# CoCoA-DVC: Consistency and Concept Aware Training for Dense Video Captioning

## Supplementary Material

## 1. Datasets

We use the ActivityNet [3] and YouCook v2 [4] datasets for evaluating our performance on the Dense Video Captioning Task, and ActivityNet [3] and Charades-STA [1] for the Moment Retrieval task.

**ActivityNet:** This dataset comprises of videos of humans performing a variety of activities. There are a total of around 20k videos with an average duration of 2 minutes and annotated with 3.7 captions along with their time ranges per video on average. The standard splits for training, validation and testing are used in this work [2, 3].

**Youcook V2:** This is a recent dataset consisting of cooking videos. It contains a total of 2k videos with an average duration of around 5 minutes and 7.7 captions per video along with time ranges. The standard splits for training, validation and testing are used in this work [2, 4].

**Charades-STA:** This dataset is used for Moment Retrieval. It has around 10k videos with activities from multiple categories. In total, Charades has around 14K clip-sentence pairs, with an average of 6.3 words per sentence.

## 2. Implementation Details

We set the batch size to be 1 and run all experiments on a single A100 GPU. For both the datasets, the weights of the different loss functions are: 1 for IP Loss, 1 for GTCRL, 2 for classification loss, 4 for giou loss and 3 for captioning loss. The number of decoder queries were set to 30, while the number of text concepts retrieved per frame was set to 25. We used the Adam optimizer with a learning rate of 0.00005 and ran the training for 30 epochs.

## References


[1] J. Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. Tall: Temporal activity localization via language query. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017. 1

[2] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13894–13904, 2024. 1

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 1

[4] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 1