AlignSAM: Aligning Segment Anything Model to Open Context via Reinforcement Learning (Supplementary Materials)

Duojun Huang[1,2] Xinyu Xiong[1] Jie Ma[1] Jichang Li[1,3] Zequn Jie[4] Lin Ma[4] Guanbin Li[1,2†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]GuangDong Province Key Laboratory of Information Security Technology
[3]The University of Hong Kong [4]Meituan

We present additional implementation details and analysis of our proposed method AlignSAM in this supplementary material.

1. Implementation Details

We illustrate the textual prompts for the implentation of explict branch on different datasets in Table 1. The "text Prompt" in the last column is utilized as textual input of the CLIP-Surgery Model.

2. Comparisons of Tuning Paradigms

This paper introduces a unified framework based on reinforcement learning to enable effective and efficient prompting for the vision foundation model SAM, without the need to access the parameters of the backbone. We here report the comparison results of our and existing paradigms for adapting SAM into downstream tasks, in terms of five desirable properties, including frozen backbone (**F**), automatic inference (**A**), gradient-free (**G**), source-free (**S**), and interpretability (**I**). Specifically, freezing the parameters (**F**) of the foundation model is required to alleviate the burden of training costs due to the giant scale of the backbone model. "**A**" means automatic inference without additional guidance in the testing phase. Gradient-free methods (**G**) do not require gradient information from the intermediate layer of the foundation model, which can be expensive to compute during training. Source-free (**S**) methods do not require training (reference) samples in the testing phase which may be inaccessible due to privacy issues. Interpretability (**I**) implies that the prediction results are explicitly and strongly connected to the provided prompts for the foundation model. As depicted in Table 2, our proposed AlignSAM combines various desirable properties that can be utilized for diverse downstream tasks.

Datasets	Target Foreground	Text Prompt
CUHK [3]	Defocus background	defocus background
SBU [4]	Shadow of any object	shadow
MSD [6]	Mirror face	glass
DUTS [5]	Visually salient objects	saliency object
Pascal-VOC [2]	Common categories	aeroplane/ bottle/...

Table 1. Summary of textual prompts for different datasets.

Methods	**F**	**A**	**G**	**S**	**I**
Manual Prompting	✓	✗	✓	✓	✓
Full Fine-tuning	✗	✓	✗	✓	✗
Adapters / LoRAs [1, 7]	✓	✓	✓	✓	✗
In-Context Learning [8]	✓	✓	✓	✗	✓
Ours	✓	✓	✓	✓	✓

Table 2. Comparison results of our and existing paradigms for adapting SAM into downstream tasks, in terms of five desirable properties.

Method	mIoU	Bottle	Car	Sheep	Cat
RL+SCLIP	58.24	45.64	62.39	64.40	85.20
RL+LAST	24.20	21.59	24.40	33.09	40.24
RL only	27.73	19.78	27.42	30.33	39.29
RL+SRM(Ours)	**62.09**	**48.09**	**66.13**	**73.12**	**85.99**

Table 3. Ablation study results of the proposed prompt labeling module SRM. The performance results are calculated by averaging IoU (%) of all the categories in Pascal-VOC 2012. "RL+X" denotes utilizing the trained RL agent to perform prompt selection and query the label of prompt from "X".

3. Analysis of Prompt Label

At each timestep, the RL agent chooses a prompt position from the action space and queries its corresponding label (foreground or background) from the output of the SRM module. Theoretically, all the selected points can be con-

Method	**Blur**		**Shadow**		**Glass**		**Saliency**		**Semantic**	
	CUHK [3]		SBU [4]		MSD [6]		DUTS [5]		Pascal VOC [2]	
	mIoU ↑	*FR*	mIoU ↑	*FR*	mIoU ↑	*FR*	E_ϕ ↑	*FR*	mIoU ↑	*FR*
Ours-w/o RL	59.75	63.80	25.62	18.30	33.41	22.80	74.19	14.60	54.06	23.20
Ours	**68.47**	70.90	**30.78**	24.80	**45.44**	43.70	**78.21**	32.90	**62.09**	36.70

Table 4. Ablation study results of the RL policy on different segmentation tasks. "Ours-w/o RL" denotes the degraded variant of the proposed approach where the RL policy is replaced with random selection for the action decision. The best performance among all approaches is highlighted in **blod**.

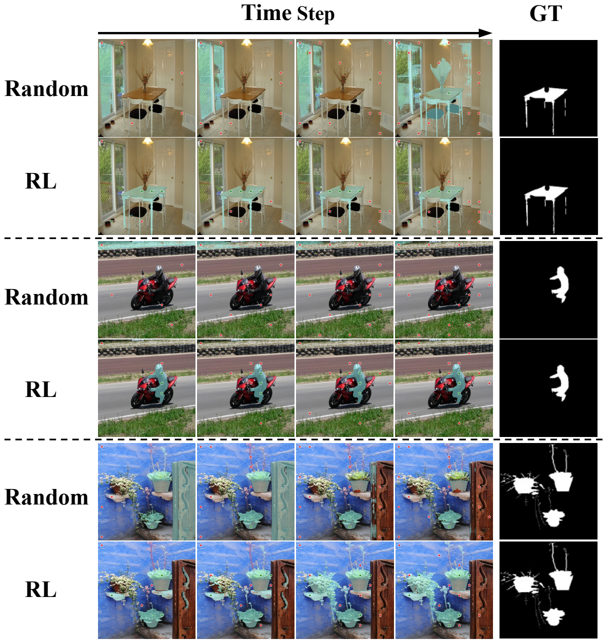


Figure 1. Representative examples to illustrate the iterative point selection and the corresponding segmentation results. The sequence progresses from left to right, showing a gradual increase in the number of point prompts. The green and red stars denote the positive and negative point prompts respectively. The input image is covered by the segmentation mask shown in a light green color. "GT" denotes the ground truth mask of the input image.

sidered positive foreground points since the agent is trained to prioritize foreground areas. However, the RL agent can only identify the target region at a coarse level due to the limited space of actions, leading to potentially unreliable labeling. Alternatively, the prompt label can be queried from the last prediction of SAM or the Vision-Language similarity map. To verify the efficacy of the reference mask for querying the prompt's label, we replace the SRM module by the above-mentioned variants. As shown in Table 3, our method utilizing the prediction of the SRM module as the label source consistently outperforms the other alternatives, demonstrating the superiority of the prompt labeling strategy.

4. Analysis of RL Policy

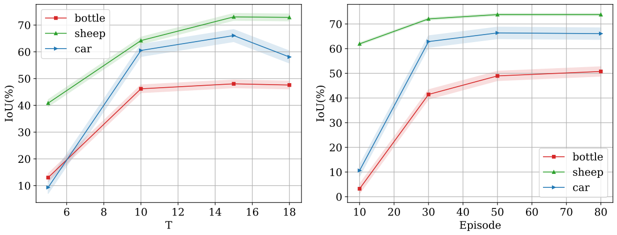
To investigate the strategy learned by the reinforcement learning agent after training, we assess the disparity in action selection between reinforcement learning strategy and random sampling. We propose a Foreground Rate (FR) to measure the ability to track the target of interest, which can be formulated as follows:

$$FR = \sum_{t\in[1,T]} \mathbb{1}\{G_I(a_t) = 1\}/T, \tag{1}$$

where $G_I(a_t)$ means querying the label of the chosen position from the ground truth mask for sample I. For both the RL and random strategies, we set the number of interaction rounds T to 15 and report *FR* averaged on all the testing samples in each dataset. As shown in Table 4, the *FR* of the RL strategy significantly surpasses that of random selection in all the reported scenarios, indicating the high-value estimation of actions in the target area by the RL network. As illustrated in Figure 1, our RL agent exhibits a greater inclination for selecting point prompts in the vicinity of the target area compared to random selection. The iterative segmentation process highlights the advantages of employing the reinforcement learning strategy in multiple aspects. First, with a limited prompting budget, the RL agent can proficiently capture the area of the target of interest, thereby facilitating the segmentation of SAM. Secondly, in scenarios with multiple disjointed target regions within an image, utilizing the RL agent for prompt selection effectively prevents the omission of target regions. In summary, our RL agent consistently selects more points surrounding the target area than random selection in diversified scenarios, thereby unlocking the potential for progressive segmentation refinement.

5. Analysis of Training Samples

To validate the robustness of AlignSAM, we compare it with other competitive state-of-the-art (SOTA) methods across various budgets of training samples. For each dataset, the training samples are randomly sampled from the training set and shared among different methods to ensure a fair comparison. As depicted in Table 5, AlignSAM consistently outperforms the second-best SOTA method in most scenarios. This implies that the RL agent and the SRM



(a) Sensitivity to T

(b) Sensitivity to E

Figure 2. Hyper-parameter sensitivity to T and E of AlignSAM.

module can synergistically align SAM to the segmentation objective even with a limited number of reference samples.

Method	Blur CUHK [3]		Glass MSD [6]		Saliency DUTS [5]	
	5-shot	20-shot	5-shot	20-shot	5-shot	20-shot
PerSAM [8]	53.21	54.69	31.03	29.50	**35.92**	40.50
Ours	**68.89**	**62.99**	**31.19**	**38.78**	33.05	**44.82**

Table 5. Comparison results of our AlignSAM and PerSAM under different numbers of training samples. The reported results denote the performance of the models evaluated by mean IoU. The best performance among all approaches is highlighted in **blod**. The second-best competitor PerSAM [8] is selected as the representative of SOTA methods.

6. Hyper-Parameter Sensitivity

We further carry out investigations to check the sensitivity of the proposed approach to the key hyper-parameters E and T. These experiments are conducted under three scenarios with varying degrees of segmentation difficulty. In Figure 2, we show the model performance of AlignSAM when T and E are respectively set to $\{5, 10, 15, 18\}$ and $\{10, 30, 50, 80\}$. As illustrated, the model trained with $T = 15$ and $E = 50$ exhibits notably superior and stable performance compared to other configurations. This emphasizes the efficacy of setting both to 15 and 50, respectively, as a good choice for the implementation.

References

[1] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 2

[3] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 1, 2, 3

[4] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016. 1, 2

[5] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3

[6] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8809–8818, 2019. 1, 2, 3

[7] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 1

[8] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hong Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1, 3