SPIKEMIND: PHYSIOLOGICALLY GROUNDED SPIKING CONCEPTS FOR EEG EMOTION RECOGNITION AND RATIONALE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) emotion recognition often relies on latent features whose physiological meaning is difficult to compare across subjects or trace through language rationales. SpikeMind uses 12 differential-entropy (DE)-grounded temporal concepts as shared variables for recognition and rationale generation. Their targets describe regional activity and contrasts computed from training-standardized DE features. A coordinate-preserving spiking encoder and fixed band–region readouts learn trajectories aligned with these targets. Reliability-weighted source-subject prototypes classify their temporal summaries and expose predicted-versus-runner-up margin evidence. A planner defines supervision and evaluation facts, while Qwen3-4B organizes the complete continuous concept record into a rationale. On SEED and SEED-IV, weighted accuracy reaches 94.0% and 86.7% under subject-dependent evaluation, and 89.1% and 79.4% under source-only leave-one-subject-out evaluation. The latter exceed EEG Conformer under the same source-prototype rule by 4.1 and 4.4 percentage points. Concept-marker correlations range from 0.748 to 0.842, and language evidence-retention F1 ranges from 0.832 to 0.894. This shared record makes marker agreement, decision sensitivity, and language evidence retention separately testable within one recognition system.

Keywords: EEG emotion recognition; spiking neural networks; physiological concepts; cross-subject generalization; language rationales.

1 INTRODUCTION

Automatic emotion recognition supports affective computing and brain-computer interfaces (Alarcão & Fonseca, 2019). EEG captures affect-related spectral and spatial patterns at high temporal resolution (Zheng & Lu, 2015), but inter-subject variability shifts their distributions (Zheng et al., 2019b; Hu et al., 2021). Beyond accuracy on unseen subjects, this raises a question about evidence: can predictions be compared through the same physiologically defined variables across people?

Graph convolutions model electrode relations (Song et al., 2020; Zhong et al., 2022), convolutional transformers integrate local and global context (Song et al., 2023), and spiking encoders provide event-driven decoding (Wang et al., 2025). These models learn discriminative representations, but their feature coordinates generally have no predefined physiological meaning. Input attribution locates influential samples, whereas comparing explanations across subjects also requires an identifiable evidence vocabulary. EEG-to-LLM systems generate language from encoder representations (Ma et al., 2026), leaving the connection between physiological markers, decision variables, and generated claims difficult to inspect.

Predefined concepts must also support discrimination. In our SEED leave-one-subject-out (LOSO) comparison, deterministic differential-entropy (DE) markers reproduce their targets exactly, yet trail SpikeMind by 14.2 percentage points in accuracy. This comparison motivates learned temporal concepts with measurable DE references and a shared record for examining decisions and rationales.

SpikeMind constructs this record from 12 time-varying DE-grounded concepts (Figure 1). A coordinate-preserving spiking encoder and fixed readouts estimate regional activity, hemispheric asymmetry, and cross-band contrasts in the standardized DE representation. Reliability-weighted source-subject prototypes classify their temporal summaries and provide signed evidence relative to the runner-up class. Qwen3-4B then organizes the prediction and all 12 concept records into a

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.

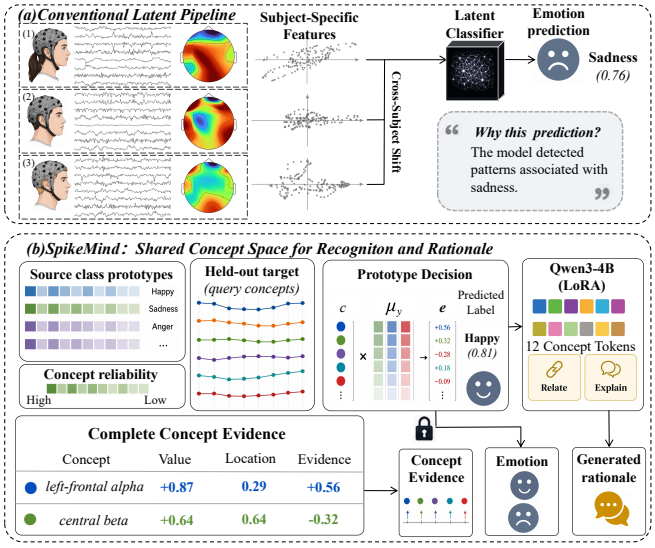

Figure 1: SpikeMind assigns predefined band–region identities to its decision coordinates and language evidence. Concept grounding is evaluated by agreement with standardized DE-derived targets across subjects.

rationale. A deterministic planner specifies the facts used for supervision and evaluation, linking generated claims to concept values, temporal locations, and decision roles.

We make three contributions. First, a dual-view spiking encoder retains temporal and scalp–frequency indices for fixed readouts aligned with standardized DE-derived targets. Second, subject-balanced prototypes and concept reliability weights perform source-only cross-subject recognition and expose decision evidence in these predefined coordinates. Third, planner-derived supervision connects Qwen3-4B rationales to the complete concept record, evaluated through classifier interventions, language evidence retention, and language organization. Experiments examine recognition, marker agreement, interventions, language retention, organization, and operation accounting.

2 Related Work

EEG Emotion Recognition and Spiking Encoders

EEG emotion recognition uses differential entropy (DE) to characterize band-wise activity (Zheng & Lu, 2015), graph networks to model electrode relations (Song et al., 2020; Zhong et al., 2022), and convolutional transformers to capture local and global dependencies (Song et al., 2023). Spiking networks support sparse computation (Roy et al., 2019) through mechanisms such as learnable membrane dynamics (Fang et al., 2021) and spike-driven attention (Yao et al., 2023). S^2M-Former combines spatial and frequency branches for auditory attention detection (Wang et al., 2025). These architectures learn predictive representations without assigning fixed DE identities to their decision coordinates. SpikeMind adapts dual-view spiking encoding to retain window, band, and scalp-location axes for predefined DE-grounded readouts.

Cross-Subject EEG Recognition

Cross-subject recognition addresses variability in EEG representations across people (Zheng et al., 2019b). Domain adaptation uses target-subject data, as in plug-and-play calibration with unlabeled EEG (Zhao et al., 2021). Source-only domain generalization instead learns from training subjects, including through denoising mixed mutual reconstruction (Wang et al., 2024). Prototypical learning summarizes support classes (Snell et al., 2017), while episodic domain generalization simulates domain shift among sources (Li et al., 2018). These learning rules specify how classes and domains are compared; physiological coordinate identities and dimension-wise stability require additional structure. SpikeMind combines subject-balanced prototypes with source-derived reliability weights in a fixed concept space for source-only LOSO prediction.

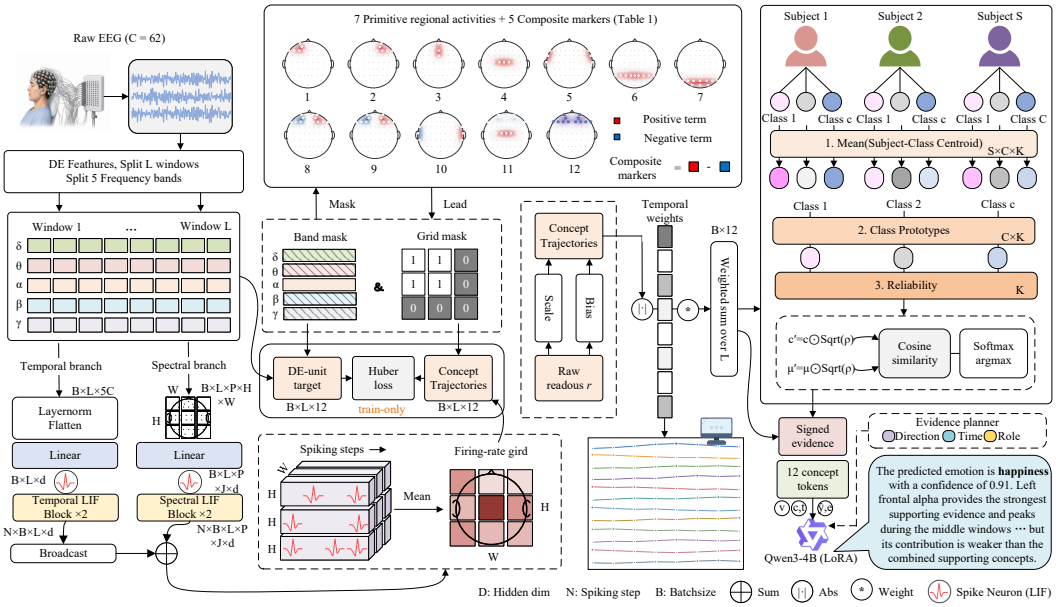

Figure 2: SpikeMind maps dual-view spikes to 12 DE-grounded concepts for prototype classification and conditions Qwen3-4B on their values, temporal locations, and predicted-versus-runner-up margin evidence.

Concept-Grounded Recognition and Language Rationales

Concept methods differ in how concepts enter a decision. TCAV probes latent concept directions (Kim et al., 2018), whereas concept bottlenecks make concepts explicit, intervenable prediction variables (Koh et al., 2020; Shin et al., 2023). Trustworthiness studies examine regional grounding (Huang et al., 2024), and V2C-CBM connects images to visual concepts through a learned tokenizer (He et al., 2025). Explanation bottlenecks predict through generated text (Yamaguchi & Nishida, 2025); EEG-to-LLM systems instead condition language on EEG embeddings (Ma et al., 2026). SpikeMind classifies DE-grounded temporal summaries and conditions language on their values, locations, and predicted-versus-runner-up margin evidence. This record makes the relationship between decision variables and generated claims directly examinable. We distinguish classifier interventions (DeYoung et al., 2020) from language retention and organization, following the separation of interpretability and faithfulness (Jacovi & Goldberg, 2020; Li et al., 2024).

3 Method

Overview

Figure 2 maps DE through spiking branches to 12 concept trajectories. Source prototypes classify their temporal summaries; prediction, values, centers, and relative evidence condition Qwen3-4B. Planner-derived targets supervise and evaluate rationales. Physiological grounding denotes supervision and measured agreement with predefined standardized DE targets.

Coordinate-Preserving Dual-View Spiking Encoder

Input $A \in \mathbb{R}^{L\times C_e\times P}$ contains DE from $C_e = 62$ electrodes and $P = 5$ bands. Frozen fold-training statistics standardize each electrode–band coordinate before constructing and standardizing concept targets within the same training boundary.

The temporal branch flattens, layer-normalizes, and projects each window to $E^{\mathrm{t}} \in \mathbb{R}^{L\times d}$; projecting $X^{\mathrm{f}} \in \mathbb{R}^{L\times P\times H\times W}$ gives $E^{\mathrm{f}} \in \mathbb{R}^{L\times P\times J\times d}$, with $J = HW$ and empty-cell mask $G \in \{0,1\}^{J}$.

Each branch repeats projected currents over N steps through two S^{2}M-Former-adapted residual leaky integrate-and-fire (LIF) blocks (Wang et al., 2025): batch normalization, $d \to 2d \to d$ membrane mixers, and temporal kernel-3 or spectral 3×3 depthwise convolutions. Outputs are

Table 1: Fixed concept dictionary computed from training-standardized DE features A. Asymmetries use right-minus-left; contrasts use first-minus-second. Concept names refer to this standardized representation.

k	Concept name ν_k	Target $m_{\ell,k}$	k	Concept name ν_k	Target $m_{\ell,k}$
Primitive regional activity			*Composite markers*		
1	Left-frontal alpha activity	$d^{\mathcal{F}_L}_{\ell,\alpha}$	8	Frontal alpha asymmetry	$d^{\mathcal{F}_R}_{\ell,\alpha} - d^{\mathcal{F}_L}_{\ell,\alpha}$
2	Right-frontal alpha activity	$d^{\mathcal{F}_R}_{\ell,\alpha}$	9	Frontal beta asymmetry	$d^{\mathcal{F}_R}_{\ell,\beta} - d^{\mathcal{F}_L}_{\ell,\beta}$
3	Frontal-midline theta activity	$d^{\mathcal{F}_M}_{\ell,\theta}$	10	Temporal gamma asymmetry	$d^{\mathcal{T}_R}_{\ell,\gamma} - d^{\mathcal{T}_L}_{\ell,\gamma}$
4	Central beta activity	$d^{\mathcal{C}}_{\ell,\beta}$	11	Central-beta/frontal-alpha contrast	$d^{\mathcal{C}}_{\ell,\beta} - d^{\mathcal{F}_B}_{\ell,\alpha}$
5	Fronto-temporal gamma activity	$d^{\mathcal{F}_T}_{\ell,\gamma}$	12	Frontal beta/theta contrast	$d^{\mathcal{F}_B}_{\ell,\beta} - d^{\mathcal{F}_B}_{\ell,\theta}$
6	Parietal alpha activity	$d^{\mathcal{P}}_{\ell,\alpha}$			
7	Occipital alpha activity	$d^{\mathcal{O}}_{\ell,\alpha}$			

$S^{\mathrm{t}} \in \{0,1\}^{N\times L\times d}$ and $S^{\mathrm{f}} \in \{0,1\}^{N\times L\times P\times J\times d}$; the supplement specifies LIF and surrogate gradients (Neftci et al., 2019).

For $x = (n, \ell, p, j)$, let $U^{\mathrm{f}}_x = W_{\mathrm{f}} S^{\mathrm{f}}_{n,\ell,p,j,:}$ and $U^{\mathrm{t}}_{n,\ell} = W_{\mathrm{t}} S^{\mathrm{t}}_{n,\ell,:}$, with $W_{\mathrm{f}}, W_{\mathrm{t}}, W_{\mathrm{o}} \in \mathbb{R}^{d\times d}$. Fusion broadcasts temporal activity over (p, j):

$$\begin{aligned} I^{\mathrm{fus}}_x &= U^{\mathrm{f}}_x + W_{\mathrm{o}}\big(U^{\mathrm{f}}_x \odot U^{\mathrm{t}}_{n,\ell}\big), \\ F_x &= G_j \,\mathrm{LIF}_n\big(G_j I^{\mathrm{fus}}_x\big). \end{aligned} \tag{1}$$

$F \in \{0,1\}^{N\times L\times P\times J\times d}$ preserves simulation-step, window, band, location, and channel indices for fixed readouts. Temporal fusion broadcasts all electrode–band context, so concept responses can integrate inputs beyond their named region or band; alignment supplies their DE reference.

Spike-to-Concept Bridge and Temporal Aggregation

Fixed readouts use left/right frontal $\mathcal{F}_{L/R}$ and temporal $\mathcal{T}_{L/R}$ regions, bilateral frontal $\mathcal{F}_B = \mathcal{F}_L \cup \mathcal{F}_R$, frontal-midline $\mathcal{F}_M$, central $\mathcal{C}$, parietal $\mathcal{P}$, occipital $\mathcal{O}$, and fronto-temporal $\mathcal{F}_T$; electrode lists appear in the supplement.

Regional means $d^{\mathcal{R}}_{\ell,p} = |\mathcal{R}|^{-1} \sum_{a\in\mathcal{R}} A_{\ell,a,p}$ define the $K = 12$ targets $m_{\ell,k} = \phi_k(A_{\ell,:,:})$ in Table 1: standardized regional activity, hemispheric asymmetry, and cross-band contrasts (Davidson, 1992; Zheng et al., 2019b; Bilucaglia et al., 2020). Electrode–band normalization can change contrast signs relative to raw DE.

Grid cell j maps to electrode $e(j)$ or $\varnothing$. Masks $M_{\mathcal{R},j} = \mathbb{I}[e(j) \in \mathcal{R}]$ define averages $\mathcal{A}(M, x) = \langle M, x\rangle / \langle M, \mathbf{1}\rangle$. Readout terms $\mathcal{T}_k = \{(a_{kq}, p_{kq}, M_{kq})\}_{q=1}^{Q_k}$ use one positive primitive term or composite signs $(+1, -1)$:

$$R_{\ell,p,j} = \frac{1}{Nd} \sum_{n,h} F_{n,\ell,p,j,h}, \quad r_{\ell,k} = \sum_{q=1}^{Q_k} a_{kq} \mathcal{A}(M_{kq}, R_{\ell,p_{kq},:}). \tag{2}$$

$R \in [0,1]^{L\times P\times J}$ retains window–band–location coordinates. Calibration $v_{\ell,k} = w_k r_{\ell,k} + b_k$, $w_k = \mathrm{softplus}(\bar{w}_k) > 0$, preserves the ordering of each signed readout $r_{\ell,k}$. The offset b_k calibrates its zero point. Targets $\widetilde{m}_{\ell,k} = (m_{\ell,k} - \mu^{\mathrm{tr}}_k)/(\sigma^{\mathrm{tr}}_k + \epsilon)$ use frozen training-fold statistics and $\epsilon > 0$; alignment is dimensionless. Calibration preserves scale-dependent affine readout relations, while concept-specific temporal weights change summary relations; the supplement derives these and standardized-target identities.

Signed trajectories receive magnitude-based weights. With $\tau_\ell = (\ell - 1)/(L - 1)$ and $\gamma_k = \mathrm{softplus}(\bar{\gamma}_k) > 0$,

$$\alpha_{\ell,k} = \mathrm{softmax}_\ell(\gamma_k |v_{\ell,k}|), \quad (c_k, t_k) = \sum_{\ell=1}^{L} \alpha_{\ell,k}(v_{\ell,k}, \tau_\ell). \tag{3}$$

Outputs are trajectories $V \in \mathbb{R}^{L\times K}$, window weights $\alpha \in [0,1]^{L\times K}$, signed summaries $c \in \mathbb{R}^K$, and weighted temporal centers $t \in [0,1]^K$.

Reliability-Weighted Prototype Classification

Concept coordinates are shared across subjects, but their stability differs. Classification uses only c, with subject-balanced prototypes and stability-based reliability.

For C classes and source subjects $\mathcal{S}$, episode $q \in \mathcal{S}$ uses supports $\mathcal{S}_q \subseteq \mathcal{S} \setminus \{q\}$. Class-specific examples $\mathcal{B}_{s,y}$ yield $\mu_{s,y} = |\mathcal{B}_{s,y}|^{-1} \sum_{i\in\mathcal{B}_{s,y}} c_i$; then

$$\mu_y^{(-q)} = \frac{1}{|\mathcal{S}_q|} \sum_{s\in\mathcal{S}_q} \mu_{s,y}, \quad \delta_k^{(-q)} = \frac{1}{C} \sum_{y=1}^{C} \mathrm{Var}_{s\in\mathcal{S}_q}(\mu_{s,y,k}). \tag{4}$$

The variance uses denominator $|\mathcal{S}_q| - 1$; $\rho^{(-q)} = \mathrm{softmax}(-\delta^{(-q)}/T_\rho)$, $T_\rho > 0$, favors stable coordinates. Equal subject weighting precedes class comparison. We omit $(-q)$ below.

For reliability-weighted cosine comparison, define $\mathcal{N}_\epsilon(x) = x/\max(\|x\|_2, \epsilon)$, $\bar{c} = \mathcal{N}_\epsilon(\sqrt{\rho} \odot c)$, and $\bar{\mu}_y = \mathcal{N}_\epsilon(\sqrt{\rho} \odot \mu_y)$. Similarities $s_y(c) = \bar{c}^\top \bar{\mu}_y$ give $\pi = \mathrm{softmax}(s(c)/T)$, $T > 0$, prediction $\hat{y} = \arg\max_y s_y(c)$, and runner-up $y^- = \arg\max_{y\neq\hat{y}} s_y(c)$:

$$\begin{aligned} g_k &= \bar{c}_k(\bar{\mu}_{\hat{y},k} - \bar{\mu}_{y^-,k}), \\ e_k &= \frac{g_k}{\sum_{j=1}^K |g_j| + \epsilon}, \qquad \sum_{k=1}^K g_k = s_{\hat{y}}(c) - s_{y^-}(c). \end{aligned} \tag{5}$$

g_k sum to the predicted-versus-runner-up similarity margin; e_k assigns each coordinate a signed relative strength.

Episodes estimate μ and ρ from supports. LOSO evaluation rebuilds them from all outer-fold source-training subjects; SD forms μ_y^{SD} from all class-y fold-training examples with uniform $\rho_k^{\mathrm{SD}} = 1/K$. Frozen normalization and these statistics exclude validation and test examples. The decision yields $(\mu, \rho, \pi, \hat{y}, e)$.

Evidence-Conditioned Rationale Generation

$\mathcal{D}_{\hat{y}} = \{(\nu_k, c_k, t_k, e_k)\}_{k=1}^K$ contains all 12 concept names, standardized values, weighted temporal centers, and relative evidence, including neutral items.

Training-only thresholds $\theta_k^c = Q_{0.25}(\{|c_{i,k}|\}_{i\in\mathcal{I}_{\mathrm{tr}}})$ and $\theta_k^e = Q_{0.25}(\{|e_{i,k}|\}_{i\in\mathcal{I}_{\mathrm{tr}}})$ use LOSO source-training or SD subject-fold records $\mathcal{I}_{\mathrm{tr}}$. Signed values and symmetric dead zones define higher/lower/near-baseline and supporting/opposing/neutral. Zero denotes the training window-marker mean before target normalization; pooled-summary means can differ. Directions use this standardized reference. Intervals $[0, 1/3]$, $(1/3, 2/3]$, and $(2/3, 1]$ define early/middle/late.

$\mathcal{P}_{\hat{y}}$ contains the prediction and all $\mathcal{P}_k = (\nu_k, \mathrm{dir}(c_k), \mathrm{bin}(t_k), \mathrm{role}(e_k))$. Its parameter-free serializer supplies complete training targets and the language reference; Qwen uses continuous evidence.

Concept-name embeddings η_k and projection $\psi : \mathbb{R}^3 \to \mathbb{R}^{d_{\mathrm{LLM}}}$ give:

$$\begin{aligned} z_k &= \psi([c_k; t_k; e_k]) + \eta_k, \\ H_0 &= [p_{1:M_p}; z_{1:K}; \eta_{\hat{y}}^{\mathrm{cls}}; \eta_r]. \end{aligned} \tag{6}$$

$p_{1:M_p}$, $\eta_{\hat{y}}^{\mathrm{cls}}$, and η_r denote prefix, predicted-class, and query embeddings. $H_0 \in \mathbb{R}^{(M_p+K+2)\times d_{\mathrm{LLM}}}$ conditions Qwen3-4B-Instruct-2507 (Qwen Team, 2025) to generate $r = (r_1, \ldots, r_M)$ from the prediction and concept tuples.

Two-Stage Training and Inference

Perception updates the encoder, fusion, calibration, and temporal weighting through support statistics. For queries $\mathcal{Q}$, $\mathcal{L}_{\mathrm{proto}}$ is mean cross-entropy; $\mathcal{L}_{\mathrm{concept}}$ averages unit-transition Huber loss on standardized residuals over examples, windows, and concepts.

Table 2: Recognition on SEED and SEED-IV (subject-level mean ± standard deviation). Bold marks the highest mean. Explanation sources: N, none; PH, post-hoc; DM, deterministic DE markers; LC, learned concepts; EE, encoder embeddings; DC, DE-grounded concepts.

SEED					
Method	Source	SD WA ↑	SD UA ↑	LOSO WA ↑	LOSO UA ↑
DE + SVM	N	84.2 ± 9.4	83.5 ± 9.7	58.7 ± 15.8	57.6 ± 16.1
DGCNN	PH	90.4 ± 8.1	89.9 ± 8.4	79.6 ± 9.1	78.7 ± 9.4
RGNN	PH	91.6 ± 7.3	91.1 ± 7.5	82.0 ± 8.8	81.2 ± 9.1
EEG Conformer	PH	92.7 ± 6.4	92.2 ± 6.6	84.2 ± 7.9	83.4 ± 8.2
EEG Conformer + source prototypes	PH	93.0 ± 6.1	92.5 ± 6.4	85.0 ± 7.3	84.2 ± 7.6
S^2M-Former	N	92.1 ± 6.6	91.6 ± 6.9	83.5 ± 8.1	82.7 ± 8.4
Direct EEG markers	DM	87.8 ± 8.7	87.0 ± 9.0	74.9 ± 10.2	73.8 ± 10.6
Generic concept bottleneck	LC	90.9 ± 7.5	90.3 ± 7.8	80.5 ± 9.0	79.6 ± 9.3
E^2-LLM	EE	92.4 ± 6.6	91.9 ± 6.9	83.7 ± 8.0	82.9 ± 8.3
SpikeMind	DC	$\mathbf{94.0 \pm 5.1}$	$\mathbf{93.6 \pm 5.4}$	$\mathbf{89.1 \pm 5.6}$	$\mathbf{88.5 \pm 5.9}$
SEED-IV					
Method	Source	SD WA ↑	SD UA ↑	LOSO WA ↑	LOSO UA ↑
DE + SVM	N	69.8 ± 11.7	68.7 ± 12.0	51.6 ± 14.4	50.4 ± 14.8
DGCNN	PH	77.5 ± 10.1	76.6 ± 10.4	64.6 ± 11.5	63.6 ± 11.9
RGNN	PH	80.1 ± 8.9	79.3 ± 9.2	72.0 ± 8.4	71.1 ± 8.8
EEG Conformer	PH	82.8 ± 7.7	82.0 ± 8.0	73.8 ± 8.2	72.9 ± 8.5
EEG Conformer + source prototypes	PH	83.2 ± 7.4	82.4 ± 7.7	75.0 ± 7.8	74.1 ± 8.1
S^2M-Former	N	81.9 ± 8.0	81.1 ± 8.3	72.7 ± 8.7	71.8 ± 9.0
Direct EEG markers	DM	73.9 ± 10.9	72.9 ± 11.3	60.8 ± 12.4	59.7 ± 12.8
Generic concept bottleneck	LC	79.4 ± 9.2	78.5 ± 9.5	69.5 ± 9.6	68.5 ± 10.0
E^2-LLM	EE	82.4 ± 7.8	81.6 ± 8.1	73.4 ± 8.4	72.5 ± 8.7
SpikeMind	DC	$\mathbf{86.7 \pm 6.7}$	$\mathbf{86.1 \pm 7.0}$	$\mathbf{79.4 \pm 7.2}$	$\mathbf{78.7 \pm 7.5}$

Language training uses source-training records $\mathcal{G}_{\mathrm{LM}}$, serializer targets r_g^*, and $N_{\mathrm{tok}} = \sum_g M_g$ tokens:

$$\mathcal{L}_{\mathrm{perc}} = \mathcal{L}_{\mathrm{proto}} + \lambda \mathcal{L}_{\mathrm{concept}},$$
$$\mathcal{L}_{\mathrm{LM}} = -\frac{1}{N_{\mathrm{tok}}} \sum_{g \in \mathcal{G}_{\mathrm{LM}}} \sum_{m=1}^{M_g} \log p_{\Theta_L}(r_{g,m}^* \mid r_{g,<m}^*, H_{0,g}). \tag{7}$$

Classification gradients reach queries and supports; alignment supervises query trajectories. The supplement specifies reductions.

Language training freezes perception, normalization, and training-rebuilt prototypes and reliability. Records use predicted classes and runners-up for e; validation controls stopping and test records supply final evaluation.

Language updates LoRA, ψ, and prefix, concept-name, predicted-class, and query embeddings from stored records with perception and Qwen's base frozen. Inference computes $(\pi, \hat{y}, e, \mathcal{D}_{\hat{y}})$ before r; implementation and energy settings appear in the supplement.

4 Experiments

Datasets

SEED records positive, neutral, and negative states from 15 subjects in three sessions, each containing 15 film clips (Zheng & Lu, 2015). Its released features provide linear-dynamical-system-smoothed differential entropy (DE) from 62 electrodes at one-second resolution in the delta, theta, alpha, beta, and gamma bands.

SEED-IV records happiness, sadness, fear, and neutrality from 15 subjects under a three-session, 62-electrode protocol with 24 clips per session (Zheng et al., 2019a). We retain its released four-second DE frames. Shared electrode and frequency-band coordinates allow the same encoder and concept dictionary, with dataset-specific class prototypes.

Table 3: (a) SEED LOSO ablations, with window-level marker correlation. (b) Concept grounding and Qwen retention against the complete 12-concept evidence plan.

(a) Ablation analysis

Variant	WA ↑	UA ↑	Corr. ↑
Full SpikeMind	89.1 ± 5.6	88.5 ± 5.9	0.793
Direct markers	74.9 ± 10.2	73.8 ± 10.6	1.000
Dense ANN bridge	87.7 ± 6.1	86.9 ± 6.4	0.776
Unconstrained readout	85.8 ± 6.8	84.9 ± 7.1	0.612
Mean temporal pooling	87.3 ± 6.3	86.5 ± 6.6	0.789
Linear concept head	86.9 ± 6.5	86.1 ± 6.8	0.781
Uniform concept reliability	88.0 ± 6.0	87.3 ± 6.3	0.788
No concept alignment	86.6 ± 6.6	85.8 ± 6.9	0.263
Direct encoder head	86.2 ± 6.4	85.4 ± 6.7	N/A

(b) Grounding and retention

Concept grounding

Setting	Corr. ↑	MAE ↓	Huber ↓
SEED SD	0.842	0.118	0.011
SEED LOSO	0.793	0.149	0.017
SEED-IV SD	0.812	0.136	0.015
SEED-IV LOSO	0.748	0.173	0.022

Qwen evidence retention

Setting	Prec. ↑	Rec. ↑	F1 ↑
SEED SD	0.918	0.872	0.894
SEED LOSO	0.896	0.839	0.867
SEED-IV SD	0.901	0.844	0.872
SEED-IV LOSO	0.872	0.796	0.832

Evaluation Protocol

Examples comprise eight non-overlapping DE frames ($L = 8$): 8 seconds on SEED or 32 on SEED-IV. Incomplete tails are discarded, trials stay intact, and temporal locations are normalized within examples. Stimulus-grouped subject-dependent (SD) evaluation uses five SEED or six SEED-IV folds (Hu et al., 2021). Leave-one-subject-out (LOSO) evaluation has 15 outer folds, each reserving 10% of source trial groups for validation.

Feature and target normalization use frozen fold-training statistics. Episode supports determine prototypes and reliability; LOSO evaluation rebuilds both from all source-training subjects. SD prototypes use all class-specific subject-fold training examples with uniform reliability. Planner thresholds use training records; validation selects checkpoints and controls language early stopping. After freezing perception, source-training records update Θ_L, and held-out subjects provide only final tests.

Weighted accuracy (WA) is overall accuracy, and unweighted accuracy (UA) is mean class recall. SD predictions are concatenated across test folds within subject; LOSO produces one score per held-out subject. We average five seeds within subject and report mean $\pm$ standard deviation across 15 subjects. Figure intervals use subject-level units.

Baselines and Recognition Performance

Baselines span DE, graph, transformer, spiking, concept-bottleneck, and EEG-to-LLM models (Song et al., 2020; Zhong et al., 2022; Song et al., 2023; Wang et al., 2025; Koh et al., 2020; Ma et al., 2026). All implementations use DE-adapted inputs and fold-local selection. Direct markers retain learned aggregation; the generic MLP bottleneck learns the same targets. Conformer with source prototypes matches episodic sampling, reliability, weighted cosine classification, temperatures, and support gradients. It uses classification supervision, while SpikeMind also uses concept alignment. The comparison evaluates complete representation-training pipelines.

Table 2 shows higher mean WA and UA for SpikeMind in every setting. Against EEG Conformer with the same source-prototype rule, LOSO WA rose by 4.1 and 4.4 points on SEED and SEED-IV; prototypes alone improved Conformer by 0.3–1.2 points. The matched-head comparison motivates the representation ablations below.

Subject shift still reduced SpikeMind's WA from SD to LOSO by 4.9 points on SEED and 7.3 on SEED-IV. WA standard deviation rose by 0.5 points on both datasets, while WA–UA gaps stayed within 0.4–0.7 points. Cross-subject variability remained larger for four-class SEED-IV.

Concept Grounding and Ablation Analysis

Table 3(a) compares spiking perception, readout, temporal weighting, prototype decision, reliability, and alignment on SEED LOSO. Direct markers retain the aggregator and prototype head; Dense ANN retains the bridge; other variants replace one mechanism.

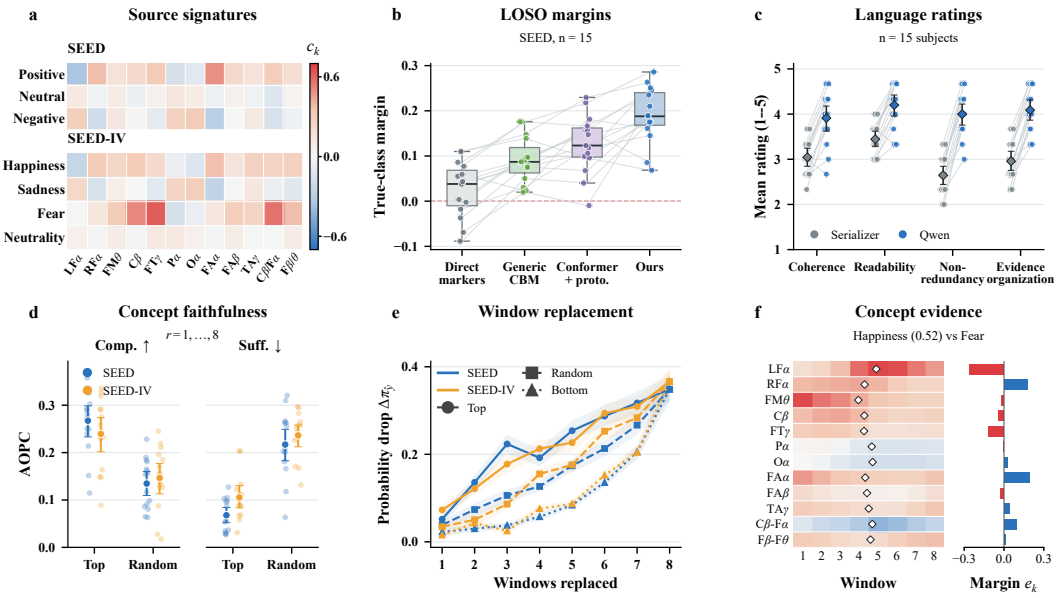

Figure 3: Cross-subject concept and language evidence. (a) Source signatures; (b) SEED LOSO true-class margins; (c) SEED LOSO blinded language ratings; (d) concept AOPC; (e) temporal replacement; (f) trajectories, weighted temporal centers, and normalized relative evidence e_k. The probability and runner-up label are classifier context. Intervals in (c–e) show 95% subject-level confidence intervals.

Grounding compares $v_{\ell,k}$ with held-out $\widetilde{m}_{\ell,k}$. Fisher-transformed test-window correlations are averaged over concepts, seeds, and subjects before inversion; MAE and Huber follow the same hierarchy. Correlation fell from 0.842 to 0.793 on SEED and from 0.812 to 0.748 on SEED-IV under subject shift (Table 3(b)). The shared dictionary fixes readout identities, while these measurements quantify held-out agreement with the standardized DE-derived targets. The readouts operate within the encoder's contextual representation.

Exact marker reproduction did not yield the strongest recognition. Direct markers reached correlation 1.000 by construction but trailed the full model by 14.2 WA points and 14.7 UA points. The unconstrained readout reduced WA by 3.3 points and correlation from 0.793 to 0.612. Removing alignment reduced WA by 2.5 points and correlation by 0.530. Together, these comparisons support the recognition utility of structured readouts and marker alignment in the evaluated system.

Temporal selection contributed beyond window-level marker agreement. Mean pooling changed correlation by only 0.004, yet reduced WA by 1.8 points and UA by 2.0 points. Dense encoding, a linear concept head, and a direct encoder head reduced WA by 1.4, 2.2, and 2.9 points, respectively. Uniform reliability incurred the smallest loss, 1.1 points, indicating an incremental benefit from stability weighting alongside subject-balanced prototypes.

Figure 3(a) summarizes source-class differences in learned, standardized DE-derived coordinates. On SEED LOSO, the true-class margin was 0.190, compared with 0.126 for Conformer plus prototypes, 0.091 for the generic bottleneck, and 0.027 for direct markers (panel (b)). This quantity, $s_y(c) - \max_{j \neq y} s_j(c)$ for ground-truth class y, measures class separation and can be negative for misclassified examples. Decision evidence e instead compares the predicted class with its runner-up.

Operation accounting. The spike-event ledger contains 5.62M SynOps and 0.58M MACs at firing rate 0.142. Continuous fusion projection W_o adds 40.63M MACs at 62 packed locations or 53.08M on the dense 9×9 grid. These listed components total 41.21M or 53.66M MACs, yielding arithmetic energy subtotals of 194.6 or 251.9 μJ at 0.9 pJ/SynOp and 4.6 pJ/MAC. Conformer's 38.60M MAC subtotal yields 177.56 μJ (Horowitz, 2014; Yao et al., 2023). The supplement specifies operator coverage, grid execution, preprocessing, and language costs.

Classifier Faithfulness

Classifier interventions test whether highly ranked concepts and windows influence the output under a common replacement operation (DeYoung et al., 2020; Zhao & Aletras, 2023; Li et al., 2024). Concepts S_r are the top-r coordinates ranked by the original $|e_k|$. The vector $c_{\setminus S_r}$ replaces selected concepts by their fold-training means, whereas c_{S_r} replaces the remainder. The prototype head stays fixed, and similarities and probabilities are recomputed. For the original predicted class $\hat{y}$,

$$\begin{aligned} \mathrm{Comp}(S_r) &= \pi_{\hat{y}}(c) - \pi_{\hat{y}}(c_{\setminus S_r}), \\ \mathrm{Suff}(S_r) &= \pi_{\hat{y}}(c) - \pi_{\hat{y}}(c_{S_r}). \end{aligned} \tag{8}$$

Higher comprehensiveness and lower sufficiency favor the selected evidence. AOPC over $r = 1, \ldots, 8$ uses 100 size-matched random controls. Thus, margin evidence supplies the ranking, while the intervention metric measures changes in the original class probability.

Temporal replacement ranks the eight original windows by $\omega_\ell = \sum_{k=1}^{12} |e_k| \alpha_{\ell,k}$. Top, bottom, and 100 random orders use the same operation for $m = 1, \ldots, 8$ replacements. Selected standardized DE frames are replaced by the fold-training electrode–band mean, zero after standardization, without shortening the example. The frozen perception path recomputes trajectories, temporal weights, concepts, and probabilities. Drops refer to the original $\hat{y}$ and include negative changes. We average examples and seeds within subject and report 95% confidence intervals.

Top concepts increased comprehensiveness AOPC over random controls by 0.133 on SEED and 0.093 on SEED-IV, while reducing sufficiency by 0.149 and 0.131. Three top-window replacements produced confidence drops of 0.178–0.224, compared with 0.086–0.106 for random and 0.025–0.037 for bottom orders. These differences show greater classifier sensitivity to highly ranked evidence under the specified replacement operation (Figure 3(d,e)).

Evidence Retention and Language Organization

Language evaluation asks whether generated rationales preserve the planned evidence and organize it into readable accounts. Retention compares parsed rationales with all 12 items in $\mathcal{P}_{\hat{y}}$ over predicted category, concept name, direction, interval, and role. Roles refer to predicted-versus-runner-up margin evidence, using fixed training-only 0.25-quantile thresholds. Precision, recall, and F1 are averaged within subject, then across seeds and subjects.

Retention F1 fell from 0.894 to 0.867 on SEED and from 0.872 to 0.832 on SEED-IV under subject shift (Table 3(b)). Precision exceeded recall by 0.046–0.076 across settings. Correctly expressed claims therefore coexisted with incomplete coverage of the reference plan. These scores measure preservation of planned facts, separately from the classifier sensitivities measured by intervention.

For SEED LOSO, three blinded raters assess 20 paired records per subject on integer 1–5 scales for coherence, readability, non-redundancy, and evidence organization. Qwen has higher subject-averaged ratings on all four criteria (Figure 3(c)). These scores describe complete outputs: Qwen's recall of 0.839 makes factual coverage relevant to the non-redundancy comparison. Retention and ratings jointly assess coverage and language organization.

Figure 3(f) pairs all 12 trajectories with weighted temporal centers and normalized relative evidence. The trajectories show within-example variation; Qwen organizes their values, locations, and evidence from 12 continuous tokens and a predicted-class embedding. Temporal centers average all eight window locations with magnitude-based weights.

5 Conclusion

SpikeMind connects EEG recognition and rationale generation through 12 DE-grounded temporal concepts. Reliability-weighted source prototypes reach LOSO accuracies of 89.1% on SEED and 79.4% on SEED-IV. Learned concepts and temporal weighting improve recognition over deterministic markers. Concept-marker correlations of 0.748–0.842 assess alignment, while interventions assess decision sensitivity. Qwen reaches retention F1 of 0.832–0.894 and higher mean language ratings than the serializer; recall quantifies evidence coverage. The shared record makes alignment, decision evidence, and language retention separately measurable. The dictionary covers standardized regional activity, asymmetry, and cross-band contrasts. Further work includes regional input specificity, composite consistency, connectivity, unseen montages, clinical interpretation, and complete perception-and-language cost measurement.

AI USE STATEMENT

AI tools assisted language editing, consistency checks, and protocol specification. The authors are responsible for the implementation, results, references, and scientific claims.

REPRODUCIBILITY STATEMENT

The supplement specifies feature selection, grouped partitions, normalization, episodic sampling, optimization, statistical units, interventions, and rationale evaluation. Configuration manifests link these procedures to checkpoints, subject-level outputs, language records, and operation counts.

REFERENCES

Soraia M. Alarcão and Manuel J. Fonseca. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2019. doi: 10.1109/TAFFC.2017.2714671.

Marco Bilucaglia, Gian Marco Duma, Giovanni Mento, Luca Semenzato, and Patrizio E. Tressoldi. Applying machine learning EEG signal classification to emotion-related brain anticipatory activity. *F1000Research*, 9:173, 2020. doi: 10.12688/f1000research.22202.3. Version 3.

Richard J. Davidson. Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1):125–151, 1992. doi: 10.1016/0278-2626(92)90065-T.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2661–2671, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Fang_Incorporating_Learnable_Membrane_Time_Constant_To_Enhance_Learning_of_Spiking_ICCV_2021_paper.html.

Hangzhou He, Lei Zhu, Xinliang Zhang, Shuang Zeng, Qian Chen, and Yanye Lu. V2C-CBM: Building concept bottlenecks with vision-to-concept tokenizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3401–3409, 2025. doi: 10.1609/aaai.v39i3.32352. URL https://ojs.aaai.org/index.php/AAAI/article/view/32352.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp. 10–14. IEEE, 2014. doi: 10.1109/ISSCC.2014.6757323.

Hao-Yi Hu, Li-Ming Zhao, Yu-Zhong Liu, Hua-Liang Li, and Bao-Liang Lu. A novel experiment setting for cross-subject emotion recognition. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6416–6419. IEEE, 2021. doi: 10.1109/EMBC46164.2021.9630314.

Qihan Huang, Jie Song, Jingwen Hu, Haofei Zhang, Yong Wang, and Mingli Song. On the concept trustworthiness in concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21161–21168, 2024. doi: 10.1609/aaai.v38i19.30109. URL https://ojs.aaai.org/index.php/AAAI/article/view/30109.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*,

volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3490–3497, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11596.

Meng Li, Haoran Jin, Ruixuan Huang, Zhihao Xu, Defu Lian, Zijia Lin, Di Zhang, and Xiting Wang. Evaluating readability and faithfulness of concept-based explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 607–625. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.36. URL https://aclanthology.org/2024.emnlp-main.36/.

Fei Ma, Han Lin, Yifan Xie, Hongwei Ren, Xiaoyu Shen, Wenbo Ding, and Qi Tian. E^2-LLM: Bridging neural signals and interpretable affective analysis, 2026. URL https://arxiv.org/abs/2601.07877.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019. doi: 10.1038/s41586-019-1677-2.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31504–31520. PMLR, 2023. URL https://proceedings.mlr.press/v202/shin23a.html.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2020. doi: 10.1109/TAFFC.2018.2817622.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Jiaqi Wang, Zhengyu Ma, Xiongri Shen, Chenlin Zhou, Leilei Zhao, Han Zhang, Yi Zhong, Siqi Cai, Zhenxi Song, and Zhiguo Zhang. S^2M-Former: Spiking symmetric mixing branchformer for brain auditory attention detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b8330f5b70b3c53172417deac6f057b1-Abstract-Conference.html.

Yiming Wang, Bin Zhang, and Yujiao Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 628–636, 2024. doi: 10.1609/aaai.v38i1.27819. URL https://ojs.aaai.org/index.php/AAAI/article/view/27819.

Shin'ya Yamaguchi and Kosuke Nishida. Explanation bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21886–21894, 2025. doi: 10.1609/aaai.v39i20.35495. URL https://ojs.aaai.org/index.php/AAAI/article/view/35495.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ca0f5358dbadda74b3049711887e9ead-Abstract-Conference.html.

Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 863–870, 2021. doi: 10.1609/aaai.v35i1.16169. URL https://ojs.aaai.org/index.php/AAAI/article/view/16169.

Zhixue Zhao and Nikolaos Aletras. Incorporating attribution importance for improving faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 4732–4745. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.261. URL https://aclanthology.org/2023.acl-long.261/.

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.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019a. doi: 10.1109/TCYB.2018.2797176.

Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 10(3):417–429, 2019b. doi: 10.1109/TAFFC.2017.2712143.

Peixiang Zhong, Di Wang, and Chunyan Miao. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2022. doi: 10.1109/TAFFC.2020.2994159.