Supplementary Materials

This supplementary document provides additional experimental results for the PAMAP2 dataset, complementing the ablation study presented in Section 6 of the main manuscript.

Section S1. Ablation Study on the PAMAP2 Dataset

To further verify the generalizability of the proposed CNN–BiGRU–Transformer model, we conducted the same ablation experiments on the PAMAP2 dataset using the same training and validation settings as in the WISDM dataset experiments. Four comparative models (A–D) were designed as follows: Model A consists of a three-branch CNN only, without temporal modeling; Model B extends Model A by introducing a BiGRU module to capture local temporal dependencies; Model C incorporates a Transformer encoder on top of the CNN to capture global temporal dependencies, without including the BiGRU module or attention mechanism; Model D further extends Model B by combining both BiGRU and Transformer modules to jointly capture short- and long-term temporal dependencies, but without the attention mechanism; finally, the proposed model integrates all modules, including the attention mechanism, to adaptively fuse the extracted spatial and temporal features.

Table S1. Performance comparison of different models on the PAMAP2 dataset.

Model	CNN	BiGRU	Transformer	Attention	Accuracy	Precision	Recall	F1 score
A	✓	×	×	×	88.50%	88.42%	88.45%	88.53%
B	✓	✓	×	×	90.59%	90.52%	90.43%	90.63%
C	✓	×	✓	×	92.68%	92.89%	92.53%	92.41%
D	✓	✓	✓	×	94.31%	94.53%	94.61%	94.59%
Proposed Model	✓	✓	✓	✓	95.62%	96.07%	95.93%	95.95%

As shown in Table S1, the performance of the model is significantly influenced by the combination of modules used. Model A, which relies solely on CNN for spatial feature extraction without temporal modeling, achieves an accuracy of 88.50%, reflecting its limited capacity. When the BiGRU module is added, Model B captures local temporal dependencies, improving its accuracy to 90.59% and F1 score to 90.63%. Model C incorporates the Transformer to model global temporal dependencies but excludes BiGRU and attention, resulting in further improvements, with an accuracy of 92.68% and an F1 score of 92.41%. Model D, which combines both BiGRU and Transformer, allows the model to capture both short- and long-term dependencies, reaching an accuracy of 94.31% and an F1 score of 94.59%. Finally, the proposed CNN–BiGRU–Transformer model, enhanced with the attention mechanism, achieves the best performance, with an accuracy of 95.62% and an F1 score of 95.95%. This demonstrates that integrating CNN, BiGRU, Transformer, and attention significantly enhances spatial–temporal feature extraction and

classification robustness. These ablation results are consistent with the findings on the WISDM dataset, further validating the effectiveness of the proposed model across different datasets.

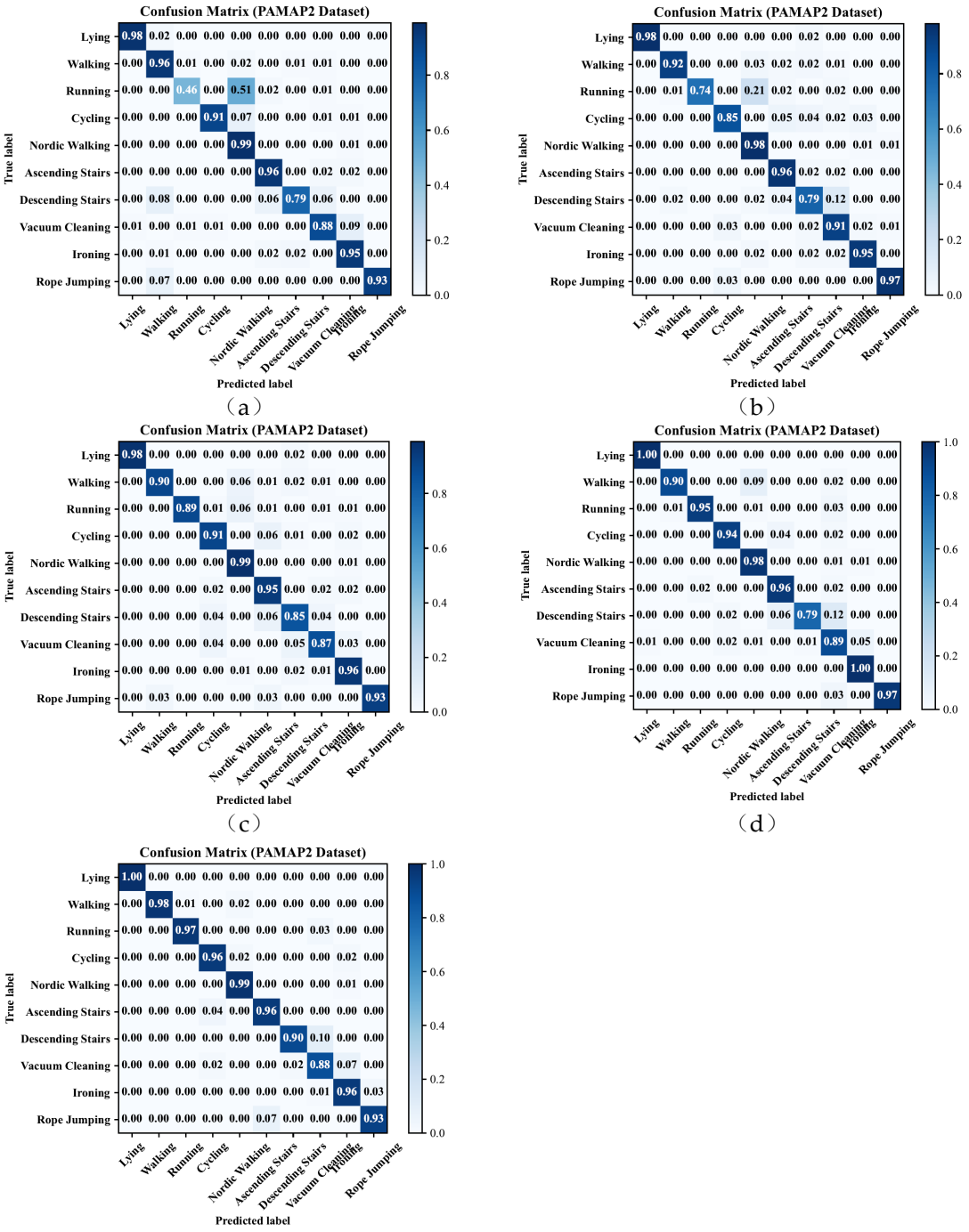


Figure S1. Confusion matrices on the PAMAP2 dataset.

Figure S1 displays the confusion matrices corresponding to the five comparative models (A–D) and the proposed CNN–BiGRU–Transformer model. Each subfigure reflects the classification performance for different activity categories. As the network structure becomes progressively

more complete, the diagonal dominance in the confusion matrices becomes more pronounced, indicating fewer misclassifications. The proposed model shows the clearest diagonal distribution and the fewest cross-category errors, confirming that the inclusion of BiGRU, Transformer, and attention modules enhances recognition accuracy and model robustness on the PAMAP2 dataset.

Overall, the ablation experiments and confusion matrix analyses on the PAMAP2 dataset further verify the effectiveness of the proposed model structure. The progressive improvement across different model variants and the clear diagonal dominance in the confusion matrices demonstrate that the integration of multi-branch CNN, BiGRU, Transformer, and attention modules effectively enhances spatial–temporal feature representation and classification robustness, confirming the rationality and validity of the model design presented in this paper.