

# 1 VISUALIZATIONS FOR R1:MDP9

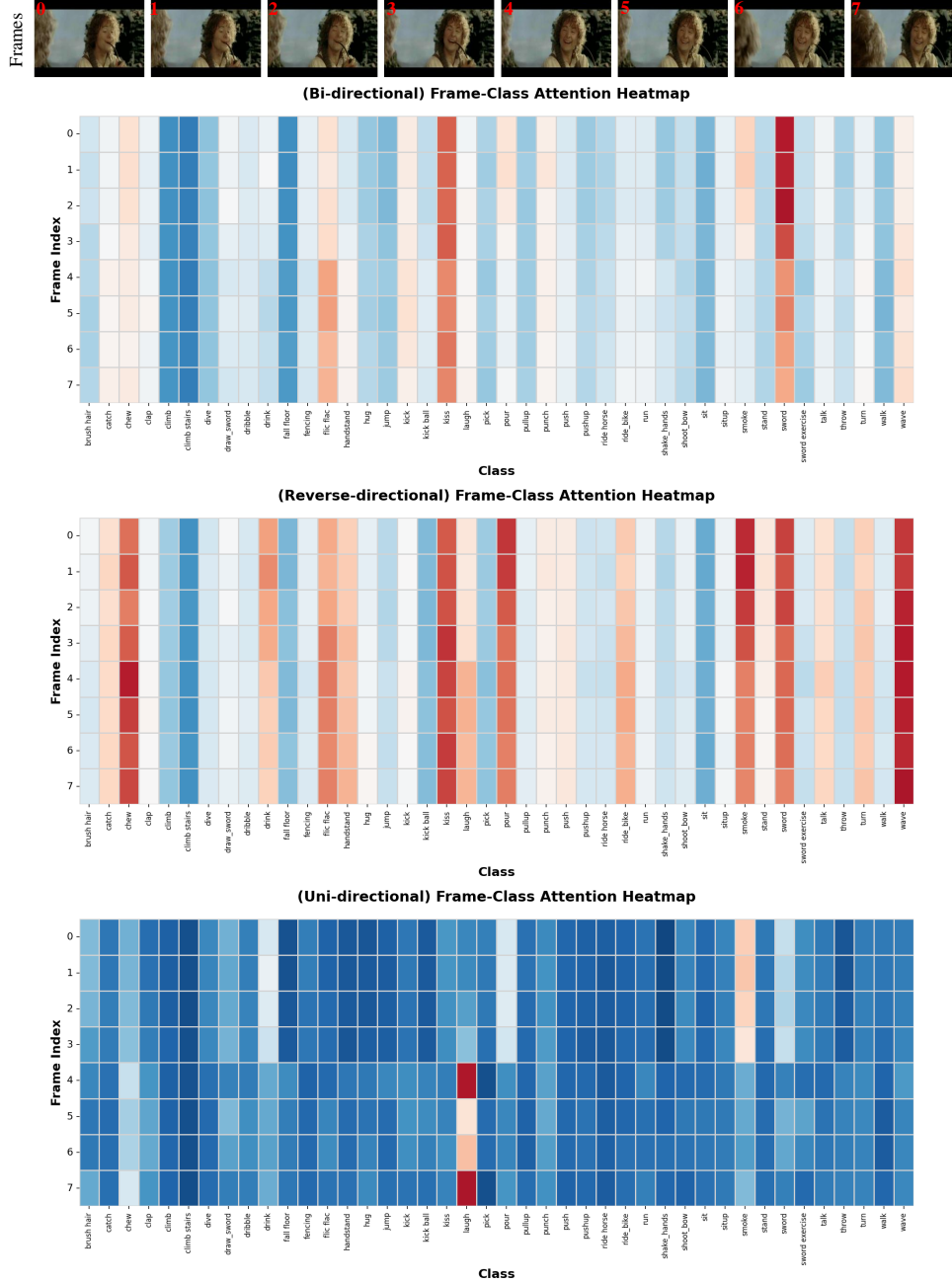


Figure 1: Qualitative comparison of frame-class attention patterns under different blending directions on HMDB51. From left to right, we visualize (i) bi-directional, (ii) reverse-directional, and (iii) our uni-directional blending for the same 8 sampled frames and action classes. All variants share the same Uni-FSAR pipeline (including LTQ, LSB, and Top- $K$  selection); only the attention direction differs. Bi-directional(1st row) and reverse-directional(2nd row) designs spread attention over many classes and often highlight non-target actions, while our uni-directional blending(3rd row) concentrates high attention on the correct class and a small subset of informative frames, consistent with its higher Top-1 accuracy.

---

## (B) Qualitative ablation of attention direction.

Each heatmap in Figure 1 visualizes **frame–class attention weights** for the same HMDB51 episode (8 frames, all classes), where the vertical axis denotes the frame index and the horizontal axis denotes the class token. Warmer colors correspond to higher attention between a frame and a class.

**Bi-directional attention.** In the bi-directional design, attention is widely spread across many classes and frames. **Several non-target classes such as *kiss*, *sword* and *flic-flac*** receive relatively strong responses across multiple frames, while the true class *laugh* does not stand out with a clear, concentrated stripe. This pattern indicates that the bi-directional interaction tends to mix visual and textual information symmetrically, but also allows irrelevant classes to keep non-negligible attention, which is consistent with the risk of prototype contamination discussed in the paper.

**Reverse-directional attention.** When we flip the direction (text  $\rightarrow$  frames only), the model **fails** to learn a stable alignment: **the heatmap shows strong activations on incorrect classes (again, *chew*, *kiss*, *pour*, *smoke*, *wave*), while the true *laugh* class** receives no distinctive peak. Attention is also concentrated on a few early frames without meaningful differentiation across the rest. This qualitatively matches the severe performance drop ( $\sim 20\%$  Top-1) observed for the reverse-directional variant and suggests that pushing information from class tokens back into frame tokens alone is not sufficient to learn reliable prototypes.

**Uni-directional attention (ours).** In contrast, the uni-directional blending (frames  $\rightarrow$  text queries) produces a **sharply focused pattern: attention is strongly concentrated on the *laugh* class and *smoke* class**, while other classes remain close to zero across all frames. This matches our Top- $K$  design, where only a few semantically most relevant frames are emphasized, and explains why the uni-directional variant achieves the best performance (82.3% Top-1 vs. 81.2% for bi-directional and  $\sim 20\%$  for reverse-directional) despite sharing the same LTQ, LSB, and Top- $K$  components.

The heatmaps thus provide qualitative evidence that uni-directional blending more effectively suppresses cross-class interference and isolates truly informative frame–class interactions, and **we observe the same tendency consistently across additional qualitative examples in Appendix A.6.**

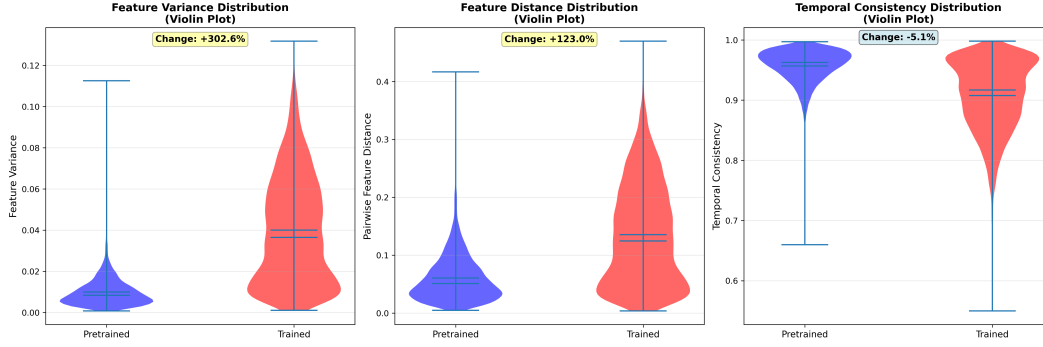


Figure 2: Feature-level statistics before and after training on HMDB51. Uni-FSAR substantially increases feature variance and pairwise feature distance, while keeping temporal consistency high, indicating that frames become more discriminative in the embedding space without introducing excessive noise.

## 2 VISUALIZATIONS FOR R2: KEE2

Fig. 2 summarizes as violin plots. The red distributions (trained) are clearly shifted towards larger variance and distance compared to the blue ones (pretrained), while the temporal-consistency distribution remains concentrated near 1.0 with only a modest shift. This visualizes that Uni-FSAR learns to *spread out* frame embeddings in a class-discriminative way without destroying temporal coherence.

These changes in feature-space variance occur together with the **Top-1 accuracy gains on HMDB51** reported in Table 1, where frame-level ambiguity is particularly severe. This and Fig.9-15 suggests that the learned increase in inter-frame dispersion is **beneficial**, helping the model focus on truly informative frames and reduce the impact of redundant/irrelevant ones, and ultimately resolve frame-level ambiguity by de-emphasizing noisy content while amplifying distinctive action cues.

Moreover, the same training strategy also improves performance on SSv2-small (Appendix A.5), where videos are temporally well trimmed and exhibit far fewer irrelevant/redundant frames, indicating that this implicit handling is **not overfitted to a specific dataset** and helps generalization