MaMoE: Leveraging Masked Autoencoders and Mixture-of-Experts for Identity-Free Micro-action Recognition

Hongkai Sui[1], Sirui Zhao[2,*], Xianglong Shi[3], Yiming Zhang[4], Jinyang Huang[5], Feng-Qi Cui[6] and Tong Xu[7]

***Abstract*— Micro-actions (MAs) are subtle and brief movements in human behavior that can effectively reflect an individual's emotional states or underlying intentions, making their recognition highly valuable for understanding nuanced human expressions. However, compared to general action recognition, micro-action recognition (MAR) is considerably more challenging attributed to the extremely small motion amplitudes of MAs and their high susceptibility to interference from irrelevant movements. To overcome these challenges, we propose a novel identity-free MAR framework, termed MaMoE, which consists of two key components: the self-supervised representation learning module (HeatmapMAE) and the hand-part-based mixture-of-experts module (Hand-MoE). Specifically, we propose HeatmapMAE with a channel masking strategy, which captures fine-grained MA representation by performing masked reconstruction on skeleton heatmap inputs. Furthermore, to mitigate the hand interference problem in MAs, we introduce Hand-MoE, which categorizes MAs into different sets based on their dependence on the hand and then performs more refined recognition through collaboration among expert models. Comprehensive experiments demonstrate that our approach achieves state-of-the-art performance on three authoritative MA datasets, exhibiting significant advantages in both recognition accuracy and robustness.**

I. INTRODUCTION

Micro-actions (MAs) are rapid and subtle behaviors, exhibiting very low intensity [1]. Unlike regular actions, MAs are not intended for any illustrative or communicational purposes but instead occur as spontaneous or involuntary bodily responses to specific stimuli, particularly negative stimuli [2]. Consequently, micro-action recognition (MAR) plays an important role in emotion recognition and psychological assessment [3]. However, because of the small amplitudes and subtle inter-class differences of MAs, they are highly susceptible to interference from other movements (especially hand interference, as shown in Fig. 1). This makes MAR a highly challenging task.

Recently, substantial research efforts have been devoted to MAR. In general, existing methods can be classified into two categories: RGB video-based [4]–[6] and skeleton-based [7]–[9]. RGB video-based approaches model fine-grained motion patterns directly from visual observations using Convolutional Neural Networks (CNNs) [4] or Transformer-based architectures [6], and have shown strong representation capacity in MAR tasks. However, their reliance on raw appearance information inevitably introduces privacy concerns, which restricts their deployment in privacy-sensitive applications such as mental health monitoring. In contrast, skeleton-based techniques, owing to their efficiency and privacy-preserving properties, have become an important research direction in MAR. These methods use only the coordinates of human skeletal joints, abstract human motion into a structured joint representation, and perform spatiotemporal feature learning based on Graph Convolutional Networks (GCN) [8], [10] or PoseConv3D [7], [11] architectures, validating their effectiveness in MAR scenarios.

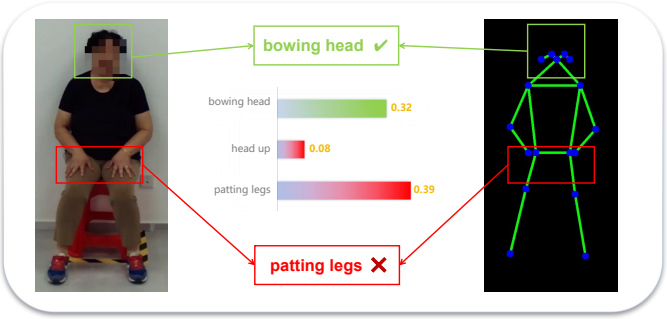

Fig. 1. An example of hand interference is shown. Although the ground-truth label is "bowing head", the model incorrectly classifies the action as "patting legs" due to the natural placement of the hands on the legs.

Despite their merits, current skeleton-based MAR methods still have inherent flaws in recognizing localized MAs. For instance, GCN-based methods have limitations in robustness and scalability [11], while PoseC3D-based methods are limited by the local receptive field mechanism of convolution, failing to model long-range joint coordination relationships. Moreover, in practice we observe that these methods still struggle to correctly identify MAs when faced with hand interference as shown in Fig. 1, because they adopt a unified holistic modeling strategy, which tends to mistake unrelated hand movements for the actions that need to be recognized. For example, on the MA-52 dataset, PoseC3D misclassifies 228 head-related MA samples as hand-related (10.23% of all head-related MA instances), despite their clearly distinct motion patterns.

To address these challenges, we propose MaMoE, an identity-free MAR framework built upon the core idea of leveraging mixture-of-experts with complementary specializations to collaboratively handle MAs with varying degrees of dependence on hand information. Specifically, MaMoE consists of two key modules: the self-supervised representation learning module (HeatmapMAE) and the hand-part-based mixture-of-experts module (Hand-MoE). HeatmapMAE transforms skeletal joint sequences into spatiotemporal

*Sirui Zhao, Hongkai Sui, Xianglong Shi, Yiming Zhang, Feng-Qi Cui and Tong Xu are with University of Science and Technology of China, Hefei, China(corresponding author. e-mail: siruit@ustc.edu.cn).

Jinyang Huang is with Hefei University of Technology, Hefei, China.

heatmap representations and learns fine-grained, robust motion features via masked reconstruction using a tailored channel masking strategy along with weighted mean squared error (WMSE) loss function. Building upon HeatmapMAE, Hand-MoE categorizes MAs according to their hand-dependency and employs type-specific experts for collaborative recognition. This design effectively suppresses non-discriminative hand interference and improves recognition robustness.

Our contributions can be summarized as follows:

- To address the challenges of MAR arising from their subtle amplitudes and high susceptibility to unrelated movements, we propose MaMoE, an identity-free MAR framework based on masked autoencoders and mixture-of-experts.
- We design a hand-part-based mixture-of-experts model (Hand-MoE), which characterizes the heterogeneous influence of hand movements across MA categories and enables targeted expert collaboration for more accurate and robust recognition.
- We introduce a self-supervised skeleton-based representation learning module (HeatmapMAE) and design a channel masking strategy and WMSE loss for it, enabling it to produce fine-grained and robust motion features to support effective expert specialization in MaMoE.
- Extensive experiments demonstrate the effectiveness of the proposed framework, and our MaMoE achieves state-of-the-art performance on three MA datasets.

II. RELATED WORK

A. Skeleton-based Action Recognition

Skeleton-based action recognition has been extensively studied due to its effectiveness in modeling human motion while preserving privacy. Liu *et al.* propose ProtoGCN [8], which dynamically decomposes the entire skeleton sequence into a combination of learnable prototypes and reconstructs them by contrasting prototypes, effectively enhancing discriminative representations of similar actions. Xu *et al.* propose LA-GCN [12], a framework that utilizes large language models to construct relationship graphs for guiding skeleton generation and provides auxiliary supervision. In the field of self-supervised learning, Sun *et al.* [13] propose a General Feature Prediction (GFP) framework that uses GCN as the backbone and replaces traditional low-level reconstruction with hierarchical high-order feature prediction for self-supervised skeleton-based action recognition. Unlike GFP, our proposed HeatmapMAE employs a ViT architecture to handle skeleton heatmap volumes, thereby providing stronger robustness and noise resistance. Through our designed mask pre-training paradigm for heatmap input, HeatmapMAE can learn compact and semantically consistent MA representations.

B. Mixture-of-Experts

Mixture-of-Experts (MoE) is a scalable learning paradigm, whose core idea is to leverage multiple experts to collaboratively solve a complex problem, rather than relying on a single all-around model. It consists of two components: a routing mechanism and experts. The routing mechanism determines which expert should handle each input, while the experts focus on handling specific types of inputs or subtasks. Through collaboration among experts, MoE can significantly improve model performance on complex tasks. For example, Min *et al.* [14] introduce a semantic prior-guided routing optimization method for Soft MoE models, significantly enhancing performance and interpretability in image classification tasks through a foreground-aware auxiliary loss and lightweight LayerScale mechanism. Dehkordi *et al.* [15] address the long-tail distribution problem in human action recognition in static images by introducing a two-stage multi-expert classification method and a graph-based category selection algorithm. Given the great success of MoE, we propose Hand-MoE, which is designed to mitigate hand interference in MAR. By explicitly modeling the dependency between MAs and hand movements, Hand-MoE achieves stable performance improvements through expert collaboration.

III. METHODOLOGY

As illustrated in Fig. 2, the proposed MaMoE mainly consists of two modules: HeatmapMAE and Hand-MoE. HeatmapMAE learns robust MA representations from skeleton heatmaps through masked autoencoder pre-training, while Hand-MoE uses two expert models to collaborate and provide more refined modeling of MA. These will be described in detail in subsequent sections.

A. Skeleton Heatmap Representation

Following the common practice in prior work [11], we encode 2D poses as a heatmap of size K $\times$ T $\times$ H $\times$ W, where K, T, H, and W denote the number of joints, the number of frames, and the height and width of each frame, respectively. For the coordinate triples (x_k, y_k, c_k) of a skeletal joint, we obtain a joint heatmap J by combining K Gaussian mappings centered at each node:

$$J_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2\sigma^2}} * c_k, \tag{1}$$

where σ controls the variance of the Gaussian mapping, and is set to 0.7. (x_k, y_k) and c_k are the position and confidence score of the k_{th} joint. We can also create a limb heatmap L:

$$L_{kij} = e^{-\frac{\mathcal{D}((i,j),\mathrm{seg}[a_k,b_k])^2}{2\sigma^2}} * \min(c_{a_k}, c_{b_k}), \tag{2}$$

where k_{th} limb lies between two joints a_k and b_k. Function $\mathcal{D}$ calculates the distance from point (i, j) to line segment [(x_{ak}, y_{ak}), (x_{bk}, y_{bk})]. Note that for the 3D skeletal coordinates (x'_k, y'_k, z'_k) extracted from the sensor, we set x_k=x'_k, y_k =y'_k and c_k=1.

B. HeatmapMAE with channel masking strategy

HeatmapMAE adopts an asymmetric encoder–decoder architecture with ViT as the backbone for heatmap volumes, which consists of a pretraining stage and a finetuning stage. During pretraining, the model performs masked reconstruction on unlabeled data to learn spatiotemporal representations

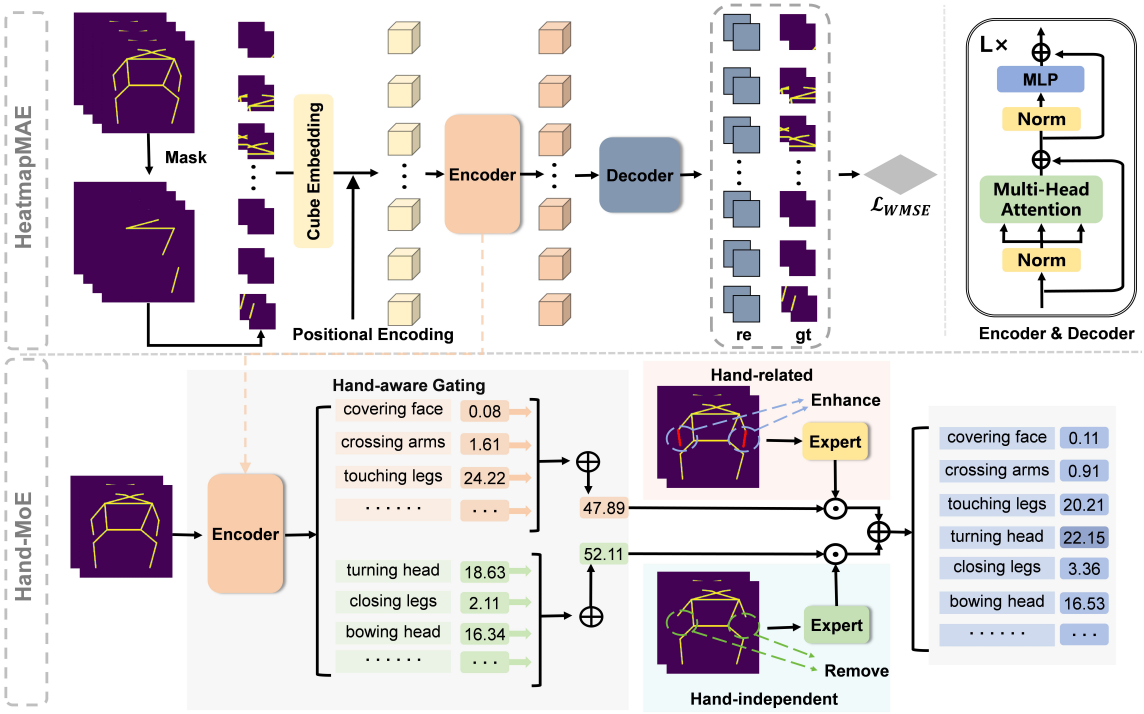

Fig. 2. Overall framework of MaMoE. The framework consists of two key modules: HeatmapMAE and Hand-MoE. We first pre-train the model by employing a channel masking strategy and WMSE loss function for the reconstruction of skeleton heatmap masks. The encoder of HeatmapMAE is then utilized, combined with a hand-aware gating mechanism and expert collaboration, to achieve more refined classification.

of skeleton heatmaps. Specifically, Given the input $X \in \mathbb{R}^{K\times T\times H\times W}$, we first adopt a joint space-time cube embedding and reshape them into L tokens of channel dimension d, denoted as $X_p \in \mathbb{R}^{L\times d}$, where $L = \frac{T\times H\times W}{2\times 8\times 8}$. Then, we add positional encoding to X_p and input them into the encoder after masking some data. The decoder reconstructs the masked heatmaps from encoded representations. During fine-tuning, only the encoder is retained and trained in a supervised manner using the full heatmap input without masking. Furthermore, to better suit heatmap input, we designed a specialized mask pre-training strategy, mainly including channel masking strategy and WMSE loss.

1) Channel Masking Strategy: Unlike conventional image or video data, each channel in skeleton heatmap volumes represents the features of a specific joint or limb. Therefore, existing masking strategies, such as random masking [16] or tube masking [17], typically apply the mask at the level of spatiotemporal blocks, which fails to capture the relationships between different joints or limbs, resulting in suboptimal representation learning. To address this issue, we propose a channel masking strategy. Specifically, we randomly select several heatmap channels and set them to zero, then input the heatmap into the encoder, and the decoder predicts the original values of the masked channels. This channel masking strategy can explicitly model dependencies between different joints or limbs, thereby learning more discriminative skeletal representations.

2) Weighted Mean Squared Error Loss: Owing to the overwhelming proportion of background pixels, the decoder tends to trivially reconstruct masked regions as zeros, which is detrimental to feature modeling. We therefore introduce a weighted mean squared error (WMSE) loss, and the formula is as follows:

$$\mathcal{L}_{WMSE} = \frac{1}{N}\sum_{i=1}^{N}(1+\alpha y_i)(y_i-\hat{y}_i)^2, \tag{3}$$

where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, α represents the weighting factor and is set α to 10 in our work. This design increases the contribution of foreground pixels, encouraging the decoder to focus on information-rich regions.

C. Hand-Part-based Mixture-of-Experts

Given the extremely small amplitude of MAs, the natural placement or unconscious motions of the hands often introduce significant noise interference to MAR, leading to recognition confusion. To address this issue, we introduce Hand-MoE, which models hand-related dependencies at the category level to enable more stable and targeted expert collaboration. Our Hand-MoE consists of a hand-aware gating strategy based on category aggregation and an expert group module.

1) Hand-aware Gating Strategy: Unlike conventional MoE frameworks that adopt sample-level dynamic gating, we propose a hand-aware category-level gating to model the structural dependency between action categories and hand movements. This is a soft routing mechanism, assigning continuous routing weights for each category so that multiple

experts can collaboratively contribute. Specifically, based on statistical analysis of the dataset, fine-grained actions are grouped into hand-related and hand-independent categories. Then we use HeatmapMAE as the base model to generate a fine-grained category distribution P_{fine}, and aggregate the probabilities of fine-grained categories within each group to compute the hand-related probability P_{hand}:

$$P^j_{\text{hand}} = \sum_{k \in \{i | f(i)=j\}} P^k_{\text{fine}}, \tag{4}$$

where $f(\cdot)$ maps each fine-grained category to its hand-dependency group j. The resulting P_{hand} serves as a category-level gating weight for expert routing.

2) Expert Group: After hand-aware gating, we input the samples into the expert group, which consists of two specialized modules: one for hand-related MAs and the other for hand-independent MAs. For hand-related samples X_r, we explicitly enhance hand features by amplifying the values of hand-related keypoint heatmaps by two times before feeding them into the expert. In contrast, for hand-independent samples X_i, we suppress potential interference from hand movements by zeroing out the hand-related keypoint heatmap prior to expert inference, as formulated in Eq. 5 and Eq. 6.

$$X'^c_r = \begin{cases} Amplify(X^c_r) & \text{if } c \in S, \\ X^c_r & \text{otherwise,} \end{cases} \tag{5}$$

$$X'^c_i = \begin{cases} Zero(X^c_i) & \text{if } c \in S, \\ X^c_i & \text{otherwise,} \end{cases} \tag{6}$$

where X^c denotes the c_{th} channel of X, S is the set of channels corresponding to hand-related joints. Then, the outputs of the two experts are combined in a weighted manner to get final output P_{MoE}:

$$P_{MoE} = Cat(P^0_{hand} \cdot E_r(X'_r), P^1_{hand} \cdot E_i(X'_i)), \tag{7}$$

where $E(\cdot)$ represent class probability distribution obtained by processing with corresponding expert model and P^0_{hand} and P^1_{hand} are the category-level gating weights obtained from the hand-aware gating module. $Cat(\cdot)$ represent concatenating the output results. Through this design, each expert focuses on the most discriminative features of a specific MA type, leading to improved recognition accuracy. Notably, all expert models share the same architecture as the base model and are trained exclusively on data from their assigned category. To mitigate potential cascading errors caused by incorrect gating, we perform a weighted fusion of the expert outputs and the base model predictions at the final inference stage, further enhancing the robustness and overall performance of MAR.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

Datasets: We conducted experiments on three popular MA datasets: MA-52 [1], iMiGUE [18], and SMG [2]. MA-52 was a large-scale full-body MA dataset, annotated at both coarse-grained (7 body-level) and fine-grained (52 action-level) categories. iMiGUE consisted of 359 press-conference videos, focusing on upper-body movements, and included 31 MA categories and one illustrative gesture class. SMG was a spontaneous full-body micro-gesture dataset containing 16 micro-gesture categories and one non-gesture class, with multimodal data including RGB, depth, silhouette, and skeleton. For skeleton data, we used HRNet to obtain 17 keypoints for MA-52, adopted 22 keypoints (upper-body joints plus 10 fingertips) for iMiGUE, and retained the original 25 keypoints for SMG.

Evaluation: For MA-52 dataset, we followed the practice of Guo *et al.* [1] and use mean F1 scores at the body and action levels to evaluate the model's performance. The formula for the mean F1 is as follows:

$$F1_{mean} = \frac{F1^{coarse}_{macro} + F1^{coarse}_{micro} + F1^{action}_{macro} + F1^{action}_{micro}}{4}. \tag{8}$$

For iMiGUE dataset and SMG dataset, we followed the practice of Chen *et al.* [2] and Liu *et al.* [18], conducting an experiment to evaluate Top-1 accuracy.

B. Experimental settings

For the input, we utilized heatmaps of size 32×K×56×56, where K denoted the number of joints or limbs. After cube embedding, we obtained 784 tokens with a channel dimension of 192. The encoder was configured with 6 transformer layers and 3 attention heads while the decoder consisted of 3 layers, with an embedding dimension of 96 and 2 attention heads. During the pre-training phase, we adopted AdamW with β_1 = 0.9, β_2 = 0.99, a base learning rate of 3e-3, a batch size of 192, and a cosine decay scheduler. During the fine-tuning phase, we set the weight decay to 0.01 and the learning rate to 1e-3. Other hyperparameters remained consistent with those used in pretraining.

C. Comparison With the State-of-the-Art Methods

We conducted a comparative analysis of state-of-the-art (SOTA) methods for skeleton-based MAR across three benchmark datasets, as shown in TABLE I and TABLE II. On the MA-52 dataset, MaMoE achieved an $F1_{mean}$ score of 64.44%, and surpassed existing SOTA models in both body and action level metrics. Notably, the significant improvement in the $F1^{action}_{macro}$ score demonstrated that MaMoE achieved more balanced and robust performance when dealing with class imbalance problems. Moreover, on the iMiGUE and SMG datasets, our MaMoE achieved Top-1 accuracies of 63.52% and 69.18% respectively, outperforming the second-best method by 2.19% and 3.28%. These results demonstrated the superiority of our method in handling MAs.

D. Ablation Studies

[1]For fair comparison, we used different numbers of skeleton keypoints, leading to result discrepancies from the original paper.

TABLE I

COMPARISON WITH THE SOTA ON THE MA-52 DATASET. **J**/**L** MEANS USING JOINT/LIMB-BASED HEATMAPS, AND **M** MEANS THE NUMBER OF ENSEMBLE MODALITIES.

Method	Publication	modality	Body	Action		Body		Action		All
			Top-1	Top-1	Top-5	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{mean}$
ST-GCN [10]	AAAI 2018	4M	73.92	55.53	85.27	64.33	73.92	42.09	55.53	58.96
ST-GCN++ [19]	MM 2022	4M	74.92	56.23	85.62	68.03	74.92	42.81	56.23	60.50
MSG3D [20]	CVPR 2020	4M	74.90	56.62	85.50	67.25	74.90	45.79	56.62	61.14
CTRGCN [21]	ICCV 2021	4M	74.45	56.64	85.09	67.52	74.45	43.43	56.64	60.51
PoseC3D [11]	CVPR 2022	2M	76.80	58.20	88.47	71.34	76.80	46.20	58.20	63.14
HD-GCN [22]	ICCV 2023	6M	75.30	57.41	88.24	65.54	75.30	45.41	57.41	60.91
FR-Head [23]	CVPR 2023	4M	75.83	58.74	88.53	67.34	75.83	44.86	58.74	61.69
BlockGCN [24]	CVPR 2024	4M	74.45	57.63	87.10	66.78	74.45	44.70	57.63	60.89
DS-GCN [25]	AAAI 2024	4M	75.19	58.02	85.73	66.46	75.19	45.06	58.02	61.18
SkateFormer [26]	ECCV 2024	4M	76.26	58.23	86.79	67.72	76.26	46.32	58.23	62.13
MMN[1] [9]	MM 2025	4M	76.96	56.28	88.13	68.56	76.96	41.67	56.28	60.87
Hyper-GCN [27]	ICCV 2025	4M	70.18	53.60	84.46	59.11	70.18	39.07	53.60	55.49
ProtoGCN [8]	CVPR 2025	6M	75.64	57.63	86.79	68.69	75.64	44.69	57.63	61.66
HeatmapMAE	Ours	J	76.28	56.73	87.92	70.01	76.28	46.45	56.73	62.37
MaMoE	Ours	J	76.48	57.51	88.79	70.89	76.48	47.28	57.51	63.04
HeatmapMAE	Ours	J+B	77.34	57.79	88.67	71.12	77.34	48.29	57.79	63.64
MaMoE	Ours	J+B	**77.59**	**58.81**	**89.78**	**72.01**	**77.59**	**49.31**	**58.81**	**64.44**

TABLE II

COMPARISON WITH THE SOTA ON THE IMIGUE AND SMG DATASETS.

Method	modality	iMiGUE		SMG	
		Top-1	Top-5	Top-1	Top-5
ST - GCN	4M	54.65	87.37	59.67	96.07
ST - GCN++	4M	56.69	88.47	59.84	94.26
MSG3D	4M	56.55	87.53	55.08	93.77
CTRGCN	4M	56.14	87.35	63.44	93.28
PoseC3D	2M	61.33	90.07	61.48	92.30
FR - Head	4M	58.94	89.72	62.95	**96.72**
BlockGCN	4M	-	-	60.00	92.95
DS - GCN	4M	51.21	85.99	60.00	93.44
SkateFormer	4M	60.92	90.05	63.28	93.77
Hyper-GCN	4M	56.44	87.86	63.11	93.11
ProtoGCN	6M	60.24	88.89	65.90	93.77
HeatmapMAE(Ours)	J	61.60	90.36	67.54	93.44
MaMoE(Ours)	J	62.52	91.03	67.54	94.59
HeatmapMAE(Ours)	J+B	63.11	91.54	68.52	95.41
MaMoE(Ours)	J+B	**63.52**	**91.60**	**69.18**	96.23

To validate the superiority of our proposed HeatmapMAE and the effectiveness of Hand-MoE, we conduct ablation experiments on MA52 dataset about them.

1) Different masking strategies: We set up five different experiments: pre-training on NTU120 without masking, and pre-training with four masking methods—random, frame, tube, and our channel masking (Fig. 3). As shown in TABLE III, mask pre-training is significantly better than NTU120-based pre-training. We attributed this gap to the domain discrepancy between NTU120 and MA datasets, which made it difficult for pre-trained models to transfer effective features. In addition, different masking methods can also affect the results, and our proposed channel masking method is more suitable for heatmap input, which outperforms the random masking strategy by 1.54% in terms of the mean F1 score.

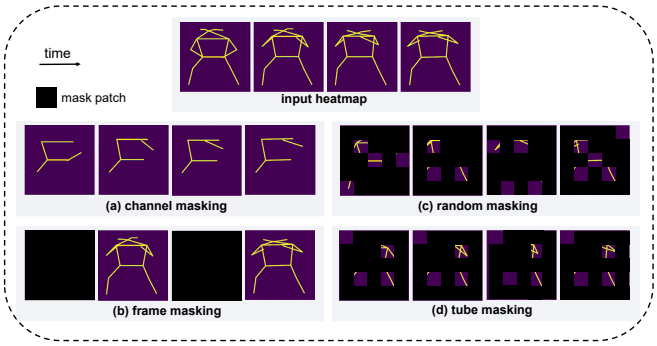

Fig. 3. Examples of different masking strategies.

2) Effectiveness of Hand-MoE: To validate the necessity of explicitly modeling hand interference, we compared three settings: a base model, a single expert model, and two expert models. As shown in TABLE IV, introducing expert models consistently improved MAR performance, with Hand-MoE achieving a 0.81% gain in mean F1 over the baseline, demonstrating the effectiveness of the proposed design.

TABLE III

ABLATION EXPERIMENTS ON MASKING STRATEGIES ON MA-52 DATASET.

pretrain/masking strategy	Top-1	Top-5	$F1_{macro}$	$F1_{micro}$	$F1_{mean}$
NTU120 dataset	40.32	80.77	32.65	40.31	36.48
tube masking	53.17	86.99	42.19	53.17	47.68
frame masking	53.37	86.18	45.38	53.37	49.37
random masking	54.35	86.82	45.75	54.35	50.05
channel masking	56.73	87.92	46.45	56.73	51.59

TABLE IV

ABLATION EXPERIMENTS ON HAND-MOE ON MA-52 DATASET.

method	Top-1	Top-5	$F1_{macro}$	$F1_{micro}$	$F1_{mean}$
HeatmapMAE(J)	56.73	87.92	46.45	56.73	51.59
+ hand-related expert	57.02	88.42	46.36	57.02	51.69
+ hand-independent expert	57.00	88.17	47.06	57.00	52.03
+ two expert	57.51	88.79	47.28	57.51	52.40

V. CONCLUSIONS

In this paper, we presented MaMoE, a novel skeleton-based identity-free MAR framework including two modules: HeatmapMAE and Hand-MoE. By converting skeletal sequences into heatmap representations and utilizing masked autoencoders for pre-training, HeatmapMAE learned more robust MA feature representations. Simultaneously, Hand-MoE leveraged different expert models based on the varying reliance on hand movements for different MAs, enabling more refined classification and mitigating the hand interference in MAs. Experimental results on several public MA datasets demonstrated that our proposed method achieved consistent improvements in both recognition accuracy and robustness, validating its effectiveness.

REFERENCES

[1] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6238–6252, 2024.

[2] Chen Haoyu, Shi Henglin, and Liu Xin, "Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis," 2023.

[3] Feng-Qi Cui, Jinyang Huang, Sirui Zhao, Kun Li, Zhi Liu, Meng Li, Ziyu Jia, Dan Guo, and Meng Wang, "Persomoni: A comprehensive video-based benchmark dataset for fine-grained personality assessment with 15 trait dimensions," *IEEE Transactions on Affective Computing*, 2026.

[4] Jian Zhou, Jingchao Yao, Nan Su, Jingchen Lu, Qingyang Yu, Yichi Zhang, and Wenqiang Hu, "Kmanet: A spatio-temporal enhancement network for micro-action recognition," *Knowledge-Based Systems*, p. 114139, 2025.

[5] Jinyang Huang, Yuanhao Feng, Feng-Qi Cui, Xiang Zhang, Zhi Liu, Xin Liu, Jianchun Liu, Fusang Zhang, and Meng Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing*, 2026.

[6] Deng Li, Bohao Xing, and Xin Liu, "Enhancing micro gesture recognition for emotion understanding via context-aware visual-text contrastive learning," *IEEE Signal Processing Letters*, vol. 31, pp. 1309–1313, 2024.

[7] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang, "Prototypical calibrating ambiguous samples for micro-action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 4815–4823.

[8] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun, "Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29248–29257.

[9] Jihao Gu, Kun Li, Fei Wang, Yanyan Wei, Zhiliang Wu, Hehe Fan, and Meng Wang, "Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 5461–5470.

[10] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[12] Haojun Xu, Yan Gao, Zheng Hui, Jie Li, and Xinbo Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2025.

[13] Shengkai Sun, Zefan Zhang, Jianfeng Dong, Zhiyong Cheng, Xiaojun Chang, and Meng Wang, "Towards efficient general feature prediction in masked skeleton modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 12212–12221.

[14] Chengxi Min, Wei Wang, Yahui Liu, Weixin Ye, Enver Sangineto, Qi Wang, and Yao Zhao, "Guiding the experts: Semantic priors for efficient and focused moe routing," *arXiv preprint arXiv:2505.18586*, 2025.

[15] Hojat Asgarian Dehkordi, Ali Soltani Nezhad, Hossein Kashiani, Shahriar Baradaran Shokouhi, and Ahmad Ayatollahi, "Multi-expert human action recognition with hierarchical super-class learning," *Knowledge-Based Systems*, vol. 250, pp. 109091, 2022.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16000–16009.

[17] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10078–10093, 2022.

[18] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying 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.

[19] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin, "Pyskl: Towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 7351–7354.

[20] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[21] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13359–13368.

[22] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10444–10453.

[23] Huanyu Zhou, Qingjie Liu, and Yunhong Wang, "Learning discriminative representations for skeleton based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10608–10617.

[24] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua, "Blockgcn: Redefine topology awareness for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2049–2058.

[25] Jianyang Xie, Yanda Meng, Yitian Zhao, Anh Nguyen, Xiaoyun Yang, and Yalin Zheng, "Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, vol. 38, pp. 6225–6233.

[26] Jeonghyeok Do and Munchurl Kim, "Skateformer: skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 401–420.

[27] Youwei Zhou, Tianyang Xu, Cong Wu, Xiaojun Wu, and Josef Kittler, "Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 12648–12658.