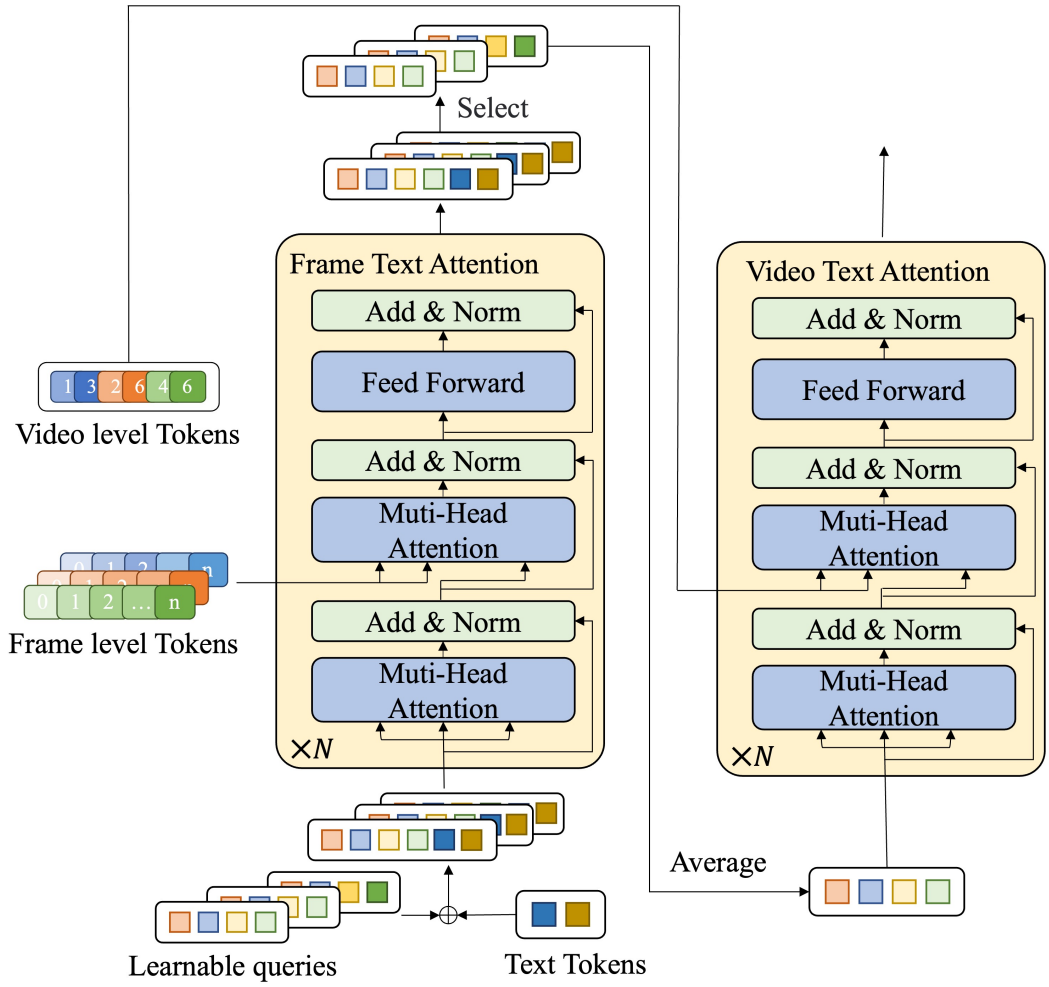

Figure 5: Details of Cross Attention Mechanism

APPENDIX

A BENEFITS OF RE-RANKING

The re-rank number K is a crucial aspect on retrieval performance. In order to investigate this relationship, we conduct experiments on the ViT-B network using the MSRVTT dataset. We vary the re-rank number from 2 to 19 and evaluate the results. As illustrated in Figure 6, we observe a gradual improvement in R@5 and R@10 Text2Video scores as K increases. This suggests that our method is effective across different re-rank numbers. However, it is important to note that excessively large values of K can lead to a decrease in retrieval efficiency. To strike a balance between efficiency and accuracy, we set $K = 15$ for all our experiments.

B CROSS ATTENTION MECHANISM

We present the details of cross attention mechanism in Figure 5. Specifically, for frame text attention, learnable queries are first concatenated with text tokens. These queries interact through self attention layer and then interact with each frame features in cross attention layer. These queries feed into video text attention after averaging. For video text attention, these queries interact with the selected video token in cross attention.

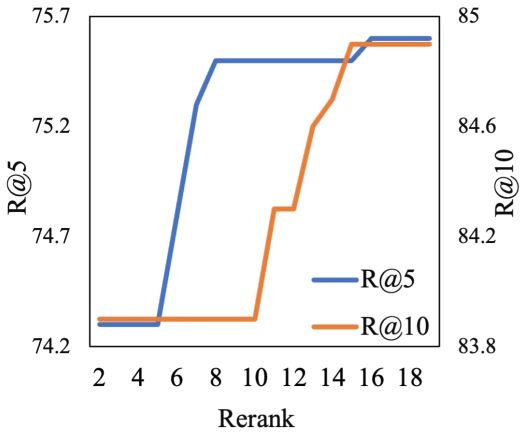

Figure 6: Different re-rank numbers.

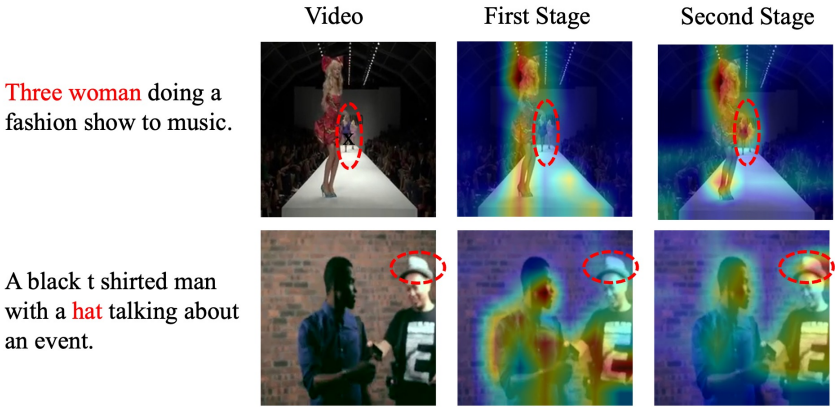

Figure 7: Fine-grained feature attention map.

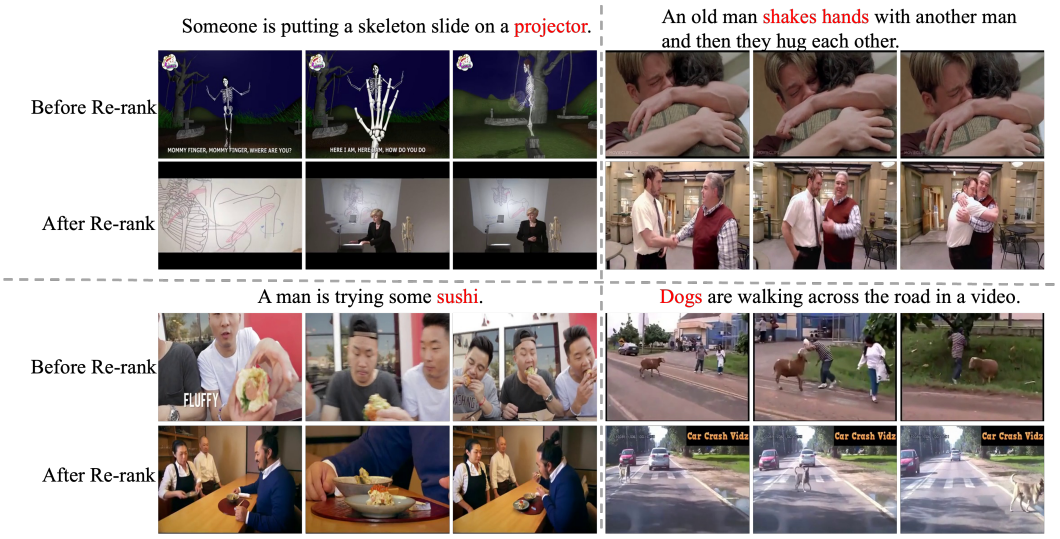

Figure 8: Text to video retrieval results.

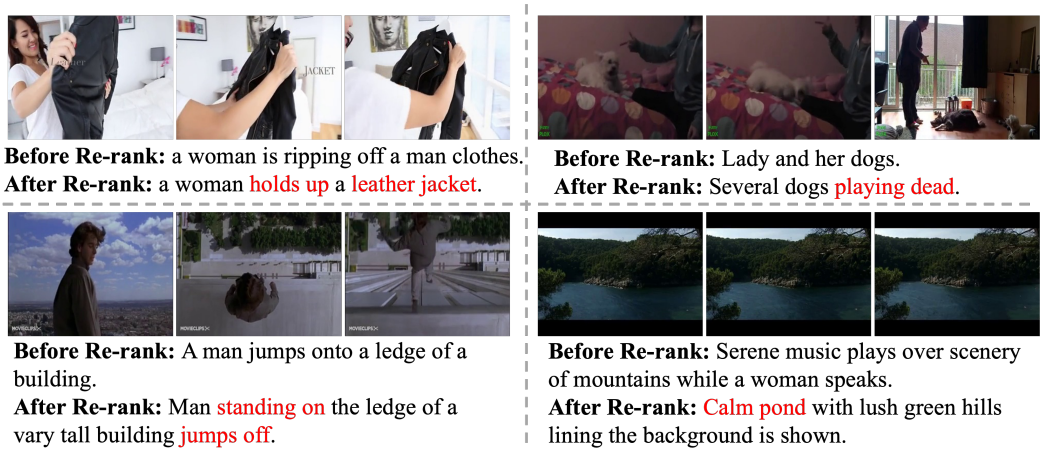

Figure 9: Video to text retrieval results.