Some in-depth analysis of the failed cases on HMDB51 in Fig. 5

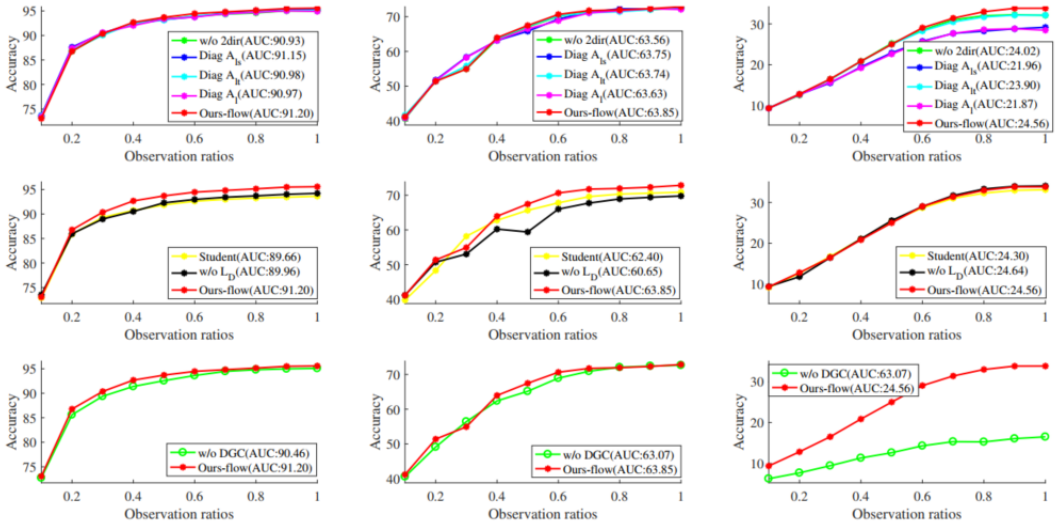

Fig. 5. Detailed results of ablative studies. (∗) denotes the performance of the corresponding method in terms of AUC. The left one denotes the results on UCF101, the middle one denotes the results on HMDB51, and the right one denotes the results on Sth-v2.

Fig. 5 indicates that for HMDB51, when the Observation ratio is approximately 0.3, Student's performance surpasses Ours-Flow.We conduct an in-depth analysis of the failed cases, which is that the "student" can correctly identify it while "Ours-flow" makes a wrong prediction. In Figure R-1, we present the detailed failed results at an observation ratio of 0.3 on HMDB51 (split3). The horizontal axis represents the action samples, while the vertical axis represents the corresponding action labels. To conduct an in-depth analysis, we have also included the predicted results of our final model (Ours(9)).

Through a thorough analysis of the results presented in Fig. R-1, we found that the primary factors contributing to the incorrect predictions of "Ours-flow" are background noise (highlighted in blue boxes), similar actions (highlighted in black boxes), and coupled actions (highlighted in purple boxes).

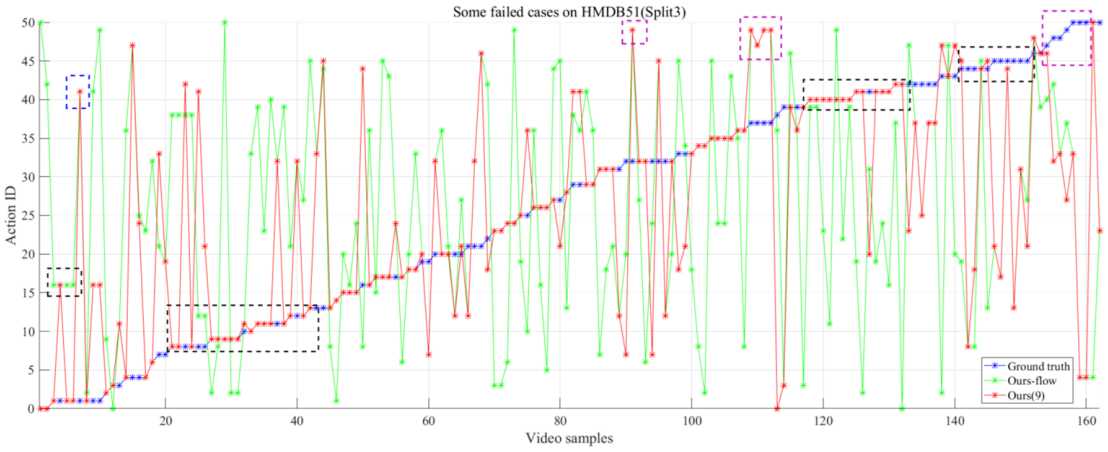

Fig. R-1 Detailed results of some failed cases

(1) Background noise

Fig. R-2 illustrates an example of incorrect predictions that may be caused by

background noise. The top video in the figure represents a false prediction, while the bottom video shows an example of the action “somersault”. As can be seen in Fig. R-2, the action “cartwheel” was falsely predicted as “somersault”. This error may have been due to two main factors: the presence of significant background noise and the short duration of the video (only 1 second). These factors contribute to significant semantic ambiguity at low observation ratios, particularly when there are actions with similar movements, such as the similar actions mentioned below. It is likely that these factors combined led to the incorrect predictions observed in this case.

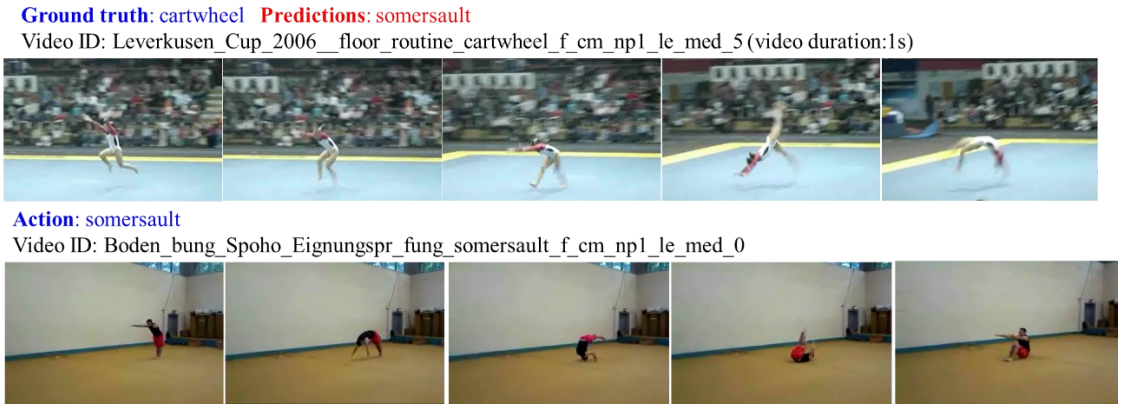

Fig. R-2 Examples of incorrect predictions by the background noise (highlighted in blue boxes Fig. R-1)

(2) Similar actions

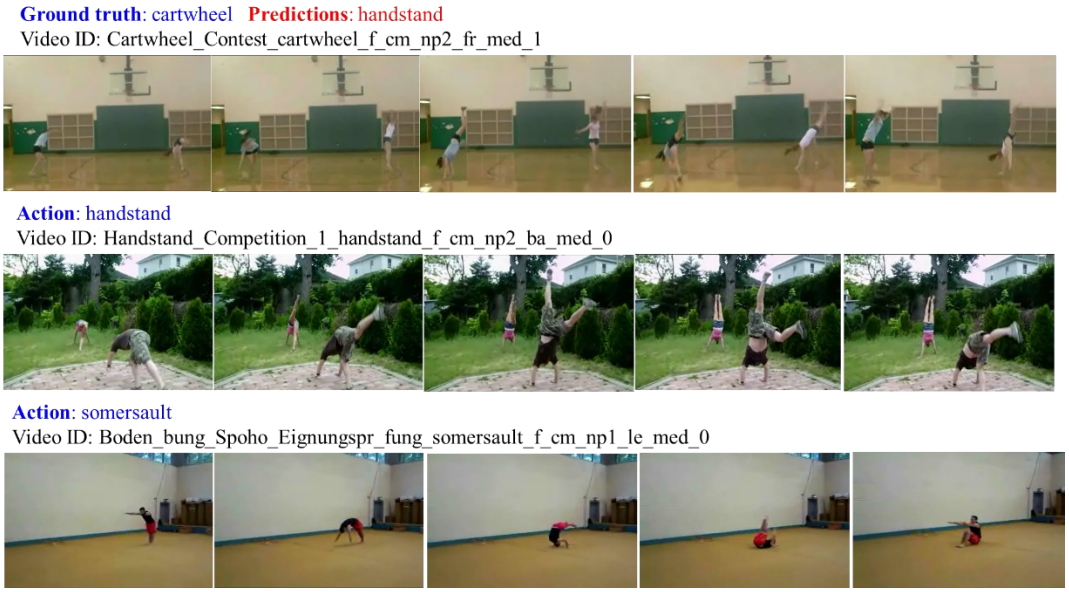

Fig. R-3 Examples of incorrect predictions by the similar actions (highlighted in black boxes)

Fig. R-3 provides examples of similar actions, including “cartwheel”, “handstand”, and “somersault”. As shown in the figure, these actions share similar movements, which can lead to incorrect predictions when using flow features for predictions without RGB features. This is because flow features mainly focus on motion features, losing the crucial spatial features that distinguish between similar actions. As highlighted in blue

boxes in Fig. R-1, most error predictions are caused by similar actions. Therefore, by fusing RGB features with flow features, many of these incorrect predictions can be corrected.

(3) Coupled actions

Fig. R-4 provides examples of coupled actions, where the same video may contain multiple actions but is labeled as only one of them. This can be attributed to two main factors: partial videos with incomplete action knowledge, leading to semantic ambiguity, and the distillation framework potentially introduces more irrelevant actions from full videos. These factors increase the semantic ambiguity of the partial video, resulting in incorrect predictions.

Fig. R-4 Examples of incorrect predictions by the coupled actions (highlighted in purple boxes)

Based on the deep analysis of failed cases, we make the three explanations for the low accuracy at an observation ratio of 0.3. (a) Background noise. With the knowledge distillation (Our-flow), the irrelevant motion background noise may be introduced. In addition, the short durations of many videos are very short, and their action information is minimal at low observation ratios, resulting in a high semantic ambiguity. Therefore, introducing motion background noise makes it even more likely to make wrong predictions, as shown in Fig. R-2. (b) Similar actions. Due to the missing scene-context information of RGB, “Our-flow” is hard to handle the highly similar actions, such as ‘handstand’, ‘cartwheel’, and ‘somersault’, especially at the low observation ratios. Through a deeper examination of our results, we found that “Our-flow” tends to predict similar actions as other very similar actions at an observation rate of 0.3, and the results can be corrected after fusing the scene-context of RGB, as highlighted in black boxes in Fig. R-1. (c) Coupled actions. There are some coupled actions on the HMDB51 dataset, as shown in Fig. R-4. For example, ‘smoke’ and ‘smile’ actions appear

simultaneously in one video, but the video is labeled as one of the actions. The knowledge distillation framework may introduce other action knowledge from full videos and thus cause wrong predictions, especially when their action semantics are ambiguous at the low observation ratios, leading to low accuracy.

We also visualize the features at the observation ratios of 0.3 and 0.4, as shown in Fig. R5 ~ Fig. R7. For the similar actions "golf (the marked green)" and "shotgun (the marked dark red)", as denoted in Fig. R6, their action semantics are ambiguous. As shown in Fig. R5 and Fig. R6, "Ours-flow" may distill more background noise or irrelevant action knowledge from the teacher network, leading to a decrease in accuracy at an observation rate of 0.3. When increasing the observation ratios, the action semantic ambiguity may decrease. "Ours-flow" will provide more helpful information from the teacher network for predictions.

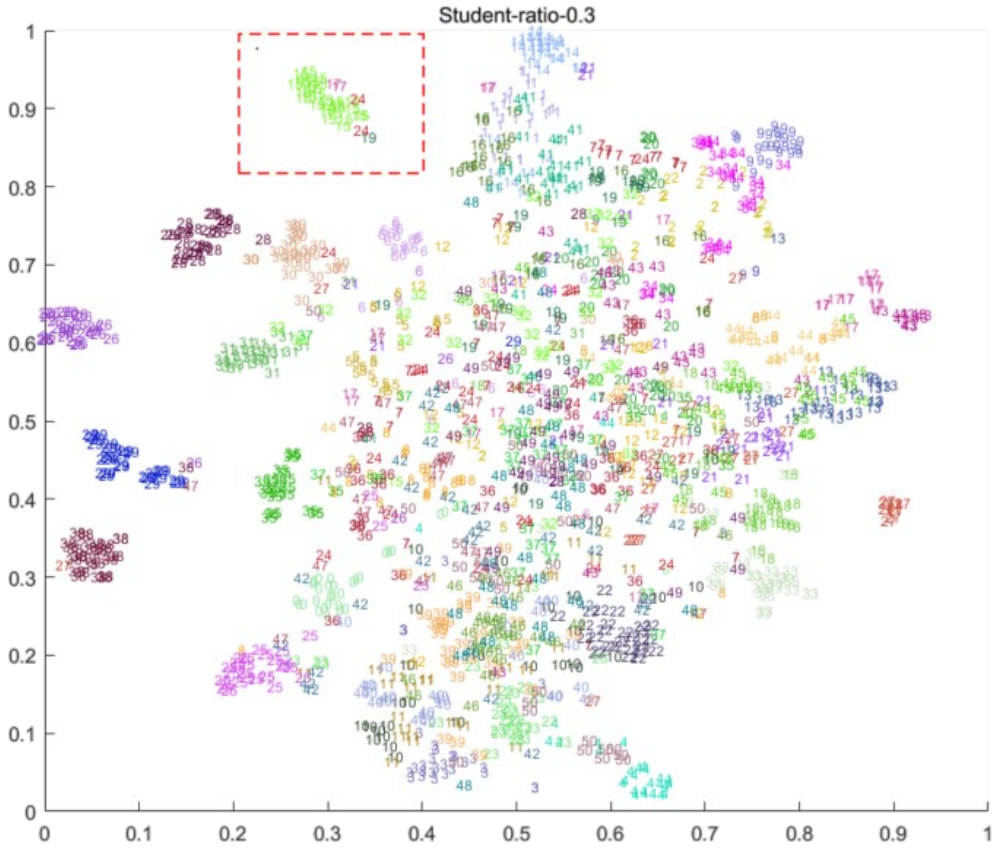

Fig.R5 *t*-SNE visualization of "Student" at an observation of 0.3

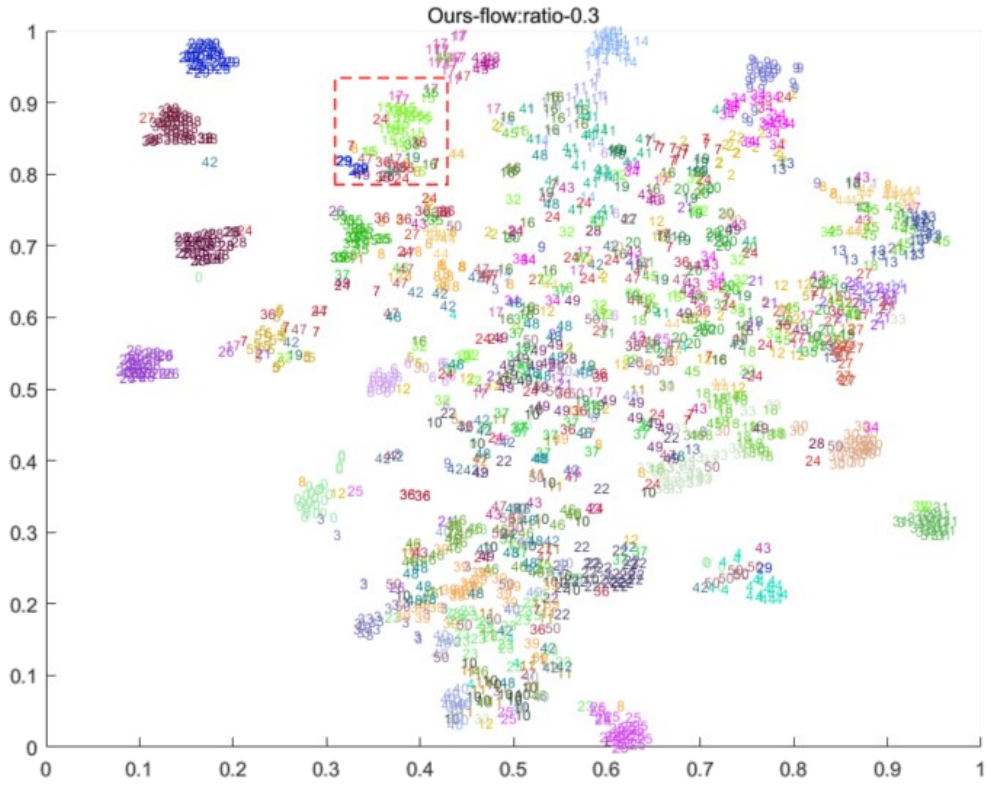

Fig.R6 *t*-SNE visualization of "Ours-flow" at an observation of 0.3

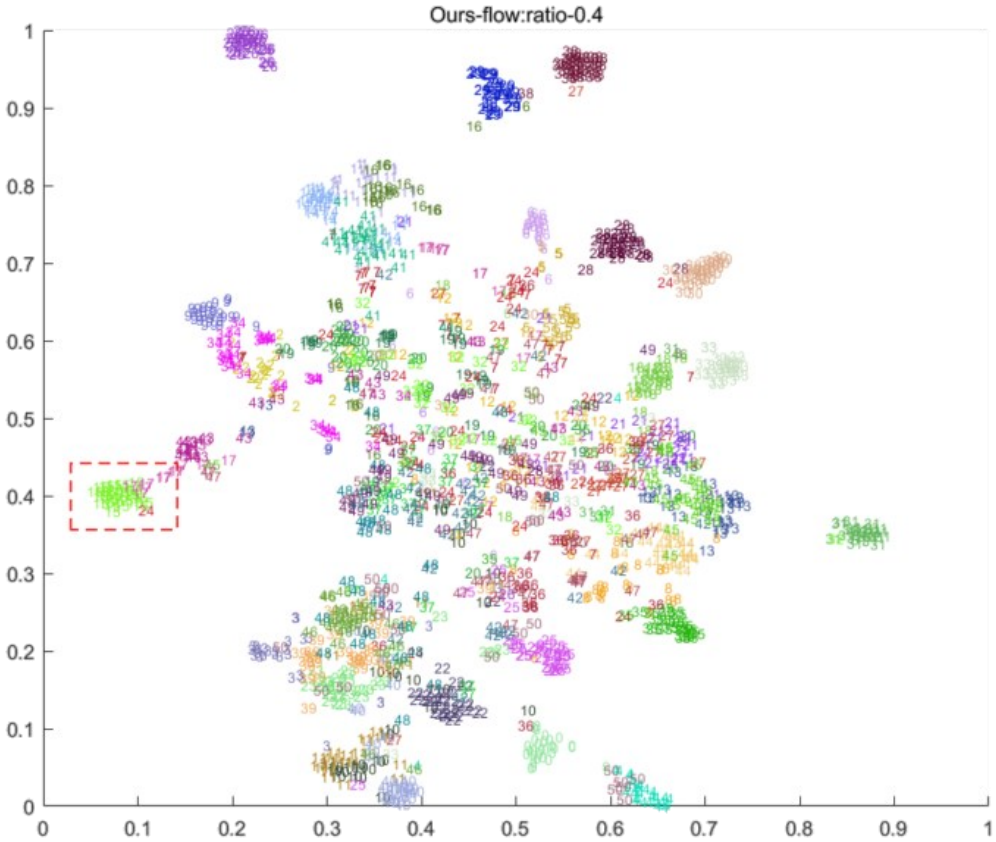

Fig.R7 *t*-SNE visualization of Student's model at observation of 0.4