

# Supplementary Materials: Calibrating Prompt from History for Continual Vision-Language Retrieval and Grounding

Anonymous Authors

## 1 DATASETS DETAILS

**MS-COCO** [1]. MS-COCO is a high quality crowd-labeled datasets. For the image-text retrieval, we follow [2] to 118k/5k training/test split. To generate separate tasks for continual learning, we use the "category" annotation in original dataset to define 12 tasks, as shown in 1.

**MSR-VTT** [3]. MSR-VTT is a large-scale dataset for the open domain video-text retrieval, which consists of 10,000 video clips and each video clip is annotated with 20 sentences. There are about 29,000 unique words in all captions. The standard splits uses 6,513 clips for training, 497 clips for validation, and 2,990 clips for testing. We evaluate the video-text retrieval on this dataset. To generate separate tasks for continual learning, we use the "category" annotation in original dataset to define 10 tasks, as shown in 2.

**RefCOCO** [4]. RefCOCO consists of 142,209 refer expressions for 50,000 objects in 19,994 images. In the RefCOCO dataset, no restrictions are placed on the type of language used in the referring expressions. We evaluate the referring expression comprehension and segmentation on this dataset via its box-level and mask-level annotations respectively. Similar to MS-COCO, we adopt the same division split in 1.

Table 1: Task division in Coco-based datasets.

Person	Vehicle	Outdoor	Animal	Accessory	Sport
Kitchen	Food	Furniture	Electronic	Appliance	Indoor

Table 2: Task division in MSR-VTT dataset.

News	Movie	Sports	Cooking	Traffic
Animation	Music	Animal	Kids	Beauty

## 2 EXPERIMENT DETAILS

**Environment Configuration.** Our HPC is implemented using PyTorch 1.9.0 with CUDA 10.0 and cudnn 7.6.5. All the experiments are conducted on a workstation with four NVIDIA GeForce RTX 2080Ti GPUs.

## 3 BROADER IMPACTS

This paper first investigates the continual learning problem for vision-language retrieval and grounding task. We introduce a novel approach HPC to calibrate prompt learning by leveraging historical information to learn task-wise and modality-wise association. As evidenced by our experiments, the integration of our HPC method with existing pre-trained multimodal models such as CLIP and GLIP yields remarkable effectiveness. This research contributes to substantial performance advancements of large-scale pre-trained models on diverse downstream multimodal tasks, thereby enhancing training stability and minimizing training costs. Furthermore, our method safeguards data security and privacy by obviating the need for storing past data. We hope our work can serve as a solid baseline in continuous learning for multi-modal application.

## 4 LIMITATIONS

Despite the simplicity and effectiveness of our approach, it has some limitations. In this work, we primarily focused on two-stream multi-modal pre-trained models and did not consider single-stream architectures. As a result, our method cannot be directly combined with these pre-trained models, which may affect its generalizability.

## REFERENCES

- [1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [3] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.
- [4] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 69–85.