An Efficient Multi-view Transformer For Emotion Recognition Of University Students

Hong Xin[1*] and Lanqiang Cong[2]

[1] Weifang Vocational College, Weifang 262737 China
[2] Shandong Vocational College of Information Technology, Weifang 261061 China
[*] Corresponding author. E-mail: hongxinwvc@163.com

Received: Feb. 07, 2026; Accepted: Apr. 01, 2026

Accurate emotion recognition for university students is essential for mental health monitoring and engagement analysis, yet the intensive computational cost of standard Vision Transformers hinders their deployment on resource-constrained edge devices. To address this challenge, we propose an Efficient Multi-view Transformer (EFFormer) designed for real-time affective computing in campus environments. EFFormer first employs a Bidirectional Mamba strategy to synthesize unified affective representations from multiple views, effectively capturing complex cross-view correlations with linear complexity. Furthermore, we introduce an instance-specific adaptive gating mechanism that dynamically executes patch pruning, attention head activation, and transformer block skipping based on the complexity of each input sample. By jointly optimizing the backbone with a resource-aware loss function and utilizing Gumbel-Softmax reparameterization, EFFormer achieves a superior trade-off between recognition accuracy and inference efficiency. Experimental results demonstrate that our framework significantly reduces computational overhead and latency while maintaining high-fidelity emotional state recognition, providing a practical and robust solution for intelligent emotion monitoring in university settings.

Keywords: Multi-view Transformer; emotion recognition; efficient and effective inference
© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202609_32.020

1. Introduction

Emotion recognition has become an essential component in modern educational technology, particularly for monitoring the mental health and learning engagement of university students [1–5]. As academic environments become increasingly digital and complex, the ability to accurately perceive students' affective states, such as stress, boredom, or engagement, can provide invaluable feedback for personalized pedagogical interventions and psychological support. Traditionally, this task has relied on single-modality observations; however, human emotion is inherently multifaceted, often requiring the integration of heterogeneous data from various perspectives to ensure robust recognition in diverse campus scenarios.

In recent years, Vision Transformers have demonstrated remarkable performance in computer vision tasks by exploiting self-attention mechanisms to model long-range dependencies within images [6, 7]. Following the success of the original ViT, several variants have been adapted for affective computing, utilizing stacked transformer blocks to extract discriminative features from multi-view inputs [8–10]. These multi-view transformer architectures allow for a more comprehensive representation of a student's state by fusing information from different sensors or camera views, significantly outperforming traditional convolutional neural networks in terms of accuracy and robustness [11–13].

Despite these advancements, a critical challenge remains: the intensive computational cost of vision transformers. The complexity of these models scales drasti-

cally as the number of patches, attention heads, and transformer blocks increases. As argued in recent research, "one-size-fits-all" networks are often computationally redundant because the complexity required to model an emotional state varies significantly between samples. For university applications, which often demand real-time inference on resource-constrained edge devices (e.g., campus smart cameras or mobile terminals), this lack of efficiency hinders practical deployment and limits the scalability of affective monitoring systems.

To address these limitations, we propose an Efficient Multi-view Transformer (EFFormer) specifically tailored for student emotion recognition under resource constraints. Our framework introduces an adaptive inference strategy that learns to derive instance-specific usage policies on the fly. Specifically, EFFormer utilizes a bidirectional Mamba (Bi-Mamba) module to synthesize a unified affective representation from multiple views, capturing intricate cross-view correlations. This fused representation then guides a light-weight gating mechanism to dynamically determine essential affective patches, active attention heads, and redundant transformer blocks for skipping. By jointly optimizing the backbone with a resource-aware loss function, EFFormer achieves a superior trade-off between recognition accuracy and inference efficiency, making it highly suitable for real-time affective monitoring in university settings.

The main contributions are summarized as follows:

- We introduce a bidirectional Mamba module to effectively synthesize heterogeneous cues from multiple student perspectives. Unlike traditional fusion methods, the Bi-Mamba strategy exploits selective state spaces to capture complex inter-view correlations with linear complexity, significantly improving the scalability of multi-view affective feature integration.
- We propose a fine-grained adaptive inference framework that operates across three dimensions: patches, attention heads, and transformer blocks. By deriving instance-specific usage policies, EFFormer dynamically prunes non-informative affective patches, deactivates redundant heads, and skips entire transformer layers on the fly, ensuring optimal resource allocation tailored to the difficulty of each input sample.
- Extensive experiments conducted on real-world datasets show EFFormer achieves the state-of-the-art performance in the emotion recognition of university students.

The remainder of this paper is organized as follows. Section 2 details the methodology of the proposed method. Section 3 presents the experimental setup, provides a comparative analysis with existing state-of-the-art methods on the FPR-3 and FPR-4 datasets, and includes a comprehensive ablation study to validate the contribution of each architectural component. Finally, Section 4 concludes the paper and discusses potential directions for future research in intelligent student monitoring.

2. Methodology

The objective of this study is to accurately recognize the emotional states of university students by leveraging heterogeneous data from multiple views under resource constraints. Given multi-view data of the i-th student $x_i = \{x_i^1, x_i^2, \ldots, x_i^V\}$, where V is the view number, it aims to find an optimal mapping function $\mathcal{F}$ that predicts the emotional state y_i from the multi-view space:

$$y_i = \mathcal{F}(\{x_i^1, x_i^2, \ldots, x_i^V\}; \Theta) \tag{1}$$

where Θ is the learnable parameters.

To achieve the objective above, an efficient multi-view transformer (EFFormer) is proposed for the emotion recognition of university students, which contains a view-specific feature extraction module, an adaptive gate selection module, and an emotion recognition module. The architecture of EFFormer is depicted in Fig. 1

2.1. View-specific feature extraction

In EFFormer, we select T2T-ViT [14] that partitions each image into a sequence of sliced patches, as the view-specific feature extractor, which captures long-range dependencies within patches from stacking multi-head self-attention and feed-forward networks to learn discriminative representations of images. Formally, given the v-th view data x_i^v of the i-th student, EFFormer initiates processing by decomposing the input into a sequence of N distinct, fixed-size patches, denoted as $x_i = [s_1^v, s_2^v, \ldots, s_N^v]$, and then leverages the linear transformation to extract view-specific patch representations $z_i = [z_1^v, z_2^v, \ldots, z_N^v]$. Meanwhile, a trainable class token z_{cls}^v is integrated with positional representations z_{pos}^v into these patch representations as the input of the transformer block:

$$h_i^v = [z_{cls}^v; z_1^v; z_2^v; \ldots; z_N^v] + z_{pos}^v \tag{2}$$

where $[\cdot]$ is the concatenation operation.

The l transformer block in the vision transformers contains a multi-head self-attention (MSA) network and a feed-forward network (FFN), where each MSA aggregates information from M parallel heads via discerning subtle emotional nuances across diverse representation subspaces:

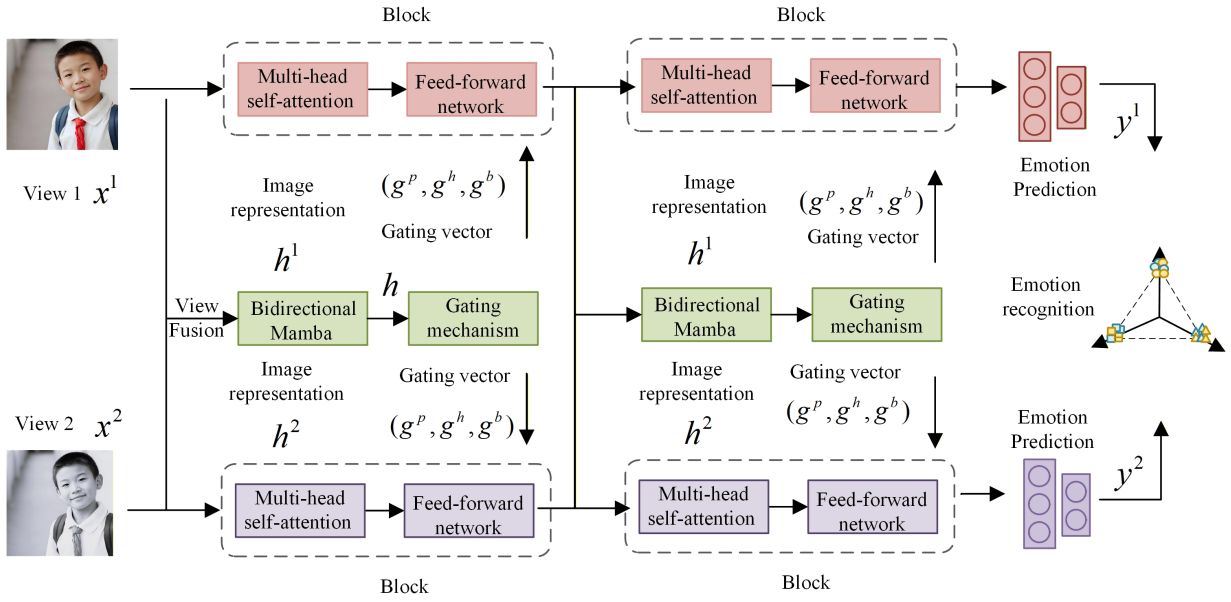


Fig. 1. The illustration of EFFormer.

$$\text{head}_{m,l}^{v} = \text{Attention}(h_i^v W_{m,l}^Q, h_i^v W_{m,l}^K, h_i^v W_{m,l}^V) \quad (3)$$

$$\text{MSA}(h_i^v) = [(\text{head}_{1,l}^{v}, \ldots, \text{head}_{M,l}^{v})] W_l^O \quad (4)$$

where $W_{m,l}^Q$, $W_{m,l}^K$, $W_{m,l}^V$, and W_l^O denote the learnable weight parameters of the m-th head in the l-th block. To refine the extracted features, the MSA output is processed by an FFN, typically structured as a dual-layer perceptron. Residual connections are strategically employed to preserve gradient integrity and mitigate feature degradation throughout the deep hierarchy:

$$\bar{h}_{i,l}^{v} = \text{MSA}(h_{i,l}^{v}) + h_{i,l}, \quad h_{i,l+1}^{v} = \text{FFN}(\bar{h}_{i,l}^{v}) + \bar{h}_{i,l}^{v} \quad (5)$$

2.2. Adaptive gate selection

Despite the impressive performance of large-scale vision transformers in emotion recognition, their computational demands escalate significantly with the expansion of patches, attention heads, and depth. For student emotion monitoring—often requiring real-time processing on edge devices—a static, "one-size-fits-all" backbone is computationally inefficient, as many emotional expressions can be identified through a fraction of the total parameters. To address this, EFFormer introduces an adaptive inference strategy. Specifically, at each transformer block, we first synthesize a unified affective representation by fusing outputs from multiple views using a bidirectional Mamba module. This fused representation then guides a lightweight gating mechanism to dynamically determine: 1) essential affective patches, 2) active attention heads, and 3) redundant transformer blocks.

Multi-view fusion via bidirectional Mamba. To effectively integrate heterogeneous cues from various student perspectives, we employ a bidirectional Mamba (Bi-Mamba) strategy. For the l-th block, let $h_{i,l}^v$ denote the v-th view representation. The Bi-Mamba strategy processes these multi-view sequences through forward and backward selective state spaces to capture cross-view correlations:

$$h_{i,l} = \text{Bi-Mamba}([h_{i,l}^1, \ldots, h_{i,l}^V]) \quad (6)$$

where $h_{i,l+1}$ represents the fusion representation for subsequent gating decisions.

Gating mechanism. Instead of a fixed computation path, a gating network is integrated before each block to derive instance-specific policies. This mechanism consists of gating heads with parameters $W_l = \{W_l^p, W_l^h, W_l^b\}$, which map the fusion representation $h_{i,l}$ into sparse decision vectors for patch selection, head activation, and block skipping. Formally, the gating policies are computed as:

$$\begin{aligned} &(g_l^p, g_l^h, g_l^b) = \sigma(\Phi(W_l, h_{i,l})) \\ &\text{s.t.} \quad g_l^p \in [0,1]^N, \; g_l^h \in [0,1]^M, \; g_l^b \in [0,1] \end{aligned} \quad (7)$$

where $\sigma(\cdot)$ denotes the sigmoid activation and Φ represents the linear gating transformation. Each element in the vectors g_l^p, g_l^h, and g_l^b represents the probability of retaining a specific computation unit.

Patch pruning. We aim to isolate the most salient affective cues while discarding background or non-informative

patches. For the l-th block, a patch is preserved only if its corresponding gate $G^p_{l,n}$ is active, sampled from g^p_l during training via Gumbel-Softmax. The input sequence is modified as:

$$h^v_i = [z^v_{cls}; G^p_{l,1} z^v_1; \ldots; G^p_{l,N} z^v_N] \tag{8}$$

The class token z^v_{cls} is unconditionally retained to preserve the global emotion descriptor.

Head activation. Multi-head attention allows the model to explore various affective subspaces. However, for "easy" samples with obvious emotional cues, fewer heads are required. We adaptively deactivate attention heads based on $G^h_{l,i}$. When a head is gated off, its complex attention map calculation is replaced by an identity mapping to save FLOPs:

$$\mathrm{MSA}(h^v_i) = [(\mathrm{head}^v_{l,m:1\to M}, \text{if } G^h_{l,m} = 1]W^O_l \tag{9}$$

Block Skipping. To further optimize the depth, the gating mechanism independently controls the two primary sub-layers: MSA and FFN. By expanding the block policy g^b_l to two dimensions, the computational flow is updated as:

$$\bar{h}^v_{i,l} = G^b_{l,0}\mathrm{MSA}(h^v_{i,l}) + h_{i,l}, \quad h^v_{i,l+1} = G^b_{l,1}\mathrm{FFN}(\bar{h}^v_{i,l}) + \bar{h}^v_{i,l} \tag{10}$$

This allows the model to bypass redundant layers when the student's emotional state has already been confidently encoded in earlier stages.

2.3. Emotion recognition

To strike an optimal balance between high-fidelity affective state recognition and reduced inference latency, the learning objective of EFFormer is designed to simultaneously maximize classification accuracy and minimize computational redundancy. The framework is optimized through a joint objective function comprising a discriminative classification loss and a resource-aware regularization term.

Specifically, given the v-th view data x^v_i of the i-th student and the corresponding ground-truth emotion label y_i, the cross-entropy loss L_{ce} is utilized to supervise the prediction produced by the transformation $\mathcal{F}$ with parameters Θ:

$$L_{ce} = \sum_{v}^{V}\sum_{i}^{I} CE(y_i(\mathcal{F}(x^v_i;\Theta))) \tag{11}$$

where $CE(\cdot)$ is the cross-entropy function. A significant challenge in training the gating mechanism is the non-differentiable nature of discrete selection decisions. To facilitate end-to-end backpropagation, we employ the Gumbel-Softmax reparameterization. For each decision entry in g, the continuous relaxation $G_{l,k}$ is computed as:

$$G_{l,k} = \frac{\exp((\log(g_{i,k}) + G_{l,k})/\tau)}{\sum_{j=1}^{K}\exp((\log(g_{l,j}) + G_{l,j})/\tau)} \quad \text{for } k \in \{1,\ldots,K\} \tag{12}$$

where τ is the temperature parameter controlling the sparsity of the distribution, and G_l denotes the Gumbel noise sampled via $G_l = -\log(-\log(U_l))$ with $U_l \sim \mathrm{Uniform}(0,1)$. K is the class number.

To enforce the model to operate within the specific computational budgets required for real-time student monitoring, we introduce a resource regularization loss L_{reg}. This term penalizes the deviation of the actual computation usage from the target retention ratios:

$$L_{reg} = \sum_{\phi\in\{p,h,b\}} \left(\frac{1}{D_\phi}\sum_{d=1}^{D_\phi}\mathbf{G}^\phi_d - \gamma_\phi\right)^2 \tag{13}$$

where D_p, D_h, D_b represent the total counts of patch gates, attention head gates, and transformer sub-layer gates across all L blocks, respectively. The hyperparameters $\gamma_p, \gamma_h, \gamma_b \in (0,1]$ serve as pre-defined sparsity constraints that dictate the desired computational intensity.

Ultimately, the total loss function L is minimized to jointly optimize the backbone parameters Θ:

$$L = L_{ce} + \lambda L_{reg} \tag{14}$$

where λ is a balancing coefficient that weights the importance of efficiency relative to recognition performance.

3. Results and discussion

3.1. Set up

Dataset: To validate the performance of EFFormer, we utilize two common datasets in the emotion recognition: FPR-3 and FPR-4. The FPR-3 dataset comprises 1,536 images of size 48×48, categorized into three distinct emotion classes, i.e., angry, happy, and fear. For our experiments, it is partitioned into a training set of 1,024 images and a testing set of 512 images. The training set is distributed with 194 samples for angry, 600 for happy, and 230 for fear. For the test set, the samples are allocated as 100 for angry, 300 for happy, and 112 for fear. The FPR-4 dataset consists of 3,000 images (48×48) across four emotion categories, i.e., angry, happy, fear, and sad, split into 2,000 samples for training and 1,000 for evaluation. Within the training set, the class distribution is perfectly balanced, with each of the four categories containing 500 samples. In contrast, the test set exhibits a degree of imbalance, comprising 180 samples for angry, 160 for happy, 230 for fear, and 430 for sad. Consistent with our multi-view learning paradigm, the second view of the input data is constructed through a comprehensive augmentation pipeline. Specifically, we apply random resized cropping followed by random horizontal flipping to introduce geometric variations. To further diversify the affective features, we implement color jittering, including

Table 1. Comparison results of EFFormer on the FPR-3 dataset.

Method	Swin	LocalViT-S	TMER	Mvgt	TMH	DAF-FER	EFFormer
Accuracy(%)	0.7265	0.7425	0.7564	0.7785	**0.8025**	0.7725	0.8008
Precision(%)	0.5168	0.5467	0.5784	0.5724	**0.6099**	0.5706	0.5997
Recall(%)	0.4274	0.4438	0.4459	0.4647	**0.4971**	0.4796	0.4815
F1-score(%)	0.4377	0.4599	0.4736	0.4829	**0.5177**	0.4912	0.5042
Gflops(G)	4.9845	4.6568	14.984	5.6712	6.8000	5.9948	**4.4340**

Table 2. Comparison results of EFFormer on the FPR-4 dataset.

Method	Swin	LocalViT-S	TMER	Mvgt	TMH	DAF-FER	EFFormer
Accuracy(%)	0.3512	0.3968	0.4154	0.4668	0.4705	**0.4778**	0.4726
Precision(%)	0.2495	0.2531	0.2638	0.2746	0.2799	0.2874	**0.2893**
Recall(%)	0.2657	0.2667	0.2768	0.2866	0.2998	0.3013	**0.3043**
F1-score(%)	0.2573	0.2597	0.2701	0.2805	0.2895	0.2942	**0.2988**
Gflops(G)	4.9845	4.7124	15.568	5.8711	6.9871	6.0125	**4.5134**

stochastic adjustments of brightness, contrast, saturation, and hue, alongside random erasing techniques.

Evaluation metrics: To conduct a rigorous and multi-faceted quantitative assessment of the proposed EFFormer, we employ four widely-adopted classification metrics: accuracy, precision, recall, F1-score, and Gflops (Giga Floating-point Operations per second). Accuracy is defined as the ratio of correctly predicted samples to the total number of evaluation instances, reflecting the overall correctness of the model's emotional state classification. Precision represents the proportion of true positive predictions among all samples identified as a specific emotion, indicating the model's ability to minimize false positive errors. Recall measures the ratio of correctly identified positive instances to the total number of actual positive samples, assessing the model's capacity to capture all relevant emotional cues without omission. The F1-score is the harmonic mean of precision and recall, providing a consolidated metric that balances the trade-off between the two, which is particularly essential for assessing performance on potentially imbalanced datasets. Finally, Gflops quantifies the total number of floating-point operations required for a single inference pass, serving as the primary indicator of the model's computational complexity and its suitability for deployment on resource-constrained edge devices.

Implementation detail: For the backbone of EFFormer, we adopt the T2T-ViT architecture, which is selected for its ability to deliver competitive recognition accuracy on large-scale datasets while maintaining a manageable computational overhead. The structural configuration of the backbone comprises $L = 19$ blocks, where each multi-head self-attention is equipped with $H = 7$ heads. The patch number is set as 196. The learning rate is set as 0.001. The batch size is set as 64. To enable adaptive inference, our proposed gate mechanism is integrated into every transformer block, commencing from the second layer onwards. For the training phase, we initialize the transformer backbone using pre-trained weights sourced from the official T2T-ViT repository to facilitate stable convergence.

3.2. Performance Comparison

Comparison methods: Six emotion recognition methods are compared to verify the performance of EFFormer, which contains Swin [15], LocalViT-S [16], TMER [7], Mvgt [10], TMH [11], and DAF-FER [12].

Comparison analysis: As shown in Table 1 and 2, EFFormer achieves a superior balance between recognition performance and computational efficiency across both datasets. The following observations can be drawn from the experimental results: First, compared to single-view architectures such as Swin and LocalViT-S, EFFormer exhibits a significant performance margin across all primary metrics (Accuracy, Precision, Recall, and F1-score). This advantage underscores the necessity of multi-view feature integration for capturing the nuanced and multi-faceted emotional states of students in complex campus environments. Second, while some heavy-weight multi-view models (e.g., TMH) show competitive accuracy on specific datasets, they often suffer from excessive computational overhead. In contrast, EFFormer maintains a state-of-the-art performance level while operating at a substantially lower computational cost. This makes our framework more suitable for real-time deployment on resource-constrained edge devices. On the more challenging FPR-4 dataset, EFFormer demonstrates even greater robustness, leading most evaluation metrics, which validates its gen-

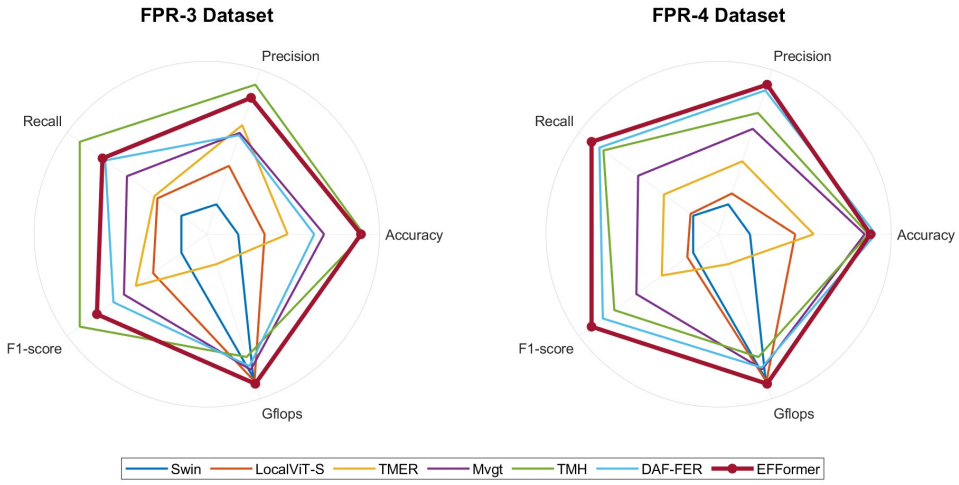


Fig. 2. The comprehensive visual comparison.

eralization capability in higher-dimensional emotion classification tasks. The performance gains and efficiency of EFFormer are mainly attributed to its architectural innovations. The bidirectional Mamba replaces the standard quadratic-complexity attention for multi-view fusion, effectively capturing intricate inter-view correlations with linear complexity. Furthermore, the adaptive gating mechanism enables a sample-dependent inference path, which intelligently prunes redundant patches and skips unnecessary transformer blocks.

The radar charts in Fig. 2 provide a comprehensive visual comparison of these methods. It is evident that EFFormer covers a more balanced and expansive area across the performance dimensions. This multi-dimensional visualization further confirms that our proposed framework successfully harmonizes the conflict between high-fidelity recognition and inference parsimony, providing an optimal solution for intelligent student monitoring.

3.3. Ablation Analysis

To evaluate the individual contribution and the synergistic effect of each proposed component in EFFormer, we conduct a comprehensive ablation study on the FPR-4 dataset by designing five model variants to isolate the effects of the Bidirectional Mamba (Bi-Mamba) module, the Adaptive Gating mechanism, and the joint loss functions: (1) Baseline, representing a standard multi-view transformer utilizing only cross-entropy loss without efficiency optimization; (2) EFFormer w/o Bi-Mamba, which replaces the Bi-Mamba fusion with standard feature concatenation; (3) EFFormer w/o Gating, employing a fixed inference path without dynamic pruning or skipping; (4) EFFormer w/o L_{reg}, optimized without the resource-aware regularization loss; and (5) EFFormer, integrating all components and the joint optimization strategy.

As summarized in Table 3, several key insights can be drawn from the results: first, the comparison between the Baseline and the "w/o Bi-Mamba" variant underscores the necessity of multi-view fusion, while the performance decline observed when replacing Bi-Mamba with simple concatenation proves its superior ability to capture complex inter-view correlations for more discriminative representation learning. Simultaneously, the decrease in performance for the "w/o Gating" variant suggests that the adaptive gating mechanism does not only reduce computational redundancy but also functions as a learned attention filter that focuses the model on essential affective cues while mitigating the impact of non-informative background noise. Furthermore, the fact that the model without L_{reg} performs inferiorly to the full model demonstrates that L_{reg} serves as more than just an efficiency constraint; by enforcing structural sparsity through Gumbel-Softmax reparameterization, it regularizes the decision-making process and enhances the model's overall generalization capability. Ultimately, this analysis confirms that the synergy of Bi-Mamba for effective feature synthesis, Adaptive Gating for inference path optimization, and the Joint Loss for balancing accuracy and sparsity is essential for the superior performance of EFFormer in student emotion recognition.

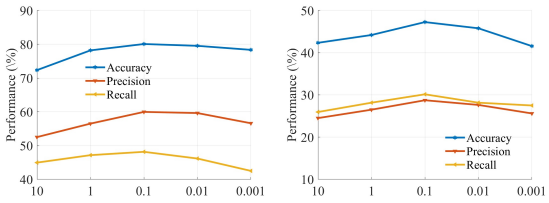
3.4. Parameter Analysis

Following [17, 18], we investigate the sensitivity of EFFormer to the balancing coefficient λ for the joint loss function (Eq. 14), which governs the trade-off between recog-

Table 3. Ablation study of different components in EFFormer on the FPR-4 dataset.

Model	Bi-Mamba	Gating	L_{reg}	L_{ce}	Accuracy(%)
Baseline				✓	0.4012
EFFormer w/o Bi-Mamba		✓	✓	✓	0.4215
EFFormer w/o Gating	✓		✓	✓	0.4658
EFFormer w/o L_{reg}	✓	✓		✓	0.4658
EFFormer	✓	✓	✓	✓	**0.4726**

nition accuracy and computational efficiency. We conduct experiments on both the FPR-3 and FPR-4 datasets by varying λ within the range $\{10, 1, 0.1, 0.01, 0.001\}$. The resulting curves for Accuracy, Precision, and Recall are illustrated in Fig. 3.



(a) FPR-3 **(b)** FPR-4

Fig. 3. Parameter analysis of λ on the two dataset.

As observed in Fig. 3(a) and 3(b), the model performance is highly sensitive to the choice of λ. When λ is set to a high value (e.g., 10), there is a noticeable decline in all evaluation metrics. This is because an excessive penalty on resource usage forces the gating mechanism to be overly sparse, leading to the loss of critical affective features. As λ decreases to 0.1, the performance reaches its peak across both datasets. This "sweet spot" indicates that a moderate regularization constraint allows the adaptive gating mechanism to effectively filter out non-informative background noise while preserving the essential emotional cues. When λ is further reduced below 0.1, the performance tends to plateau or slightly decrease, as the model reverts toward a dense architecture that lacks the beneficial "feature denoising" effect of the sparse gating strategy. Consequently, $\lambda = 0.1$ is selected as the default parameter to ensure an optimal balance between high-fidelity recognition and inference efficiency.

4. Conclusion

In this paper, we proposed EFFormer, an Efficient Multi-view Transformer specifically designed for real-time emotion recognition of university students in resource-constrained campus environments. To address the high computational complexity of standard Vision Transformers, EFFormer introduces two key innovations. First, a Bidirectional Mamba (Bi-Mamba) module is employed to synthesize unified affective representations from multiple student perspectives. By exploiting selective state spaces, this strategy captures complex cross-view correlations with linear complexity, significantly improving the scalability of multi-view fusion. Second, we developed an instance-specific adaptive gating mechanism that dynamically prunes non-informative patches, deactivates redundant attention heads, and skips transformer blocks on the fly based on the difficulty of each input sample. Extensive experiments on the FPR-3 and FPR-4 datasets demonstrate that EFFormer achieves state-of-the-art performance while maintaining a significantly lower computational footprint (Gflops) compared to existing single-view and multi-view methods. The comprehensive visual analysis further confirms that our framework effectively harmonizes the conflict between high-fidelity recognition and inference parsimony. By combining efficient feature synthesis with dynamic resource allocation, EFFormer provides a robust and practical solution for intelligent emotion monitoring on edge devices, paving the way for personalized pedagogical interventions and mental health support in university settings.

5. Acknowledgement

This work was supported by the Humanities and Social Sciences and Special Project of Weifang Vocational College: <Research on Dynamic Monitoring and Early Warning Model of College Students' Mental Health Based on Multi-view Emotion Recognition>.

References

[1] W. Zhang and J. Wang, (2024) *"English text sentiment analysis network based on CNN and U-Net"* **IFS/ACM Transactions on Machine Learning *1***(1): 13–18. DOI: 10.70891/JSE.2024.100009.

[2] M. Liu, Y. Li, Z. Chen, and Y. Lin, (2026) *"Information-Driven Complementarity and Consistency Mining for Multi-View Clustering"* **IEEE Signal Processing Letters *33***: 216–220. DOI: 10.1109/LSP.2025.3639380.

[3] A. Xiang, Z. Qi, H. Wang, Q. Yang, and D. Ma. "A multimodal fusion network for student emotion recognition based on transformer and tensor product". In: *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*. 2024, 1–4. DOI: 10.1109/ICSECE61636.2024.10729485.

[4] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2024) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE Internet of Things Journal *11***(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[5] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems *35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[6] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, (2021) *"Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition"* **Advances in neural information processing systems *34***: 11960–11973.

[7] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, (2023) *"Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild"* **IEEE Transactions on Circuits and Systems for Video Technology *34***(5): 3192–3203. DOI: 10.1109/TCSVT.2023.3312858.

[8] D. Kim and B. C. Song. "Emotion-aware multi-view contrastive learning for facial emotion recognition". In: *European Conference on Computer Vision*. 2022, 178–195. DOI: 10.1007/978-3-031-19778-9_11.

[9] J. Chen, S. Dey, L. Wang, N. Bi, and P. Liu, (2024) *"Attention-based multi-modal multi-view fusion approach for driver facial expression recognition"* **IEEE Access**: DOI: 10.1109/ACCESS.2024.3462352.

[10] Y.-J. Cui, X.-H. Liu, J. Liang, and Y.-M. Fu. "Mvgt: A multi-view graph transformer based on spatial relations for eeg emotion recognition". In: *International Conference on Neural Information Processing*. 2025, 3–17. DOI: 10.1007/978-981-95-4378-6_1.

[11] Y. Wei, S. Yuan, R. Yang, L. Shen, Z. Li, L. Wang, and M. Chen. "Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, 5240–5252. DOI: 10.18653/v1/2023.acl-long.287.

[12] X.-B. Nguyen, H.-T. Nguyen, T.-H. Nguyen, N.-T. Do, and Q. V. Dinh. "Emotic masked autoencoder on dual-views with attention fusion for facial expression recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 4784–4792. DOI: 10.1109/CVPRW63382.2024.00481.

[13] G. Hou, Y. Shen, W. Zhang, W. Xue, and W. Lu. "Enhancing emotion recognition in conversation via multi-view feature alignment and memorization". In: *Findings of the association for computational linguistics: EMNLP 2023*. 2023, 12651–12663. DOI: 10.18653/v1/2023.findings-emnlp.842.

[14] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. "Tokens-to-token vit: Training vision transformers from scratch on imagenet". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, 558–567. DOI: 10.1109/ICCV48922.2021.00060.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, 10012–10022. DOI: 10.1007/s11042-024-19615-9.

[16] Y. Li, K. Zhang, J. Cao, R. Timofte, M. Magno, L. Benini, and L. Van Goo. "LocalViT: Analyzing locality in vision transformers". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2023, 9598–9605. DOI: 10.1109/IROS55552.2023.10342025.

[17] Z. Zhan, X. Mao, H. Liu, and S. Yu, (2025) *"STGL: Self-Supervised Spatio-Temporal Graph Learning for Traffic Forecasting"* **Journal of Artificial Intelligence Research *2***(1): 1–8. DOI: 10.70891/JAIR.2025.040001.

[18] W. Liu, (2024) *"Channel Reorganization for Few-Shot Segmentation"* **Journal of Artificial Intelligence Research *1***(1): 36–40. DOI: 10.70891/JAIR.2024.100025.