Enhancing Micro-gesture Classification via Global-Aware Importance Estimation in Vision Transformer

Xin Hu[1,2], Chenyang Pu[1], Yunan Li[1,2,3,*], Yulang Xu[1,2], Kun Xie[1,2] and Qiguang Miao[1,2,3]

[1]School of Computer Science and Technology, Xidian University, China
[2]Xi'an Key Laboratory of Big Data and Intelligent Vision, China
[3]Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, China

Abstract

Micro-gesture classification presents persistent challenges due to the subtle nature of gestures, short duration, and high susceptibility to background interference. While Vision Transformers (ViTs) have shown strong potential in modeling spatio-temporal dependencies, their uniform treatment of all patch tokens often causes attention to be diluted across static or irrelevant background regions. This "information averaging" effect becomes particularly problematic when the foreground signals are weak and the background remains visually dominant. To address this, we propose the Global-Aware Importance Estimation (GAIE) module, which analyzes token-level semantic contributions and guides the ViT to focus more effectively on fine-grained and meaningful gesture regions. GAIE preserves contextually valuable background cues while compressing redundant information, thus enhancing the model's sensitivity to subtle foreground movements. Our method achieved second place in Track 1 of the MiGA 2025 Challenge, demonstrating its effectiveness in real-world micro-gesture classification scenarios.

Keywords

Micro-gesture classification, Video Vision Transformer, Background redundancy

1. Introduction

Micro-gesture [1, 2] classification in video presents numerous challenges, primarily due to the subtlety of gestures, their short duration, and susceptibility to interference from large amounts of irrelevant background information. In recent years, Transformer-based models—particularly Vision Transformers (ViTs)—have demonstrated strong potential in modeling spatio-temporal dependencies [3, 4]. However, conventional ViTs treat all patch tokens equally, ignoring their semantic importance differences. This often causes the model's attention to be dispersed across background regions, thereby weakening its focus on critical gesture features.

In the specific context of micro-gesture classification, models face unique difficulties: target regions (e.g., subtle finger motions or local facial expressions) typically occupy very small spatial areas and exhibit low motion amplitude, while large portions of the video background remain static or highly repetitive. Although such background stability may appear "non-distracting" on the surface, it often leads to an "information averaging" phenomenon in globally modeled architectures like ViTs.

More specifically, since ViTs default to treating all tokens equally [5], when foreground changes are subtle and the background is dominant in size, the model may over-rely on background tokens. As a result, discriminative signals from the foreground can be diluted or even overwhelmed. This issue is especially pronounced in micro-gesture classification, leading to problems such as shifted attention, weak discriminative features, and insensitivity to subtle variations.

Moreover, while certain background regions may not directly contain gesture-related motion, they can still provide semantic or temporal context that supports gesture interpretation. Simply discarding these tokens risks losing potentially useful information. Therefore, there is a pressing need for a mechanism

MiGA@IJCAI25: International IJCAIWorkshop on 3rd Micro-gesture Analysis for Hidden Emotion Understanding, August 28, 2025, Guangzhou, China.
*Corresponding author.
xinhu123@stu.xidian.edu.cn (X. Hu); chengyangpu@stu.xidian.edu.cn (C. Pu); yunanli@xidian.edu.cn (Y. Li); yulangxu@stu.xidian.edu.cn (Y. Xu); xiekun@xidian.edu.cn (K. Xie); qgmiao@xidian.edu.cn (Q. Miao)
0000-0001-7316-4354 (Y. Li); 0000-0002-2872-388X (Q. Miao)

that can identify key tokens while effectively integrating redundant background information, rather than eliminating it entirely.

To this end, we propose the Global-Aware Importance Estimation (GAIE) Module, an attention-based method for analyzing information contribution. GAIE aims to preserve foreground sensitivity while adaptively compressing and integrating background tokens, enabling ViTs to better focus on fine-grained and meaningful gesture regions.

Our method achieved second place in Track 1 of the MiGA 2025 Challenge, demonstrating the effectiveness of the GAIE module in enhancing micro-gesture classification performance in real-world scenarios. The main contributions of our method are summarized as follows:

- We propose GAIE, a Global-Aware Importance Estimation module that enhances ViT's focus on subtle and discriminative gesture regions by adaptively weighting token importance.
- Our method achieves 68.70% Top-1 accuracy on the iMiGUE test set, ranking 2nd in Track 1 of the MiGA 2025 Challenge.

2. Related Works

2.1. Micro-gesture Classification

Micro-gesture (MiG) [1, 2] classification aims to automatically recognize and categorize subtle, low-amplitude movements occurring on human facial regions or body parts, such as transient facial twitches and subconscious hand gestures. In recent years, several methods have made significant progress in this task and achieved leading performance in the MiGA Challenge. For example, Ensemble Mode [6] employs a multi-scale heterogeneous ensemble network with residual connections and group training strategy to enhance micro-gesture representation. M2HEN [7] integrates cross-modal fusion and prototypical refinement modules to improve feature discriminability. VCLIP [8] adopts a CLIP-based distillation framework that leverages 3D heatmaps and textual features for cross-modal interaction. These approaches have all achieved excellent results in micro-gesture recognition within the MiGA Challenge.

2.2. Token Merging

Most token merging methods are designed for Vision Transformers (ViTs) performing image classification tasks. For example, ToMe [5] merges tokens based on bipartite soft matching, and DiffRate [9] performs adaptive token compression by jointly pruning and merging tokens through a differentiable compression rate. However, these methods perform dynamic merging over all tokens without explicitly preserving the most critical ones.

3. Proposed Method

3.1. Overall network framework

As shown in Figure 1, we adopt ViT [3, 4] as the backbone network to process the input video. The video is represented as $V \in \mathbb{R}^{T\times H\times W\times C}$, where T denotes the number of frames (temporal dimension), H and W represent the height and width of each video frame, and C is the number of channels.

To efficiently extract spatio-temporal features from the video, we employ the *tubelet embedding* method. This approach divides the input video V into a series of non-overlapping spatio-temporal patches [4], each with a size of $t \times h \times w$, where t denotes the temporal length and h, w denote the spatial dimensions. Each tubelet is linearly projected into a D-dimensional feature space, and cosine positional encodings are added to form the initial tokens $H_0 \in \mathbb{R}^{N\times D}$.

Feature extraction. After obtaining tokens H_0 are fed into a stack of L Transformer encoder layers. Each encoder layer consists of Layer Normalization (LN), Multi-Head Self-Attention (MSA), and a Multi-Layer Perceptron (MLP).

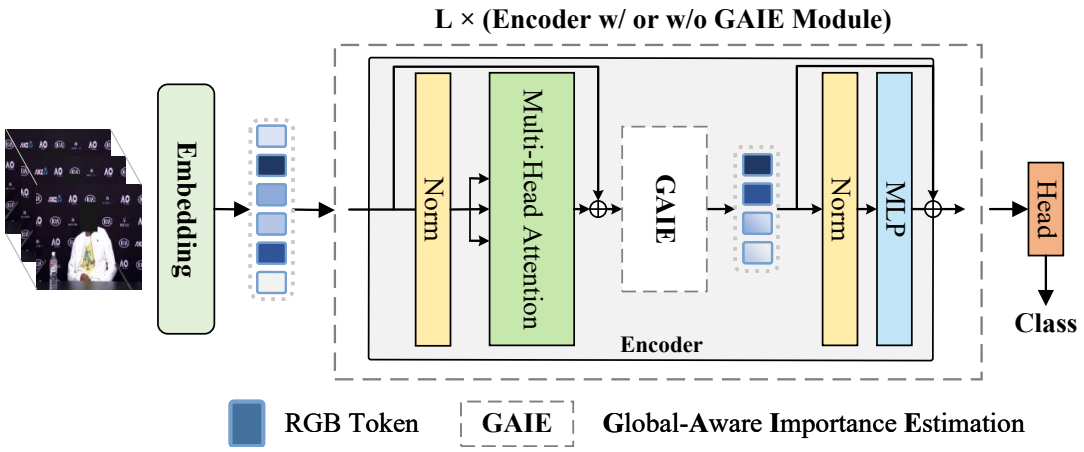

Figure 1: Overview of the proposed framework.

The computation at the i-th layer is defined as follows:

$$H_i' = \text{MSA}(\text{LN}(H_{i-1})) + H_{i-1} \tag{1}$$

$$H_i = \text{MLP}(\text{LN}(H_i')) + H_i' \tag{2}$$

It is worth noting that we insert the proposed *Global-Aware Importance Estimation (GAIE)* module into selected encoder layers to suppress redundant background information. Finally, the mean of the output tokens from the last layer, $\text{Mean}(H_L)$, is fed into a linear classification head to produce the final predictions.

3.2. Global-Aware Importance Estimation

To address the issue that Vision Transformer (ViT) treats redundant background regions indiscriminately in micro-gesture recognition tasks, we propose the Global-Aware Importance Estimation (GAIE) module. This module is designed to identify and retain regions that are highly relevant to foreground gestures based on each token's contribution within the global context, while suppressing background information that may interfere with gesture discrimination.

ViT's multi-head self-attention mechanism exhibits strong global modeling capabilities, and notably, it begins to distinguish between foreground and background regions even at shallow and intermediate layers. Motivated by this observation, the GAIE module introduces a token-level global importance estimation strategy to guide the subsequent feature fusion process more effectively.

Global Importance Score. Specifically, we first introduce a *Global Importance Score* to quantify the significance of each token within the overall spatio-temporal context. For the i-th token in the input sequence, let $\mathbf{q}_i$ denote its query vector, and $\mathbf{K}$ represent the key matrix composed of all tokens' key vectors. The attention weights $\mathbf{w}_i$ assigned by the i-th token to all others are computed as:

$$\mathbf{w}_i = \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{K}^\top}{\sqrt{d_k}}\right) \tag{3}$$

Then, the *global-aware importance score* for the j-th token is defined as the average degree to which it is attended to across all attention distributions:

$$\text{score}_j = \frac{1}{N} \sum_{i=1}^{N} w_{i,j} \tag{4}$$

where N is the total number of tokens. This score allows us to assess each token's potential contribution to micro-gesture recognition, while retaining its global contextual representation. Foreground regions

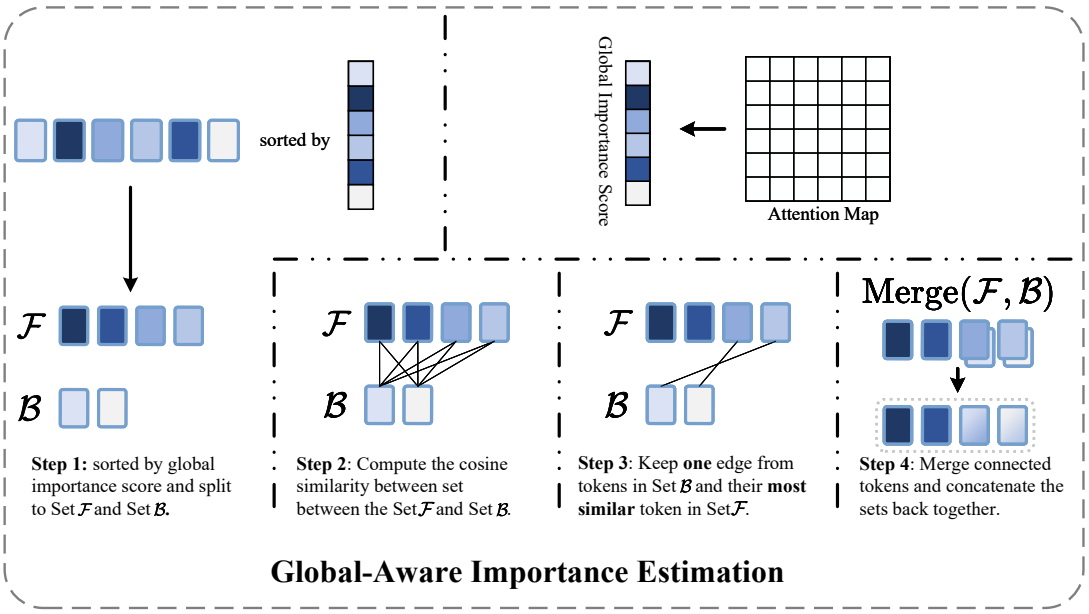

Figure 2: Global-Aware Importance Estimation. Darker tokens indicate stronger micro-gesture relevance foreground information, while lighter tokens represent unrelated background information.

often have higher scores, while some background areas—such as objects involved in the gesture or occluding elements—may also yield moderate or high importance due to their contextual relevance.

Aggregation Strategy. Based on the global importance estimation, the GAIE module aggregates semantic information from low-importance background tokens into the most relevant foreground tokens via *cosine similarity*.

The overall aggregation strategy of GAIE is illustrated in Figure 2 and includes the following steps:

1. **Token Selection:** All tokens are ranked based on their Global Importance Scores. The top K tokens are selected to form the foreground candidate set $\mathcal{F}$, while the remaining $N - K$ tokens constitute the background set $\mathcal{B}$. The number of selected tokens is controlled by a keeping ratio $\rho = \frac{K}{N}$.
2. **Bipartite Graph Construction:** A fully connected bipartite graph is built [5] between $\mathcal{F}$ and $\mathcal{B}$ using cosine similarity to capture the inter-token relationships.
3. **Redundancy Mapping:** For each token in $\mathcal{B}$, only the most similar connection to a token in $\mathcal{F}$ is preserved. These links indicate spatio-temporal background redundancy and are used to guide semantic aggregation.
4. **Token Fusion:** For each foreground token, the features of its connected background tokens are averaged and merged into it, thus suppressing redundant background information.

By integrating the GAIE module, we effectively compress redundant background content while preserving critical spatio-temporal context, thereby enhancing the model's expressiveness and recognition accuracy in micro-gesture regions.

When GAIE is inserted into the i-th encoder layer, the original set of N tokens is re-evaluated and refined. After importance scoring and background-token fusion, only K tokens are retained ($K < N$), forming the new token set H_i. This mechanism not only reduces token redundancy but also strengthens the model's attention to gesture-related regions, facilitating fine-grained representation and improved discriminative performance in subsequent layers.

4. Experimental Setup

4.1. Datasets

iMiGUE Dataset. The iMiGUE dataset [2] contains 32 micro-gestures and an additional non-micro-gesture class, collected from post-match press conference videos of professional tennis players. The challenge follows an interdisciplinary evaluation protocol, where 72 subjects are divided into a training set consisting of 37 subjects and a test set consisting of 35 subjects.

For the micro-gesture classification track, a total of 12,893, 777, and 4,562 MG clips from iMiGUE are used for training, validation, and testing, respectively.

4.2. Implementation Details

Our model is implemented using Python and the PyTorch framework, and trained on two NVIDIA RTX 4090 GPUs. The backbone network is the vanilla ViT-Base with joint spatio-temporal attention. The tubelet size is set to $t \times h \times w = 2 \times 16 \times 16$.

GAIE modules are inserted after the MSA components of the 4th, 7th, and 10th encoder layers, with a keeping ratio $\rho = 0.7$.

Input video frames are first resized to a spatial resolution of 256×256. During training, frames are randomly cropped to 224×224, and during inference, center cropping to 224×224 is applied. Temporally, 16 frames are randomly sampled during training and uniformly sampled during inference.

We use the AdamW optimizer with a weight decay of 0.05 and an initial learning rate of 2×10^{-5}. The learning rate follows a cosine decay schedule with a minimum value of 2×10^{-6}. The model is trained for 30 epochs on the iMiGUE dataset.

The batch size is set to 2 per GPU, resulting in a total batch size of 4. Model initialization is performed using ViT-Base weights pretrained and fine-tuned on the K400 dataset [10] via VideoMAE [11]. It is important to note that our training set includes the validation set.

4.3. Results and Analysis

4.3.1. Comparison with other entries

Table 1 presents the final Top 3 results of the The 3rd MiGA-IJCAI Challenge Track 1. We compare our method with the top-performing entries on the leaderboard. As shown in Table 1, our method achieved 2nd place with a Top-1 Accuracy of 68.70It is worth noting that our method only utilizes the RGB modality, while other methods may have exploited additional modalities or fused multiple sources of information. Despite this, our approach demonstrates strong generalization and robustness, achieving competitive performance purely from RGB data. This highlights the effectiveness and efficiency of our design, especially in scenarios where additional modalities are unavailable or impractical.

Table 1
Final ranking in The 3rd MiGA-IJCAI Challenge Track 1 on the iMiGUE test set.

User	Rank	Top-1(%)
gkdx2	1	73.21
Awuniverse(Ours)	2	68.70
Lonelysheep	3	67.01

4.3.2. Comparison with State-of-the-art Methods

As shown in Table 2, Our method achieves an accuracy of 68.70% using the RGB modality, significantly outperforming other methods that rely solely on RGB inputs, such as TSM (58.77%), VideoSwin

(61.73%), and ViViT (67.84%). This demonstrates the superior representational capacity of our proposed architecture in visual modeling.

For the Skeleton modality, methods like ST-GCN (46.38%) and AAGCN (54.73%) show relatively lower performance, while the 3D convolution-based PoseC3D achieves 61.11%, still falling short of our RGB-only method. This suggests that the Skeleton modality alone may suffer from limited information in the context of micro-gesture recognition.

Moreover, although multimodal methods such as Ensemble Mode (70.25%), M2HEN (70.19%), and VCLIP (68.90%) leverage both RGB and Skeleton inputs, our method—using RGB only—achieves results comparable to some of these multimodal approaches.

Table 2
Comparison with state-of-the-art methods on the iMiGUE dataset.

Method	**Modality**	**Top-1(%)**
TSM [12]	RGB	58.77
VideoSwin [13]	RGB	61.73
ViViT [4]	RGB	67.84
Ours	RGB	**68.70**
ST-GCN [14]	Skeleton	46.38
AAGCN [15]	Skeleton	54.73
PoseC3D [16]	Skeleton	61.11
Ensemble Mode [6] (MiGA'24 1st)	RGB + Skeleton	70.25
M2HEN [7] (MiGA'24 2nd)	RGB + Skeleton	70.19
VCLIP [8] (MiGA'24 3rd)	RGB + Skeleton	68.90

4.3.3. Effectiveness of the proposed method

As shown in Table 3, our proposed method significantly reduces computational cost (MACs from 101.848G to 65.820G) and inference latency (from 21.96 ms to 13.74 ms), while achieving improved recognition accuracy (from 67.84% to 68.70%), all under a similar parameter scale. This clearly demonstrates that our model enhances the ability to model critical micro-gesture features while maintaining computational efficiency.

Table 3
Efficiency and performance comparison.

Method	MACs (G)	Params (M)	Latency (ms)	Accuracy (%)
ViViT	101.848	65.000	21.96	67.84
Ours	**65.820**	**65.00**	**13.74**	**68.70**

5. Conclusion

In this paper, we propose a novel network for micro-gesture recognition, featuring a Global-Aware Importance Estimation Module (GAIE) designed to suppress irrelevant and redundant background information. Our method achieved 2nd place in The 3rd MiGA-IJCAI Challenge Track 1, demonstrating strong performance on micro-gesture recognition tasks using only RGB modality. Experimental results on the MiGA Track 1 dataset validate the effectiveness of our approach, highlighting its competitive advantage even against multimodal methods. The proposed network offers a lightweight yet powerful solution for fine-grained spatiotemporal micro-gestures understanding.

Acknowledgments

The work is jointly supported by the National Natural Science Foundation of China under grants No.62472342, and 62272364, the National Science and Technology Major Project under grant No.2022ZD0117103, the provincial Key Research and Development Program of Shaanxi under grant No.2024GH-ZDXM-47, the Research Project on Higher Education Teaching Reform of Shaanxi Province under grant No.23JG003, the Fundamental Research Funds for the Central Universities under grant No.QTZX25037.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, International Journal of Computer Vision 131 (2023) 1346–1366.

[2] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10631–10642.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021).

[4] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836–6846.

[5] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, J. Hoffman, Token merging: Your ViT but faster, in: International Conference on Learning Representations, 2023.

[6] H. Huang, Y. Wang, L. Kerui, Z. Xia, Multi-modal micro-gesture classification via multiscale heterogeneous ensemble network, MiGA-IJCAI workshop (2024).

[7] G. Chen, F. Wang, K. Li, Z. Wu, H. Fan, Y. Yang, M. Wang, D. Guo, Prototype learning for micro-gesture classification, MiGA-IJCAI workshop (2024).

[8] Y. Wang, Z. Dong, P. Li, Y. Liu, A multimodal micro-gesture classification model based on clip, MiGA-IJCAI workshop (2024).

[9] M. Chen, W. Shao, P. Xu, M. Lin, K. Zhang, F. Chao, R. Ji, Y. Qiao, P. Luo, Diffrate: Differentiable compression rate for efficient vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 17164–17174.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).

[11] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, in: Advances in Neural Information Processing Systems, 2022.

[12] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7083–7093.

[13] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[14] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[15] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020) 9532–9545.
[16] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2969–2978.