Action Modifers Supplementary Material

This supplementary document allows us to examine a sample of the embedding space (Appendix A). We present two additional ablations. First, per adverb results across modalities, giving additional insight into the learning of adverbs (Appendix B). We also investigate alternative choices for the query Q to the scaled dot-product attention (Appendix C).

A. Embedding Space

While visualizing the high-dimensional embedding space is difficult, we provide t-SNE projections of this space for a sample, to show the learning achieved. We consider all videos of the narrated action 'cook', and show the embedding space before (*i.e.* from I3D features) and after training. We highlight in two figures adverb-antonym pairs 'completely'/'partially' (Fig. 2) and 'quickly'/'slowly' (Fig. 3) and fade out points corresponding to other adverbs for ease of viewing. In each case, we show that our training successfully separates the embedding space based on the adverb. The figures also visualize a couple of video examples in each case, with 3 videos correctly embedded within the corresponding ground-truth and one incorrect prediction 'slowly'→'quickly'.

Similarly, we plot the t-SNE projections of the embedding for videos narrated with the action 'spread' (Fig. 4) before and after training. From this we can see that despite having far fewer examples of the action, the method is still able to successfully separate adverb-antonym pairs.

B. Per Adverb Results

In Figure 1 we show the effect of different modalities on the results per adverb. Firstly, we observe that considering all adverbs (All), the inclusion of both RGB and Flow is better than either modality separately. However, modalities perform differently across individual adverbs. For example, 'finely' is retrieved significantly more successfully with RGB than with Flow. Unsurprisingly, 'quickly' and 'slowly' benefit from the inclusion of Flow features alongside RGB.

C. Choice of Q

As noted in the main manuscript, we use a query Q to attend to the relevant parts of the video, for weakly-supervised embedding. We have chosen the embedding of the action, $g(a)$, as the query to our scaled dot-product attention (Eq. 6). Our attention is calculated by the compatibility of the query Q with the key K (a linear projection of the video segment features), therefore the choice of Q is integral to the weakly-supervised embedding. Here, we

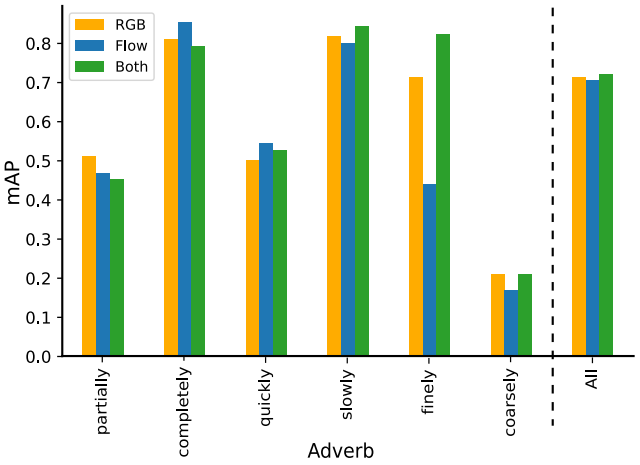

Figure 1. Video-to-adverb retrieval mAP per adverb with different modalities.

	Q	P@1
Action	$g(a)$	**0.774**
	One-hot Vector	0.736
Adverb	GloVe	0.702
	Vec(W_m)	0.731
Both	$O_m(g(a))$	0.728

Table 1. Comparison of the choice of Q.

compare $Q = W^Q g(a)$ to several alternatives, including incorporating the adverb into the query. We report the results in Table 1. For this ablation, we do not use the two-stage optimization, and thus the performance matches that of 0.774 in Table 3 in the paper.

First, we compare the action's embedding $g(a)$ to a one-hot vector of the action. The embedding offers a better query. Second, we test using the adverb as a query. In this case, we use a single adverb from each antonym pair (*e.g.* 'slowly'/'quickly'→'quickly'). This offers an understanding of the type of adverb we are after, so as to pick relevant video segments to this action manner. We compare the GloVe representation to a flattened representation of the learned action modifier. Again, while this allows the method to focus on segments relevant to the type of action manner, using the embedding of the action performs best. Finally, we test the full action-adverb embedding $O_m(g(a))$. This showed a drop in performance compared to using the action's embedding alone. This is potentially related to the fact that adverbs are not mutually exclusive as described in the paper's results.

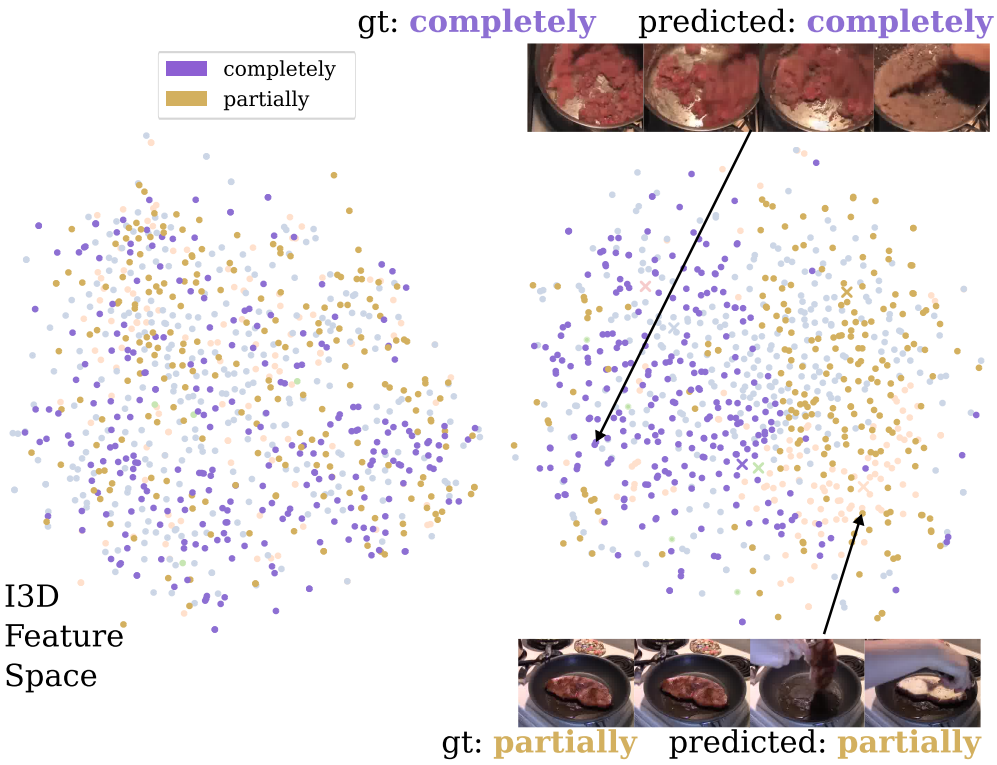

Figure 2. Comparison between the feature space for the action 'cook' before and after training highlighting antonym pairs. We highlight the 'completely'/'partially' pair with the other adverbs faded out.

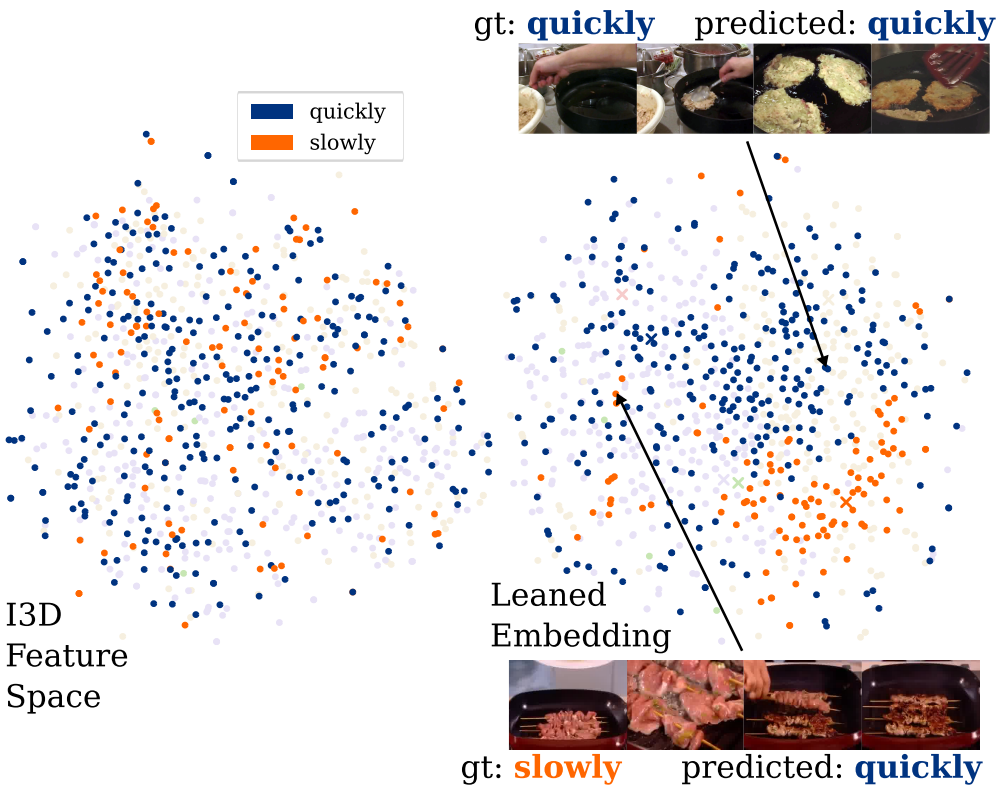

Figure 3. Comparison of the features spaces before and after training for the antonym pair 'quickly'/'slowly' in the action 'cook'. We fade out adverbs which are not 'quickly' or 'slowly'.

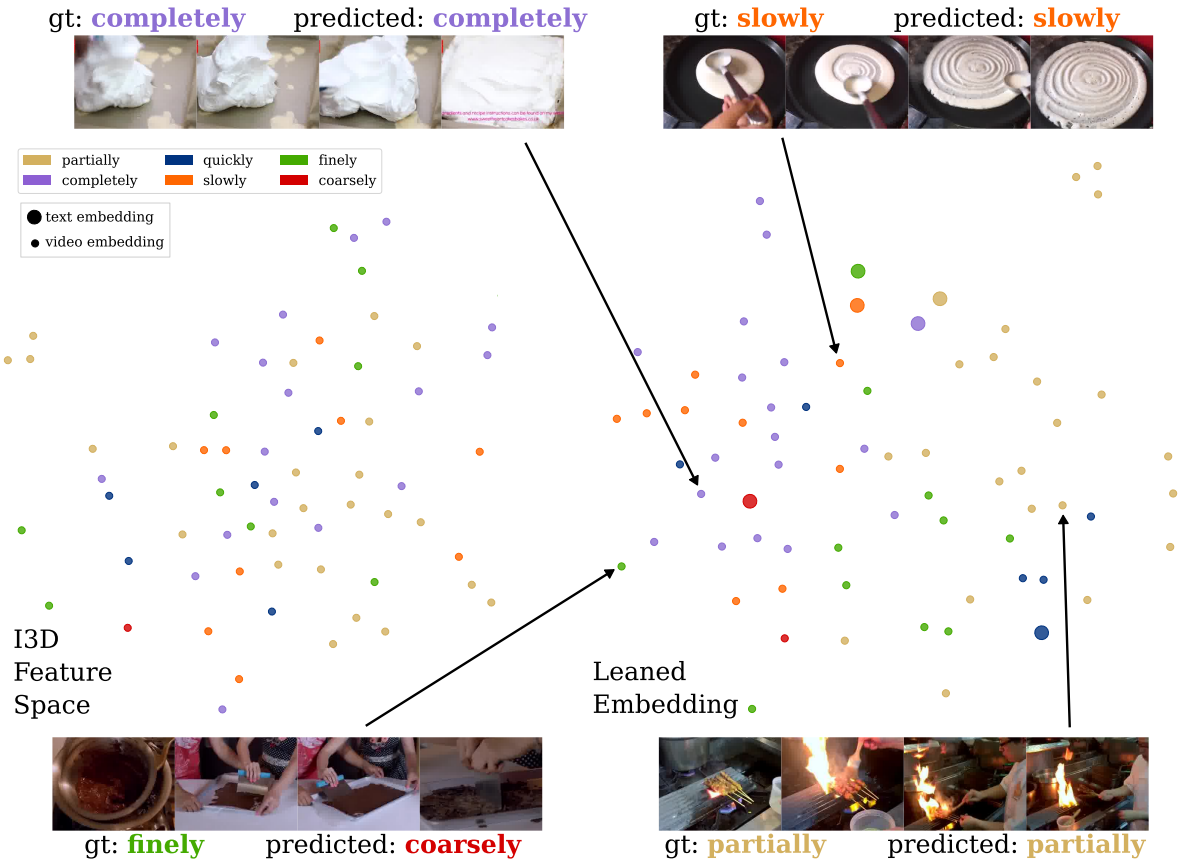

Figure 4. Comparison of the features spaces before and after training for the action ‘spread’. All adverbs are shown.