

# FreqMixFormer – Supplementary Material

Wenhan Wu

University of North Carolina at  
Charlotte  
Department of Computer Science  
Charlotte, NC, USA  
wwu25@uncc.edu

Ce Zheng

Carnegie Mellon University  
Robotics Institute  
Pittsburgh, PA, USA  
cezheng@andrew.cmu.edu

Zihao Yang

Microsoft Corporation  
One Inventory Organization  
Redmond, WA, USA  
mattyang@microsoft.com

Chen Chen

University of Central Florida  
Center for Research in Computer  
Vision  
Orlando, Florida, USA  
chen.chen@crcv.ucf.edu

Srijan Das

University of North Carolina at  
Charlotte  
Department of Computer Science  
Charlotte, NC, USA  
sdas24@uncc.edu

Aidong Lu

University of North Carolina at  
Charlotte  
Department of Computer Science  
Charlotte, NC, USA  
aidong.lu@uncc.edu

In this supplementary material, we provide the following items:

- Partial DCT vs full DCT algorithms.
- Evaluation of the number of DCT coefficients.
- Evaluation on UAV-Human dataset.
- Additional results.
- Implementation details.
- More visualizations.
- Limitations and future work.

## 1 Partial DCT vs Full DCT Algorithms

In FreqMixFormer, we utilize DCT in Frequency-aware Attention Block (FAB) to extract skeletal frequency features. As illustrated in Fig. 2 and 3 in the main paper, only *Query* matrix  $Q$  and *Key* matrix  $K$  are processed with DCT and IDCT modules for attention score, *Value* matrix  $V$  is only processed with linear transformation, the methodology can be found in Algorithm 1 as Partial DCT Algorithm. Moreover, we also investigate the Full DCT Algorithm, where DCT and IDCT process  $V$ , and the methodology is shown in Algorithm 2. However, the full DCT algorithm performs poorly in the experiment: the full DCT algorithm only achieves 87.7% on the NTU-60 X-Sub setting, while the partial DCT algorithm achieves 91.5% accuracy. The overview of the FreqMixFormer with full DCT algorithm is illustrated in Fig. 2. The Spatial Attention Block (SAB) in this experiment is shown in Fig. 1 (a) and the Frequency-aware Mixed Former (FAB) with full DCT algorithm is shown in Fig. 1 (b).

We hypothesize that the primary issues contributing to this gap are: 1) Applying DCT to  $Q$  and  $K$  can effectively highlight key frequency features and improve the model accuracy by matching relevant features during the computation of attention scores. 2) By excluding  $V$  from the frequency domain, the original temporal-spatial information is retained. This retention may help preserve more detailed and dynamic information in the final representation, enhancing the model’s ability to utilize these details for action recognition. 3) Recognizing actions relies not only on the frequency characteristics of movements (such as the speed and rhythm) but also on the specifics of how the actions are performed (like the swinging of an arm). Processing  $Q$ ,  $K$ , and  $V$  in different domains may allow the model to balance these needs.

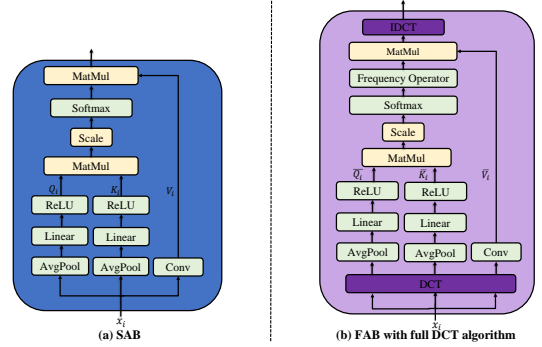


Figure 1: (a) SAB utilized in this experiment. (b) FAB with full DCT algorithm.

## 2 Evaluation of the number of DCT coefficients

In order to explore the frequency operator in-depth, we conduct an evaluation of the number of enhanced DCT coefficients. Table 2 shows the extra ablation study on the number of DCT coefficients  $N_c$  that we set as high-frequency coefficients (the rest are set as low-frequency coefficients). The high-frequency coefficients are enhanced by a frequency operator coefficient  $\varphi$  discussed in Section 3.4 of the main paper. For a fair comparison, we keep  $\varphi = 0.5$  during the experiments. As shown in the table, with the number of the enhanced DCT coefficient  $N_c = 12$ , the model achieves the best performance on NTU-60 (91.5% in X-Sub and 96.0% in X-View) dataset, and further increasing does not result in improvements.

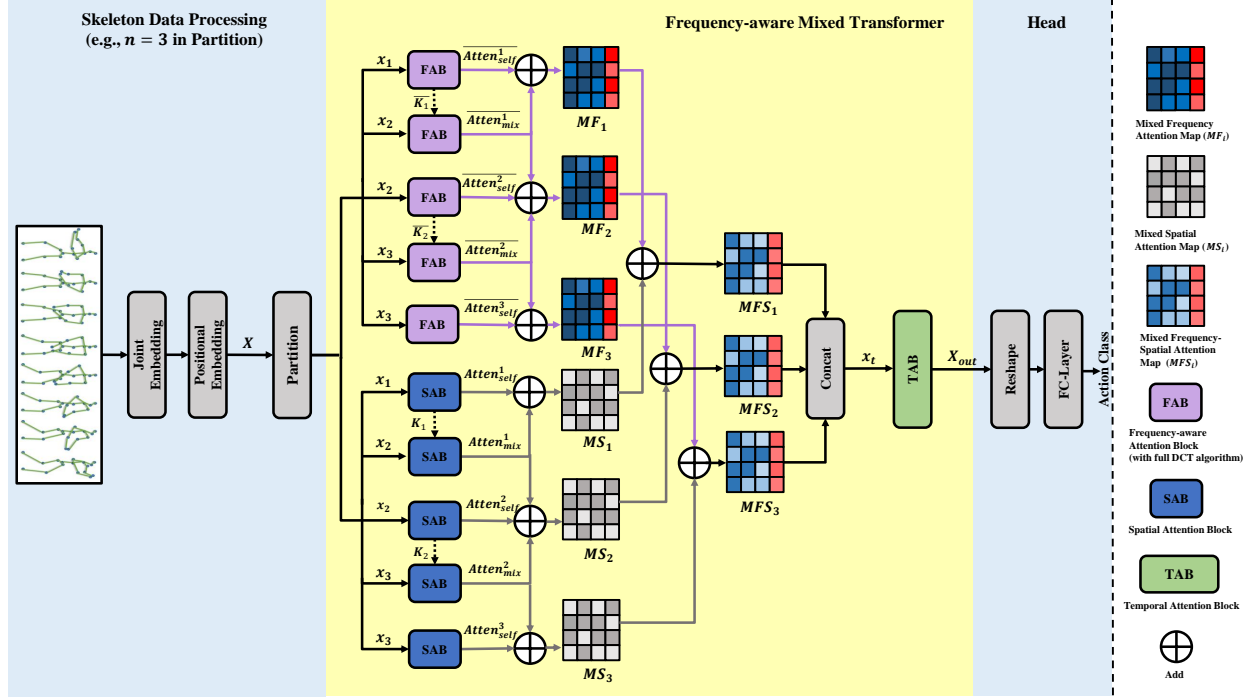
### Algorithm 1 Partial DCT

**Input:** the skeleton sequence is processed with joint embedding and positional embedding as the initial input  $X$ , where  $X \in \mathbb{R}^{C \times F \times J}$ .

**Init:**  $W_Q$ ,  $W_K$  and  $W_V$  are the learnable weight matrices.

**Output:** the partial DCT attention score

- (1)  $\bar{X} = DCT(X)$
- (2)  $Q = XW_Q$ ,  $K = XW_K$ ,  $V = XW_V$
- (3)  $\bar{Q} = \bar{X}W_Q$ ,  $\bar{K} = \bar{X}W_K$
- (4)  $Atten(\bar{Q}, \bar{K}, V) = IDCT \left( Softmax \left( \frac{\bar{Q}\bar{K}^T}{\sqrt{d}} \right) \right) V$



**Figure 2: Overview of the proposed FreqMixFormer with full DCT algorithm. The overall structure is similar to the method introduced in Section 3.2 of the main paper. The detailed structures of SAB and FAB are illustrated in Fig. 1.**

(5)  $Atten_1 = Atten(\bar{Q}, \bar{K}, V)$

**Return :**  $Atten_1$

---

#### Algorithm 2 Full DCT

---

**Input:** the skeleton sequence is processed with joint embedding and positional embedding as the initial input  $X$ , where  $X \in \mathbb{R}^{C \times F \times J}$ .

**Init:**  $W_Q, W_K$  are the learnable weight matrices.

**Output:** the full DCT attention score

- (1)  $\bar{X} = DCT(X)$
- (2)  $Q = XW_Q, K = XW_K, V = XW_V$
- (3)  $\bar{Q} = \bar{X}W_Q, \bar{K} = \bar{X}W_K, \bar{V} = \bar{X}W_V$
- (4)  $Atten(\bar{Q}, \bar{K}, \bar{V}) = \left( Softmax \left( \frac{\bar{Q}\bar{K}^T}{\sqrt{d}} \right) \right) \bar{V}$
- (5)  $Atten_2 = IDCT(Atten(\bar{Q}, \bar{K}, \bar{V}))$

**Return:**  $Atten_2$

---

## 3 Evaluation on UAV-Human Dataset

### 3.1 UAV-Human Dataset

UAV-Human [6] is an action recognition dataset comprising 22,476 video clips with 155 classes. The dataset was collected via a UAV across various urban and rural settings, both during daytime and nighttime. It extracts action data from 119 distinct subjects engaged in 155 different activities across 45 diverse environmental locations.

For evaluation (X-Sub, 17 joints in each subject), 89 subjects are selected for training and 30 for testing.

### 3.2 Experiment Settings

The hardware configurations are the same as the experiments reported in the main paper. The model is trained with 100 epochs, and the batch size is 128. We set a warm-up at the first 5 epochs. The weight decay is set as 0.0005, and the basic learning rate is 0.2. There is a 0.1 reduction at the 50th epoch.

### 3.3 Comparison Results

As Table 3 shows, we compare our performance with the state-of-the-art methods on the UAV-Human dataset. Our FreqMixFormer outperforms all the existing methods and achieves the new state-of-the-art results on this benchmark.

## 4 Additional Results

### 4.1 Accuracy Difference Results

We further analyze the Top-1 Accuracy Difference (%) between the proposed FreqMixFormer and the baseline method SkeMixFormer[10] with the joint input modality on NTU RGB+D 120 X-Sub. As illustrated in Fig. 6, the most significant improvements typically appear in confusing actions with subtle movements. For instance, our model achieves an improvement of **35.09%** for "make OK sign", **21.55%** for "make victory sign", and **18.56%** for "counting money". These results underscore FreqMixFormer's performance in recognizing actions that are visually confusing by extracting the frequency-spatial features.

**Table 1: Comparison with recent Frequency-based methods. The best performance is highlighted in bold.**

Method	Frequency Transformation	NTU-60 X-Sub (%)	NTU-60 X-View (%)	Param (M)	GFLOPS
SLnL-rFA	Fast Fourier Transform (FFT)	89.7	95.4	9.46	7.78
DCE-CRL	Discrete Cosine Encoding (DCE)	90.6	96.6	2.92	39.2
WDCE-Net	Discrete Wavelet Transform (DWT)	93.0	97.2	-	-
FreqMixFormer(ours)	Discrete Cosine Transform (DCT)	<b>93.6</b>	<b>97.4</b>	<b>2.04</b>	<b>2.40</b>

**Table 2: Search for the best number of DCT coefficient  $N_c$ .**

$N_c$	NTU-60	
	X-Sub (%)	X-View (%)
3	91.3	95.6
6	91.0	95.2
9	91.2	95.5
12	<b>91.5</b>	<b>96.0</b>
15	91.4	95.7

**Table 3: Comparison with the SOTA on UAV-Human dataset. The best performance is highlighted in bold. T indicates the Transformer-based method.**

	Method	Source	UAV-Human X-Sub (%)
GCN	ST-GCN [11]	AAAI'18	30.3
	DGNN [8]	CVPR'19	29.9
	2s-AGCN [13]	CVPR'19	34.8
	HARD-Net [5]	ECCV'20	37.0
	Shift-GCN [12]	CVPR'20	42.9
	MS-G3D [4]	CVPR'20	43.4
	CTR-GCN [1]	ICCV'21	43.4
	ACFL [9]	ACMMM'22	45.3
T	SkeMixFormer [10]	ACMMM'23	48.9
	<b>FreqMixFormer (ours)</b>		<b>49.6</b>

## 4.2 Comparison with Frequency-based Results

We provide an extra comparison with the previous frequency-based methods in skeleton action recognition. As shown in Table 1, our FreqMixFormer outperforms all the existing methods utilizing frequency analysis on the NTU-60 X-Sub dataset. Moreover, our model also plays a significant role in efficiency, as it has the least parameters (2.04M) and the best GFLOPs (2.40) among the frequency-based methods.

## 5 Implementation Details

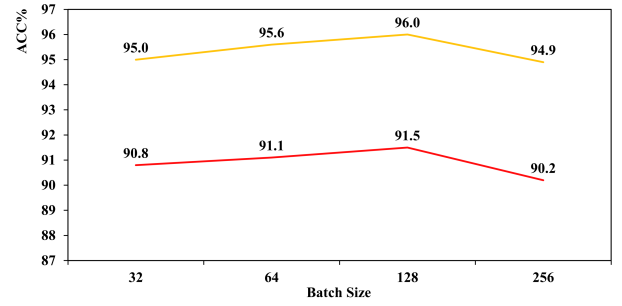
### 5.1 Multi-stram Fusion Strategy

The comparison is made with three ensembles of different modalities (joint only, 4-stream ensemble, 6-stream ensemble. We denote the stream as S for convenience) following the setting of InfoGCN [2]: S1:  $k = 1$ , motion = False; S2:  $k = 2$ , motion = False; S3:  $k = 8$  ( $k = 6$  for NW-UCLA and UAV-Human datasets), motion = False; S4:  $k = 1$ , motion = True; S5:  $k = 2$ , motion = True; S6:  $k = 8$  ( $k = 6$  for NW-UCLA and UAV-Human datasets), motion = True, where  $k$  indicates  $k$  value of  $k$ -th mode representation of the skeleton. And 4-stream = S1+S2+S4+S5, 6-stream = S1+S2+S3+S4+S5+S6. For a

fair comparison, experiments using the baseline method are also conducted with this ensemble strategy.

### 5.2 Evaluations of the Batch Size

Fig. 3 illustrates the impact of batch size during the training. We take the experimental results on the NTU-60 dataset as an example. As we can see, increasing the batch size from 32 to 128 enhances performance. However, a higher batch size (256) is not better because it requires more memory and leads to convergence issues. Thus, we choose 128 as our default batch size.

**Figure 3: The batch size settings.**

### 5.3 Channel Transformation in Temporal Attention Block

As we mentioned in Section 3.4 of the main paper, we adopt some tricky strategies in the baseline method [10] as our temporal channel transformation  $CT(\cdot)$ , which is stacked with two modules: 1) Channel Reforming Model. An improving model derived from SENet[3], which enhances the feature separation between groups and reduces noise, it is essential to reorganize the channel relationships within each group. 2) Multiscale Convolution Module. The first part of the Temporal MixFormer in [10], which is a simple optimization from MS-G3D [7] of maintaining a fixed filter while adjusting dilation, enabling the acquisition of more diverse multiscale temporal information and reducing computational costs. We simply adopt this combination as the  $CT(\cdot)$  operation.

## 6 More Visualizations

In this section, we exhibit more attention maps, the same as the visualization results illustrated in Section 4.5 of the main paper. Since we have provided the action "eat a meal" from the Hard set, we give more visualization results from the Medium set (headache) and Easy set (kicking) as examples. All the skeletons and attention

maps are generated by the NTU-60 dataset. As shown in Fig. 4 and Fig. 5, our proposed Frequency-aware Mixed attention maps (extracted by FAB modules) contain more detailed information and joint correlations compared with the spatial maps (extracted by SAB).

## 7 Limitations and Future Work

Despite the high accuracy of our model, it still has some limitations. Firstly, our model is still not efficient and lightweight enough. As we discussed in the ablation study from the main paper, there is a gap between our method and the recent GCN-based methods such as HD-GCN [4] (1.68M parameters vs 2.04M, 1.60 GFLOPS vs 2.40 GFLOPS), and we have no remarkable advantages of efficiency over the recent transformer-based methods. Secondly, we keep all the high-frequency coefficients during the training, which is not robust to noisy joint information. The more efficient way is to enhance the high-frequency coefficients selectively instead of the whole coefficients. Our future work will focus on finding the best trade-off point between efficiency and accuracy.

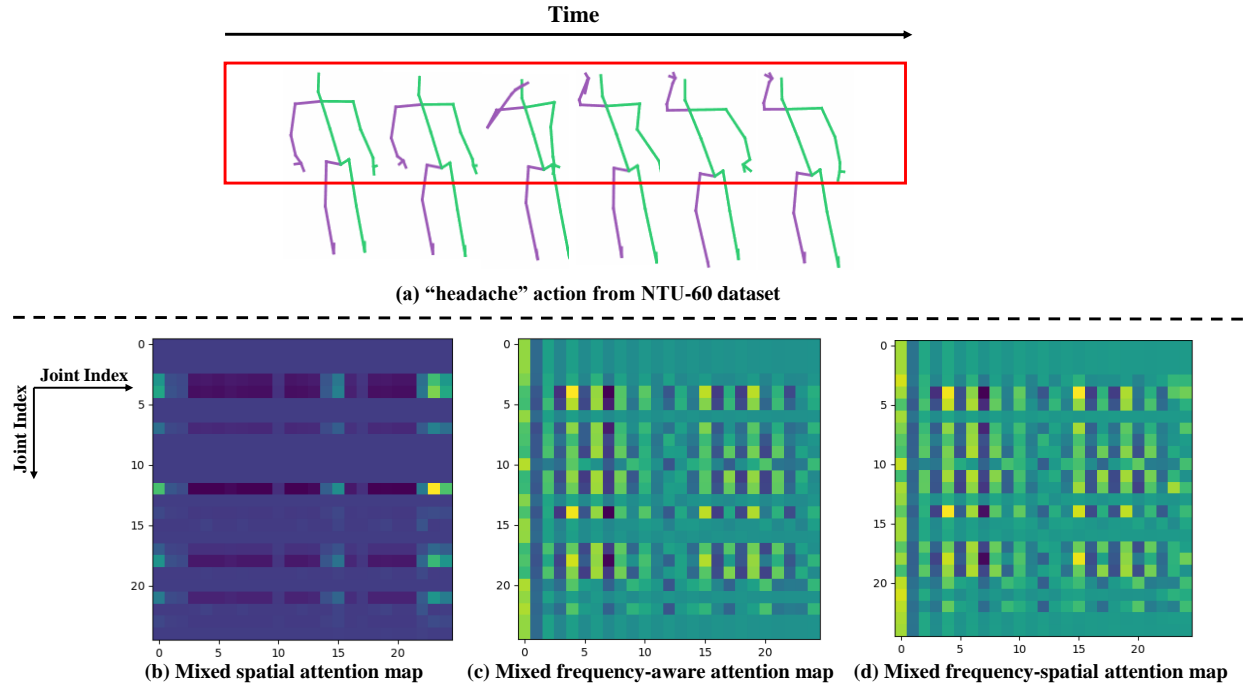


Figure 4: The visualization results of "headache" action from the NTU-60 dataset. (a) is the skeleton sequence, the red box indicates the attention area with stronger correlations. (b) is the mixed spatial attention map. (c) is the mixed frequency-aware attention map. (d) is the mixed frequency-spatial attention map. In this example, FAB focuses more on correlations of arms and legs, while SAB only focuses on the correlations of the right hand.

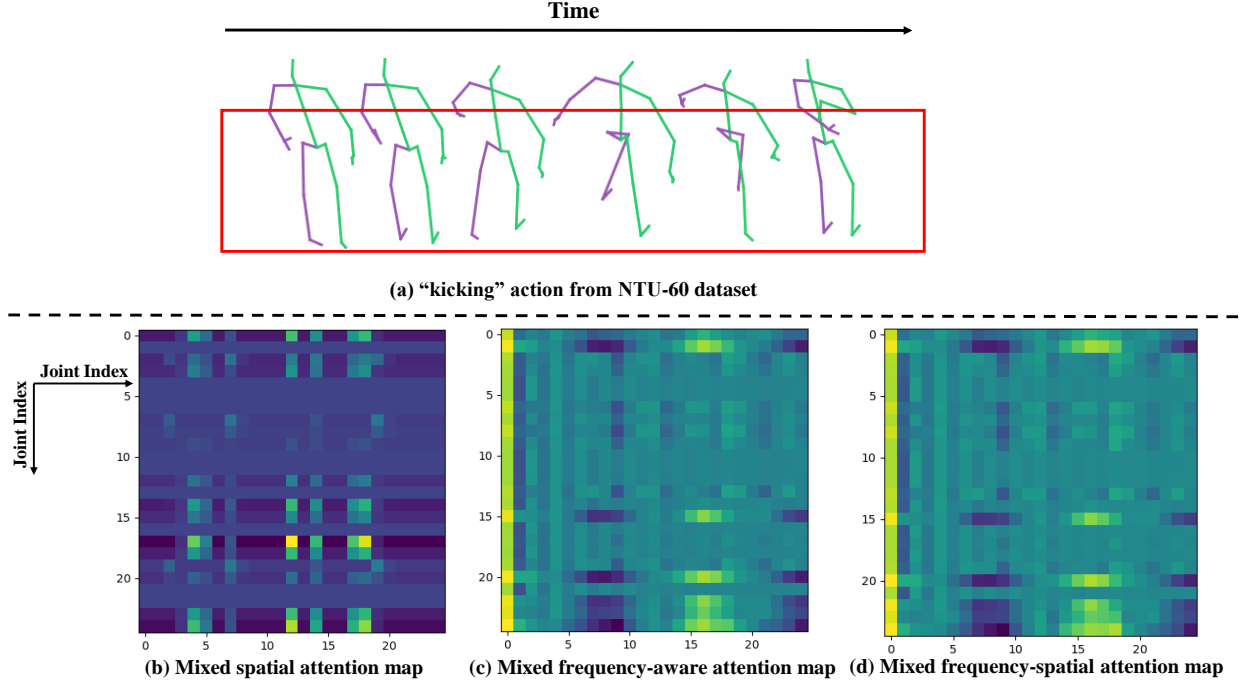


Figure 5: The visualization results of "kicking" action from the NTU-60 dataset. (a) is the skeleton sequence, the red box indicates the attention area with stronger correlations. (b) is the mixed spatial attention map. (c) is the mixed frequency-aware attention map. (d) is the mixed frequency-spatial attention map. In this example, FAB focuses more on the correlations of the spine and right leg, while SAB only focuses on the correlations of the left hip and right ankle.

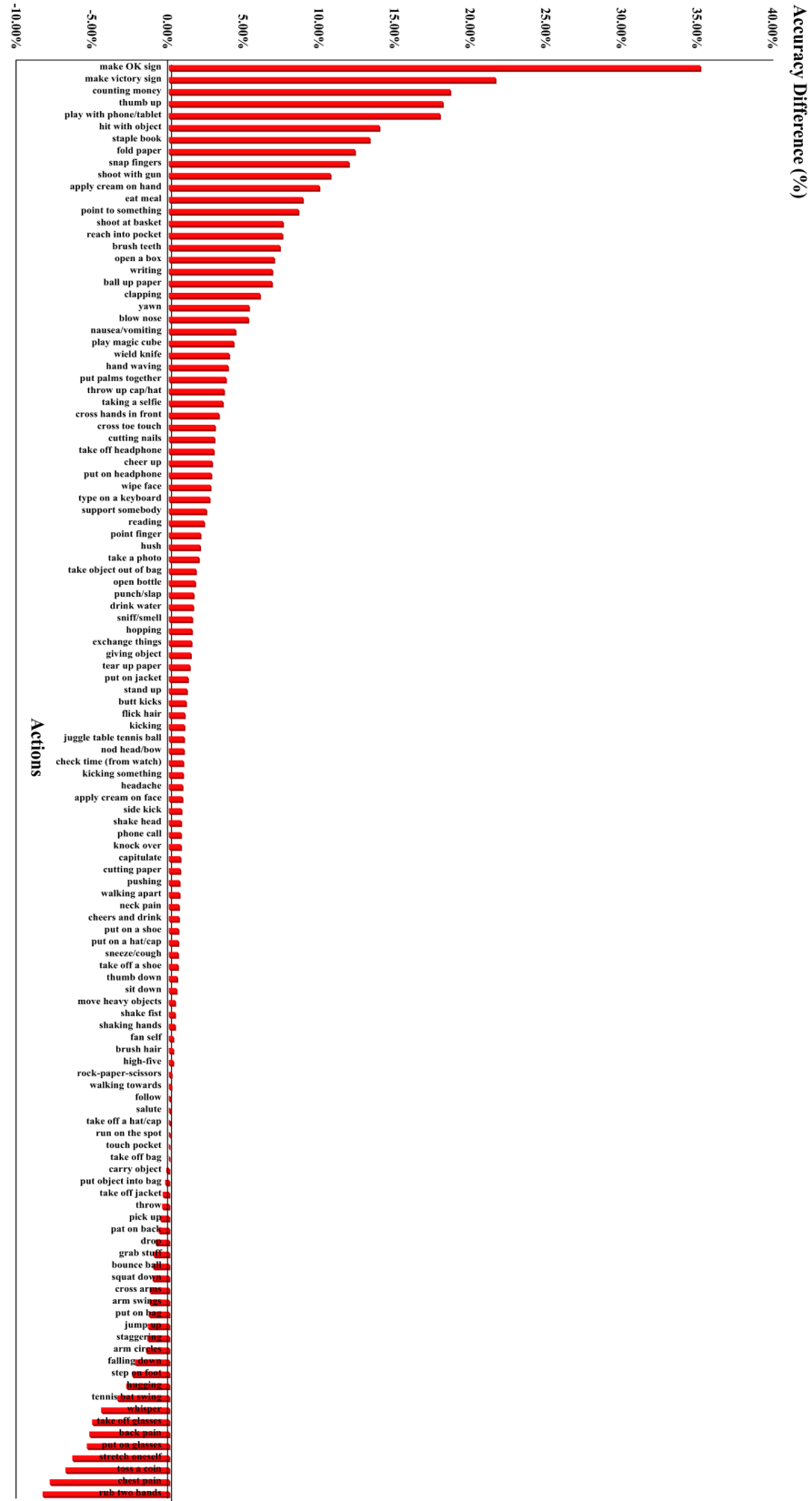


Figure 6: Top-1 Accuracy Difference (%) between the proposed FreqMixFormer and the baseline method SkeMixFormer with the joint input modality on NTU RGB+D 120 X-Sub.

## References

- [1] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13359–13368.
- [2] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20186–20196.
- [3] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [4] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. 2023. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10444–10453.
- [5] Tianjiao Li, Jun Liu, Wei Zhang, and Lingyu Duan. 2020. Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 420–436.
- [6] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16266–16275.
- [7] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 143–152.
- [8] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7912–7921.
- [9] Xuanhan Wang, Yan Dai, Lianli Gao, and Jingkuan Song. 2022. Skeleton-based action recognition via adaptive cross-form learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1670–1678.
- [10] Wentian Xin, Qiguang Miao, Yi Liu, Ruyi Liu, Chi-Man Pun, and Cheng Shi. 2023. Skeleton MixFormer: Multivariate Topology Representation for Skeleton-based Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2211–2220.
- [11] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [12] Yuheng Yang, Haipeng Chen, Zhenguang Liu, Yingda Lyu, Beibei Zhang, Shuang Wu, Zhibo Wang, and Kui Ren. 2023. Action recognition with multi-stream motion modeling and mutual information maximization. *arXiv preprint arXiv:2306.07576* (2023).
- [13] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. 2023. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10608–10617.