Collaborative Vision-Text Representation Optimizing for Open-Vocabulary Segmentation (*Supplementary material*)

Siyu Jiao[1,2⋆], Hongguang Zhu[1,2⋆], Jiannan Huang[1,3], Yao Zhao[1,2], Yunchao Wei[1,2], and Humphrey Shi[3,4]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Peng Cheng Laboratory
[3] Georgia Institute of Technology
[4] Picsart AI Research (PAIR)
cici579869@gmail.com

In the supplementary material, we first introduce the dataset settings in Sec. 1. Then, the Prompt engineering techenic is introduced in detailed in Sec. 2, including the template-based Prompt and the description-based Prompt. Moreover, we provide additional qualitative results in Sec. 3.

1 Dataset

We follow [4, 8–10] to conduct experiments on the popular benchmarks of open-vocabulary *semantic* and *panoptic* settings, COCO-Stuff [2], COCO-Panoptic [5], Pascal-VOC [3] ADE20K [11], and Pascal-Context [6] to evaluate the performance of MAFT+.

- **COCO-Stuff**: COCO-Stuff is a large-scale semantic segmentation dataset that contains 164K images with 171 annotated classes, which are divided into the training set (118K images), validation set (5K images), and testing set (41K images). In our experiments, we use the full 118K training set as the training data to train the *semantic* models.
- **COCO-Panoptic**: COCO-Panoptic shares the same training images with COCO-Stuff. These images are labeled into 133 categories. In our experiments, we use COCO-Panoptic to train the *panoptic* models.
- **Pascal-VOC**: Pascal-VOC includes 1,449 images for testing with 20 annotated classes. In the open-vocabulary *semantic* segmentation, all 20 classes are used for evaluation (dubbed as PAS-20).
- **ADE20K**: ADE20K is a large-scale scene understanding dataset comprising 2k images for validation with two types of annotations: one with 150 classes featuring panoptic annotations and another with 847 classes featuring semantic annotations. For the open-vocabulary *semantic* segmentation, we evaluate our method on two settings of ADE20K: 150 classes (dubbed as A-150) and 847 classes (dubbed as A-847). In the open-vocabulary *panoptic* segmentation, we use the setting with 150 class annotations for evaluation.

[⋆] Equal contribution

- **Pascal-Context** is a dataset for semantic understanding which contains 5K validation images. Two versions are used for open-vocabulary *semantic* segmentation, one with 59 frequently used classes (dubbed as PC-59) and another with the whole 459 classes (dubbed as PC-459).

2 Template-based Prompt & Description-based Prompt

Prompt engineering has been proven to be beneficial for open-vocabulary segmentation. In our default setting, we follow the common practice of using the template-based prompt to augment class names into sentences. In addition, we also explore using GPT [1] or Llama [7] to apply description-based prompts.

Template-based Prompt. Following established approaches [4, 8–10], we use multiple templates to integrate the class names into sentences. These sentences are then fed into CLIP-T, and the resulting outputs are averaged to generate the text embedding for each class. The templates are listed in Tab. 1.

Description-based Prompt. We assume that the detailed descriptions of one class name contain additional valuable information that helps to optimize CLIP-T. To investigate this, we design description-based prompts, leveraging Large Language Models (LLMs) to generate descriptions, including using GPT-3.5 [1] to generate description sentences, and use the open-source LLM, Llama-2 [7], to generate descriptive text embeddings. Through experimental verification, we selected a few prompts suitable for LLMs to generate descriptions. The prompts and responses are shown in Tab. 2a and Tab. 2b, respectively.

The results indicate that some descriptions provide valuable visual attributes, facilitating the alignment of vision-text representations in the CLIP feature space. However, they may introduce noise. *e.g.*, both *cat* and *chair* have descriptions that include the sentence “four legs”.

3 Similarity Map & Visualize results

We provide more qualitative results, including similarity maps (Fig. 1), and visualize results in Pascal-VOC, COCO-Stuff, ADE20K datasets (Fig. 2, 3).

Similarity map. Fig. 1 presents the normalized similarity maps between text and image embeddings in A-150 and A-847 datasets. We choose 200 categories in A-847 for visualization. It is evident that the elevated similarity values of fine-tuned CLIP’s similarity map are mainly located on the main diagonal, indicating the fine-tuned CLIP achieves a better alignment of vision-text representation.

Qualitative Analysis. Fig. 2, 3 show segmentation results on Pascal-VOC, COCO-Stuff, ADE20K. The frozen CLIP results may contain background noise (1st and 2nd rows in Fig. 2) or misclassify when there are many objects in one image (3rd row in Fig. 2). The fine-tuned CLIP generates better results compared to the frozen CLIP, which can even correct misclassified areas in ground-truth (*fence* in the 4th row in Fig. 2).

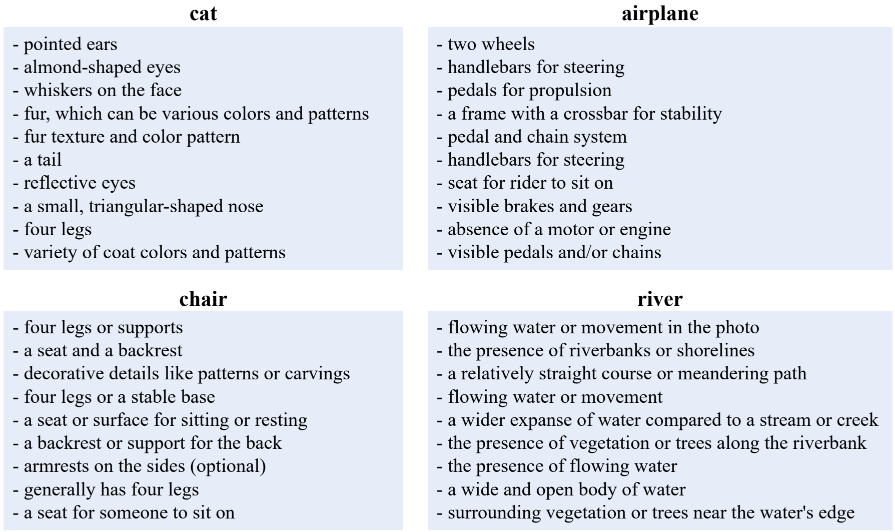
Table 1: Prompt templates used in our method.

Templates
"a photo of a { }."
"This is a photo of a { }"
"There is a { } in the scene"
"There is the { } in the scene"
"a photo of a { } in the scene"
"a photo of a small { }."
"a photo of a medium { }."
"a photo of a large { }."
"This is a photo of a small { }."
"This is a photo of a medium { }."
"This is a photo of a large { }."
"There is a small { } in the scene."
"There is a medium { } in the scene."
"There is a large { } in the scene."

Table 2: Description-based prompts and responses

Description prompts
"Please describe the appearance of { }. Please characterize it briefly."
"Describe the physical attributes of { }. Please characterize it briefly."
"What can you tell me about the appearance of the category of { }? Please characterize it briefly."
"Tell me about the outward features of the category of { }. Please characterize it briefly."
"Briefly outline the visual traits of the category of { }."
"Can you provide details about what the category of { } looks like? Please characterize it briefly."
"I'm curious about the visual characteristics of the category of { }. Please characterize it briefly."
"Provide a description of the visual aspects of { }. Please characterize it briefly."
"Q: What are visual features of distinguishing a smartphone? A: - a touchscreen Q: What are features for distinguishing a { }? A: -"

(a) Description prompts used in our method.



cat

- pointed ears
- almond-shaped eyes
- whiskers on the face
- fur, which can be various colors and patterns
- fur texture and color pattern
- a tail
- reflective eyes
- a small, triangular-shaped nose
- four legs
- variety of coat colors and patterns

airplane

- two wheels
- handlebars for steering
- pedals for propulsion
- a frame with a crossbar for stability
- pedal and chain system
- handlebars for steering
- seat for rider to sit on
- visible brakes and gears
- absence of a motor or engine
- visible pedals and/or chains

chair

- four legs or supports
- a seat and a backrest
- decorative details like patterns or carvings
- four legs or a stable base
- a seat or surface for sitting or resting
- a backrest or support for the back
- armrests on the sides (optional)
- generally has four legs
- a seat for someone to sit on

river

- flowing water or movement in the photo
- the presence of riverbanks or shorelines
- a relatively straight course or meandering path
- flowing water or movement
- a wider expanse of water compared to a stream or creek
- the presence of vegetation or trees along the riverbank
- the presence of flowing water
- a wide and open body of water
- surrounding vegetation or trees near the water's edge

(b) LLMs responses of category *cat*, *airplane*, *chair*, and *river*.

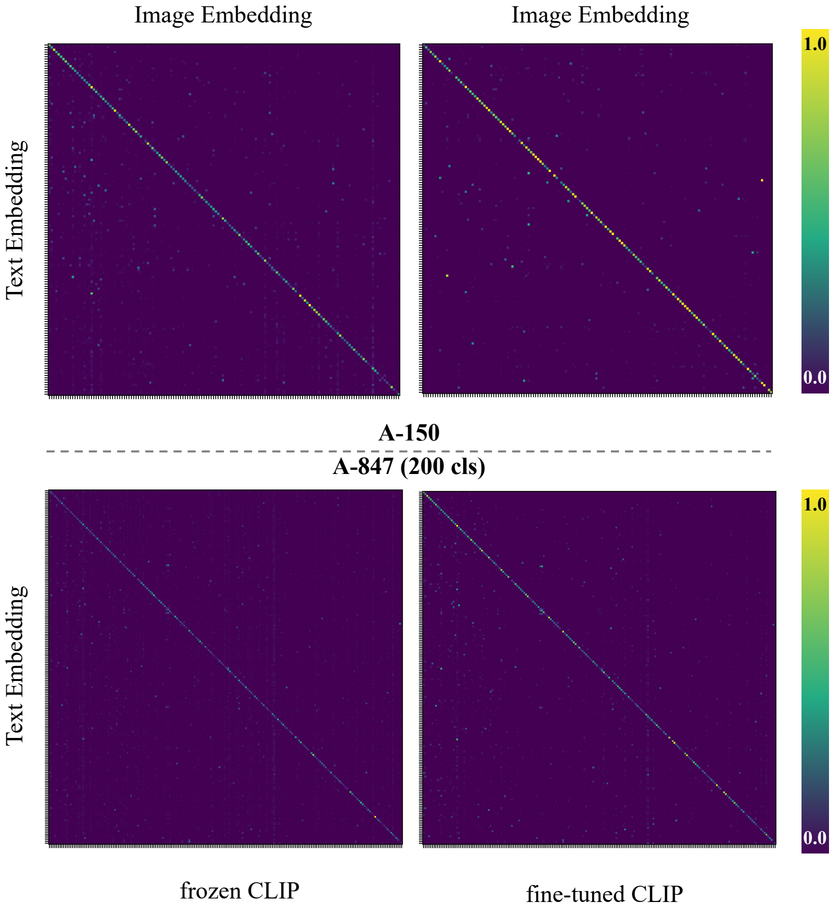


Fig. 1: Normalized cosine similarity on A-150 and A-847, we choose 200 categories in A-847 for visualization.

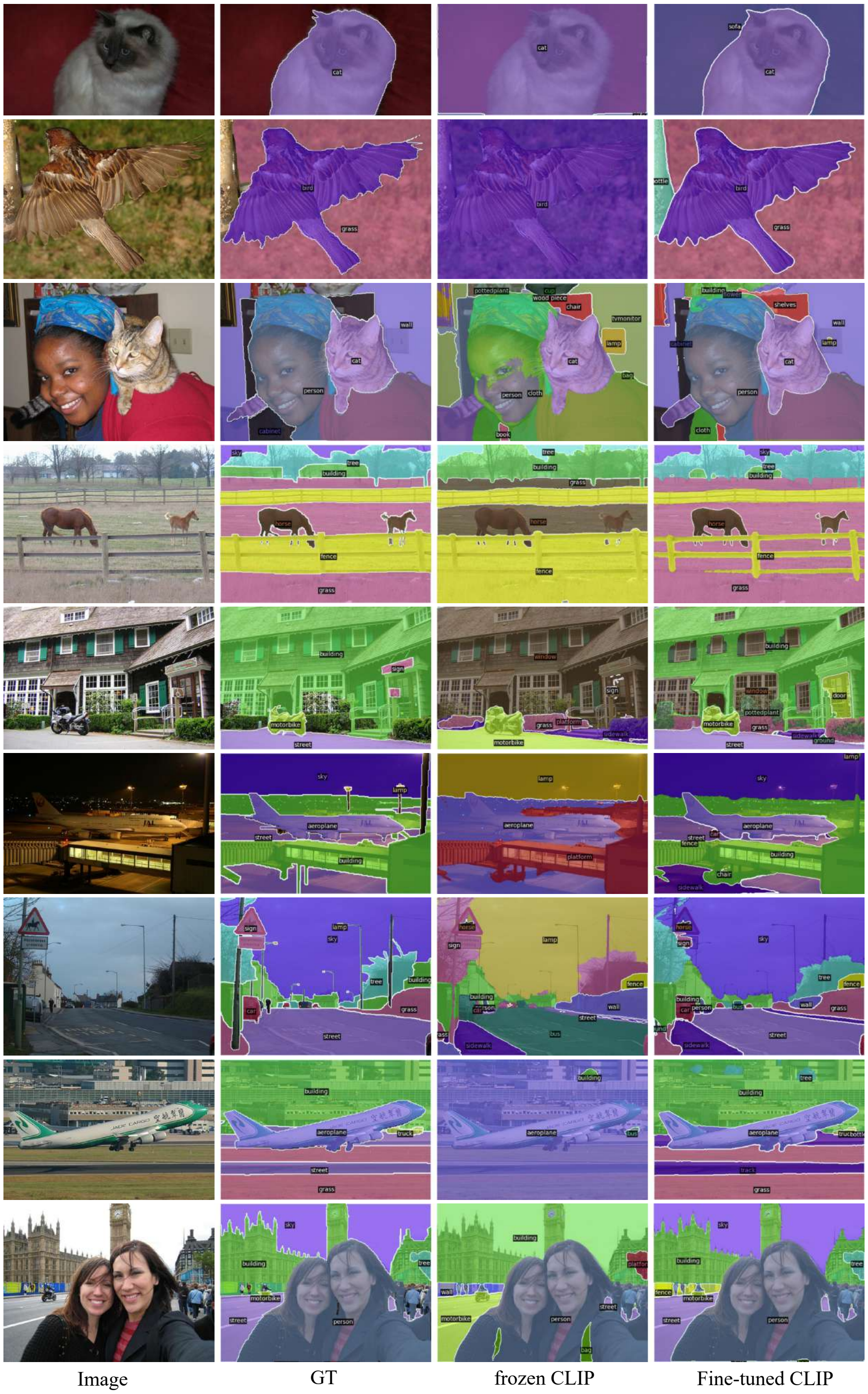


Fig. 2: Qualitative results.

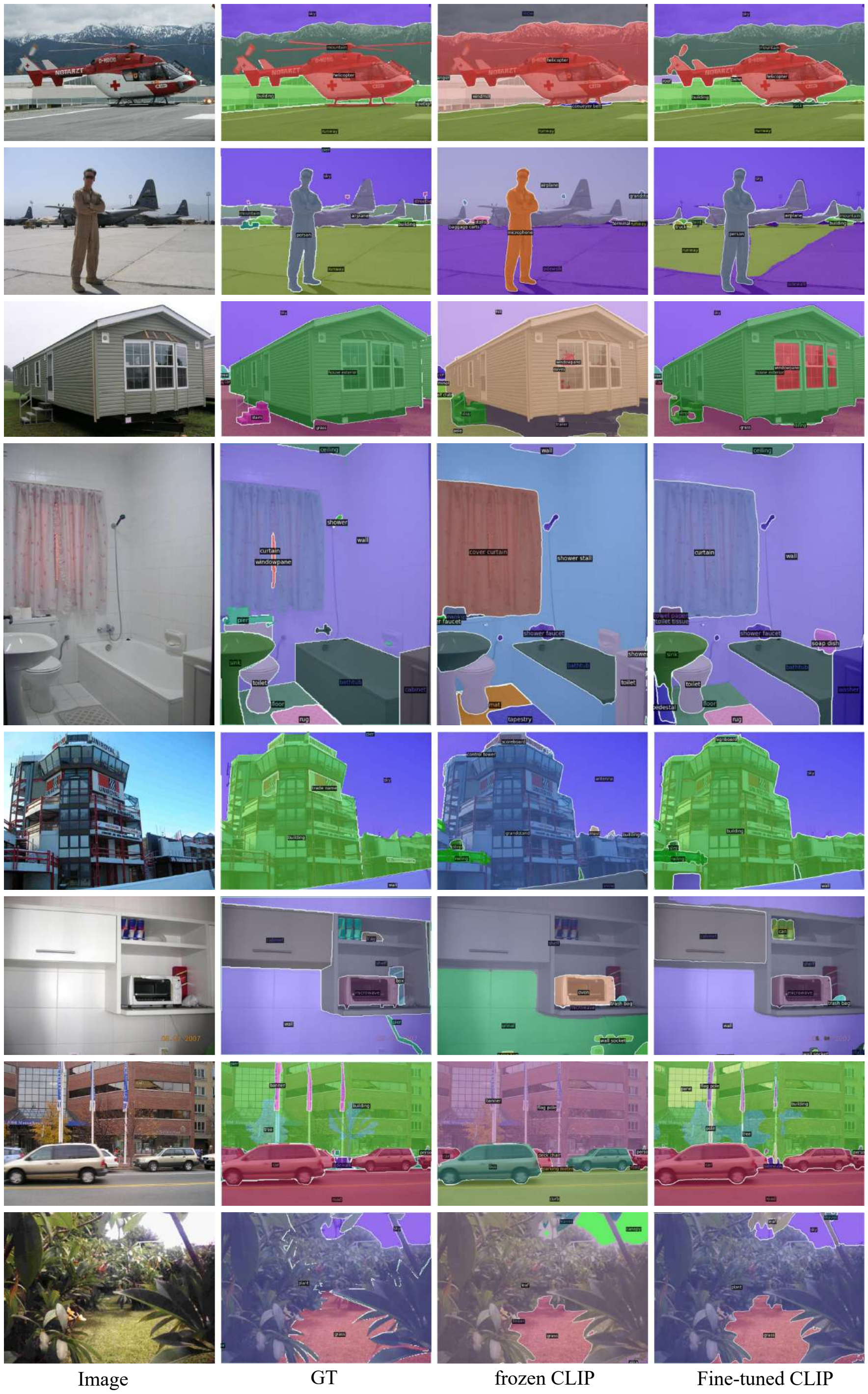


Fig. 3: Qualitative results.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
3. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)
4. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems **36** (2023)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
6. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2014)
7. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
8. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
9. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
10. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36** (2023)
11. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017)