Supplementary Material

1 Predictions Visualization

Our best configuration of the ASFormer is to stack three decoders after the encoder. In this section, we visualize the predictions from the encoder and each decoder in 50Salads dataset, as shown in Fig. 1. By comparing the output predictions of different decoders, we can clearly observe an iterative refinement. The action label for each color is shown in Fig. 2, as well as the example frame for each action.

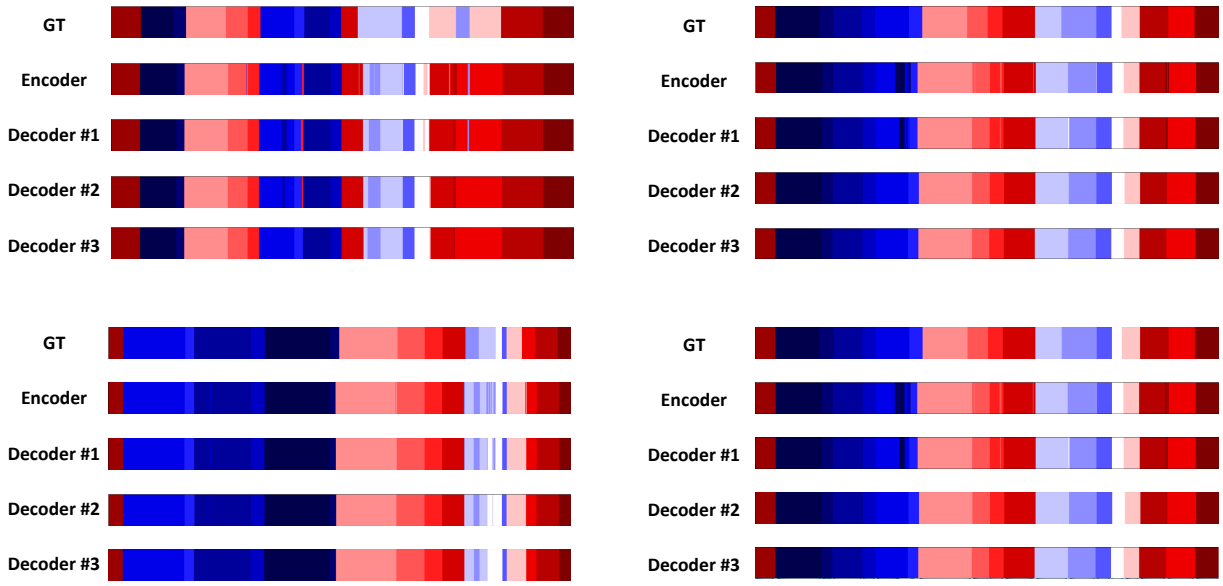


Figure 1: The visualization of predictions of our ASFormer on 50Salads dataset.

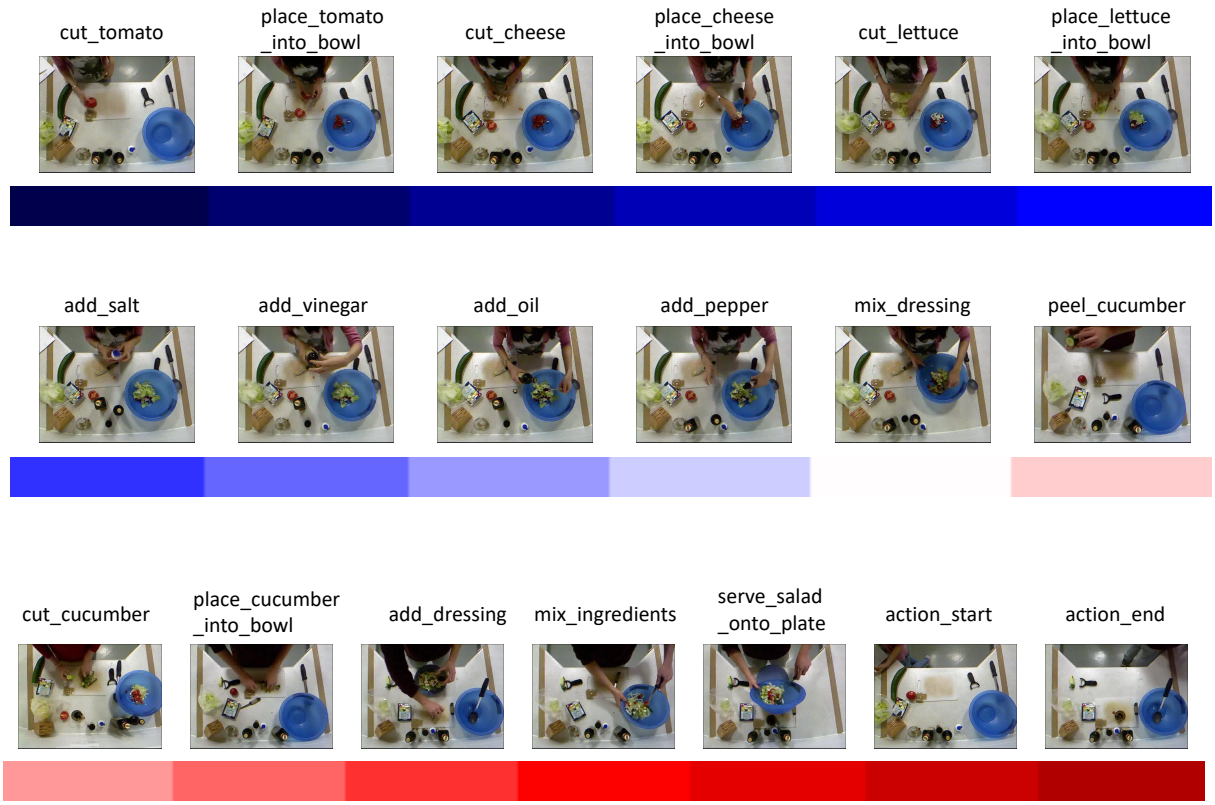
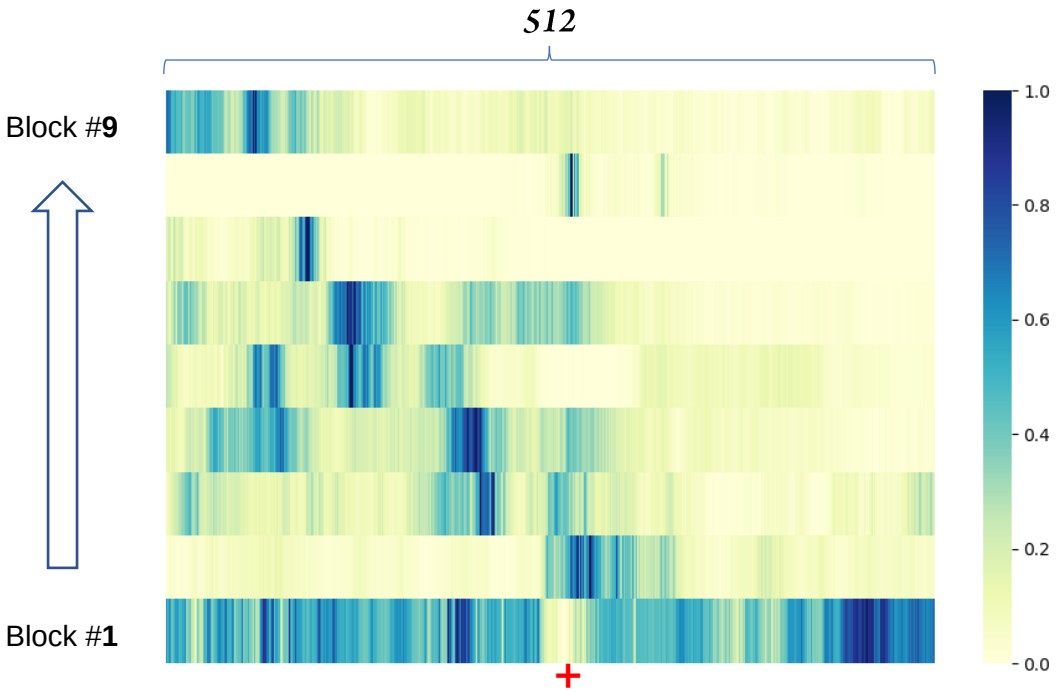


Figure 2: The action label and the example frame for each color in 50Salads dataset.

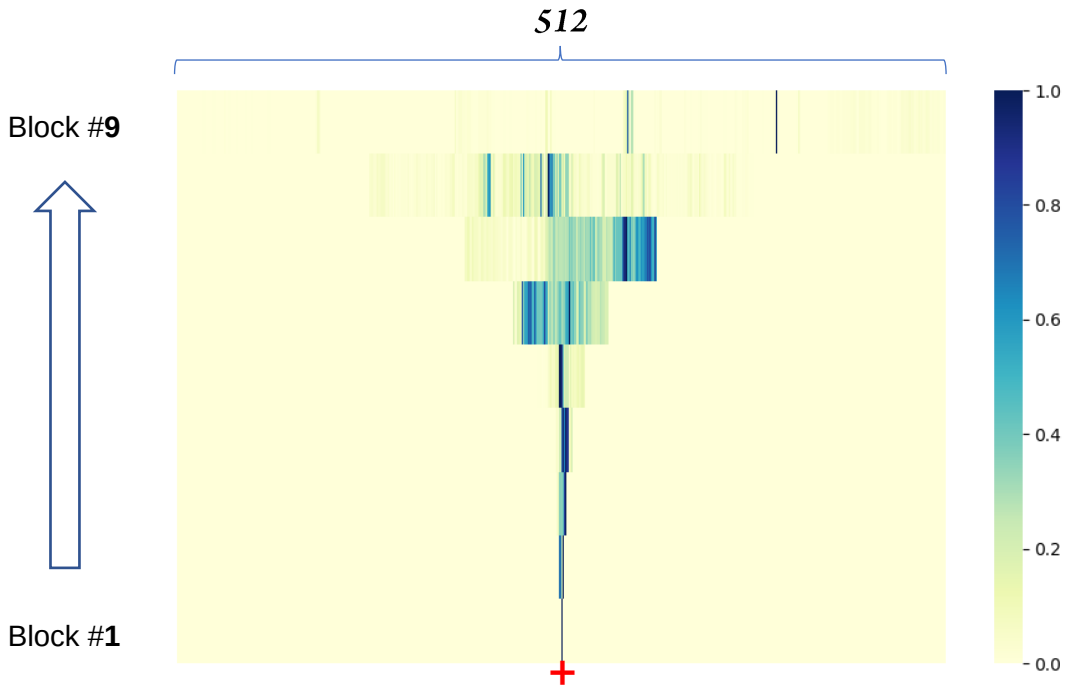
2 More Visualization of Attention Weights

In this section, We show more visualization of attention weights for an anchor frame (red +) in each encoder block on 50Salads dataset. In each picture, we show its corresponding video file.

rgb-01-2

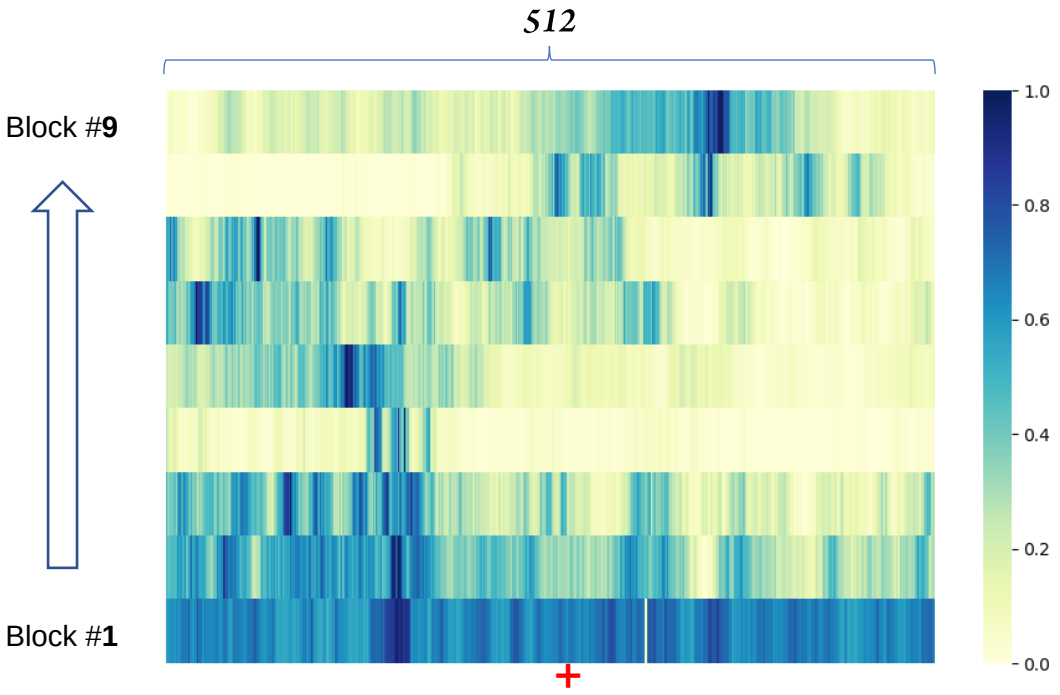


(a) non-hierarchical

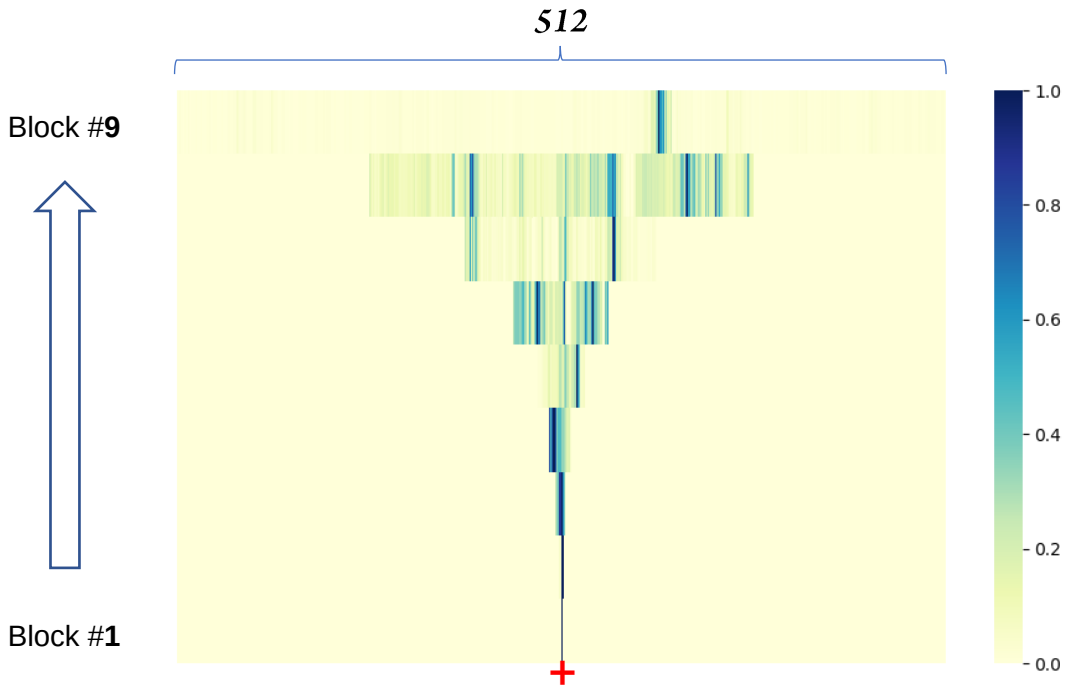


(b) hierarchical

rgb-03-1

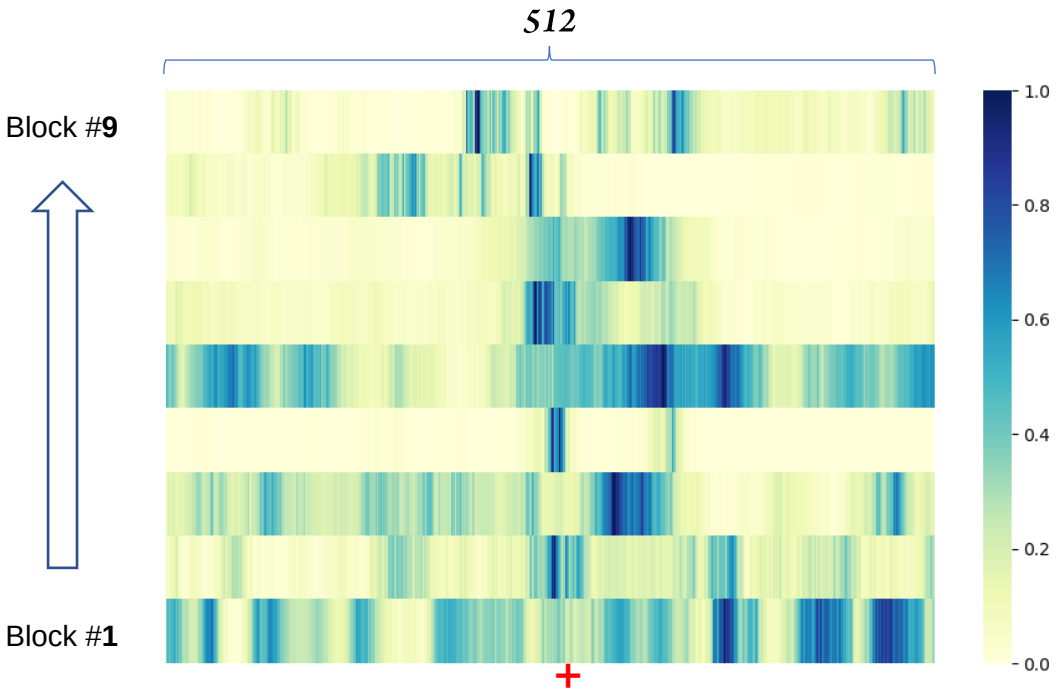


(a) non-hierarchical

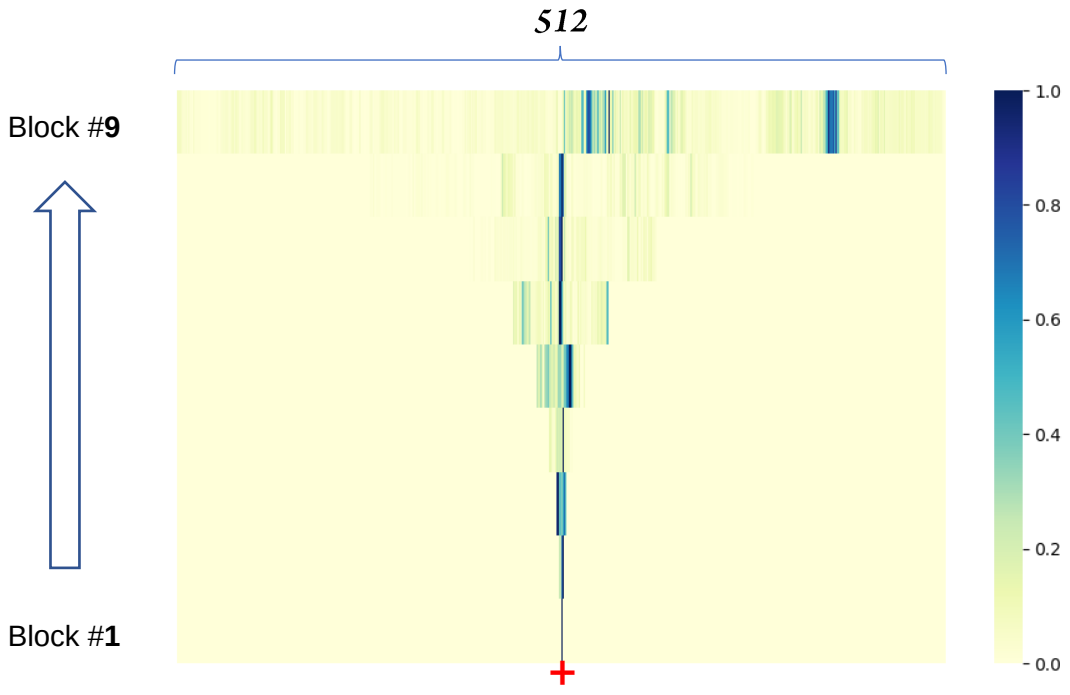


(b) hierarchical

rgb-04-1

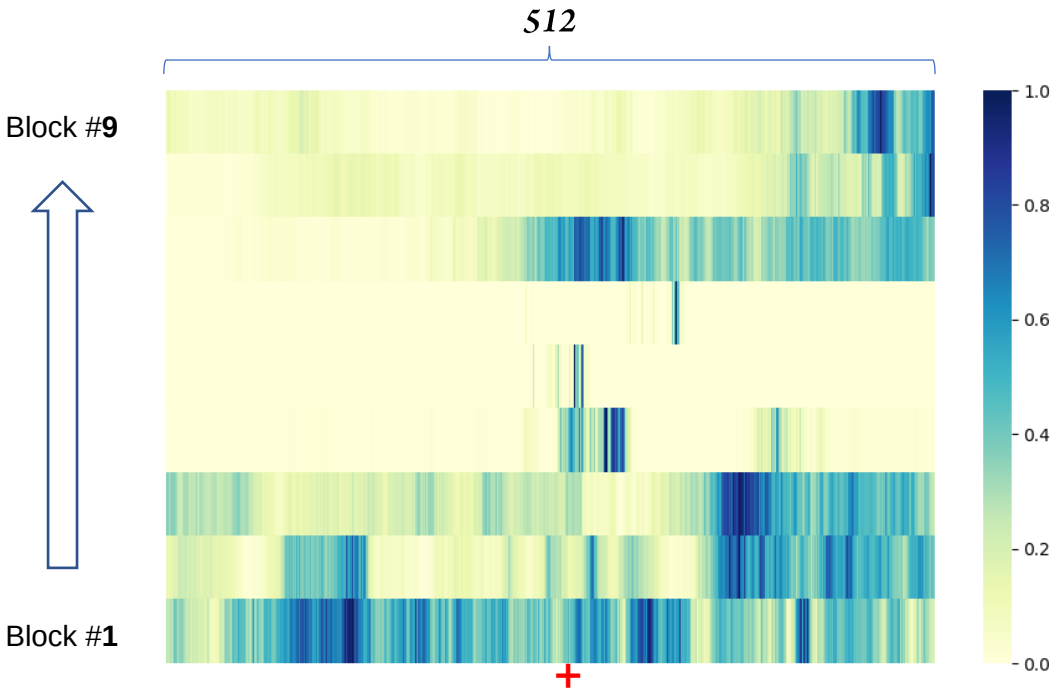


(a) non-hierarchical

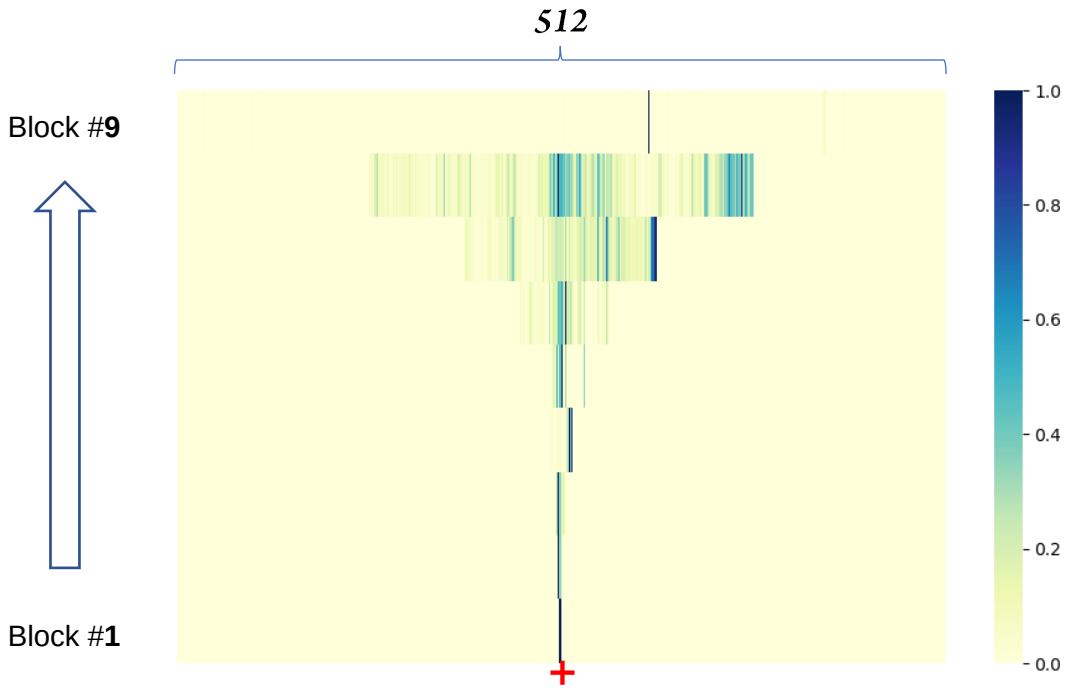


(b) hierarchical

rgb-05-1



(a) non-hierarchical



(b) hierarchical