/2}. \tag{4}$$

When $L \le d_h$, the first product forms an $L \times L$ local token-interaction matrix; otherwise, the first product forms a $d_h \times d_h$ feature-interaction matrix. The selection is determined by the tensor dimensions rather than the sample values. Both branches produce the same output before finite-precision effects, while the adaptive strategy chooses the lower-cost order and smaller intermediate. The output mixes information across positions within each block and across learned feature dimensions. These feature dimensions encode multichannel EEG information rather than corresponding directly to individual electrodes. The scaling factor α is applied after either multiplication order.

The accumulated outputs are converted back to binary spikes by HELIF after concatenation along the temporal dimension. We then merge heads and use a projected residual update:

$$\mathbf{Y} = \mathbf{Z} + \mathrm{BN}\left[\mathrm{Merge}\Big(\phi(\mathrm{Concat}_i \mathbf{O}^{(i)})\Big)\,\mathbf{W}_O\right]. \tag{5}$$

Here, "full-spike" refers to the binary query, key, value, and post-attention readout activations. The accumulated products, scaling, output projection, batch normalization, and residual stream can remain real-valued. The spike-based MLP then transforms the representation before the next block. Appendix C gives the tensor indexing and intermediate shapes.

Computational cost. Let $M = \lceil T'/L \rceil$ be the number of blocks after padding. Excluding the shared query, key, and value projections, the two core matrix products in Equation 4 require

$$C_{\mathrm{STFSAtten}} = 2N_h M L d_h \min(L, d_h) \tag{6}$$

MACs, with an intermediate size of $\mathcal{O}(N_h M \min(L^2, d_h^2))$ when all blocks are processed in parallel. Ignoring padding gives the approximation $\mathcal{O}(N_h T' d_h \min(L, d_h))$, which is linear in T' for fixed L and d_h. Global attention requires $2N_h T'^2 d_h$ core MACs. Under the reported setting $T' = 100$, $L = 3$, and $d_h = 64$, padding gives $M = 34$, and the adaptive implementation selects the token-first branch with a 3×3 intermediate and 39,168 core MACs per head. This removes quadratic dependence on the global sequence length, but does not imply a uniform advantage over conventional block-local attention. Binary query, key, and value operands support spike-gated accumulation on suitable implementations. The model-level computational and hardware comparisons in Section 4.5 evaluate the complete architecture.

4 EXPERIMENT

4.1 DATASETS AND EVALUATION

We evaluate SEED (Zheng & Lu, 2015), SEED-V (Liu et al., 2022), and DREAMER (Katsigiannis & Ramzan, 2017). SEED and SEED-V use discrete emotion labels. For DREAMER, valence and arousal ratings are binarized independently using a threshold of 3.0: ratings ≥ 3.0 are labeled high (1), and ratings < 3.0 are labeled low (0). The main comparison uses subject-dependent evaluation. For each subject, complete trials are first assigned to training, validation, and test partitions in an 8:1:1 ratio. Each trial is segmented into fixed-length windows only after this assignment, and every window inherits the partition of its parent trial. Consequently, windows originating from the same trial cannot occur in different partitions. We also report performance on unseen subjects for SEED and DREAMER using strict leave-one-subject-out (LOSO) folds. In each fold, one subject is kept entirely unseen during training, validation, and model selection and is used only for testing. For SEED, all sessions from the held-out subject remain together in the test fold. Table 1 averages five random-seed runs, with each seed generating a new trial-level 8:1:1 split. In Tables 1 and 2, accuracy and macro-F1 are computed at the window level without trial-level aggregation; standard deviations are calculated across seeds and LOSO folds, respectively.

Experiments use PyTorch on one NVIDIA RTX 4090. We select checkpoints by validation performance. The reported defaults are $L = 3$, $d_h = 64$, decay span $\Delta_\gamma = 0.16$, shared decay $\gamma_{\text{shared}} = 0.89$, downsampling factor $s = 2$, refractory decay $\lambda = 0.90$, and $K = 5$ branches. Appendix D reports the associated sensitivity studies; Appendix A describes augmentation.

4.2 COMPARISON WITH EXISTING METHODS

Table 1: Subject-dependent performance (mean ± standard deviation, %). Adapted architecture baselines are reimplemented from the cited original models for EEG emotion recognition. Bold indicates the highest mean per task and metric.

Method	SEED	SEED-V	DREAMER-V	DREAMER-A
		Accuracy		
DCNN (Atwood & Towsley, 2016)	75.39 ± 3.21	66.19 ± 4.82	63.49 ± 5.34	70.24 ± 4.67
EEGNet (Lawhern et al., 2018)	84.36 ± 2.45	72.44 ± 3.87	62.55 ± 4.95	61.17 ± 5.03
GAT (Veličković et al., 2018)	92.48 ± 1.78	81.45 ± 3.12	79.89 ± 3.45	80.12 ± 3.38
MMP-GCN (Song et al., 2022)	82.13 ± 2.67	79.32 ± 3.25	77.83 ± 3.56	77.24 ± 3.61
Brain-GRL (Li et al., 2022c)	88.08 ± 2.12	84.32 ± 2.87	84.31 ± 2.95	85.27 ± 2.78
Spikformer (Zhou et al., 2023)	83.24 ± 2.54	80.43 ± 3.08	78.90 ± 3.42	78.31 ± 3.49
SpikingGCN (Zhu et al., 2022)	92.01 ± 1.82	84.46 ± 2.79	82.12 ± 3.12	83.38 ± 3.05
SCNN (Wei et al., 2025)	88.44 ± 2.18	81.54 ± 3.14	79.32 ± 3.48	80.68 ± 3.41
TCJA-SNN (Zhu et al., 2025)	86.89 ± 2.31	78.12 ± 3.42	76.34 ± 3.67	76.78 ± 3.73
FSTA-SNN (Yu et al., 2025)	89.12 ± 2.05	79.56 ± 3.29	77.12 ± 3.62	77.98 ± 3.56
STFSFormer (Ours)	**92.54 ± 1.68**	**87.63 ± 2.34**	**85.91 ± 2.56**	**85.47 ± 2.61**
		F1		
DCNN (Atwood & Towsley, 2016)	74.97 ± 3.45	63.94 ± 5.17	66.03 ± 5.08	65.80 ± 5.22
EEGNet (Lawhern et al., 2018)	84.04 ± 2.68	72.20 ± 4.12	63.23 ± 4.78	62.69 ± 4.91
GAT (Veličković et al., 2018)	81.86 ± 2.95	81.23 ± 3.34	79.45 ± 3.67	79.78 ± 3.52
MMP-GCN (Song et al., 2022)	81.86 ± 2.89	78.94 ± 3.48	77.46 ± 3.79	76.87 ± 3.84
Brain-GRL (Li et al., 2022c)	87.84 ± 2.34	80.41 ± 3.42	83.10 ± 3.18	**87.05 ± 2.41**
Spikformer (Zhou et al., 2023)	82.94 ± 2.76	80.09 ± 3.31	78.57 ± 3.65	77.95 ± 3.72
SpikingGCN (Zhu et al., 2022)	**91.99 ± 1.85**	84.54 ± 2.81	82.87 ± 2.96	83.82 ± 2.88
SCNN (Wei et al., 2025)	90.09 ± 1.96	81.00 ± 3.37	79.05 ± 3.71	80.61 ± 3.53
TCJA-SNN (Zhu et al., 2025)	85.78 ± 2.56	77.67 ± 3.65	76.12 ± 3.89	76.45 ± 3.95
FSTA-SNN (Yu et al., 2025)	88.67 ± 2.28	79.23 ± 3.52	77.68 ± 3.48	77.68 ± 3.63
STFSFormer (Ours)	90.18 ± 1.92	**87.25 ± 2.47**	**85.57 ± 2.69**	85.11 ± 2.74

Table 1 compares the adapted architecture baselines under the same preprocessing, trial-level partitioning, window generation, and evaluation protocol. For each cited model, we retain its defining

architectural operator and configure the task-specific input interface and classifier for the corresponding EEG dataset. STFSFormer has the highest mean accuracy on all four tasks. On SEED, its 92.54% accuracy is close to GAT's 92.48%, while its F1 of 90.18% is below SpikingGCN's 91.99%. On SEED-V, STFSFormer reaches 87.63% accuracy and 87.25% F1, exceeding the evaluated baselines on both metrics. DREAMER results show a similar distinction between metrics: STFSFormer has the highest accuracy for both valence and arousal, whereas Brain-GRL has the highest arousal F1 (87.05%, compared with 85.11%). The results support competitive recognition across tasks, with the clearest joint accuracy and F1 advantages on SEED-V and DREAMER-Valence.

4.3 Cross-Subject Evaluation

Table 2: Strict LOSO cross-subject results (mean ± standard deviation across held-out-subject folds, %). Each entry reports ACC / Macro-F1. Bold denotes the highest mean for each task and metric.

Method	SEED	DREAMER-Valence	DREAMER-Arousal
EEGNet (Lawhern et al., 2018)	70.81 ± 10.62 / 69.09 ± 11.81	$\mathbf{66.24 \pm 8.33}$ / 63.13 ± 8.41	70.22 ± 12.00 / $\mathbf{64.51 \pm 9.17}$
Spikformer (Zhou et al., 2023)	71.85 ± 10.94 / 70.63 ± 11.28	63.89 ± 8.00 / 61.84 ± 7.70	70.52 ± 10.95 / 63.09 ± 8.39
STFSFormer (Ours)	$\mathbf{72.44 \pm 11.75}$ / $\mathbf{71.63 \pm 11.77}$	66.18 ± 7.63 / $\mathbf{63.63 \pm 7.50}$	$\mathbf{71.21 \pm 9.48}$ / 63.21 ± 8.01

Table 2 compares the three models under the same strict LOSO protocol. STFSFormer obtains the highest mean ACC and Macro-F1 on SEED. On DREAMER-Valence, its ACC is 0.06 percentage points below EEGNet, while its Macro-F1 is the highest among the compared methods. On DREAMER-Arousal, STFSFormer achieves the highest ACC, whereas EEGNet has the highest Macro-F1. The fold-wise standard deviations indicate substantial variation across held-out subjects, particularly on SEED. This experiment uses source-only evaluation on unseen subjects; no target-subject data are used for adaptation, validation, or model selection.

4.4 Component Ablations

Table 3: Performance Analysis of Ablation Experiments [%].

Case	Spiking Neuron	Attention	Accuracy				F1-Score			
			SEED	SEED-V	Valence	Arousal	SEED	SEED-V	Valence	Arousal
(Exp.1)	HELIF	STFSAtten	**92.54**	**87.63**	**85.91**	**85.47**	**90.18**	**87.25**	**85.57**	**85.11**
(Exp.2)	LIF	STFSAtten	90.32	84.78	82.95	82.53	87.89	84.12	82.49	82.06
(Exp.3)	HELIF	SSA	87.45	81.23	79.42	79.00	84.56	80.34	79.01	78.58
(Exp.4)	LIF	SSA	83.24	80.43	78.90	78.31	82.94	80.09	78.57	77.95

Table 3 separates the neuron and attention contributions. Exp. 1 combines HELIF and STFSAtten; Exp. 2 substitutes LIF; Exp. 3 substitutes spatial-only spiking self-attention (SSA); and Exp. 4 substitutes both. Relative to Exp. 4, the complete model improves accuracy by 9.30, 7.20, 7.01, and 7.16 percentage points on SEED, SEED-V, DREAMER-Valence, and DREAMER-Arousal, respectively. With attention fixed, replacing HELIF with LIF reduces accuracy by roughly 2–3 points. With HELIF fixed, replacing STFSAtten with SSA produces larger reductions of approximately 5–7 points. Both components contribute, with the attention substitution producing the larger change in these comparisons. Detailed per-task F1 results remain in the table.

4.5 Model Complexity and FPGA Evaluation

Table 4: Model complexity comparison under a standard input setting.

Method	Params (M)	MACs (G)	FPS (4090)
EEGNet	0.10	0.02	2100
SpikingGCN	2.42	0.85	520
SCNN	3.17	1.10	480
STFSFormer	1.80	0.46	900

Table 5: FPGA resource utilization and estimated energy per EEG segment under a consistent synthesis-based evaluation protocol.

Method	LUTs (%)	BRAM (%)	DSP (%)	Power (W)	Energy (mJ)
EEGNet	58.9	22.4	48.0	2.3	4.6
SpikingGCN	62.7	31.6	18.5	1.8	3.0
SCNN	55.3	27.8	12.0	1.9	2.7
STFSFormer	45.3	24.0	8.0	1.2	1.5

Table 4 uses batch size 1, $C = 62$ channels, and $T = 200$ time steps (one second at 200 Hz). STFSFormer has 1.80M parameters, requires 0.46G MACs, and reaches 900 FPS on an RTX 4090. EEGNet is smaller and faster, while STFSFormer achieves higher SEED accuracy (92.54% versus 84.36%). Relative to SpikingGCN and SCNN, STFSFormer uses fewer parameters and MACs and achieves higher throughput. These comparisons characterize a trade-off rather than a minimum-cost model.

The FPGA evaluation in Table 5 uses a common clock constraint and streaming inference setting. Power is estimated at steady state, and energy per segment is computed from achieved segment throughput. STFSFormer has estimated power of 1.2 W and a *whole-model* energy cost of 1.5 mJ per EEG segment, with 45.3% LUT, 24.0% BRAM, and 8.0% DSP utilization. The energy estimate is lower than those of the implemented baselines. Appendix F separately reports spike activity and an operation-level energy estimate; that model excludes memory and peripheral costs and is distinct from the FPGA estimate.

Additional analyses. The sensitivity studies in Appendix D examine temporal block size, encoder decay profiles, downsampling, refractory decay, and branch count. Appendix E contains confusion matrices and convergence curves. Protocol details for both subject-dependent evaluation and strict cross-subject LOSO evaluation are provided in Appendix A.

5 Conclusion

In this work, we presented STFSFormer, an energy-efficient spiking Transformer for EEG emotion recognition. The model combines HELIF dynamics for transient-sensitive spike generation, parallel decay branches for complementary temporal responses, and STFSAtten for local spike-based feature interaction. Its dimension-aware multiplication rule further adapts the attention computation to the relative sizes of the temporal block and feature dimension.

Across SEED, SEED-V, and DREAMER, STFSFormer achieves competitive subject-dependent recognition performance, while the ablation results confirm the complementary contributions of the neuron and attention designs. The computational and FPGA evaluations show that the model provides a practical accuracy–efficiency trade-off, with 1.80M parameters, 0.46G MACs, and an estimated whole-model energy cost of 1.5 mJ per EEG segment. A strict LOSO evaluation reports performance on unseen subjects without target-subject adaptation. The current results support the use of jointly designed spike dynamics and local attention for resource-constrained EEG sensing; broader subject-independent settings and adaptive temporal grouping remain useful directions for future study.

AI Use Statement

We used generative AI tools for English translation and manuscript drafting and revision, including language editing, organization, LaTeX formatting, and reference-format and submission checks. The tools were not used to generate experimental data or reported numerical results. The authors manually reviewed the AI-assisted text, equations, citations, and formatting against the underlying manuscript, experimental records, and source publications, and take responsibility for the final content of this work.

REFERENCES

James Atwood and Don Towsley. Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Xin Chen, Youyong Kong, Hongli Chang, Yuan Gao, Zidong Liu, Jean-Louis Coatrieux, and Huazhong Shu. MGSN: Depression EEG lightweight detection based on multiscale DGCN and SNN for multichannel topology. *Biomedical Signal Processing and Control*, 92:106051, 2024.

Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018.

Qunxi Dong, Hongxin Cai, Zhigang Li, Jingyu Liu, and Bin Hu. A multiview brain network transformer fusing individualized information for autism spectrum disorder diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4854–4865, 2024. doi: 10.1109/JBHI.2024.3396457.

Cunhang Fan, Hongyu Zhang, Wei Huang, Jun Xue, Jianhua Tao, Jiangyan Yi, Zhao Lv, and Xiaopei Wu. DGSD: Dynamical graph self-distillation for EEG-based auditory spatial attention detection. *Neural Networks*, 179:106580, 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2024.106580. URL https://www.sciencedirect.com/science/article/pii/S0893608024005045.

Hongxiang Gao, Xingyao Wang, Zhenghua Chen, Min Wu, Zhipeng Cai, Lulu Zhao, Jianqing Li, and Chengyu Liu. Graph convolutional network with connectivity uncertainty for eeg-based emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(10):5917–5928, 2024. doi: 10.1109/JBHI.2024.3416944.

Peiliang Gong, Pengfei Wang, Yijia Zhou, and Daojun Tian. A spiking neural network with adaptive graph convolution and LSTM for EEG-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1440–1450, 2023. doi: 10.1109/TNSRE.2023.3253560.

Luyao Jin, Yonghao Song, Huan Zhao, Junyi Cao, Vincent C. K. Cheung, and Wei-Hsin Liao. Frequency-aware spatial-temporal attention explainable network for EEG decoding. *IEEE Journal of Biomedical and Health Informatics*, 29(10):7175–7185, 2025. doi: 10.1109/JBHI.2025.3576088.

Stamos Katsigiannis and Naeem Ramzan. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.

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.

Qi Li, Yunqing Liu, Yujie Shang, Qiong Zhang, and Fei Yan. Deep sparse autoencoder and recursive neural network for EEG emotion recognition. *Entropy*, 24(9):1187, 2022a.

Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. EEG-based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57, 2022b.

Yang Li, Gonzalo Mateos, and Zhengwu Zhang. Learning to model the relationship between brain structural and functional connectomes. *IEEE Transactions on Signal and Information Processing over Networks*, 8:830–843, 2022c.

Jingyu Liu, Weigang Cui, Yipeng Chen, Yulan Ma, Qunxi Dong, Ran Cai, Yang Li, and Bin Hu. Deep fusion of multi-template using spatio-temporal weighted multi-hypergraph convolutional networks for brain disease analysis. *IEEE Transactions on Medical Imaging*, 43(2):860–873, 2023.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997.

Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.

Xuegang Song, Feng Zhou, Alejandro F Frangi, Jiuwen Cao, Xiaohua Xiao, Yi Lei, Tianfu Wang, and Baiying Lei. Multicenter and multichannel pooling gcn for early ad diagnosis based on dual-modality fused brain network. *IEEE Transactions on Medical Imaging*, 42(2):354–367, 2022.

Enze Su, Siqi Cai, Longhan Xie, Haizhou Li, and Tanja Schultz. STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG. *IEEE Transactions on Biomedical Engineering*, 69(7):2233–2242, 2022.

Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural networks*, 111:47–63, 2019.

Fuze Tian, Lixian Zhu, Qiuxia Shi, Rui Wang, Lixin Zhang, Qunxi Dong, Kun Qian, Qinglin Zhao, and Bin Hu. The three-lead EEG sensor: Introducing an EEG-assisted depression diagnosis system based on ant lion optimization. *IEEE Transactions on Biomedical Circuits and Systems*, 17(6):1305–1318, 2023.

Fuze Tian, Haojie Zhang, Yang Tan, Lixian Zhu, Lin Shen, Kun Qian, Bin Hu, Björn W. Schuller, and Yoshiharu Yamamoto. An on-board executable multi-feature transfer-enhanced fusion model for three-lead EEG sensor-assisted depression diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 29(1):152–165, 2025. doi: 10.1109/JBHI.2024.3487012.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Shaolong Wei, Shu Jiang, Mingliang Wang, Liang Sun, Haonan Rao, Weiping Ding, and Jiashuang Huang. Scnns: spike-based coupling neural networks for understanding structural-functional relationships in the human brain. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6524–6532, 2025.

Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2973–2990, 2025. doi: 10.1109/TPAMI.2025.3530246.

Kairong Yu, Tianqing Zhang, Hongwei Wang, and Qi Xu. FSTA-SNN: Frequency-based spatial-temporal attention module for spiking neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22227–22235, 2025. doi: 10.1609/aaai.v39i21.34377.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Wei-Long Zheng and Bao-Liang 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.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=frE4fUwz_h.

Rui-Jie Zhu, Malu Zhang, Qihang Zhao, Haoyu Deng, Yule Duan, and Liang-Jian Deng. TCJA-SNN: Temporal-channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5112–5125, 2025. doi: 10.1109/TNNLS.2024.3377717.

Zulun Zhu, Jiaying Peng, Jintang Li, Liang Chen, Qi Yu, and Siqiang Luo. Spiking graph convolutional networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2434–2440, 2022. doi: 10.24963/ijcai.2022/338. URL https://www.ijcai.org/proceedings/2022/338.

A PREPROCESSING AND TRAINING DETAILS

Tables 6 and 7 summarize the settings needed to reproduce the reported experiments. Each EEG recording is filtered offline with a zero-phase 0.5–50 Hz band-pass filter. Complete trials (films for DREAMER) are first assigned to subject-dependent training, validation, and test sets in an 8:1:1 ratio. Window generation is then performed separately inside each partition, so windows from the same trial or film cannot occur in different sets. The same split-before-windowing rule is used for all datasets, while window length follows the dataset-specific configuration. Windows are non-overlapping, and neuron states reset at the beginning of every window.

Table 6: Dataset preprocessing and window construction. Window length and stride are given in samples.

Dataset	Channels C	Sampling rate	Window T	Stride	Overlap
SEED	62	200 Hz	200	200	None
SEED-V	62	250 Hz	250	250	None
DREAMER	14	128 Hz	256	256	None

Channel-wise z-score normalization is applied independently within each trial. During training, augmentation is applied with probability 0.3 by randomly choosing additive Gaussian noise ($\sigma = 0.05$), amplitude scaling in $[0.95, 1.05]$, temporal jitter within ± 20 ms, or Mixup with $\alpha = 0.2$ (Zhang et al., 2018). No augmentation is used for validation or inference.

Table 7: Model, optimization, and inference settings used in the reported configuration.

Item	Setting
Token dimensions	T' is the sequence length after temporal downsampling by $s = 2$; $D = 256$ is the token dimension. Thus $T' = 100$, 125, and 128 for SEED, SEED-V, and DREAMER, respectively.
Attention	$N_h = 4$ heads, $d_h = D/N_h = 64$, temporal block size $L = 3$, and two stacked attention–MLP blocks. The final incomplete block is zero-padded and its padded outputs are removed.
Spike encoder	$K = 5$ parallel branches; $\gamma_k = 0.81 + (k-1)\Delta_\gamma/(K-1)$ with $\Delta_\gamma = 0.16$, giving $\{0.81, 0.85, 0.89, 0.93, 0.97\}$.
Neuron dynamics	Encoder membrane decay $\eta = 0.89$, refractory decay $\lambda = 0.90$, shared downstream decay $\gamma_{\mathrm{shared}} = 0.89$, and threshold $V_{\mathrm{th}} = 0.3$.
Learnability	γ_k, γ_{shared}, η, and λ are fixed hyperparameters throughout training; they are not initialized as trainable tensors. Linear/convolutional projections, normalization parameters, and classifier weights are learned end to end using surrogate gradients.
Optimization	AdamW, initial learning rate 10^{-3}, weight decay 0.05, batch size 64, at most 100 epochs, label smoothing 0.1, and early-stopping patience 20. The checkpoint with the highest validation performance is evaluated.
Inference	Batch size 1 for throughput measurement; deterministic preprocessing, no augmentation, and zero neuron states at the start of each EEG window.

The symbols therefore have distinct roles: T' indexes downsampled temporal tokens, D is the full token width, N_h is the number of attention heads, d_h is the width of one head, and K is the number of parallel HELIF encoding branches. The five γ_k values specify five temporal response profiles and do not impose five predefined EEG frequency bands.

Cross-subject protocol. For cross-subject evaluation, we use strict leave-one-subject-out (LOSO): each subject is held out in turn, producing 15 folds for SEED and 23 folds for DREAMER. In every fold, all sessions and trials from the held-out subject are assigned exclusively to the test set; none of its data are used for training, validation, early stopping, or model selection. For SEED, the three sessions of each subject are kept together and leave the source set as a unit. The remaining subjects form the source-domain pool. Source training and validation are constructed only from these subjects, with complete trials assigned before windowing so that trial isolation is preserved. Models are trained end to end with AdamW, an initial learning rate of 10^{-3}, weight decay 0.05, and batch size 64 for at most 60 epochs. Early stopping uses source-validation accuracy with a

patience of 20 epochs, and the checkpoint with the highest source-validation accuracy is evaluated on the held-out subject. We use random seed 42 and report the mean and sample standard deviation across held-out-subject folds. EEGNet, Spikformer, and STFSFormer use identical LOSO folds and preprocessing. No target-subject adaptation or support set is used.

B Expanded Neuron Dynamics

The conventional LIF recurrence is

$$\begin{aligned} v[t] &= \eta v[t-1] + x[t] - V_{\mathrm{th}} s[t-1] \\ s[t] &= H_{\mathrm{sg}}(v[t] - V_{\mathrm{th}}) \end{aligned} \tag{7}$$

where $\eta = e^{-\Delta t/\tau_m}$ sets the membrane decay, V_{th} is the threshold, and H_{sg} uses a step forward pass and a surrogate derivative. Its integration of past inputs supplies temporal context. HELIF instead drives the membrane with a recurrent difference signal:

$$f[t] = \gamma f[t-1] + (x[t] - x[t-1]), \tag{8}$$

The input difference suppresses constant inputs and emphasizes temporal changes, while γ controls the persistence of this response. The membrane and refractory updates are

$$\begin{aligned} v[t] &= \eta v[t-1] + f[t] - V_{\mathrm{th}} r[t-1], \\ s[t] &= H_{\mathrm{sg}}(v[t] - V_{\mathrm{th}}), \\ r[t] &= \lambda r[t-1] + s[t]. \end{aligned} \tag{9}$$

The refractory state accumulates spike history, adding self-inhibition to subsequent membrane updates. These recurrences define the two added states $f[t]$ and $r[t]$; their decay factors control response persistence and inhibition. They are applied independently within each input window.

C Attention Tensor Operations

Let the projected encoder tokens be denoted by $\mathbf{X} \in \mathbb{R}^{T' \times D}$ in this appendix. All operations additionally carry a batch dimension, omitted below. Spiking projections give

$$\begin{aligned} \mathbf{Q} &= \phi(\mathbf{X}\mathbf{W}_Q), \\ \mathbf{K} &= \phi(\mathbf{X}\mathbf{W}_K), \\ \mathbf{V} &= \phi(\mathbf{X}\mathbf{W}_V), \end{aligned} \tag{10}$$

with $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times D}$. The multi-head shapes are

$$\begin{aligned} \mathbf{Q}, \mathbf{K}, \mathbf{V} &\in \{0,1\}^{N_h \times T' \times d_h}, \\ d_h &= \frac{D}{N_h}, \end{aligned} \tag{11}$$

Let $M = \lceil T'/L \rceil$. When T' is not divisible by L, the sequence is zero-padded to length ML before block extraction:

$$\begin{aligned} \mathbf{Q}^{(i)} &= \mathbf{Q}[:,\, iL : (i+1)L,\, :], \\ \mathbf{K}^{(i)} &= \mathbf{K}[:,\, iL : (i+1)L,\, :], \\ \mathbf{V}^{(i)} &= \mathbf{V}[:,\, iL : (i+1)L,\, :], \end{aligned} \tag{12}$$

Since no softmax or other nonlinearity occurs between the two products, matrix associativity permits two equivalent evaluation orders. STFSAtten uses a dimension-aware adaptive multiplication strategy to select the lower-cost order and smaller intermediate according to L and d_h:

$$\mathbf{O}^{(i)} = \alpha \begin{cases} \big(\mathbf{Q}^{(i)}(\mathbf{K}^{(i)})^\top\big)\mathbf{V}^{(i)}, & L \le d_h, \\ \mathbf{Q}^{(i)}\big((\mathbf{K}^{(i)})^\top \mathbf{V}^{(i)}\big), & L > d_h, \end{cases} \quad \mathbf{O}^{(i)} \in \mathbb{R}^{N_h \times L \times d_h}, \tag{13}$$

where the implementation uses $\alpha = 1/\sqrt{d_h}$ after either branch. The token-first branch forms an $L \times L$ intermediate, whereas the feature-first branch forms a $d_h \times d_h$ intermediate. Concatenation followed by removal of padded positions restores the original temporal axis:

$$\mathbf{O}_{\text{cat}} = \mathrm{Trim}_{T'}\Big(\mathrm{Concat}\big(\{\mathbf{O}^{(i)}\}_{i=0}^{M-1}\big)\Big) \in \mathbb{R}^{N_h \times T' \times d_h}. \tag{14}$$

A spiking readout and head merge then produce

$$\mathbf{O}_s = \phi(\mathbf{O}_{\text{cat}}) \in \mathbb{R}^{N_h \times T' \times d_h},$$
$$\mathbf{U} = \text{Merge}(\mathbf{O}_s) \in \mathbb{R}^{T' \times D}. \tag{15}$$

followed by the projected residual update

$$\text{STFSAtten}(\mathbf{X}) = \text{BN}(\mathbf{U}\mathbf{W}_O),$$
$$\mathbf{Y} = \mathbf{X} + \text{STFSAtten}(\mathbf{X}), \tag{16}$$

Here, "full-spike" denotes the binary query, key, value, and post-attention readout activations. The accumulated products, scaling, output projection, batch normalization, and residual representation can be real-valued. Excluding the query, key, and value projections, the two matrix products require

$$\mathcal{C}_{\text{STFSAtten}} = 2N_h M L d_h \min(L, d_h) \tag{17}$$

MACs. Processing all blocks in parallel stores an intermediate of size $\mathcal{O}\big(N_h M \min(L^2, d_h^2)\big)$; sequential block processing reduces this working storage to $\mathcal{O}\big(N_h \min(L^2, d_h^2)\big)$. Ignoring padding, the compute is approximately $2N_h T' d_h \min(L, d_h)$ and therefore scales linearly with T' for fixed L and d_h. With $T' = 100$, $L = 3$, and $d_h = 64$, the padded length is 102 and $M = 34$; the adaptive rule selects the token-first branch, whose 3×3 intermediate requires 39,168 core MACs per head. This design removes the quadratic dependence of global attention on sequence length, but it does not claim a uniform computational advantage over conventional block-local attention.

D HYPERPARAMETER SENSITIVITY

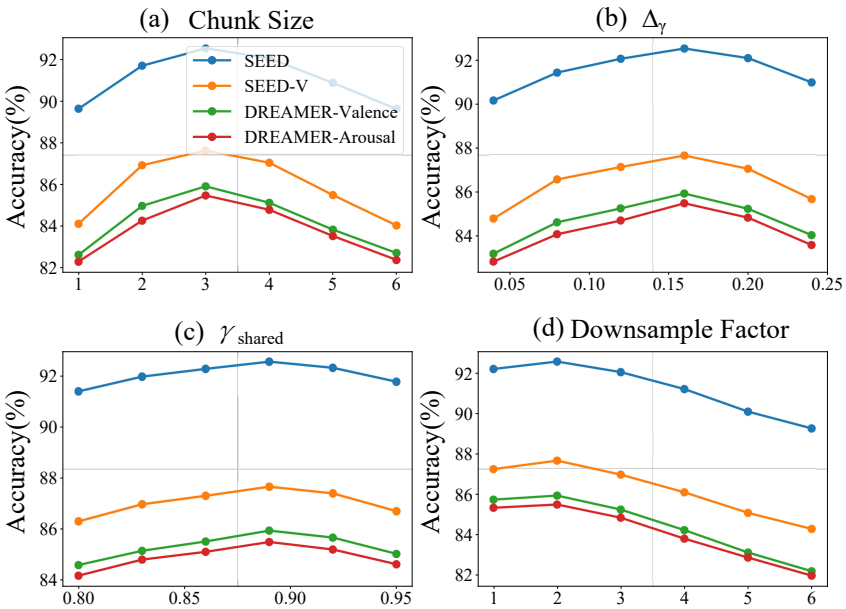

Figure 3: Hyperparameter sensitivity of STFSFormer on four EEG emotion recognition tasks. We sweep (a) the attention chunk size L, (b) the multi-branch decay span Δ_γ in the spike encoder, (c) the shared decay γ_{shared} for later spiking neurons, and (d) the temporal downsampling factor s. Across datasets, the best performance consistently occurs at L=3, Δ_γ=0.16, γ_{shared}=0.89, and s=2.

Figure 3 varies temporal block size L, multi-branch decay span Δ_γ, shared decay γ_{shared}, and downsampling factor s. Across the four tasks, the reported optima are $L = 3$, $\Delta_\gamma = 0.16$, $\gamma_{\text{shared}} = 0.89$, and $s = 2$. These settings govern temporal context, response persistence, and token resolution. Performance decreases with more aggressive downsampling beyond $s = 2$ in the reported sweep. The curves describe configuration sensitivity; they do not directly identify the physiological origin of the changes.

Table 8: Sensitivity to the refractory decay λ in HELIF. Accuracy is reported in %.

λ	SEED	SEED-V	DREAMER-V	DREAMER-A
0.70	91.98	86.91	85.21	84.84
0.75	92.21	87.19	85.47	85.05
0.80	92.39	87.43	85.68	85.26
0.85	92.50	87.56	85.82	85.39
0.90	**92.54**	**87.63**	**85.91**	**85.47**
0.95	92.31	87.28	85.62	85.14

Table 9: Sensitivity to the number of encoder branches K. Accuracy is reported in %.

K	SEED	SEED-V	DREAMER-V	DREAMER-A
1	91.62	86.31	85.03	84.71
2	92.05	86.84	85.41	85.09
3	92.33	87.21	85.63	85.27
4	92.47	87.46	85.80	85.36
5	**92.54**	**87.63**	**85.91**	**85.47**
6	92.49	87.52	85.86	85.41

D.1 Refractory Decay and Branch Count

Table 8 varies the refractory decay λ. Accuracy peaks at $\lambda = 0.90$ on each task and decreases at 0.95. In the neuron recurrence, increasing λ extends spike-history inhibition, so this sweep tests the influence of that persistence on recognition.

Table 9 varies the number of parallel encoder branches. Accuracy improves from $K = 1$ to $K = 5$, with a small decrease at $K = 6$. The reported model uses five branches. This result supports the chosen branch count in the evaluated setting; the branches are learned temporal response profiles, not prescribed bands.

E Prediction and Optimization Diagnostics

E.1 Confusion Matrices

Figure 4 reports row-normalized confusion matrices for SEED, SEED-V, and the two DREAMER tasks. They show class-specific recognition and error patterns beyond aggregate accuracy. For both DREAMER valence and arousal, ratings ≥ 3.0 are labeled high (1), and ratings < 3.0 are labeled low (0). The matrices describe prediction errors but do not by themselves establish a neurophysiological cause for those errors or isolate the contribution of an individual model component.

E.2 Convergence

Figure 5 shows the supplied training and validation accuracy and loss trajectories. The displayed runs converge, with validation performance leveling off as training progresses. The best validation epoch is used for checkpoint selection. HELIF, surrogate gradients, residual connections, and normalization are trained together; these trajectories alone do not separate their effects on optimization stability.

F Spike Firing Rates and Operation-Level Energy

We define firing rate as the ratio of spike events to neuron-time activations:

$$\mathrm{FR} = \frac{\sum_\ell \sum_t \sum_i \mathbf{1}[s_i^{(\ell)}(t) = 1]}{\sum_\ell N_\ell T}. \tag{18}$$

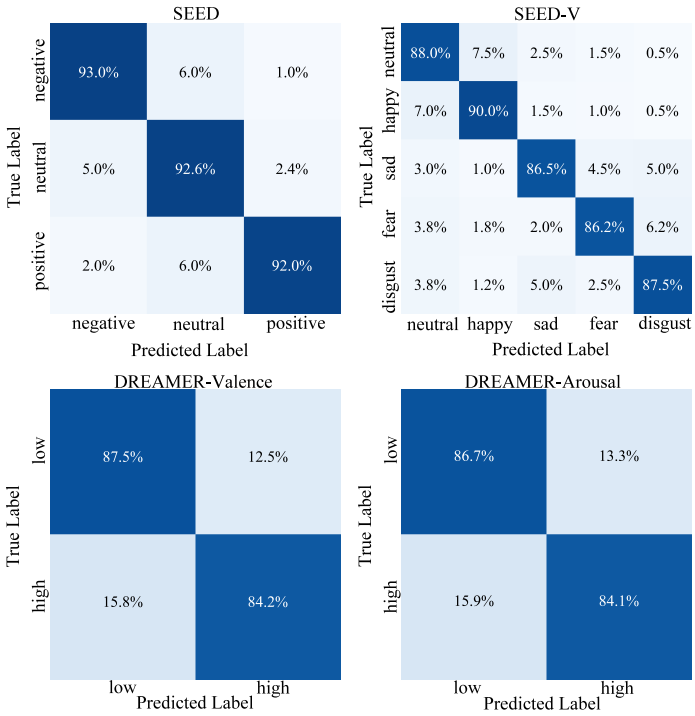

Figure 4: Row-normalized confusion matrices (%) of STFSFormer on SEED, SEED-V, and DREAMER (valence/arousal), showing class-wise recognition and misclassification rates.

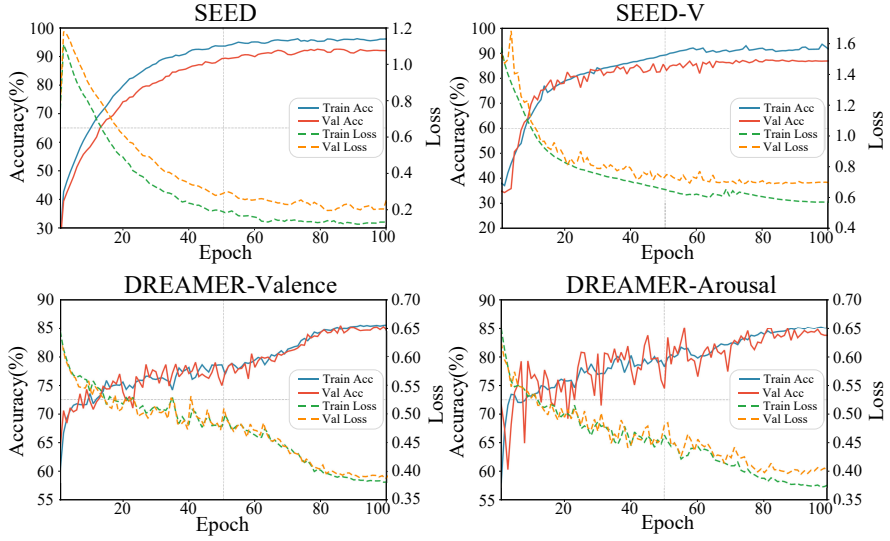

Figure 5: Training and validation accuracy and loss curves of STFSFormer on SEED, SEED-V, and DREAMER (valence/arousal), showing stable convergence and well-behaved optimization across all benchmarks.

Here N_ℓ is the number of neurons in layer ℓ and T is the number of evaluated time steps, under the common-duration convention of this expression. Table 10 reports overall average firing rates

Table 10: Spike firing rates (%) and theoretical energy consumption across different tasks.

Component	DREAMER		SEED	SEED-V
	Valence	Arousal		
Φ_{mf} (γ_1=0.81)	18.34	17.26	21.47	20.15
Φ_{mf} (γ_4=0.97)	8.75	8.12	11.64	10.42
STFSAtten	15.24	14.56	17.83	16.47
SpikeMLP	16.18	15.42	18.96	17.62
Overall Avg.	14.62	13.85	17.48	16.14
Energy Ratio	2.87%	2.72%	3.43%	3.16%

of 14.62%, 13.85%, 17.48%, and 16.14% on DREAMER-Valence, DREAMER-Arousal, SEED, and SEED-V. The listed lower-decay encoder branch ($\gamma_1 = 0.81$) fires more frequently than the higher-decay branch ($\gamma_4 = 0.97$). STFSAtten and SpikeMLP rates remain below 25% in this table. The observations quantify sparse activity; actual operation skipping depends on the execution implementation.

Using the original operation-level model in which accumulation consumes approximately $5.1\times$ less energy than a MAC in 45 nm CMOS gives

$$\frac{E_{\mathrm{SNN}}}{E_{\mathrm{ANN}}} \approx \frac{\mathrm{FR}}{5.1}. \tag{19}$$

The table reports ratios from 2.72% to 3.43%. This estimate excludes memory access and peripheral overhead and is separate from the whole-model FPGA energy estimate of 1.5 mJ per segment in Table 5.