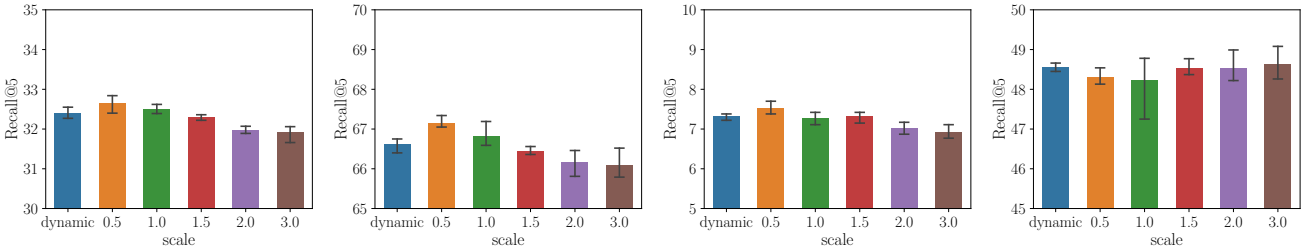
Multi-event Video-Text Retrieval Supplementary Materials

Gengyuan Zhang [1,2] **Jisen Ren** [1] **Jindong Gu** [3] **Volker Tresp** [1,2]
[1] LMU Munich, Munich, Germany
[2] Munich Center for Machine Learning, Munich, Germany
[3] University of Oxford, Oxford, United Kingdom
zhang@dbs.ifi.lmu.de jindong.gu@outlook.com

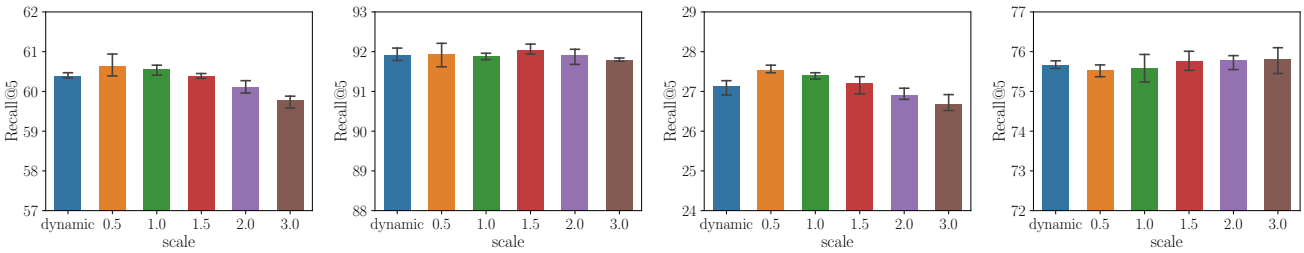
1. Continued Figures

We extend Fig.5-6 in the main paper with metrics on $k = 10$ and $k = 50$. Similarly, we find that the dynamic weighting strategy is a good tradeoff between Video-to-Text and Text-to-Video tasks and manifests stability in different trials.



(a) Average (b) One-Hit (c) All-Hit (d) Text-to-Video

Figure 1: We compare the model performance with different weighting strategies in the MeVTR loss on $Recall@10$ on the Video-to-Text task for ActivityNet Captions.



(a) Average (b) One-Hit (c) All-Hit (d) Text-to-Video

Figure 2: We compare the model performance with different weighting strategies in the MeVTR loss on $Recall@50$ on the Video-to-Text task for ActivityNet Captions.