Improving the Generalization of Segmentation Foundation Model under Distribution Shift via Weakly Supervised Adaptation

Supplementary Material

In this supplementary material, we first investigate the impact of the amount of weak supervision on adaptation performance. We further demonstrate that after weakly supervised adaptation we can further improve the performance of one-shot segmentation. Additional analysis of hyper-parameters and qualitative results are presented as well.

1. The Impact of Weak Label Numbers on Performance

In this experiment, we aim to demonstrate the cost-effectiveness of utilizing weak labels. To facilitate comparison, we incrementally choose weakly labeled images of 50, 100, 200, 400, 800, 1600, 3200, and 4246 in COCO dataset for adaptation. We adapt the model using three different types of weak labels and evaluate their performance using three different prompts. We make the following observations from Fig. 1 (a-c). First, when training/adapation weak supervision is the same with testing prompt, we observe the most effective generalization of SAM. Moreover, the generalization improves upon more weak supervision except for adaptation with point label and testing with box and polygon prompts. This suggest the mask decoder is still sensitive to the shift of prompt used for training and testing.

2. One-Shot Personalized Segmentation

We further investigate the effectiveness of adapting SAM for one-shot personalized segmentation. Specifically, PerSAM [2] is a training-free personalization approach for SAM that enables it to achieve one-shot segmentation based on visual cues. Given only a single image with a reference mask, PerSAM first localizes the target concept by a location prior in the test image, and then segments the target object through target-guided attention, target-semantic prompting, and cascaded post-refinement.

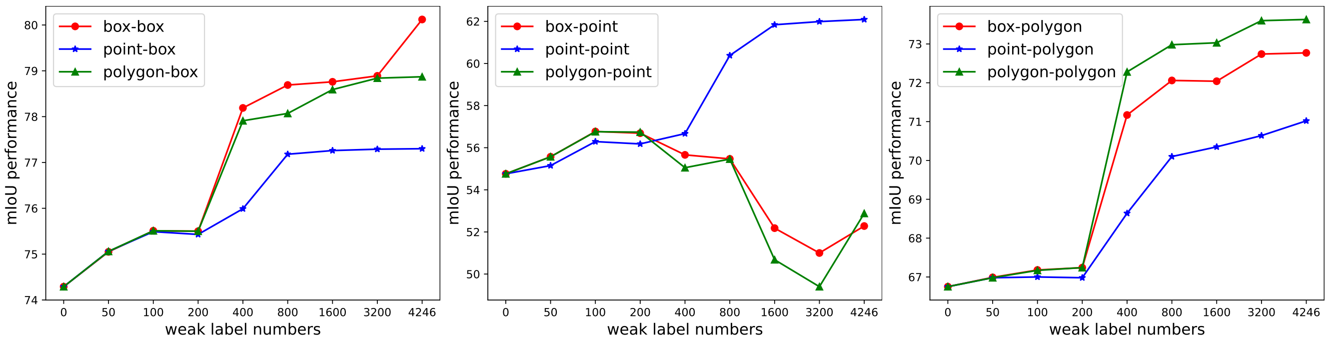
In the main text, we have demonstrated that SAM performs poorly when facing significant domain shift. Similarly, in one-shot tasks, there may also be downstream tasks with significant domain shift. To demonstrate the effectiveness of weakly supervised adaptation, we use the PerSAM framework and conduct one-shot experiments on the ISIC dataset for medical image segmentation. In particular, we compare three alternative designs for PerSAM. First, we directly use PerSAM with one-shot reference image and test on the testing set of ISIC. We further evaluate PerSAM-F which finetunes the mask weight on one-shot reference image. Finally, our method adapts SAM with weakly labeled data and then use the adapted SAM for one-shot PerSAM. We further improved PerSAM by sampling 30 feature points on the reference image using Kmeans to diversify the reference prototypes. The reference prototypes are used to query the positive point on the testing image as prompt for SAM. We denot the improved version as PerSAM*.

Table 1. Oneshot experimental results on the ISIC dataset. * indicates that the reference prototypes are collected by Kmeans algorithm.

Method	IoU	Acc
PerSAM	38.16	87.48
PerSAM-F	**41.06**	82.73
PerSAM + OURS	40.00	**87.93**
PerSAM*	43.96	70.57
PerSAM-F*	40.97	57.98
PerSAM + OURS*	**49.86**	**72.84**

As shown in Tab. 1, we observe that without any modifications to PerSAM, our adapted SAM achieves better one-shot segmentation performance than the original PerSAM model in terms of both IoU and Accuracy (Acc). With the improved PerSAM*, our weakly supervised adaptation is much more effective in localizing the target objects outperforming both the original PerSAM and the improved PerSAM with 6-10% IoU. We also visualize the one-shot segmentation results in Fig. 2. The green and red stars refer to the positive and negative point prompt on the testing image. The original PerSAM tends to either ignore the foreground partially or over estimate the foreground. While our improved PerSAM*, thanks to the multiple point prompt, achieves better localization of the foreground object.

The above experiment was conducted on medical image segmentation with a large domain gap, e.g. ISIC Dataset. To enhance the persuasiveness, especially that our method improves model generalization, not just task-specific adaptation. We supplemented experiments on one-shot segmentation in natural images in Tab. 2. We observe that i) there is a clear improvement with mask mixture finetuning (PerSAM-F); and ii) PerSAM + OURS is still better than PerSAM-F. The improvement of PerSAM-F suggest the segmentation granularity is still a concern for applying SAM for in-distribution downstream task. When specific information on segmentation target is considered (PerSAM + OURS) the performance is substantially better (31.89 v.s. 21.22 in mIoU). This suggests segmentation specific information (e.g. spatial extent) is also crucial for adapting SAM to in-distribution downstream tasks.



(a) Using box as testing prompt. (b) Using point as testing prompt. (c) Using polygon as testing prompt.

Figure 1. Annotation cost vs. performance. 1. Performance of different numbers of weak labels on performance. 2. Performance of three weak labels under the same prompt verification.

Table 2. Oneshot segmentation results on COCO dataset.

Method	mIoU	Acc
PerSAM	21.22	32.33
PerSAM-F	23.43	47.86
PerSAM + OURS (w/ box weak supervision)	**31.89**	**70.55**

Table 3. Experimental results of Hyper-Parameter sensitivity analysis on COCO dataset. The bold text indicates values used in our method.

Hyper-Param.		box	point	poly
Temp. τ	0.1	79.05	63.65	72.98
	0.3	**79.89**	**63.86**	**73.01**
	0.5	79.19	63.46	72.96
LR	$1e-3$	78.42	**63.90**	72.91
	$1e-4$	**79.89**	63.87	**73.01**
	$1e-5$	79.01	63.81	72.99
$\lambda_{stu}^{dice} : \lambda_{tea}^{dice}$	1.0 : 0.0	67.19	41.64	64.4
	0.7 : 0.3	72.98	42.11	68.62
	0.5 : 0.5	**79.89**	**63.82**	**73.01**
	0.3 : 0.7	79.76	63.60	72.68
	0.0 : 1.0	79.26	63.16	72.62

3. Hyper-Parameter Sensitivity

In this section, we evaluate the sensitivity to different hyper-parameters. For Anchor loss, the coefficients of two dice losses are denoted as λ_{stu}^{dice} and λ_{tea}^{dice} , For the Anchor loss, the coefficients of the two dice losses are denoted as λ_{stu}^{dice} and λ_{tea}^{dice}, respectively, and are set as follows: 1.0:0, 0.7:0.3, 0.5:0.5, 0.3:0.7, 0:1.0. For Contrast loss, we set the temperature τ to 0.1, 0.3, 0.5. For model finetuning, we use the Adam optimizer with learning rates set to 0.001, 0.0001, and 0.00001, respectively. As shown in Tab 3, out proposed weakly supervised adaptation method is relatively stable to the choice of hyper-parameters.

4. Additional Experiment

Preferring "Shared Weights" over "EMA" In our approach, the teacher and student models share the same weights. Another approach is using EMA for teacher model weights. Comparative experiments with "EMA weight" in Tab. 4 show weight sharing's superior performance.

EWC regularization Regularizing the model weights is subject to the difference in scale and size of model weights. We adopt anchor regularization in self-training. For other regularization methods, such as Elastic Weight Consolidation[1], we adapt this approach for SAM adaptation, and the results "EWC reg." in Tab. 4 suggest our proposed regularization is still optimal.

Table 4. The additional experiments

Method	COCO 2017			ISIC		
	box	point	poly	box	point	poly
Direct	74.29	55.06	65.64	66.74	53.42	62.82
EMA weights	78.14	55.03	73.22	78.12	63.41	73.74
EWC reg.	76.44	52.52	71.33	78.87	**66.89**	75.88
OURS	**80.12**	**64.39**	**73.72**	**80.26**	63.90	**76.59**

5. Visualization Examples on Various Downstream Domains

Finally, we present more qualitative results on multiple downstream segmentation datasets. Specifically, COCO illustrations are shown in Fig. 3, and ones of ISIC are in Fig. 4, and ones of OCID are in Fig. 5, and ones of CAMO are in Fig. 6, and ones of COCO-C are in Fig. 7,8. The observations suggest SAM after weakly supervised adaptation achieves much superior segmentation quality on all types of weak supervision and testing prompts.

References

[1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017. 2

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

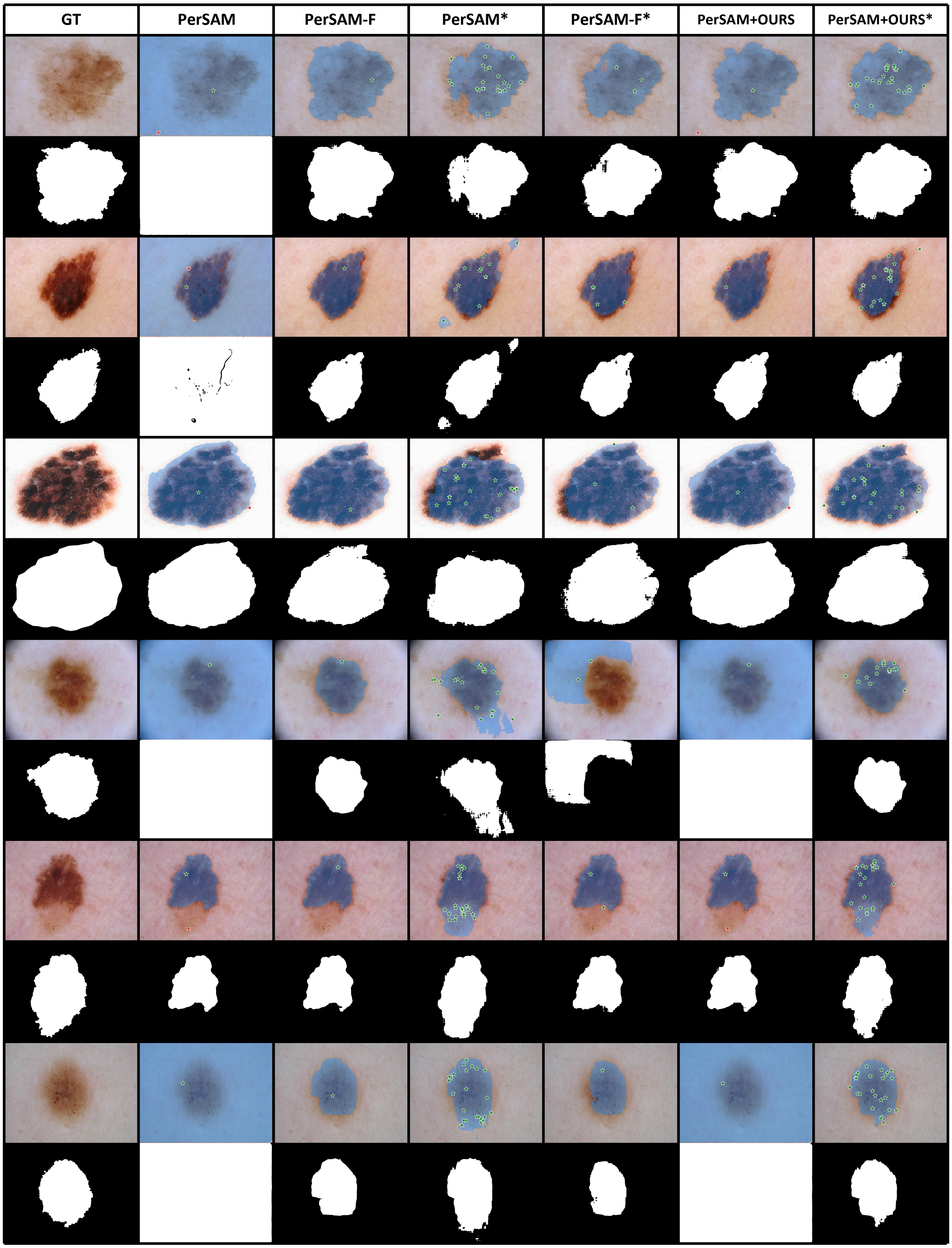


Figure 2. Qualitative results for One-Shot PerSAM segmentation.

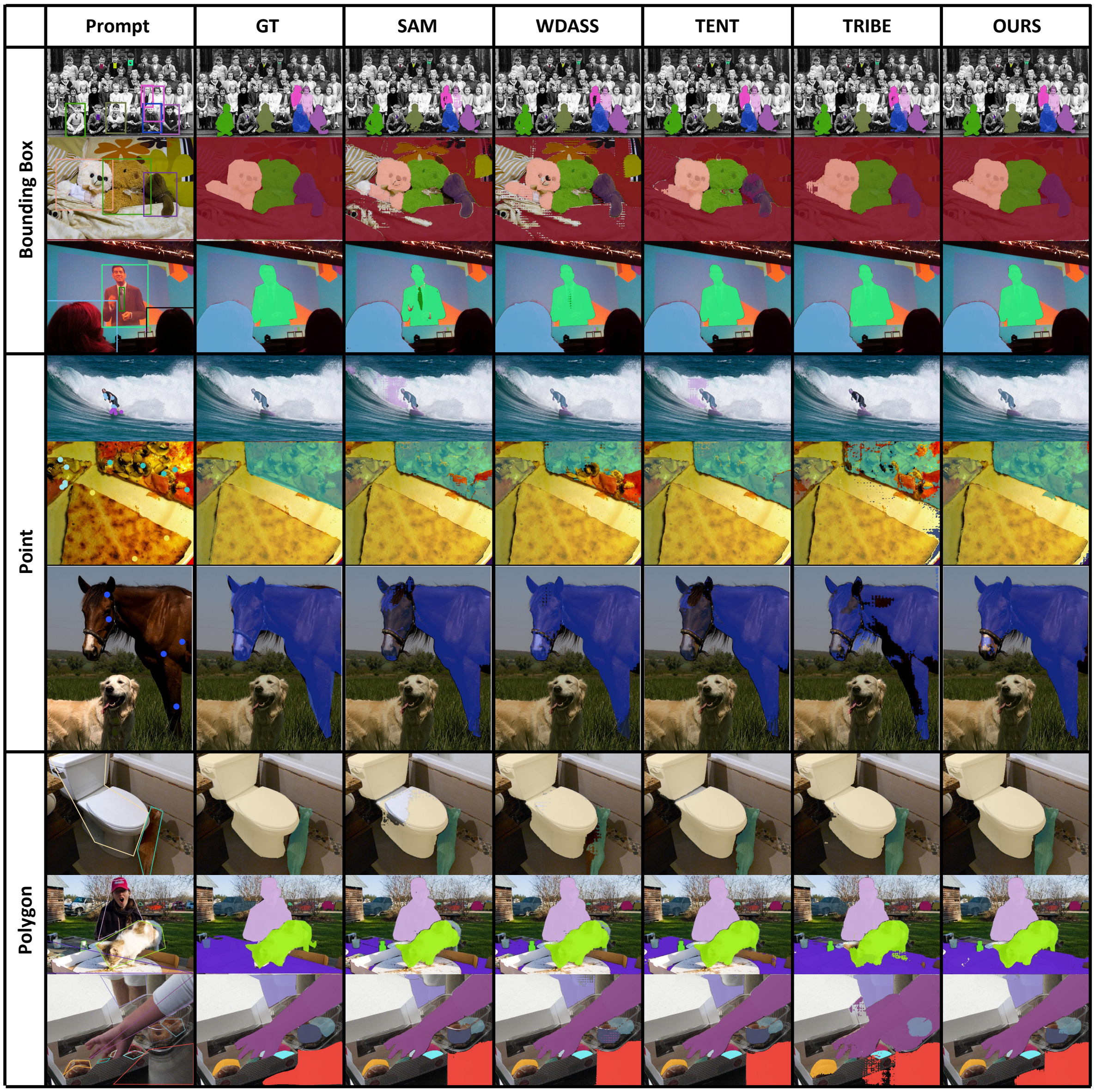


Figure 3. Visualization examples on COCO dataset.

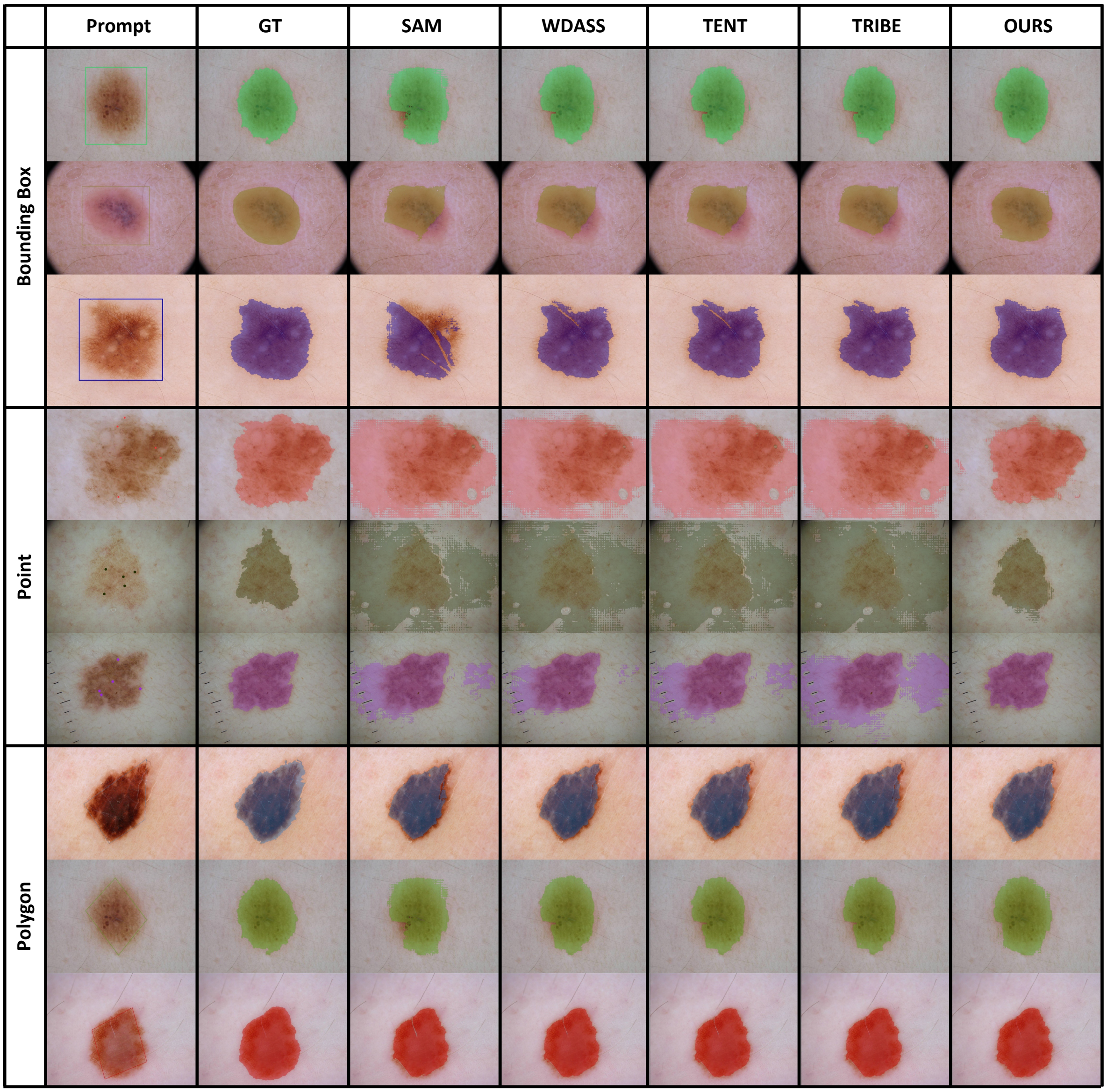


Figure 4. Visualization examples on ISIC dataset.

Figure 5. Visualization examples on OCID dataset.

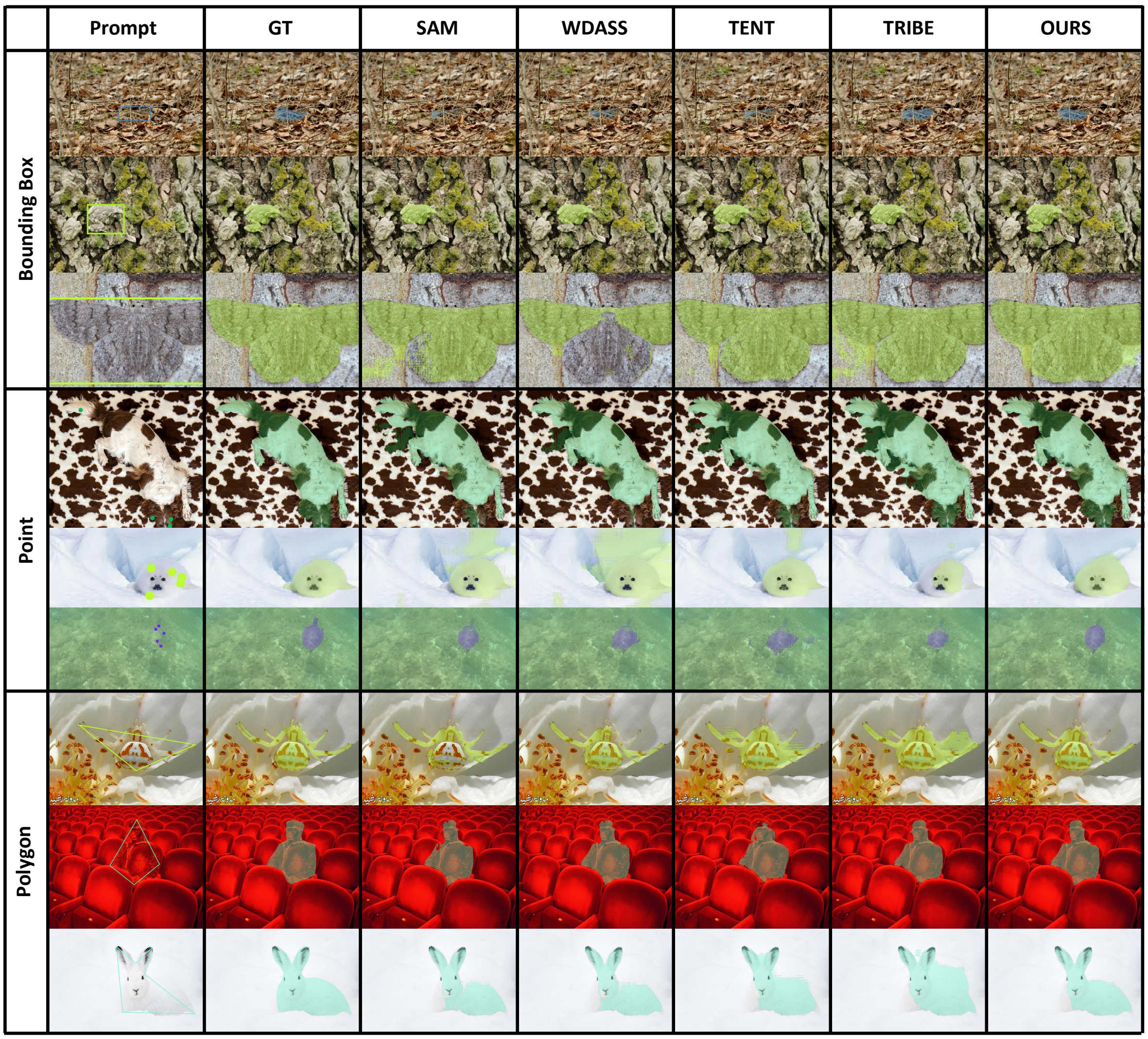


Figure 6. Visualization examples on CAMO dataset.

Figure 7. Visualization examples on COCO-C dataset.

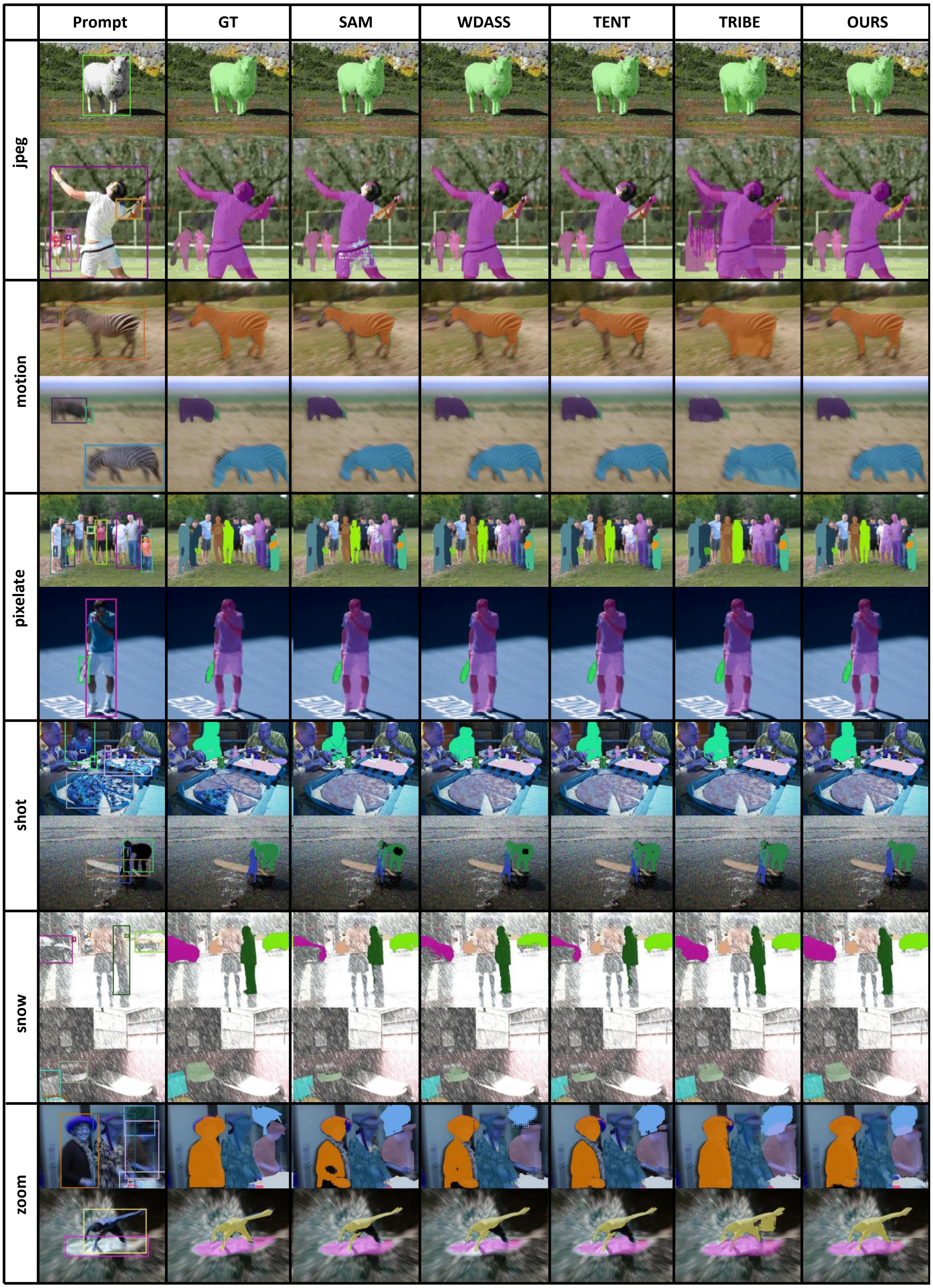


Figure 8. Visualization examples on COCO-C dataset.