

---

# Appendix for Mitigating Semantic Collapse in Partially Relevant Video Retrieval

---

Anonymous Author(s)

Affiliation

Address

email

## A Datasets

As described in the main manuscript, we evaluate our method on four untrimmed text-video datasets. Below, we provide a brief overview of each. QVHighlights[3] is a collection of news and vlog-style videos, recently reorganized for PRVR[5]. Each video is paired with an average of 3.3 text queries describing semantically diverse segments. TVR [4] is built from scenes across six popular TV shows, with each video annotated by five text queries targeting different segments. The training set contains 17,435 videos and 87,175 queries, while the evaluation set includes 2,179 videos and 10,895 queries. ActivityNet Captions [2] is sourced from YouTube videos, with an average of 3.7 text queries per video. The dataset includes 10,009 videos for training and 4,917 for evaluation. Charades-STA [1] extends the original Charades dataset by adding sentence-level annotations for specific temporal segments. It consists of 13,898 video-sentence pairs for training and 4,233 for evaluation.

## B Positive and Negative Societal Impacts

**Positive Impact.** By enabling retrieval based on partial content descriptions within long, untrimmed videos, the proposed method enhances the user experience in video search and navigation. This is particularly valuable in domains such as education, where lengthy untrimmed videos are commonly utilized.

**Negative Impact.** The ability to isolate specific video contexts and retrieve segments based on partial descriptions could be misused in surveillance settings (e.g., CCTV), enabling the tracking of individuals or the extraction of sensitive behaviors without consent. Such misuse raises potential concerns regarding privacy and ethical deployment.

## References

- [1] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.
- [2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Nieves. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.
- [3] J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.
- [4] J. Lei, L. Yu, T. L. Berg, and M. Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.
- [5] W. Moon, C.-H. Cho, W. Jun, M. Shim, T. Kim, I. Lee, D. Wee, and J.-P. Heo. Prototypes are balanced units for efficient and effective partially relevant video retrieval. *arXiv preprint arXiv:2504.13035*, 2025.