# Supplementary Materials: Low-rank Prompt Interaction for Continual Vision-Language Retrieval

Anonymous Authors

## 1 METHOD

### 1.1 Training and Inference

*1.1.1 Training.* After the training is completed, we employ the visual encoder to extract the visual features from all the training images without prompts. Subsequently, we apply K-Means clustering to these visual features, selecting five visual centroids as task-specific keys. For the image-text retrieval task, we extract five textual centroids the same as visual centroids for task identity prediction.

*1.1.2 Inference.* During the inference stage, we first predict which task the image/caption belongs to by extracting its visual/textual features without prompts. We then use the K-NN (k-nearest neighbors) algorithm to find the task-specific keys closest to the image/caption features. The task associated with this task-specific keys is identified as the task to which the predicted image/caption belongs. For the image-text retrieval task, we should predict image task identity and caption task identity independently while for the referring expression comprehension task, we only need to predict the task identity according to the image. Subsequently, we select the prompts and interaction module associated with this task to perform target prediction.

## 2 EXPERIMENT

### 2.1 Dataset and Task division

We conduct experiments on two visual-language retrieval tasks, Image-text Retrieval and Referring Expression Comprehension. For the image-text retrieval task, we select the **MS-COCO** [3] dataset, while for the referring expression comprehension task, we chose the **RefCOCO** [6], **RefCOCO+** [6], and **RefCOCOg** [4] datasets.

All four datasets are based on the same MSCOCO2014 image dataset but come with different annotations, with each image corresponding to multiple captions. Thus, we adhere to the same task division criteria. We divide all data into 12 categories according to the super category, which are appliance, sports, outdoor, electronic, accessory, indoor, kitchen, furniture, vehicle, food, animal and person.

For the referring expression comprehension task, we select three datasets, each with distinct characteristics. RefCOCO and RefCOCO+ include subsets for train, val, testA, and testB; while RefCOCOg comprises subsets for train, val, and test. Throughout our experimental process, we exclusively use the train dataset to train our models and subsequently evaluate their performance on both the test and val subsets.For RefCOCO and RefCOCO+, the val subsets provide a sufficient number of training samples for each task. TestA is predominantly human-centered; therefore, for tasks except for person, there are fewer training samples available. As for testB, this subset does not contain images with people, leading us to define only 11 tasks for evaluation on testB.

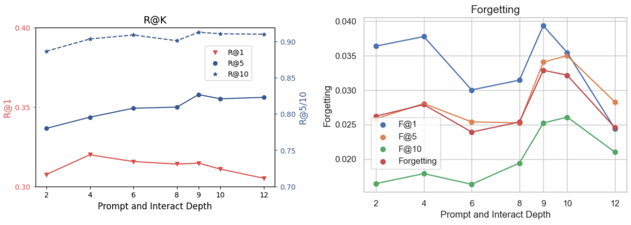
### 2.2 Implementation Detail

For the image-text retrieval task, we employ CLIP as the backbone. CLIP extracts features from images and text, after which we directly compute logits using visual and textual features. These logits are then utilized as scores for retrieval purposes. Regarding the task of referring expression comprehension, we use GLIP as the backbone. GLIP comprises a vision encoder, a language encoder and a Region Proposal Network (RPN). We use the vision encoder and language encoder to extract the visual and textual features. Then input the visual and textual features to the RPN to predict bounding boxes. We freeze these networks and only update the prompts and the cross-modal prompt fusion module.

## 3 ABLATION STUDY

We add more ablation experiments in this section.

### 3.1 Prompt Depth



R@K

Forgetting

**Figure 1: Metrics for Different Prompt Depth.**

We investigate the impact of prompt depth. Specifically, with prompt depth set to $i$, we employ decomposition prompting and the cross-modal prompt fusion module in the first $i$ layers of the encoder. The experimental results are shown in Figure 1.

The performance of $R@1$ initially increases and then decreases, achieving optimal performance at a depth of 4. As the depth increases, $R@5$ and $R@10$ experience a gradual increase. The $Forgetting$ first decreases and then increases when the depth is less than 9, followed by a decreasing trend at greater depths.

### 3.2 Qualitative Analysis

In this section, we conduct a qualitative analysis of GLIP [2], S-Prompts [5], MaPLe [1], and LPI-P. Fig. 2 and 3 present the visualization results of different models. In the caption, the red font indicates **positive tokens**. For the inference results, we select the top five predicted bounding boxes for visualization, namely the $R@5$ candidate boxes. These are sorted by confidence scores, from highest to lowest, with the bounding boxes colored in **BLUE**, **CYAN**, **RED**, **GREEN**, and **PINK**, respectively. The targets are annotated

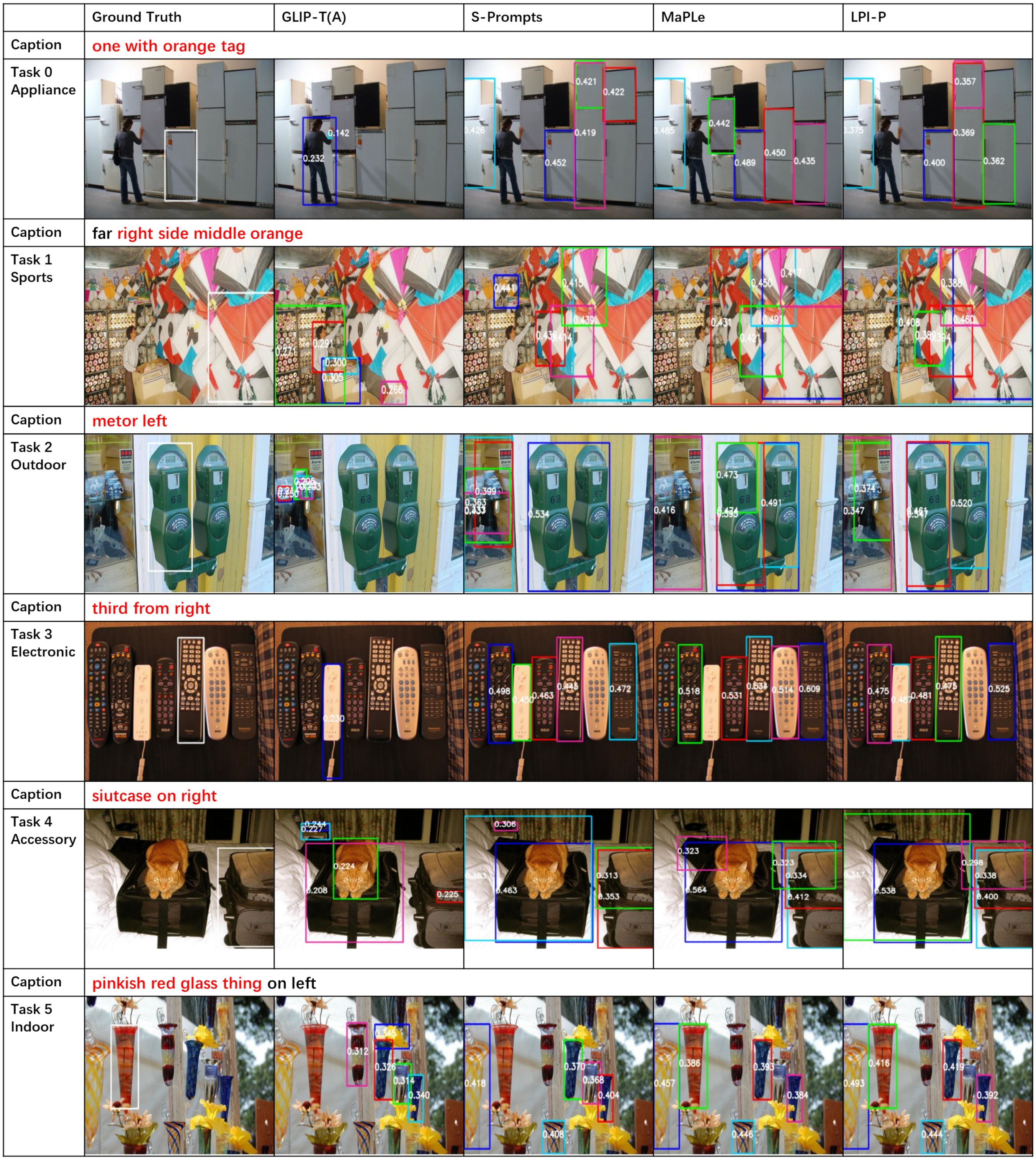


**Figure 2: Qualitative Analysis of Task 0 to 5.**

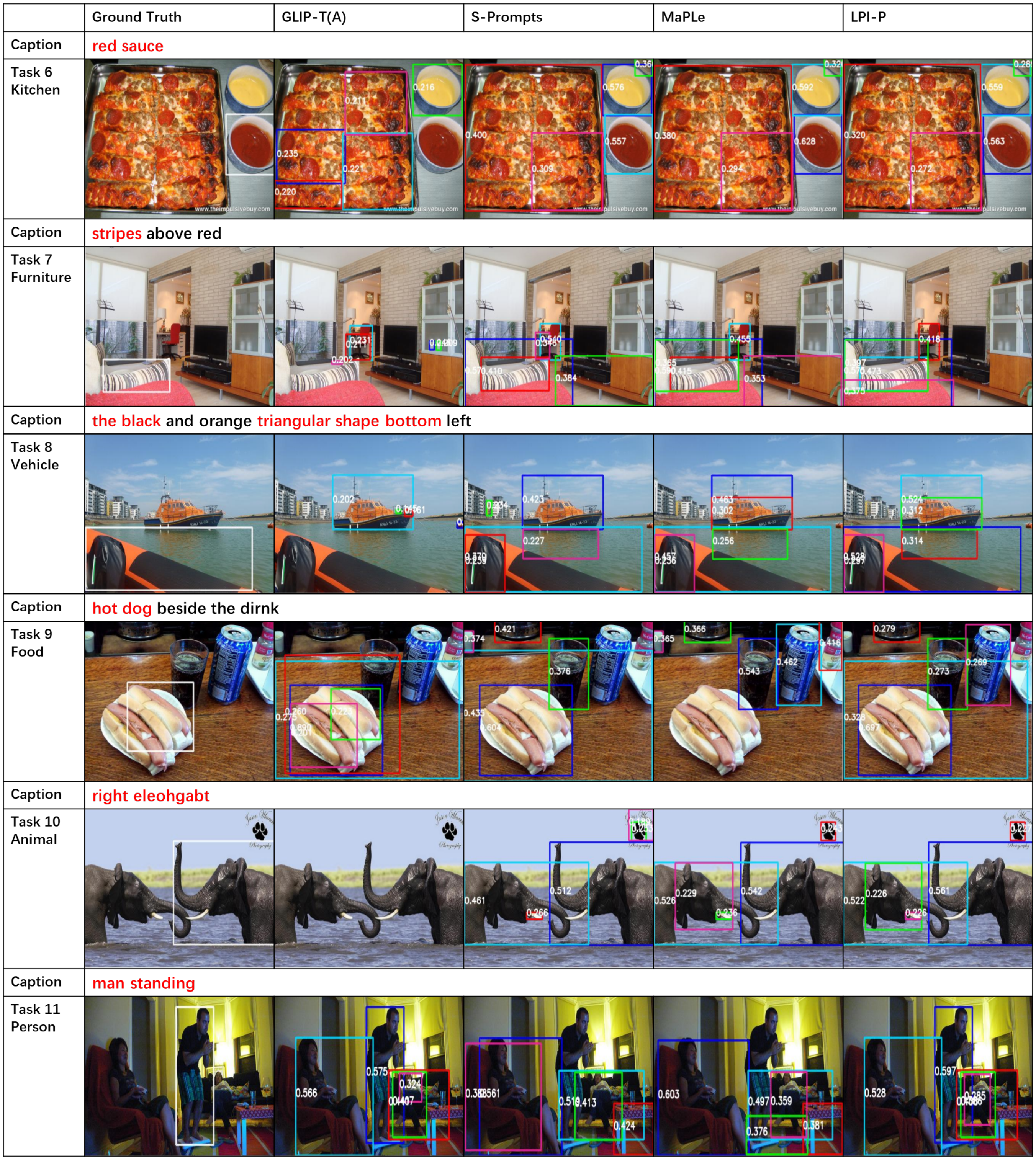


**Figure 3: Qualitative Analysis of Task 6 to 11.**

within white bounding boxes in the Ground Truth column. From these images, we can draw the following conclusions:

- The pre-trained GLIP model is capable of identifying objects in captions, but it exhibits weaker recognition abilities for positional words such as "middle," "left," "right," and colors like "orange" and "red,".
- For objects with unclear references, GLIP struggles to accurately identify them, as seen in tasks 3 and 5.
- When it comes to captions with spelling errors, GLIP's accuracy significantly drops, as observed in tasks 2, 4, and 10.
- In the case of complex captions, LPI-P performs optimally, as shown in task 8.
- Maple tends to overfit, meaning it underperforms compared to GLIP on simple tasks like task 9 and 11.

In summary, the GLIP model struggles to understand positional words and colors in captions. S-Prompts using prompt techniques can enhance the understanding of captions while it falls short in aligning text and image content accurately. MaPLe and LPI-P demonstrate superior performance in most tasks. However, due to a lack of constraints in learning prompts, MaPLe is prone to overfitting, failing to achieve desired results in simple tasks. Our model LPI-P employs contrastive learning with additional constraints for improved performance and is capable of understanding more complex texts.

# REFERENCES

[1] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19113–19122.

[2] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10965–10975.

[3] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.* Springer, 740–755.

[4] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2022. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems* 35 (2022), 5682–5695.

[6] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. arXiv:1608.00272 [cs.CV]