Supplementary material
CascadeFormer: A Family of Two-stage Cascading Transformers for Skeleton-based Human Action Recognition

Anonymous submission

More Ablation Studies

We conduct further ablation studies to examine the effects of input representations, decoder architectures, and backbone freezing strategies.

Input Data Representation

We investigate alternative input data representations beyond the original joint coordinates using **CascadeFormer 1.0**, as shown in Table 1. Surprisingly, the use of raw joint coordinates yields the highest accuracy of **94.66%** on the Penn Action dataset. To explore the effectiveness of bone-based representations, we develop three variants. The first approach constructs each bone by subtracting the coordinates of one joint from its adjacent joint, resulting in an accuracy of 92.32%. The second method concatenates the coordinates of two consecutive joints, achieving 93.16% accuracy. The third approach linearly parameterizes each bone segment using its slope and intercept, which attains 93.91% accuracy. Overall, although all bone-based variants perform competitively, the original joint representation proves to be the most effective for our model on this task.

Data Representation	Accuracy
Joints	**94.66%**
Bones (Subtraction)	92.32%
Bones (Concatenation)	93.16%
Bones (Parameterization)	93.91%

Table 1: **Comparison of different input data representations on Penn Action**. Using the original joint coordinates achieves the highest accuracy (94.66%), outperforming all three alternative bone-based variants.

Decoder Architecture

We conduct an ablation study to compare alternative decoder architectures for masked pretraining on the *CascadeFormer 1.0* using the Penn Action dataset. As shown in Table 2, our default choice—a simple linear layer to reconstruct masked joints—achieves the highest accuracy of **94.66%**. We then evaluate an MLP decoder composed of two linear layers with a ReLU activation in between, which yields a reduced accuracy of 92.51%. Finally, we test an MLP decoder with a residual connection to facilitate gradient flow, resulting in an even lower accuracy of 91.20%. These results suggest that the linear decoder not only provides the most effective reconstruction but also generalizes better, likely due to its lower risk of overfitting on small-scale datasets. Based on this finding, we adopt the linear decoder for all experiments throughout this work.

Decoder Architecture	Accuracy
linear	**94.66%**
MLP	92.51%
MLP + residual	91.20%

Table 2: **Comparison of decoder architectures during masked pretraining on Penn Action**. A simple linear decoder outperforms both MLP and MLP with residual connection, indicating that increased decoder complexity may lead to overfitting on smaller datasets.

Backbone Freezing Decision

In this ablation study, we examine the impact of parameter-freezing strategies for the transformer backbone during the cascading finetuning stage. Using *CascadeFormer 1.0* on the Penn Action dataset, we present our findings in Table 3. Our primary approach involves fully finetuning the entire backbone, allowing all transformer parameters to be updated during training. This strategy yields the best performance with an accuracy of **94.66%**. In contrast, freezing the entire backbone results in a significant drop in accuracy to 85.11%. We also explore partial finetuning by training only the final transformer layer, which achieves 88.39% accuracy. These results suggest that full backbone finetuning is crucial for effective downstream adaptation in action recognition. Consequently, we adopt full backbone finetuning for all experiments.

Backbone Freezing Decision	Accuracy
fully finetune	**94.66%**
fully freeze	85.11%
finetune the last layer	88.39%

Table 3: **Effect of different backbone freezing strategies during cascading finetuning on Penn Action.** Fully finetuning the transformer backbone yields the highest accuracy, while freezing all layers significantly degrades performance.