Learning to Focus and Precise Cropping: A Reinforcement Learning Framework with Information Gaps and Grounding Loss for MLLMs

Supplementary Material

8. Training Dynamics Analysis

In this section, we visualize and analyze evolution of several key metrics during the first and second stages of training.

8.1. Stage-I

To analyze the model's behavior during the first stage of training, we present the evolution of four key metrics for both the BaseLine and our Stage-I Model with Info Gap in Figure 5. These metrics are: the total reward, the number of tool calls (Tool Call Num), the Intersection over Union (IoU) of predicted crop regions with ground truth, and the overlap ratio (Overlap), which measures the proportion of ground-truth region covered by the model's predicted crops.

Reward. As shown in the 'Reward' plots, the BaseLine model (Figure 5 (a)) exhibits a faster initial increase in reward. We attribute this to uncompressed images in its training set, which present a simpler learning task. Critically, both models eventually converge to a similar reward level. This demonstrates that the introduction of the information gap compels Stage-I to achieve a comparable performance level on more challenging (compressed) data, demonstrating the model's effective utilization of the crop.

Tool Calls. The 'Tool Call Num' for both models increases rapidly and then stabilizes. Notably, the Stage-I Model (Figure 5 (b)) converges to a higher number of tool calls. We hypothesize this is because the compressed input image necessitates more exploratory crops to gather sufficient information to answer the question. This behavior confirms that the Stage-I Model learns to actively utilize the tool to attend to relevant image regions.

IoU. The 'IoU' metric, which evaluates the precision of the cropped regions, shows no significant improvement for either model throughout the training process. This suggests that without an explicit grounding reward, the model is not incentivized to refine the precision of its crops. The consistently higher IoU of the BaseLine is likely due to the uncompressed images making box prediction an easier task.

Overlap. The 'Overlap' reveals a distinct difference between the training progress of BaseLine and Stage-I. The BaseLine's overlap remains relatively stable without a clear upward trend. In contrast, the Stage-I's overlap shows a significant increase before plateauing. This indicates that while image compression initially makes it difficult for the model to identify question-relevant regions, the model learns a stronger region identification capability, driven by the accuracy reward, to correctly answer the question.

8.2. Stage-II

The Figure 6 illustrates the evolution of four key metrics during the second training stage, with and without the use of the grounding reward. These metrics are: IoU, overlap, the ratio of the predicted crop box area to the image area (Bbox Ratio), and the visual token number of the predicted cropped region (Token Num).

IoU. Comparing the IoU evolution in Figures 6 (a) and (b), we observe a distinct upward trend in (b) during Stage 2 training. This indicates that the introduction of the grounding reward enables the model to localize key regions with greater precision.

Bbox Ratio and Token Num. Comparison for the Bbox Ratio reveals that this metric remains largely constant in (a), whereas it decreases significantly in (b), leading to a subsequent reduction in the Token Num. This demonstrates that the grounding reward encourages smaller crop regions, thereby improving the model's inference efficiency.

Overlap. As shown in (b), the Overlap metric exhibits a moderate increase. This suggests that the model is not indiscriminately shrinking the crop area; rather, it is learning to prune redundant, non-critical parts of the region. Consequently, the resulting crop contains less distracting information, making it more conducive for the model to answer questions based on the visual content.

Setting	HR-Bench 8K	HR-Bench 4K	V^*
prediction	41.7	47.0	55.6
GT	50.0	53.0	60.3

(a) GT Test.

Setting	HR-Bench 8K	HR-Bench 4K	V^*
prediction	100.0	100.0	100.0
RandNoise	100.0	80.8	100.0

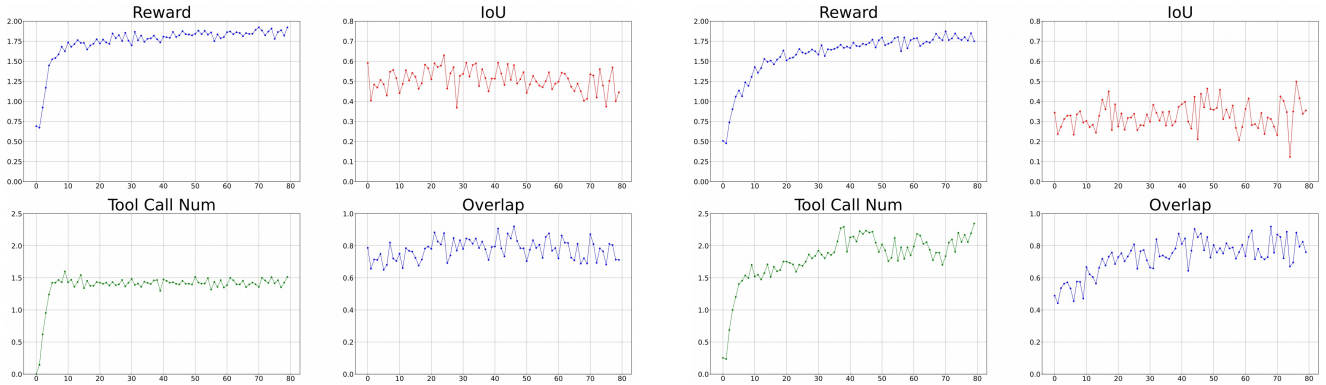
(b) Noise test.

Table 7. Results for more rigorous preliminary analysis.

9. More Preliminary Analysis

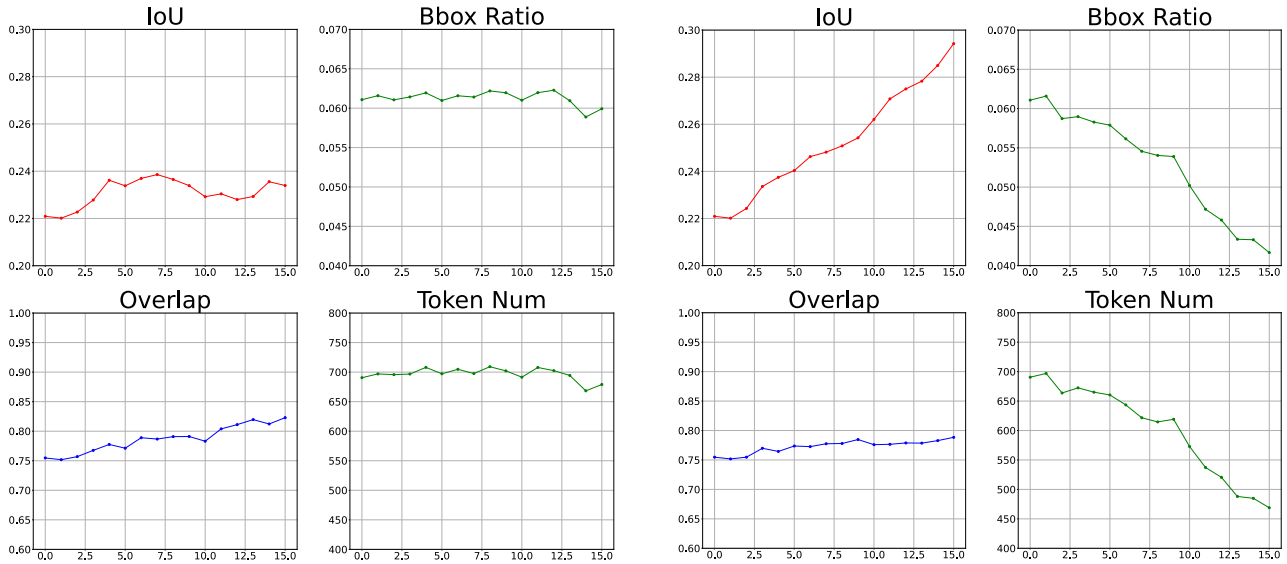
To more rigorously demonstrate DeepEyes's insufficient attention to cropped regions, we also conducted the following experiments. All experiments in this section are conducted with the maximum number of visual tokens set to 1,024.

GT test. We isolate samples from benchmarks where regions predicted by DeepEyes poorly cover GT ($Overlap \leq$



(a) Traing progress of BaseLine (b) Traing progress with Info Gap

Figure 5. Training progress of BaseLine and Stage-I model.



(a) Training progress without grounding reward (b) Traing progress with grounding reward

Figure 6. Training progress of Stage-II.

0.2). As shown in Table 7a, Replacing the regions of these samples with GT only yields returns within 10% (significant growth be expected if model attend to these regions). This provides a more rigorous proof that DeepEyes cannot fully utilize cropped regions.

Noise test. We select samples where DeepEyes fails when forced to answer directly (without cropping pattern) but succeeds with cropping. As whon in Table 7b, when replacing the cropped region with noise, the model still answers most of these samples correctly. This demonstrates that the performance gain is attributed to the cropping pattern (structural cues) rather than the enhanced visual information within the crops.

10. More Ablation Studies

Data Utilization Strategy. As shown in the Table 8, we compare two data usage strategies. The first strategy, 'Mix Data', involves training the model in a single stage by mixing our collected data with the Visual Probe data at a 1:1 ratio. The second strategy, which we term 'Stage-II', is the two-stage approach adopted. In the first stage, the model is trained exclusively on our collected data. Subsequently, in the second stage, we train the model with a small amount of Visual Probe data, incorporating a grounding reward. The results demonstrate that our proposed two-stage strategy achieves superior performance.

Resolution Selection Strategy. To demonstrate the effectiveness of our resolution selection strategy in the first stage, we design two simple baselines, 'Hard' and 'Random', and

Model	HR-Bench 8k			HR-Bench 4k			V^*		
	FSP	**FCP**	**Overall**	**FSP**	**FCP**	**Overall**	**Attr**	**Spatial**	**Overall**
Mix Data	77.5	59.0	68.3	85.0	61.0	73.0	82.6	73.7	79.1
Stage-II	83.3	60.8	72.1	86.8	63.7	75.3	82.6	77.6	80.6

Table 8. Comparison between different data utilization strategies. Mix Data: Using the mixture of our collected data and Visual Probe data to train the model in a single stage. Stage-II: First, train the model on our collected data, and then train it with a minimal amount of Visual Probe data with grounding reward.

Model	HR-Bench 8k	HR-Bench 4k	V Star
Hard	60.5	63.2	71.1
Random	59.0	64.0	72.4
Answer	62.0	63.5	75.0

Table 9. Comparison of accuracy between different resolution selection strategies in the multi-region case.

Design	FSP		FCP		Acc
	Acc	IoU	Acc	IoU	
IoU Reward	81.8	39.3	63.5	47.4	72.7
- IoU Thresh	79.0	29.3	64.2	39.0	71.6
- IoU Reward	79.0	30.6	63.7	37.7	71.2

Table 10. Comparison between different reward designs in Stage-II on HR-Bench 8k.

compare them against our method on the multi-target subset (FCP / Spatial) of the dataset. This subset is chosen because it is more challenging than the single-target subset (FSP / Attr), and thus better highlights the efficacy of the information difference mechanism in our first stage. In the 'Hard' baseline, we consistently select the most heavily downsampled image (where $max(h, w) = 224$). In the 'Random' baseline, we randomly select one image from the pool of all images generated by downsampling the original image at various scales. As shown in the Table 9, our method outperforms both of these simple baselines.

Reward Design in Stage-II. We compare different reward designs in Stage-II on the HR-Bench 8k. As shown in the Table 10, the first row presents our design where an IoU reward is employed during the second-stage training (without the L_1 reward). In the second row, we remove the condition that the IoU reward is only applied when the overlap exceeds a certain threshold. The experimental results demonstrate that retaining the conditional IoU reward (i.e., applying it only when the overlap is above the threshold) leads to higher IoU and accuracy. This suggests that this condition effectively guides the model to perform more precise cropping. Furthermore, the models trained with an IoU reward consistently outperform those without it in terms of accuracy and IoU. This indicates that the performance gain from the Stage-II training is primarily attributed to the IoU reward itself, rather than simply the introduction of additional training data.

11. Benchmarks and Metrics Details

Our method is evaluated on three benchmarks. The first, **HR-Bench 8K** with an average resolution of 7680, which consists of two sub-tasks: Fine-grained Single-instance Perception (FSP) and Fine-grained Cross-instance Perception (FCP). The 8K images are cropped around the objects in question to produce **HR-Bench 4K**. The third, $\boldsymbol{V^*}$, with an average resolution of 2246x1582, features sub-tasks in attribute recognition (Attr) and spatial reasoning (Spatial). We evaluate our model on three datasets: hr-8k, hr-4k, and vstar. For all three benchmarks, the evaluation metric is **accuracy** (Acc), defined as the number of questions answered correctly. Additionally, to assess the cropping precision of our model, we also employ the **Intersection over Union** (IoU). This is calculated as the IoU between the model's predicted cropping box and the GT box of the question-relevant region.

Method		**Time (A100-hours)**
DeepEyes		480.0
Ours	Stage-I	320.0
	Stage-II	96.0
	Offline Res. Selection	2.5
	Total	**418.5**

Table 11. Comparison of Training Time.

12. Training Details

We show the related hyper-parameters we use in Table 12. We also compare our training time with that of DeepEyes as shown in Table 11. Our total training time is shorter than that of DeepEyes. This is because we reduce the number

Parameter	Value
train batch size	256
rollout num per sample	16
ppo mini batch size	32
ppo micro batch size per gpu	2
rollout log prob micro batch size per gpu	4
ref log prob micro batch size per gpu	4
single response max tokens	2048
max turns	5
kl loss coef	0.0
entropy coeff	0.0
nodes num	1
gpus num per node	8
learning rate in Stage-I	1e-6
learning rate in Stage-II	5e-7
overlap threshold (τ)	0.9
acc reward weight (r_{acc})	0.8
format reward weight (r_{format})	0.2
tool call weight (r_{tool})	1.2

Table 12. The hyperparameters we used in the training pipeline.

of input visual tokens in the first stage by compressing the original images, accelerating the model rollout process.

13. Visualization Analysis

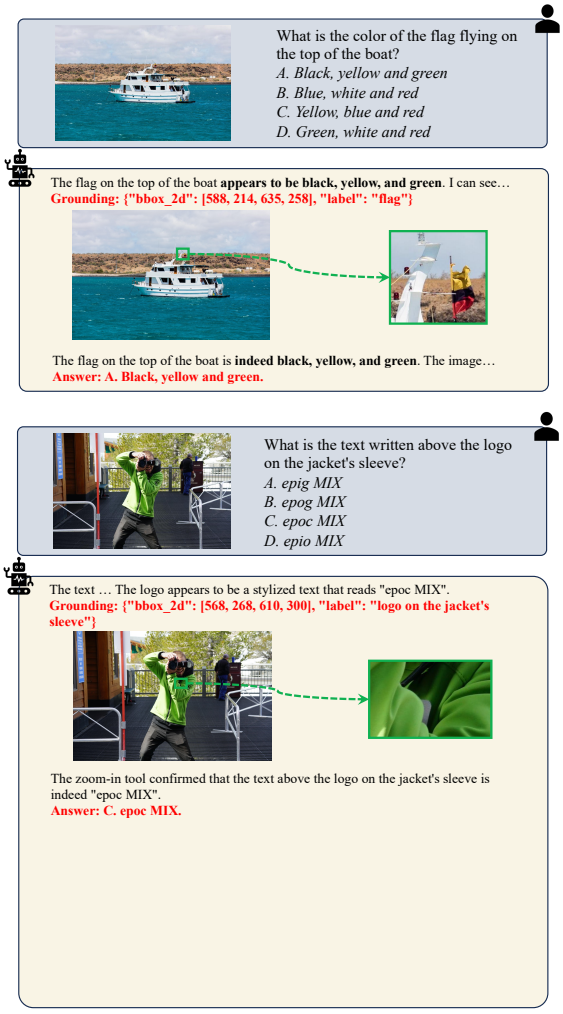
In Figure 7, we visually analyze the inference processes of DeepEyes and Stage-I. In the first example, while both models correctly crop the flag, DeepEyes provides an incorrect answer, whereas Stage-I arrives at the correct one. In the second example, both models initially fail to crop the jack's sleeve. However, DeepEyes proceeds to answer the question even with the jack's sleeve absent from the cropped region. In contrast, Stage-I identifies the absence of the jack's sleeve in the initial crop, performs a second cropping action, and ultimately succeeds in locating it and answering the question correctly. In Figure 8, we visualize and analyze the inference processes of the Stage-I and Stage-II models. In the first crop, both models fail to capture the target car and subsequently enlarge their cropping regions in the second attempt. However, the Stage-I model's second crop includes excessive redundant areas, causing it to fail again in identifying the car. In contrast, the Stage-II model accurately crops the entire road area, leading to a correct answer.

14. Limitations and Future Works

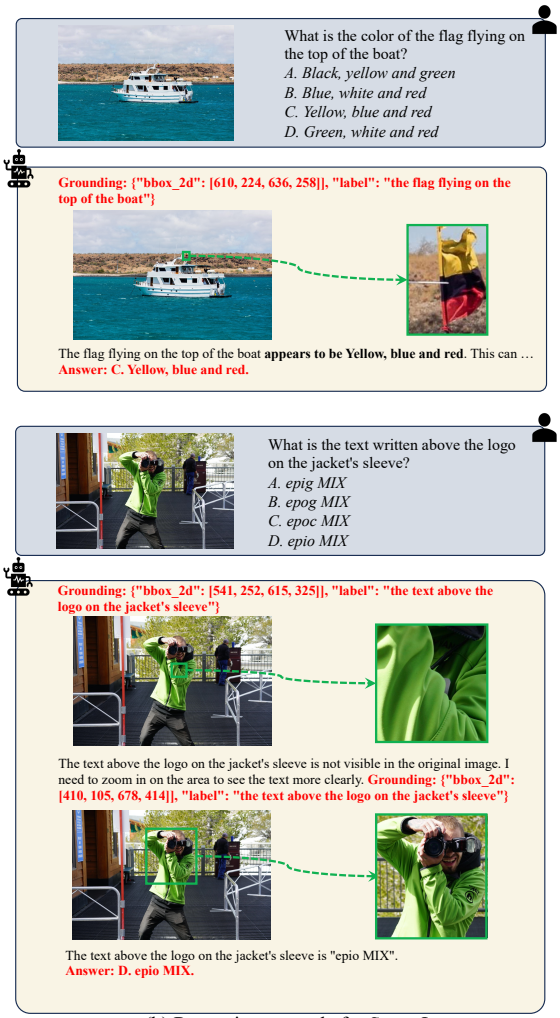
In this work, we first identify a critical issue in existing agent-based workflows for complex image understanding: the sub-optimal tool invocation that stems from a rigid formalization of the cropping tool. We address this by proposing an information gap mechanism. Building upon this, we further enhance model's cropping precision and, consequently, its overall performance by manually annotating a small set of bboxes and introducing a grounding reward.

However, the performance gain from the second stage of training is relatively limited, which we attribute to the small number of annotated bboxes. This suggests that leveraging synthetic data [8, 10], could address the data scarcity problem in high-resolution VQA. For instance, a large-scale, high-resolution VQA dataset with bbox annotations could be synthesized by cropping objects from general-purpose VQA datasets like GQA [1] (which are low-resolution but rich in annotations including bbox) and pasting them onto high-resolution backgrounds.

Furthermore, current methods only contains cropping tool, whose effectiveness is constrained by the inherent grounding ability of the MLLM and the fact that cropped regions contain background distractors. For future, a promising direction is to integrate segmentation models—such as open vocabulary semantic segmentation [3, 9, 11], referring expression segmentation [2, 6], or in-context segmentation [4, 5, 7] —as callable tools for the agent.

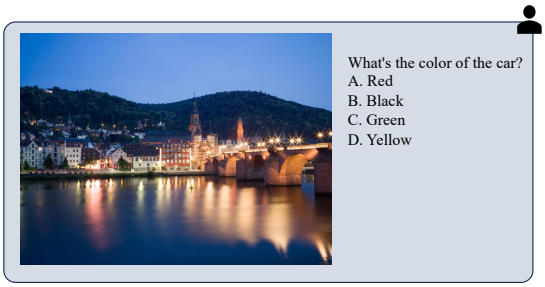


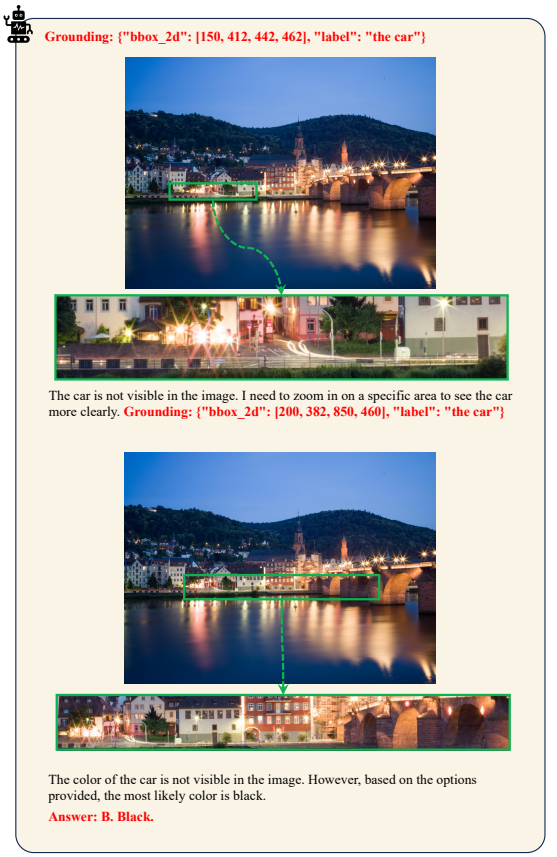
(a) Reasoning example for DeepEyes.



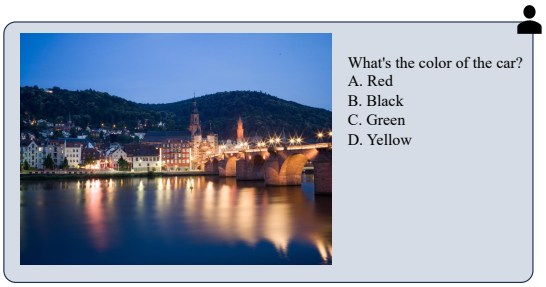
(b) Reasoning example for Stage-I.

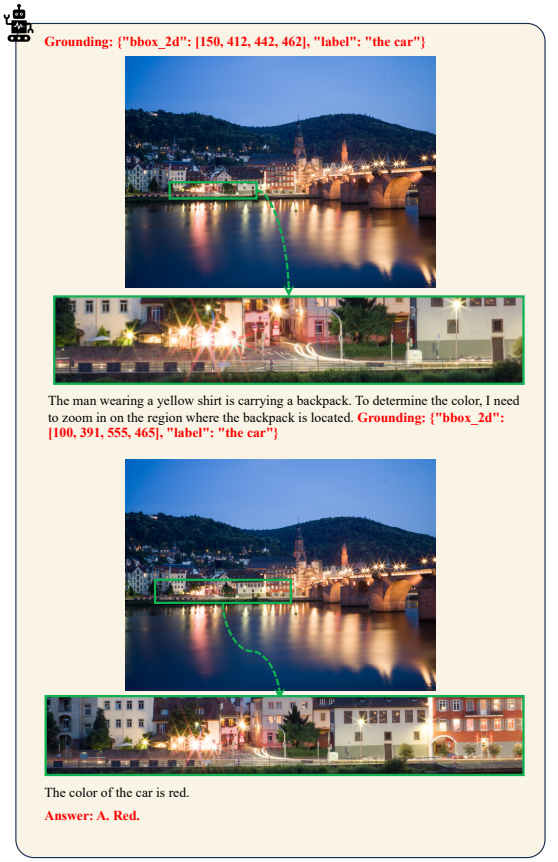
Figure 7. Comparison between DeepEyes and Stage-I.





(a) Reasoning example for Stage-I.





(b) Reasoning example for Stage-II.

Figure 8. Comparison between Stage-I and Stage-II.

References

[1] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 4

[2] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024. 4

[3] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 4

[4] Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. Towards more unified in-context visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13362–13372, 2024. 4

[5] Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Tao Gong, Bin Liu, Jing Han, Wenbin Tu, Shengwei Xu, et al. Unicl-sam: Uncertainty-driven in-context segmentation with part prototype discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20201–20211, 2025. 4

[6] Hao Tang, Chenwei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. *arXiv preprint arXiv:2503.01342*, 2025. 4

[7] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 4

[8] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023. 4

[9] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023. 4

[10] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pages 42098–42109. PMLR, 2023. 4

[11] Xuanpu Zhao, Dianmo Sheng, Zhentao Tan, Zhiwei Zhao, Tao Gong, Qi Chu, Bin Liu, and Nenghai Yu. Training-free open-vocabulary semantic segmentation via diverse prototype construction and sub-region matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10474–10482, 2025. 4