Supplementary Material for "Unified Open-World Segmentation with Multi-Modal Prompts"

Yang Liu[1]* Yufei Yin[2]* Chenchen Jing[3] Muzhi Zhu[1] Hao Chen[1]† Yuling Xi[1]
Bo Feng[4] Hao Wang[4] Shiyu Li[4] Chunhua Shen[1]

[1] Zhejiang University [2] Hangzhou Dianzi University [3] Zhejiang University of Technology [4] Apple

A. Discussion and Limitation

Closed-Set Segmentation. To enhance open-world generalization, COSINE sacrifices some performance in closed-set scenarios. For example, on COCO, COSINE achieves a 50.6 PQ and 42.0 AP, while OpenSeeD obtains 59.5 PQ and 53.2 AP, and DINOv achieves 57.7 PQ and 50.4 AP. However, COSINE outperforms these models on unseen datasets. We argue that pre-trained foundation models capture a broader range of visual knowledge. Fine-tuning these models on a limited segmentation dataset can lead to performance improvements in closed-set scenarios, but it may reduce their generalization ability in unseen scenarios. Therefore, unlike existing methods, which train all model parameters, COSINE uses frozen foundation models. We believe that the model's ability to generalize to open-world scenarios is more critical.

Model Pool. The Model Pool explores a limited set of foundation models. In our preliminary experiments, we investigated the impact of the SAM encoder but did not observe significant performance improvements. Additionally, it introduced greater computational cost and constrained the input image resolution.

CLIP Embeddings. While COSINE leverages CLIP embeddings to enable open-vocabulary segmentation, its reliance on CLIP alone restricts the capacity to capture fine-grained visual attributes, such as texture, material properties, or nuanced object conditions. To address this limitation, COSINE introduces a flexible framework designed to facilitate multimodal collaboration across diverse foundation models. This architecture allows seamless integration of more expressive backbones, such as multimodal large language models, in future extensions, thereby enhancing the model's ability to support fine-grained, open-world segmentation. We envision COSINE as a step forward in bridging vision-language representations and structured segmentation tasks. Its generality and extensibility offer a promising foundation for future research aimed at fine-grained understanding and long-tail concept segmentation in complex, open-world scenarios.

Discrepancy with [1]. Unlike [1], which suggests that multimodal in-context learning is predominantly driven by textual signals with minimal contribution from the visual modality, COSINE demonstrates clear cross-modal synergy, as shown in Table 5. We attribute the discrepancy with [1] to the use of weaker image encoders in their framework. As highlighted in [15, 16], strong visual backbones (e.g., DINOv2) are essential for enabling effective image-conditioned reasoning in multimodal models.

Relation to Other Paradigms. In addition to open-vocabulary and in-context segmentation, another important paradigm in open-world segmentation focuses on anomaly or out-of-distribution (OOD) detection. These methods identify pixels or instances from unseen categories as anomalies, which can then be incrementally learned [3, 14]. COSINE offers a multimodal approach that effectively addresses these challenges, serving as a promising direction for advancing open-world segmentation.

Limitations. Although our experiments validate that foundation models, such as DINOv2 and CLIP, exhibit complementary information, this work does not explore more advanced models with alternative training strategies, such as MLLMs [5, 17, 19] and diffusion models [10, 13]. Furthermore, while COSINE leverages multiple foundation models to achieve complementary information and enhance generalization in open-world scenarios, it inevitably introduces higher computational costs. One potential solution is to distill the knowledge from different models into a single model. These challenges will be the focus of our future work.

Broader Impacts. Our approach is built upon open-source foundation models and only trains a lightweight decoder, which significantly reduces both training costs and carbon emissions. We do not anticipate any significant ethical or social concerns now.

B. Implementation Details

Training Details. We establish the frozen Model Pool by leveraging DINOv2 (ViT-L) [9] and CLIP (ConvNeXt-

*Equal contribution. †HC is the corresponding author.

Model		Prompt		LVIS-92^i		ADE20K		
DINOv2	CLIP	vision	text	1-shot	5-shot	PQ	AP	mIoU
✓		✓		24.3	27.7	-	-	-
✓	✓	✓		24.5	27.8	-	-	-
	✓		✓	-	-	6.6	2.3	26.8
✓	✓		✓	-	-	13.2	7.6	30.2
✓	✓	✓	✓	27.7	32.1	17.7	8.1	30.4

Table S1. Effect of different models and training branches. All models are trained for 10k steps.

Large) [7, 12] as foundational models, while only training lightweight SegDecoder modules. Specifically, the single-scale and multi-scale variants of the SegDecoder contain 25M and 32M trainable parameters, respectively. The Image-Prompt Aligner has one block and the Multi-Modality Decoder has six blocks. All training data is converted to instance masks, and stuff classes are treated as single-instance categories. So we train only for instance segmentation, and merge instances by class at inference for semantic segmentation. We optimize COSINE for 50K steps with a batch size of 64 using the Adam optimizer [8] ($\beta_1 = 0.9, \beta_2 = 0.999$). A linear learning rate scheduler is employed with a base learning rate of $1e{-}4$ and a 100-step warmup phase. The weight decay is set to 0.05. For COCO and Objects365, we apply random horizontal flipping and large-scale jittering (LSJ) [2] with a random scale sampled from range 0.1 to 2.0, followed by a fixed-size crop to 896×896 for DINOv2. For CLIP, the images are resized to 1024×1024 before being inputted. For referring segmentation datasets, we only resize the images without flipping and cropping operations.

Evaluation. For one-shot semantic segmentation, the in-context examples are from the support sets. Like [6], we simply concatenate diverse image examples to accommodate the few-shot learning scenario. For few-shot instance segmentation, we randomly select 10 samples (or all available samples if fewer than 10 are present) for each category. We integrate the representations of image prompts and text prompts to form the token features. We enhance the classification score by pooling CLIP features using the predicted masks, thereby improving the generalization capability of the model. Our method can seamlessly adopt the approach of [18] to perform open-vocabulary tasks. For VOS, we select the first frame of the video as the image example and deploy a memory mechanism to store intermediate results, following [6]. For referring segmentation, we adhere to the evaluation pipeline of LISA [4].

C. Additional Results

Effect of different models and training branches. We investigate the impact of different foundation models across various training branches. As shown in Table S1, DINOv2 and CLIP are commonly used foundation models for in-context and open-vocabulary segmentation tasks, respectively. The introduction of additional models further enhances performance on these tasks. When different models are jointly used for multi-modal training, complementary information is shared, enabling the models to collaborate more effectively and achieve stronger generalization performance.

Temporal consistency assessment. To comprehensively assess the temporal stability of video segmentation methods, we report the STB [11] scores on the DAVIS 2017 dataset. The results are as follows: Painter achieves an STB score of 0.91, while both SegGPT and COSINE reach 0.96, indicating significantly higher temporal consistency.

Visualizations. As shown in Fig. S1, we visualize the segmentation results under in-context settings, including example-based semantic segmentation, example-based instance segmentation and video object segmentation. As shown in Fig. S2, we visualize the open-vocabulary segmentation and referring segmentation. These results demonstrate that COSINE achieves highly accurate predictions across various modalities and granularities, highlighting its strong potential for open-world generalization.

References

[1] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1

[2] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[3] Brian KS Isaac-Medina, Yona Falinie A Gaus, Neelanjan Bhowmik, and Toby P Breckon. Towards open-world object-based anomaly detection via self-supervised outlier synthesis. In *Eur. Conf. Comput. Vis.*, 2024. 1

[4] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Adv. Neural Inform. Process. Syst.*, 2023. 1

[6] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. *arXiv preprint arXiv:2410.04842*, 2024. 2

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2:

Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, 2023. 1

[11] Mingyang Qian, Yi Fu, Xiao Tan, Yingying Li, Jinqing Qi, Huchuan Lu, Shilei Wen, and Errui Ding. Coherent loss: A generic framework for stable video segmentation. *arXiv preprint arXiv:2010.13085*, 2020. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 2

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[14] Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-world semantic segmentation including class similarity. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1

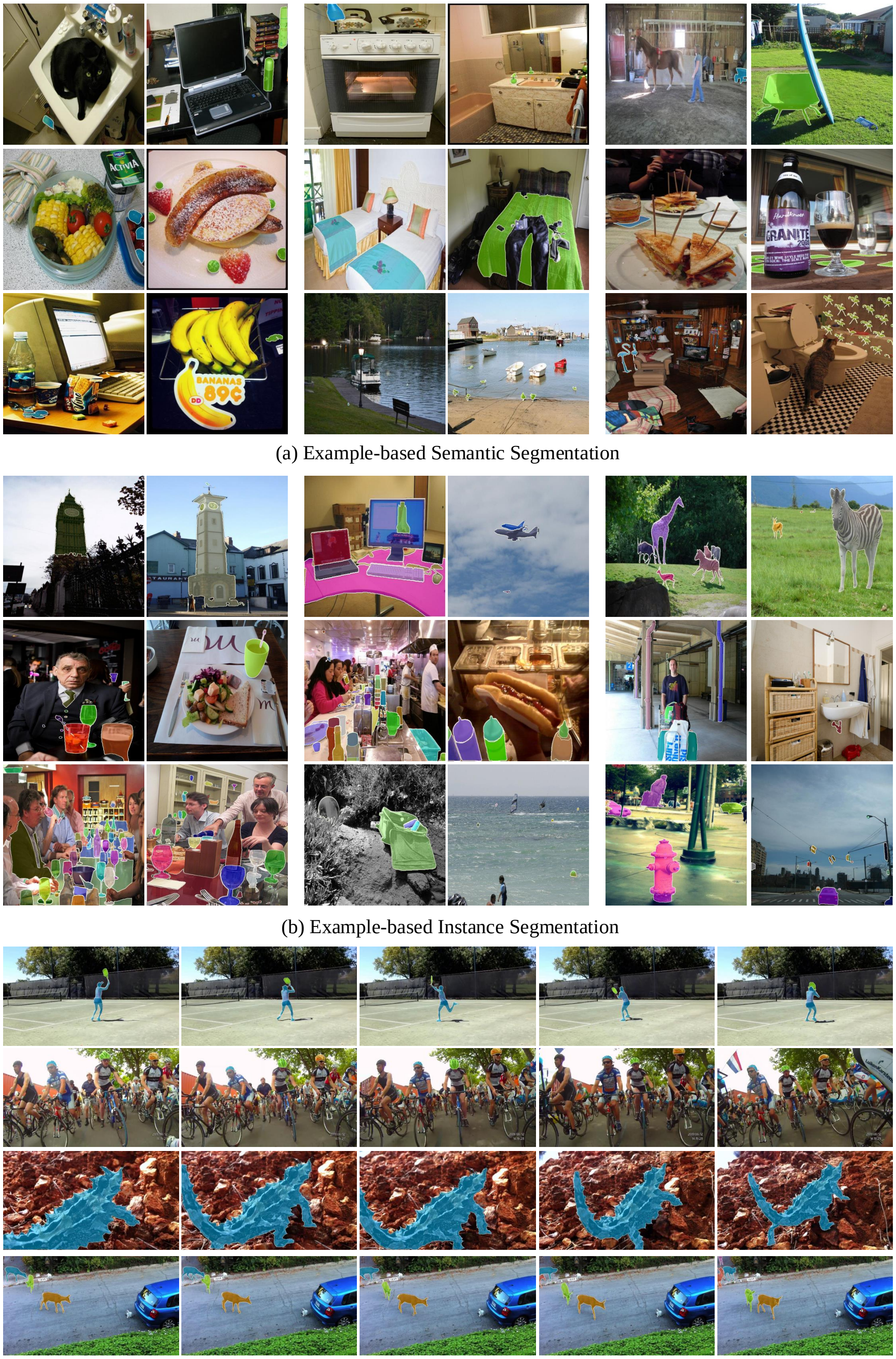
[15] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Adv. Neural Inform. Process. Syst.*, 2024. 1

[16] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1

[17] 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. 1

[18] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Adv. Neural Inform. Process. Syst.*, 2023. 2

[19] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

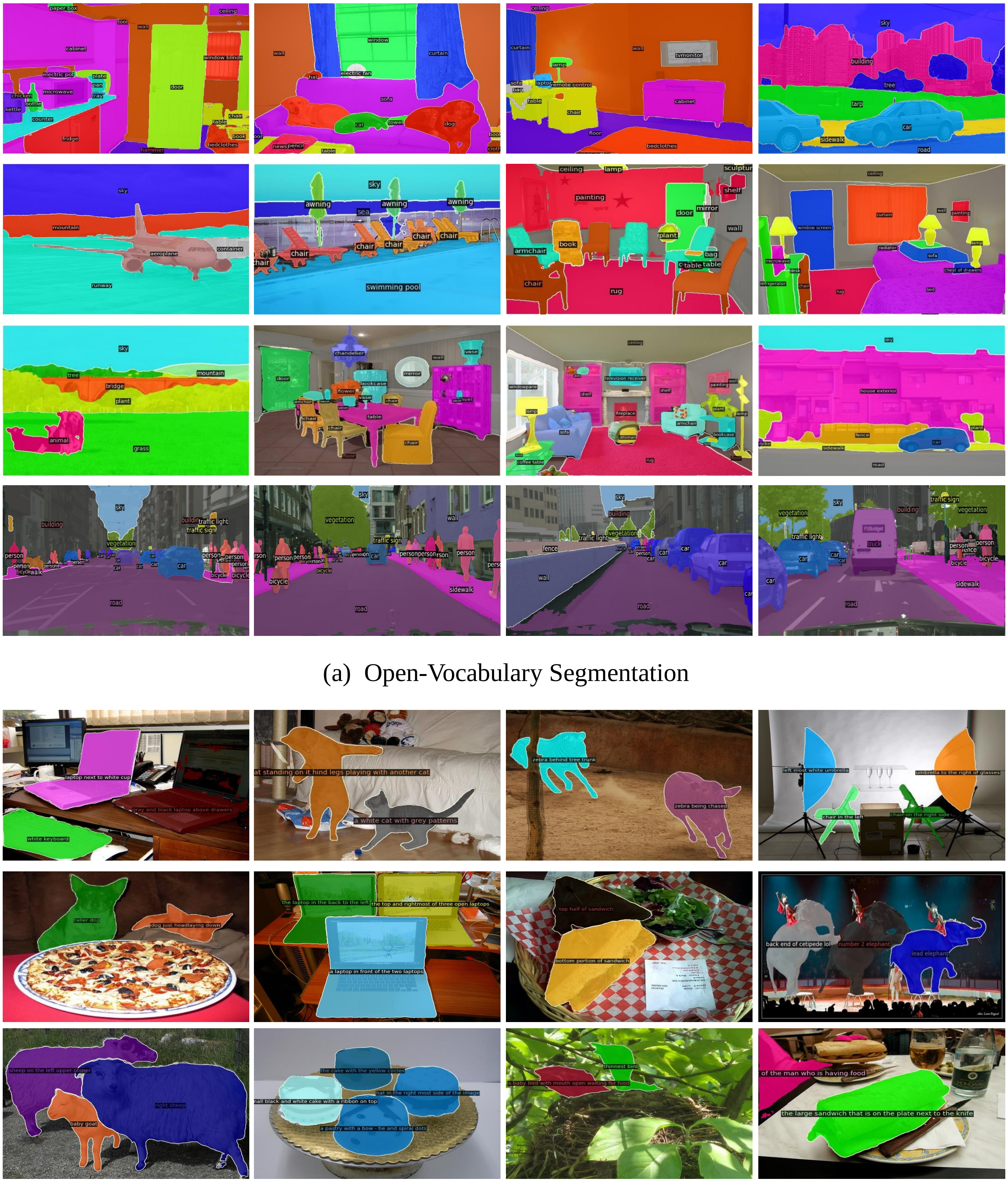


(a) Example-based Semantic Segmentation

(b) Example-based Instance Segmentation

(c) Video Object Segmentation

Figure S1. Visualizations of in-context segmentation tasks. (a) Example-based semantic segmentation on LVIS dataset. The left image with the blue mask is the image example, and the right image with the green mask is the result. (b) Example-based instance segmentation on LVIS dataset. We will obtain instance outputs sharing the same classes with the given image prompt. (c) Video object segmentation on the YouTuBe-VOS 2019 dataset.



(a) Open-Vocabulary Segmentation

(b) Referring Segmentation

Figure S2. Visualizations of open-vocabulary segmentation and referring segmentation.