The Power of Prior: Training-Free Open-Vocabulary Semantic Segmentation with LLaVA

Supplementary Material

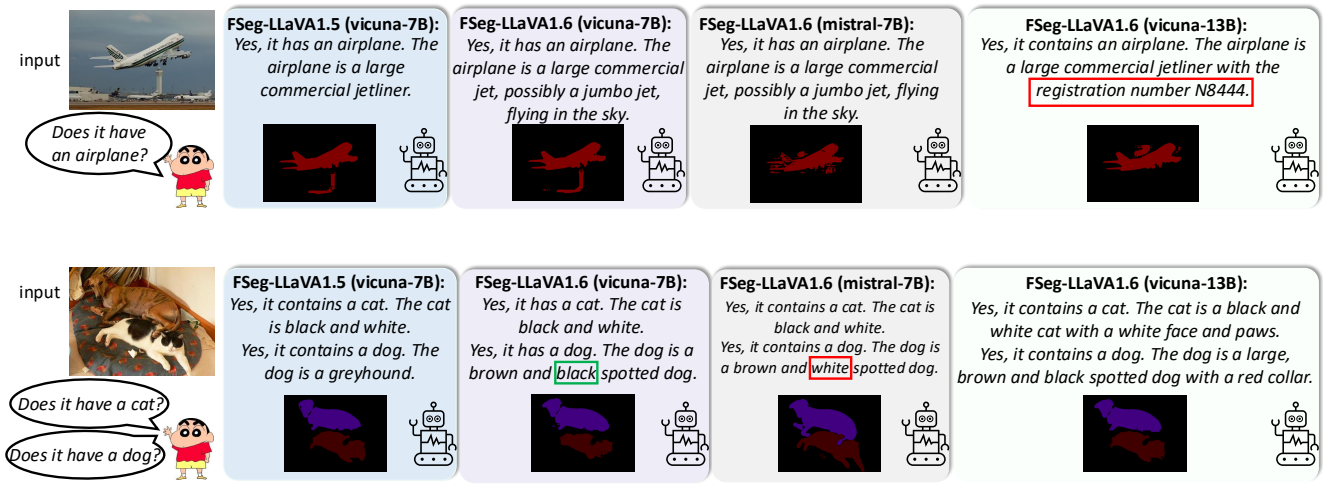

Figure 1. The prediction text response of our FSeg-LLaVA with different backbones and the corresponding segmentation masks. "Red box" denotes a wrong text response. "Green box" denotes a comparison case. It can be found that our method can predict more accurate and comprehensive descriptions of the targets with a larger model size. While there are also some mistakes under different model sizes, the corresponding segmentation results remain unaffected.

1. Overview

In this supplementary material, we demonstrate more implementation details in sec. 2; the text response, including correct and noisy predicted results, in sec. 3; more qualitative results in sec. 4; more ablation study on hyperparameters in sec. 5; and the limitation of FSeg-LLaVA in sec. 6.

2. More Implementation Details

Prompt settings. First, our predefined category setting follows standard OVSS practice, where CLIP-based methods also query each class to determine its presence. Moreover, QAP can be implemented by feeding a batch of category prompts as a single sentence, allowing the model to generate responses in a forward pass. Second, we use the same prompt template across all datasets. This ensures that performance gains originate from LLaVA's prior knowledge rather than prompt engineering.

Inference time. On VOC21, inference is 2× slower than CLIP-based methods. The overhead mainly comes from extracting intermediate LLM features, which precludes using off-the-shelf LLM inference engines (*e.g.*, vLLM). For memory usage, a 7B model fits in a 24 GB GPU; a 14B model requires a 48 GB GPU.

3. Text Response

Fig. 1 illustrates the text responses and corresponding segmentation results produced by our FSeg-LLaVA. We observe that the method can generate highly accurate and detailed descriptions of the target objects, whether in simple cases (single object, *e.g.*, the first example) or in more complex scenarios involving multiple similar objects (*e.g.*, the second example). However, in both scenes, certain misdescriptions still occur, as highlighted by the red boxes. These inaccuracies may stem from differences in the LLaVA backbones. For instance, the simple case shows an error in FSeg-LLaVA1.6 (vicuna-13B), whereas in the complex case, smaller backbones such as FSeg-LLaVA1.6 (mistral-7B) may produce incorrect textual outputs. Moreover, FSeg-LLaVA1.5 (vicuna-7B) and even predicts the totally opposite description, compared to FSeg-LLaVA1.6 (mistral-7B), as marked in the green box and the red box of the second row. Despite these textual inaccuracies, the final segmentation results remain unaffected. Moreover, this phenomenon indicates that the capabilities of multimodal large language models (MLLMs) are worth further exploration.

4. Qualitative Results

Fig. 2 depicts more qualitative results of our FSeg-LLaVA from different datasets, especially in complicated scenes. It

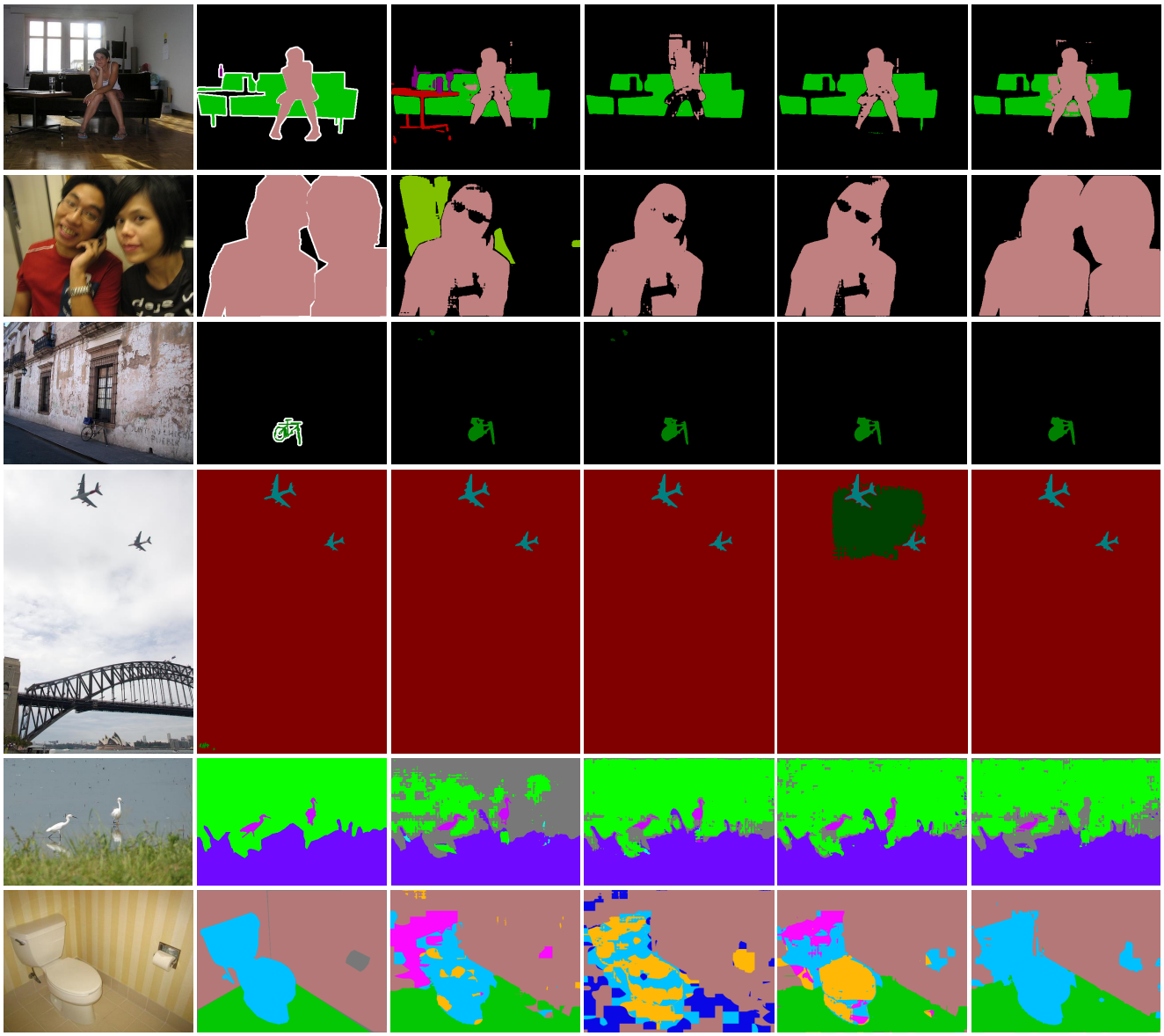

Figure 2. More qualitative results of our method under different backbones. It can be observed that our method can predict better results with larger model sizes. Moreover, in complex scenes, a larger model size can predict wide categories.

Table 1. Performance on VOC21 using different α.

α	0.25	0.5	0.75	0.85	0.95
FSeg-LLaVA1.5	15.2	60.5	**68.0**	67.3	67.3

Table 2. Performance on VOC21 using different β.

β	0.12	0.16	0.19	0.2	0.25
FSeg-LLaVA1.5	47.8	65.3	**68.0**	67.2	58.1

can be observed that in simple cases (*i.e.*, rows 3 and 4), our method with different backbones can all predict quite good segmentation masks. However, the boundary can be further optimized. Sometimes, the FSeg-LLaVA1.6 (mistral-7B) will introduce noise in the prediction, like in row 4. On the other hand, our method can also handle the multi-class case (*i.e.*, rows 1 and 2). The prediction improves with an increase in the size of the backbone. Such results suggest that complex scenarios require more comprehensive MLLM features as guidance for segmenting different ob-

jects. Besides, for more complex cases (*i.e.*, rows 5 and 6), a larger model size will predict relatively accurate segmentation masks for both foreground objects and background. Therefore, our method verifies that existing MLLMs have the ability to segment, and we believe our method can lead to a new research direction of training-free open-vocabulary semantic segmentation.

5. More Ablation Study on Hyperparameters

The hyperparameters α and β are shared across LLaVA configurations, but differ for LLaMA and Mistral, reflecting their distinct pre-training knowledge and internal representations. We also add ablations for α and β in Tab. 1 and Tab. 2. α must be relatively large to ensure accurate activation, while β constrains to avoid overly small activations that are hard to distinguish from noise.

6. Limitation

Our method has three possible improvements. First, it produces only one class mask for each inference, requiring multiple queries for full segmentation, which is inefficient when the class set is huge. Second, it relies on SAM [2] for mask generation, which requires spatial prompts, whereas object classes often admit clear prompts (*e.g.*, dog, cat). In contrast, stuff classes (*e.g.*, sky, grass) lack localized cues, making it inaccurate to segment them. Third, we only leverage LLaVA [3, 4] to tackle OVSS, stronger alternatives like Qwen-VL [5] or InternVL [1] remain unexplored.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 3

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pages 34892–34916, 2023. 3

[4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 3

[5] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3