BiMamba-TTA

Lado Turmanidze

February 3, 2026

Jia, Z., Du, T., Tian, Z., Li, H., Zhang, Y., & Liu, C. (2025).
"A Multimodal BiMamba Network with Test-Time Adaptation for Emotion Recognition Based on Physiological Signals".
In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*.

Paper Overview & Problem Formulation

The goal of this paper is to recognize emotions (valence and arousal) from multimodal physiological signals (EEG, EOG, EMG, ECG, GSR) using deep learning, while remaining robust to missing sensor data at test time. This is cast as a **multi-class classification** problem over time-series data. The input is a collection of M modality signals, each recorded over L time steps across C_i channels:

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}, \qquad \mathbf{x}_i \in \mathbb{R}^{C_i \times L}, \quad i \in \{1, \ldots, M\}$$

The ground-truth label y is converted to a one-hot vector $\mathbf{p}(y|\mathbf{x}) \in \mathbb{R}^N$, where N is the number of emotion classes. The model outputs a predicted probability distribution over classes:

$$\hat{\mathbf{y}} = p(\hat{y} \mid \mathbf{x}) \in \mathbb{R}^N$$

Two core challenges remain:

1. How to model long-range temporal dependencies *within* each modality and correlations *across* modalities simultaneously
2. How to adapt at test time when one or more modalities go missing, amplifying distribution shift.

Background: State Space Models (SSMs)

A **state space model** maps an input sequence $x(t)$ to an output $y(t)$ through a latent hidden state $\mathbf{h}(t)$. The state evolves according to learned dynamics and input matrices $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$ and the output is read via output and feedthrough matrices $C \in \mathbb{R}^{1\times N}$, $D \in \mathbb{R}^{N\times 1}$.
Continuous-time form (ODE):

$$\frac{d\mathbf{h}(t)}{dt} = A\,\mathbf{h}(t) + B\,x(t)$$
$$y(t) = C\,\mathbf{h}(t) + D\,x(t)$$

Discrete-time form (recurrence): Discretized with step size Δ (e.g., via Bilinear Transform) to allow computation on digital hardware:

$$\mathbf{h}_t = \bar{A}\,\mathbf{h}_{t-1} + \bar{B}\,x_t$$
$$y_t = C\,\mathbf{h}_t$$

where the discretized matrices are:

$$\bar{A}_\Delta = (I - \tfrac{\Delta}{2}A)^{-1}(I + \tfrac{\Delta}{2}A), \quad \bar{B}_\Delta = (I - \tfrac{\Delta}{2}A)^{-1}\Delta B$$

Note: *D is omitted from the discrete-time form as the selective mechanism and hidden state dynamics are sufficient for the modeling task.*

Mamba: The Selective State Space Model

In a vanilla SSM, A, B, C are time-invariant parameters (same for all time steps). Mamba makes B, C, and the step size Δ **input-dependent**, so the model learns to selectively gate information at every time step.

Step 1 - Input-dependent parameters: separate linear layers project x_t into dynamic B_t, C_t, Δ_t:

$$B_t = \text{Linear}_B(x_t), \quad C_t = \text{Linear}_C(x_t), \quad \Delta_t = \text{Softplus}(\text{Linear}_\Delta(x_t))$$

where $\text{Softplus}(x) = \ln(1 + e^x)$.
Step 2 - Gated state update: Δ_t controls the update magnitude (large Δ_t = "focus here"; small Δ_t = "skip"):

$$\mathbf{h}_t = \bar{A}_{\Delta_t} \mathbf{h}_{t-1} + \bar{B}_{\Delta_t} x_t$$

Step 3 - Output readout: the hidden state is projected to the output via the dynamic matrix C_t:

$$y_t = C_t \, \mathbf{h}_t$$

BiMamba-TTA: Intra-Modal BiMamba Module

Vanilla Mamba is autoregressive (past only). Emotion evolves in *both* temporal directions, so we use **bidirectional** state-space modeling. For each modality i, after an initial encoder produces h_i:

1. Gating: a SiLU-activated ($\sigma(x) = \frac{x}{1+e^{-x}}$) gate suppresses noise and highlights emotion-relevant features:

$g_i = \sigma\left(W_i^g\, h_i + b_i^g\right)$, where W_i^g, b_i^g are learned weight matrix and bias.

2. Forward and backward SSM passes (Rev_t flips along time, $\odot$ is element-wise product):

$$h_i^{\rightarrow} = g_i \odot \mathrm{SSM}^{\rightarrow}\Big(\sigma\big(\mathrm{Conv1D}^{\rightarrow}\big(W_i^h h_i + b_i^h\big)\big)\Big)$$

$$h_i^{\leftarrow} = g_i \odot \mathrm{SSM}^{\leftarrow}\Big(\sigma\big(\mathrm{Conv1D}^{\leftarrow}\big(\mathrm{Rev}_t\big(W_i^h h_i + b_i^h\big)\big)\big)\Big)$$

3. Merge + residual: The output u_i is the final intra-modal feature. The **intra-modal** module processes each signal independently to capture long-range temporal dependencies within that modality:

$$u_i = h_i + W_i^o\left(\frac{h_i^{\rightarrow} + \mathrm{Rev}_t(h_i^{\leftarrow})}{2}\right) + b_i^o$$

BiMamba-TTA: Inter-Modal BiMamba Module

Different modalities are physiologically linked (e.g. EEG arousal correlates with GSR conductance peaks and ECG heart-rate variability drops). The inter-modal module captures these **high-order cross-modal correlations** via shared hidden states.

1. Concatenate & transpose: stack all u_i along the channel axis and swap time and channel dimensions so that BiMamba sweeps *across modalities*:

$$m = \text{Transpose}\big(u_1 \circ u_2 \circ \cdots \circ u_M\big) \in \mathbb{R}^{\sum_{i=1}^{M} C_i' \times L'}$$

where $\circ$ denotes channel-wise concatenation and C_i' is the output channel count of modality i.

2. BiMamba along the channel dimension:

$$H = \text{BiMamba}(m) \in \mathbb{R}^{\sum_{i=1}^{M} C_i' \times L'}$$

In the **forward** pass hidden states built from modalities $\{1, \ldots, i-1\}$ enrich modality i; in the **backward** pass, modalities $\{M, \ldots, i+1\}$ do so. Every modality enriches every other.

Auxiliary Task & Training Objective

Each intra-modal BiMamba branch also has its own classifier to:

1. balances training across modalities
2. prevents single-modality overfitting
3. produces the unimodal entropies $\mathrm{Ent}(x_i)$ that are later needed for TTA filtering.

Unimodal prediction (per modality i, N = number of classes):

$$p(\hat{y}_i \mid x_i) = \mathrm{softmax}\big(W_i\, u_i + b_i\big) \in \mathbb{R}^N$$

Per-modality cross-entropy loss (n = batch size):

$$\mathcal{L}_i = -\frac{1}{n}\sum_{j=1}^{n} p(y \mid x)^{(j)}\ \log\ p(\hat{y}_i \mid x_i)^{(j)}$$

Total training objective ($\mathcal{L}_{\text{task}}$ = main multimodal cross-entropy, α_i = per-modality auxiliary weight):

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{task}} + \sum_{i=1}^{M} \alpha_i\, \mathcal{L}_i$$

TTA Step 1: Two-Level Entropy-Based Sample Filtering

Missing sensors amplify distribution shift unevenly. We select only samples that are **confident** (low multimodal entropy $\Rightarrow$ close to source domain) and **multimodally rich** (high unimodal entropy $\Rightarrow$ genuinely needs several modalities).

Multimodal and unimodal entropies (N = number of classes):

$$\mathrm{Ent}(x) = -\sum_{c=1}^{N} p(\hat{y}{=}c \mid x) \log\, p(\hat{y}{=}c \mid x)$$

Adaptive thresholds (smoothing factor $\beta_t = \beta_{t-1} + \frac{t}{\#\mathrm{iter}}(1-\beta_{t-1})$, β=0.2, $\#\mathrm{iter}$=7):

$$\gamma_m = \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \mathrm{Ent}(x)^{(j)} + \gamma'_m \cdot \beta_t$$

$$\gamma_u = \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \sum_{i=1}^{M} \mu_i\, \mathrm{Ent}(x_i)^{(j)} - \gamma'_u \cdot \beta_t$$

Selection criterion (μ_i = unimodal weight for modality i):

$$S(x) = \left\{ x \;\middle|\; \mathrm{Ent}(x) \le \gamma_m \;\text{ and }\; \textstyle\sum_{i=1}^{M} \mu_i\, \mathrm{Ent}(x_i) \ge \gamma_u \right\}$$

TTA Step 2: Mutual Information Sharing Across Modalities

When a modality is corrupted its prediction degrades. Intact modalities still carry useful signal. We align all modality predictions so that informative ones guide the corrupted ones, without letting a bad modality drag down the good ones.

Complementary probability of modality i (average prediction of all *other* modalities):

$$p'(\hat{y}_i \mid x_i) = \frac{\sum_{j=1}^{M} p(\hat{y}_j \mid x_j) - p(\hat{y}_i \mid x_i)}{M-1}$$

Mutual-information-sharing loss: minimize KL divergence between each modality's prediction and a *stabilized* target - the average of its complementary probability and the full multimodal prediction $p(\hat{y} \mid x)$. Including $p(\hat{y} \mid x)$ prevents a severely corrupted modality from pulling intact ones off course:

$$\mathcal{L}_{\text{mis}}(x) = \sum_{i=1}^{M} D_{\text{KL}}\Big(p(\hat{y}_i \mid x_i) \,\Big\|\, \tfrac{1}{2}\Big(p'(\hat{y}_i \mid x_i) + p(\hat{y} \mid x) \Big) \Big)$$

TTA: Final Loss & Optimization

The two TTA components are combined into a single loss, weighted by sample confidence and gated by the two-level filter.

Sample confidence weight (Ent_0 is predefined normalisation factor):

$$\alpha(x) = \frac{1}{\exp\left(\mathrm{Ent}(x) - \mathrm{Ent}_0\right)}$$

Total TTA loss:

$$\mathcal{L}_{\text{test}}(x) = \alpha(x)\, \mathbf{1}_{\{x \in S(x)\}} \left(\mathrm{Ent}(x) + \lambda\, \mathcal{L}_{\text{mis}}(x)\right)$$

Optimization: only a small subset $\hat{\Theta} \subset \Theta$ of parameters is updated ("surgical fine-tuning" - first conv layer per encoder, first FC layer of the inter-modal module, all batch-norm layers):

$$\min_{\hat{\Theta} \in \Theta} \; \mathcal{L}_{\text{test}}(x)$$

This keeps the core model stable while adapting to the target domain.

Experimental Results

Without missing data - Accuracy

Method	DEAP Val	DEAP Aro	MAHNOB Val	MAHNOB Aro
TSception	0.613	0.635	0.633	0.599
LGGNet	0.618	0.636	0.632	0.609
VSGT	0.631	0.628	0.613	0.599
MambaFormer	0.621	0.587	0.588	0.619
BiM-TTA	**0.673**	**0.641**	**0.650**	**0.635**

With missing data - Avg. improvement (%)

Method	DEAP Val	DEAP Aro	MAHNOB Val	MAHNOB Aro
Tent	−0.162	0.101	−0.043	0.086
EATA	−0.061	0.134	0.171	0.217
READ	0.781	0.372	0.058	−0.146
2LTTA	0.858	0.429	0.124	0.017
BiM-TTA	**1.172**	**1.309**	**1.089**	**0.506**

Note: *Val - valence, Aro - arousal; DEAP and MAHNOB are datasets.*

How can we use BiMamba-TTA?

Our Task (possibly): Personalized backchannel detection from multi-person group interactions with test-time adaptation to individual participants.

BiMamba-TTA Architecture Adaptation:

- **Intra-modal BiMamba:** Model temporal evolution within video (e.g., prolonged gaze $\Rightarrow$ nod) and audio (e.g., prosodic buildup $\Rightarrow$ "mhm") independently
- **Inter-modal BiMamba:** Capture cross-modal synchrony (head nods aligned with vocal backchannels, facial affect with para-verbal cues)
- **Test-Time Personalization:** Train on 77 participants, adapt to 1 target person:
  - Filter samples with confident predictions (low multimodal entropy) and rich multimodal cues (high unimodal entropy)
  - Handle missing/noisy modalities (occluded face, overlapping speech) via mutual information sharing
  - Surgical fine-tuning: only batch-norm + first encoder layers $\Rightarrow$ preserve general backchannel patterns while adapting to individual expression style

Key Advantage: Unsupervised personalization - no labels needed from target person.