A. Implementation Details

Evaluation metrics We evaluate our method and other state-of-the-art methods using Top-1 and Top-5 accuracy scores, where the predictions are considered to be correct if the top 1 or top 5 highest probability answers match the actual label.
Implementation Details All models in this paper are trained using NVIDIA GeForce 2080Ti GPUs and NVIDIA RTX A5000 GPUs. The initial learning rate is set at 0.1 for training from scratch and 0.05 for initializing with Kinetics pretrained weights. Stochastic Gradient Descent (SGD) is used as the optimizer with 0.0005 weight decay and 0.9 momentum. We use cosine/poly annealing for learning rate decay and multi-class cross entropy loss to constrain the final predictions. Unless further specified, the videos are decoded as a single clip and all the frames are randomly scaled and center cropped to the size 224×224 during training. During testing, we scale the shorter spatial side to 256 and take 3 crops of 224×224 to cover the longer spatial axis. We average the scores for all individual predictions.

B. Incorporate with different recognition backbone models

We further demonstrate that our method could be used with different recognition backbone models to improve the accuracy. We compare the results using our PMI Sampler as well as uniform sampling and MG Sampler on 3 different backbones: SlowOnly-R50 [11], I3D [3] and X3D [10]. Results on Table 9 show that our method consistently outperforms other methods and brings accuracy improvement across different backbone models.

Method	Frames	Backbone	Top-1 Acc (%)	Top-5 Acc (%)
Uniform	8	I3D [3]	59.2	89.9
MG Sampler [48]	8	I3D [3]	55.2	87.6
Ours	8	I3D [3]	**61.8**	**91.7**
Uniform	8	SlowOnly-R50 [11]	60.0	90.3
MG Sampler [48]	8	SlowOnly-R50 [11]	57.1	88.4
Ours	8	SlowOnly-R50 [11]	**63.1**	**93.5**
Uniform	16	X3D [10]	73.5	95.1
MG Sampler [48]	16	X3D [10]	74.6	95.0
Ours	16	X3D [10]	**81.3**	**97.7**

Table 9. Evaluate our method with different recognition backbones on Diving48. PMI Sampler can be incorporated with any recognition backbone models to improve the accuracy.

Method	Frames	Backbone	Top-1 (%)	Top-5 (%)
Uniform	8	X3D-M [10]	74.0	95.3
MG Sampler [48]	8	X3D-M [10]	74.6	95.9
Ours	8	X3D-M [10]	**75.5**	**96.0**

Table 10. Our propposed PMI Sampler can be used in dense clip sampling for improved accuracy. We demonstrate an relative improvement in top-1 accuracy over baseline method by 2% and 1.2% over SOTA.

C. Use in dense clip sampling

Our proposed PMI Sampler could also be used in dense clip sampling during training. We uniformly sample the videos into 10 clips and for each clip, we adaptively select 8 frames. The baseline method is to uniformly select those frames in each clip. We evaluate the performance of our proposed method along with the current state-of-the-art MG Sampler in dense clip sampling scenario. As shown in Table 10, our method achieves a relative improvement over the baseline method by 2% and 1.2% over SOTA.

D. Analysis

In aerial videos, the human actor occupies less than 10% resolutions and the rest pixels belong to the background. When the camera is moving, the overall background deviations are much larger than the actual motion changes. Therefore, pixel-wise RGB difference used in MG Sampler [49] will be dominated by background noises and fails to map the motion distribution for both videos in Figure 1. However, our proposed patch mutual information is more robust because of the inherent advantage of mutual information. Mutual information measures the image similarity only by considering the overall pixel value distribution in the two images, see Eq 4. Thus, it is more robust to outliers and noises. However, it ignores spatial information between pixels and that is very important for action recognition. Patch mutual information avoids such issues by dividing the frames into patches and measuring the mutual information of small patches. In this way, the spatial information within the patches can be conserved. Because of the robustness of PMI, we can further employ the shifted Leaky ReLu to make the motion-salient frames easier to distinguish. As shown in Figure 1, PMI Sampler ensures that the sampled frame comprehensively covers all the essential segments with high motion salience, so that key information about the somersault in Diving48 may not be missed. Also, PMI Sampler is robust to background noises. It can identify the motion static period even when the camera is shaking in UAV videos, see Figure 1, selecting more frames from the motion salient period and fewer frames from the motion static period.

E. More Visualization Results

We generate more visualization results between our method and current state-of-the-art method, MG Sampler [48], on the three datasets: UAV-Human [24], NEC-Drone [4] and Diving48 [25] in Figure 5,7,6,8.

As mentioned in Section 3.3, our proposed method quantifies the motion information contained in adjacent frames based on the similarity measure between corresponding frame patches. As for trimmed videos, human actors are performing scripted actions in the same scene and the backgrounds are always similar in the same video. Therefore, our

patch similarity guided frame selection strategy is more robust to background noises. Moreover, since the backgrounds are similar, the unsimilarity is dominated by human actions, thus our method yields to better motion information representations. As shown in Figure 5, 6, when the camera is moving and no salient motion exists, MG Sampler [48] suffers from the pixel value changes corresponding to the backgrounds. However, our proposed PMI Sampler can accurately identify the motion static periods.

Our method also gives more accurate motion information distribution for aerial videos and makes it much easier to distinguish the motion salient frames, see Figure 7. As mentioned in Section D, this is attributed to our proposed patch mutual information score, which considers both the pixel distribution and spatial relationships inside the patches. Our method can select more frames from the motion salient periods and fewer frames from the motion static periods.

F. Limitations

Our proposed method may have two limitations. First, as shown in Figure 8, when the motion is consistent and smooth (such as swing the racket, drink, rub hands) during the whole video, our method will perform just like uniform sampling. Second, in the case where the action label is highly associated with the gesture during motion static periods, our method may be less effective due to the reduced number of sampled frames during these periods. For instance, as depicted in the second video in Figure 8 with label "all clear", the action is primarily determined by the static gesture between frames 12 and 28. However, our approach samples fewer frames during this period, which might hinder the performance. To mitigate this issue, we can adjust the value of α in the shifted Leaky ReLu to achieve a smoother motion information distribution, enabling the selection of more frames during such periods. Nonetheless, further investigation is necessary to comprehensively address this concern.

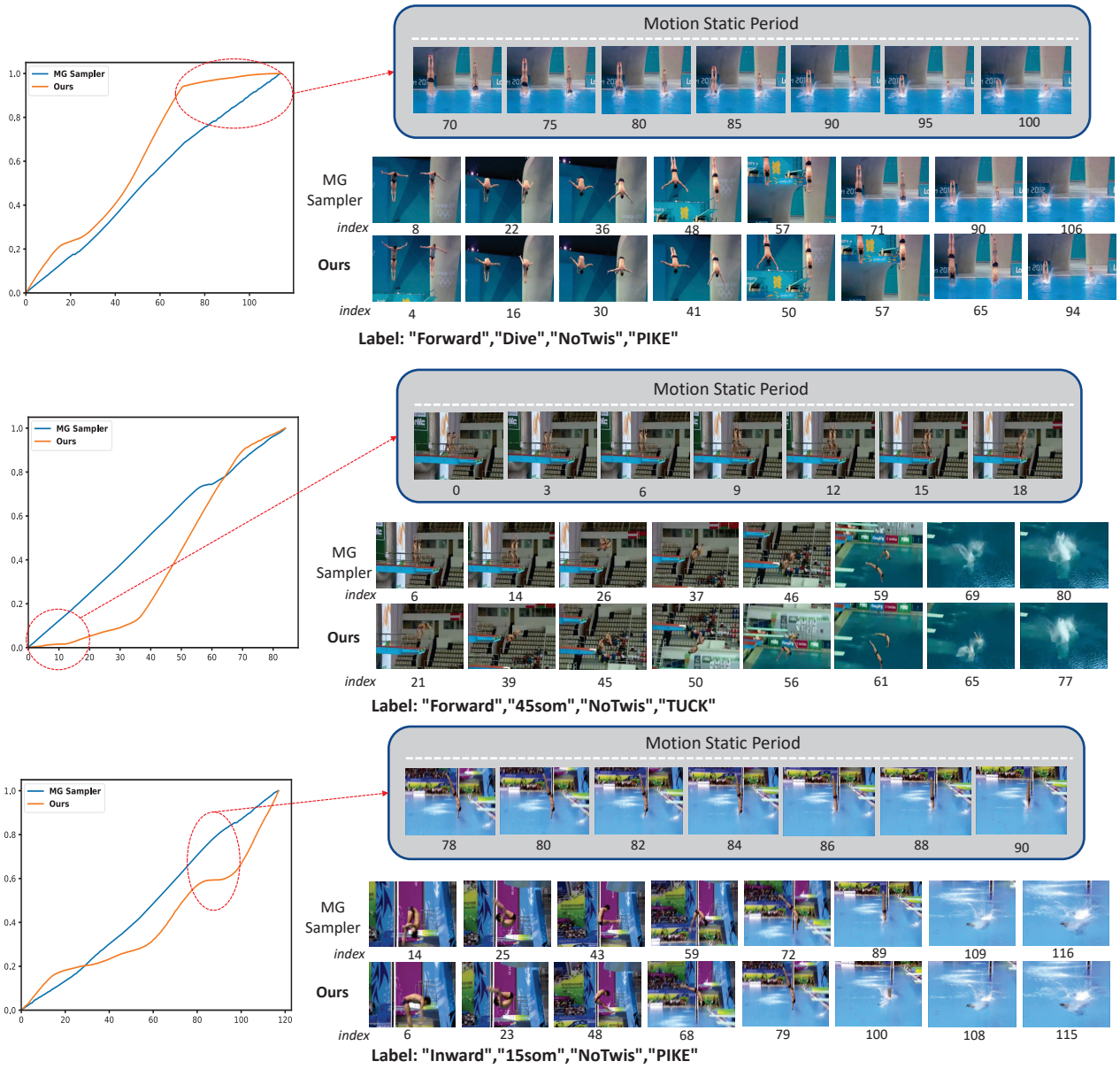

Figure 5. Comparison between our method and MG Sampler [48] on typical videos from Diving48 [25]. As shown above, MG Sampler fails to measure the motion information between frames and cannot reflect the motion distribution of the video because of the background changes caused by the camera moving. However, our method is more robust to background noises and could accurately identify the motion static periods in the start, middle and end of the videos.

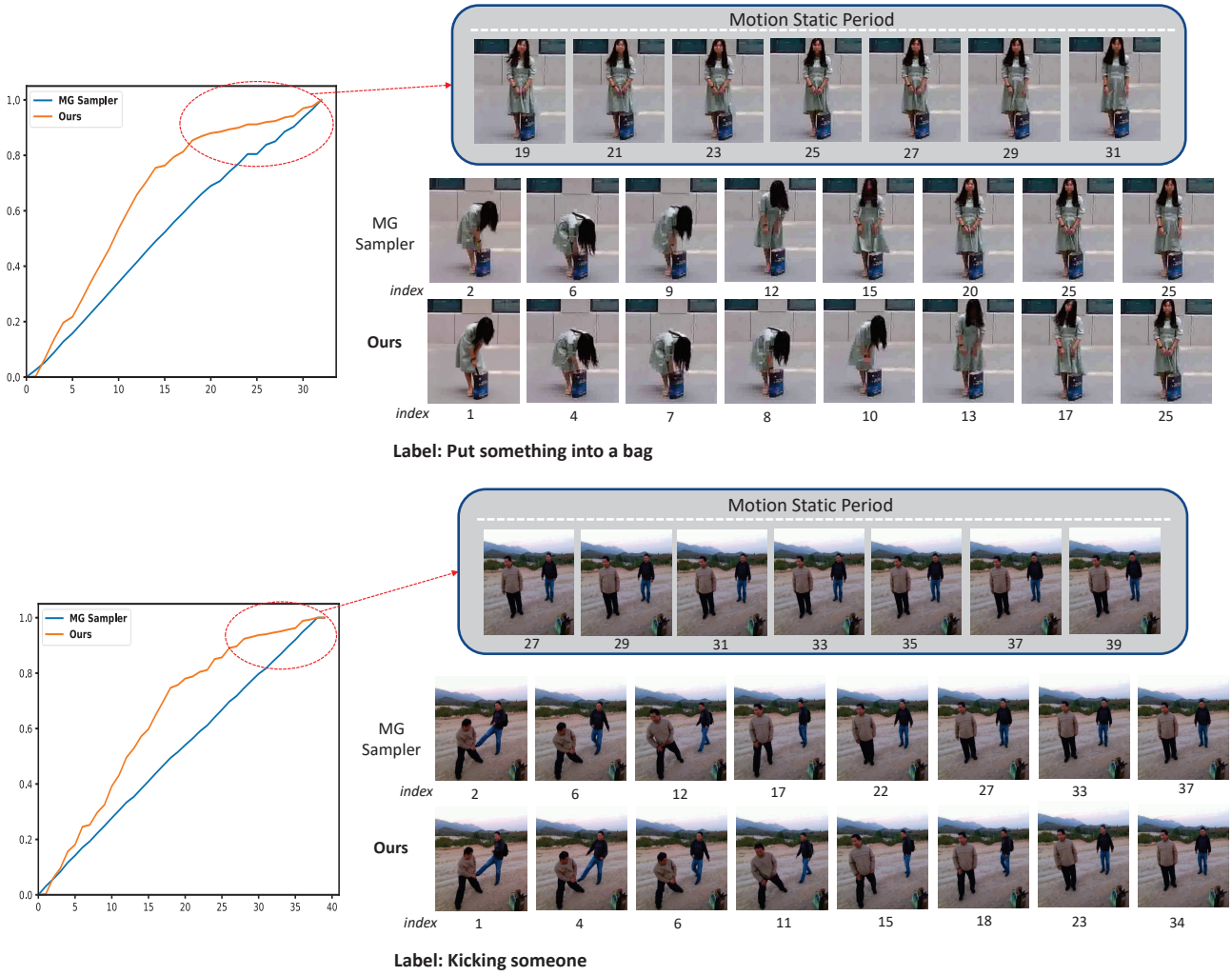

Figure 6. Comparisons between our method and MG Sampler [48] on typical videos from UAV-Human [24] and NEC-Drone [4]. Compare to Diving48 [25], UAV videos are more shaky and most pixels are corresponding to backgrounds(frames in the figure are cropped for better visualization). Therefore, they contain more background noises. MG Sampler fails to handle such challenges from UAV videos. However, due to the robustness of our proposed patch mutual information, our method could accurately distinguish the motion static period.

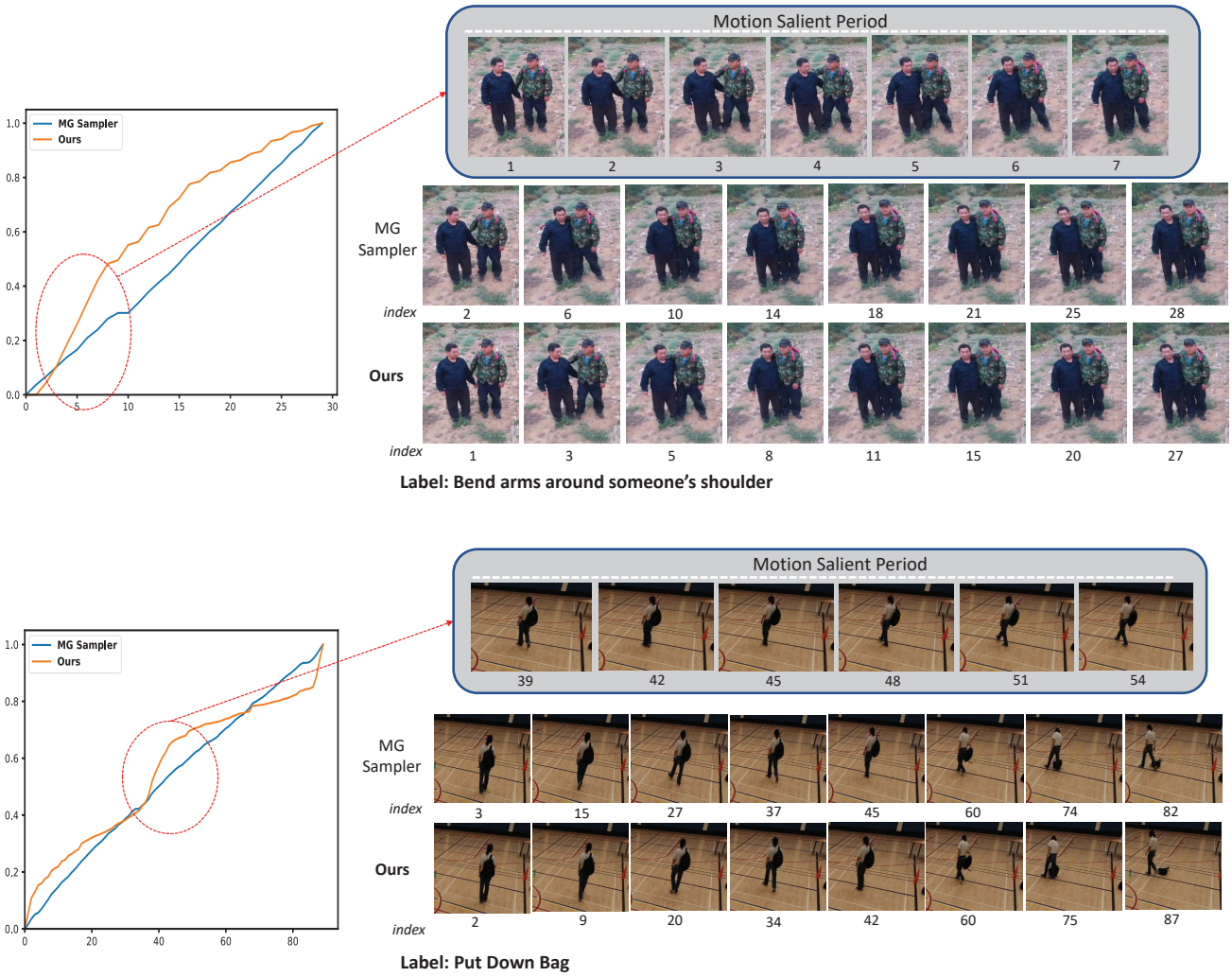

Figure 7. More comparisons between our method and MG Sampler [48] on typical videos from UAV-Human [24] and NEC-Drone [4]. Our method makes it easier to distinguish the motion salient frames. Our method selects more frames from the motion salient periods and less frames from motion static period, so that sampled frames contain more useful motion information.

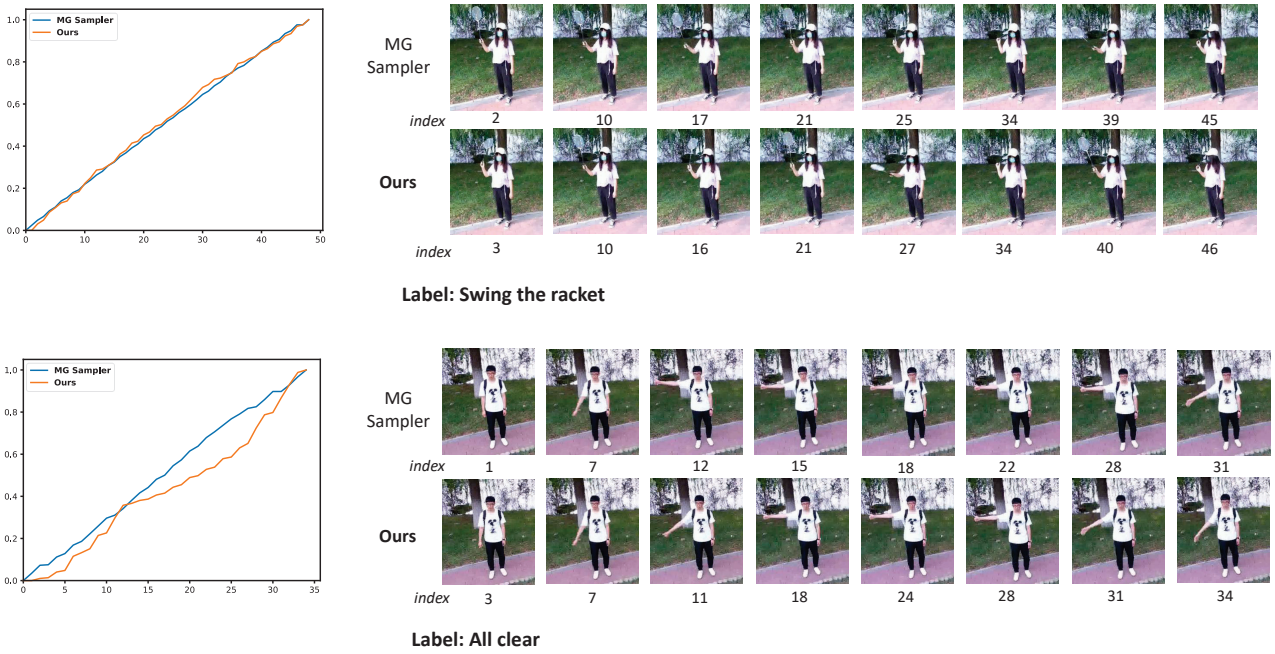

Figure 8. There are two limitations of our method. First, as shown in the first video (with label: swing the racket), when motion is consistent and smooth during the whole video, our method will perform just like uniform sampling. Second, in instances where the action label is highly associated with the gesture during motion static periods, our method may be less effective due to the reduced number of sampled frames during these periods. As shown in the second video above (with label: all clear), the action is primarily determined by the static gesture between frame 12 to frame 28. However, our method samples fewer frames from such period. To mitigate this issue, we can adjust the value of α in the shifted Leaky ReLu to achieve a smoother motion information distribution, enabling the selection of more frames during such periods. Nonetheless, further investigation is necessary to comprehensively address this concern.