PromptGAR: Flexible Promptive Group Activity Recognition

Supplementary Material

Method	Prompt Types				**Head Inputs**		Re-train	Top1	Mean
	RGB	Bbox	Kpt	InstID	$\widehat{X}_{gar}$	$\widehat{F}_p$			
PromptGAR	✓	✓	✓	✓	✓	✓	×	**96.0**	**96.3**
	✓		✓	✓	✓	✓		95.7	95.8
	✓	✓		✓	✓	✓		94.1	94.3
	✓	✓	✓	✓	✓			95.8	96.0
	✓		✓	✓	✓			95.5	95.7
	✓	✓		✓	✓			94.0	94.2
	✓				✓			93.9	94.4
	✓	✓	✓	✓		✓		95.2	95.6
	✓		✓	✓		✓		95.2	95.6
	✓	✓		✓		✓		93.4	94.0

Table 10. **Ablation on Head Input Tokens**.

Method	Train				Test				Top1	Mean
	RGB	Bbox	Kpt	InstID	RGB	Bbox	Kpt	InstID	Acc	Acc
PromptGAR	✓	✓	✓	✓	✓	✓	✓	✓	**96.0**	**96.3**
					✓		✓	✓	95.7	95.8
	✓	✓		✓	✓	✓		✓	95.4	95.8
	✓				✓				94.3	94.7

Table 11. **Low Reliance on Skeletons**.

O	Memory	Top1 Acc	Mean Acc
16	40 *GB*	95.8	96.0
48	72 *GB*	**96.0**	**96.3**
56	*OOM*	-	-

Table 12. **# of Pooled Prompts** O.

ϵ	Top1 Acc	Mean Acc
0	95.4	95.6
5	95.6	95.9
10	**95.8**	**96.0**

Table 13. **Relative Instance Scale** ϵ.

1. More on Implementation Details

Data Augmentations. Following MViTv2 in action recognition [27], we apply a comprehensive set of augmentations for group activity recognition (GAR): Random Augment [10], Random Resized Crop, Flip, Random Erasing [53], CutMix [49], and MixUp [51].

Volleyball Dataset [20]. Since Volleyball only contains annotations with central 16 frames, we replace the depth-wise prompt pooling with a single MLP layer projecting channels from $T \times 20$ to the number of pooled prompts O.

NBA Dataset [45]. We regenerated annotations due to inaccuracies in SAM [45] and MP-GCN [28], such as background audience false positives, player ID switching, and ID reassignment upon reappearance. We used MOTIP [14] for robust player tracking, optimized for sports data, and RTMPose [21] for accurate skeleton generation, handling occlusions effectively.

2. Additional Ablation Studies

Head Inputs. Tab. 10 indicates demonstrates the significance of prompt token inputs within the head, even when dealing with diverse visual prompts. We compared performance using a model trained on full prompts, varying the head inputs during testing. Under full prompt inputs, using only the GAR class token $\widehat{\mathbf{X}}_{gar}$ as the head input resulted in a minor 0.2% performance drop, while using only the prompt tokens $\widehat{\mathbf{F}}_p$ led to a 0.8% drop. This shows that both inputs are crucial for PromptGAR's high performance. Even with reduced prompt inputs, both $\widehat{\mathbf{X}}_{gar}$ only and $\widehat{\mathbf{F}}_p$ only still yielded reasonable results. Notably, when using only prompt tokens $\widehat{\mathbf{F}}_p$, the performance of full prompt inputs was similar to the one with skeleton-only inputs. This suggests that when RGB features don't directly influence the final prediction, bounding box information is effectively captured within the skeleton data. In summary, this table showcases the flexibility of head inputs across various visual prompts and underscores the importance of both the RGB feature representation $\widehat{\mathbf{X}}_{gar}$ and the prompt feature representation $\widehat{\mathbf{F}}_p$ for optimal performance.

Low Reliance on Skeletons. Tab. 11, our model has competitive performance even when trained without skeletal data. This outcome indicates that our model does not heavily rely on skeletal information.

Number of Pooled Prompts. Tab. 12 explores the impact of the number of pooled prompt channels (O) on performance. The prompt encoder's output has the shape $D \times N \times O$, where D, N, O are the feature dimension, the number of instances, and the pooled prompt channels after pooling along temporal and prompt type dimensions, repectively. A smaller O value reduces the representativeness and expressiveness of each actor's key information. As shown in the table, increasing O from 16 to 48 improves top1 accuracy by 0.2%. However, this increase quadratically raises the computational cost and CUDA memory usage of the subsequent recognition decoder module. Specifically, the CUDA memory usage for $O = 48$ is 32 GB higher than for $O = 16$. Further increasing O to 56 would exceed the 80 GB limit of an A100 GPU's CUDA memory, resulting in an out-of-memory error. Therefore, while a higher O value improves performance, a balance must be struck to avoid excessive computational demands and memory limitations.

Relative Instance Scale. Tab. 13 details the process of determining the optimal relative instance scale, denoted as ϵ. This value serves as a coefficient for the relative instance weights, which are added to the standard attention

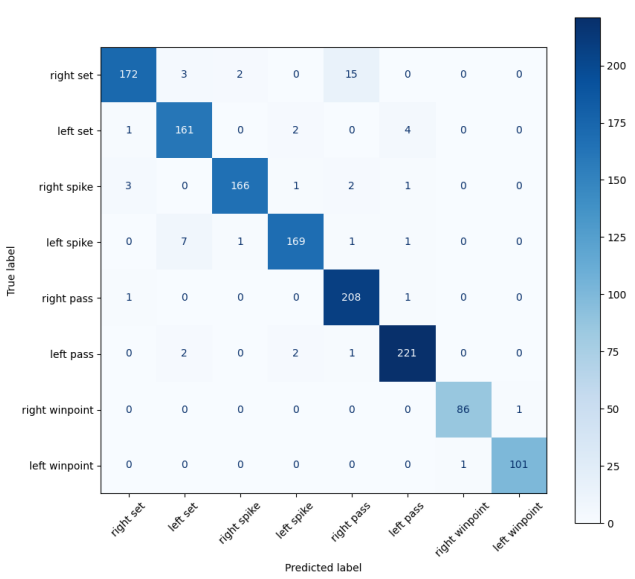

Figure 8. **Volleyball Confusion Matrix**

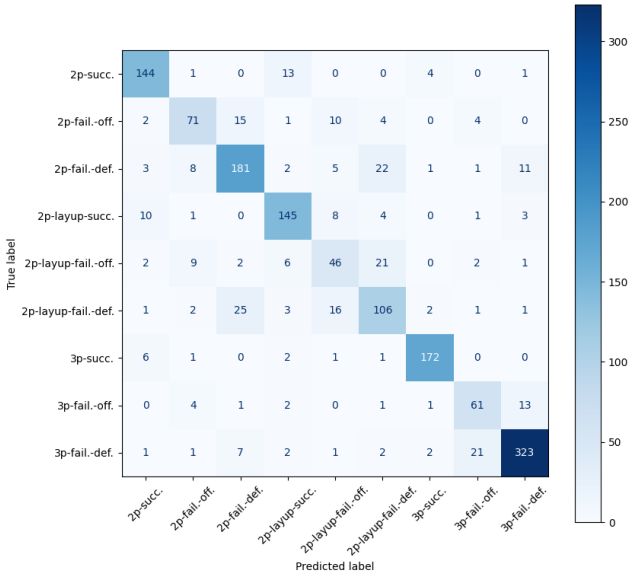

Figure 9. **NBA Confusion Matrix**

weights. Due to the inherent sparsity of the relative instance weight matrix $E^{(rel)}$, a small ϵ value would diminish the impact of instance *ID*s during training. Conversely, an excessively large ϵ would overpower the standard attention weights, hindering the GAR class token's ability to learn effectively from the prompt tokens. The table reveals that setting ϵ to 10 yields a 0.4% improvement in top-1 accuracy compared to setting it to 0 (effectively disabling relative instance attention) and a 0.2% improvement compared to setting it to 5. Therefore, we selected 10 as the final ϵ value. Notice that these experiments are conducted with a pooled prompt channel count (O) of 16.

3. Analysis

Quantitative Results. Fig. 8 - 9 shows the confusion matrix of Volleyball and NBA, respectively. (a) For Volleyball, while PromptGAR demonstrates high accuracy on the Volleyball dataset, a common misclassification occurs where '*right-set*' is mistaken for '*right-pass*'. This error is attributed to the inherent similarity in player actions and positions, echoing findings in previous research [54]. (b) For NBA, errors highlight the challenge of distinguishing between highly similar actions. Specifically, differentiating '*offensive*' and '*defensive*' requires nuanced visual analysis of player uniform differences and rebound outcomes. Similarly, distinguishing '*2p-fail*' from '*2p-layup-fail*' demands attention to shooting position and posture near the rim.