ConformalSAM: Unlocking the Potential of Foundational Segmentation Models in Semi-Supervised Semantic Segmentation with Conformal Prediction

Supplementary Material

A. More Experimental Details. In this section, we detail the experimental setup. In Stage I, the model is trained in a fully supervised manner using pseudo-labels for unlabeled data, with a total dataset size of $L + U$ and a batch size of 4. In Stage II, a naive self-training approach is applied, where pseudo-labels are generated without confidence filtering, and a weighted unsupervised loss is used to mitigate the impact of low-quality pseudo-labels. The batch size remains 4 (2 labeled and 2 unlabeled samples). To ensure a consistent labeled-to-unlabeled ratio across annotation proportions, labeled data is duplicated as needed, resulting in an effective dataset size of $2 * U$ in Stage II.

Learning_rate	Batch_size	1/2(732)	Batch_size	1/2(732)
0.03	2	3.49	4	7.39
0.01	2	75.87	4	78.94
0.003	2	78.59	4	79.10
0.001	2	78.63	4	77.62
0.0003	2	76.58	4	76.83
0.0001	2	75.70	4	75.21

Table 7. Experimental results on the original PASCAL VOC 1/2 (732) dataset with different initial learning rates and batch sizes. The best result is colored red.

B. Ablation on Initial Learning Rate. As shown in Tab. 7, we experimented with several commonly used initial learning rate and batch sizes on PASCAL VOC 1/2(732). The results show that the hyperparameters we used are optimal.

ConformalSAM			1/128(158)	1/64(316)	1/32(632)	1/16(1263)
SEEM	CP	SR				
✓			20.64	20.53	21.01	21.45
✓		✓	25.97	29.53	32.81	34.30
✓	✓		20.72	20.82	21.19	21.59
✓	✓	✓	26.21	30.02	33.33	34.64

Table 8. Ablation study on components of ConformalSAM on the ADE20K dataset. The best results are colored red.

C. Ablation on ADE20K Dataset. As shown in Tab. 8, we validate the effectiveness of ConformalSAM on the ADE20K dataset.The experimental results demonstrate that our method achieves an average improvement of 0.40 mIoU across four different splits (Rows 2 and 4). Furthermore, in Fig. 5, we present the masks generated by SEEM and the results of calibrated inference. Notably, in the absence of a background class, CP can filter out some of SEEM's misclassifications, mitigating the impact of incorrect pseudo-labels on the model.

D. Ablation on Constant Confidence Threshold. We also test a simple confidence thresholding instead of using CP.

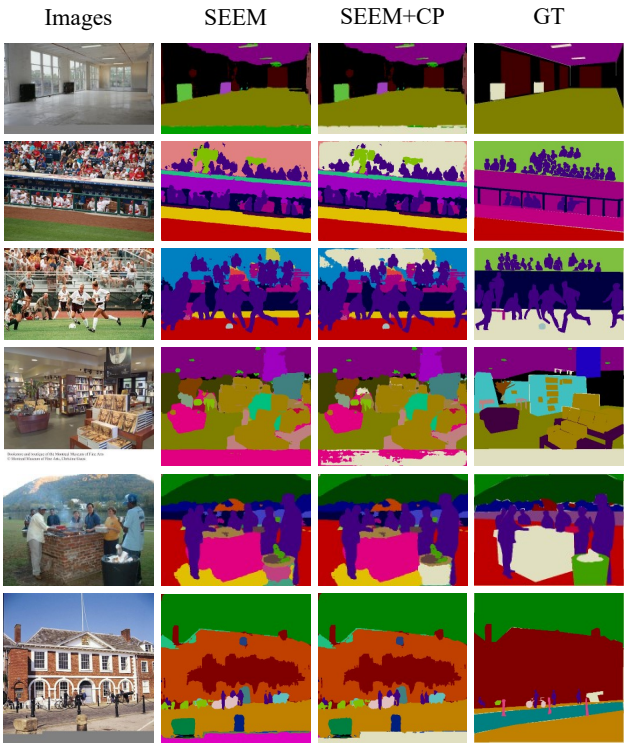

Figure 5. Segmentation masks from the ADE20K dataset.

(a) PASCAL VOC

Method	1/16(92)	1/8(183)	1/4(366)	1/2(732)
Confidence=0.7	43.67	41.84	45.78	48.66
SEEM+CP	**78.09**	**78.23**	**78.32**	**79.10**

(b) ADE20K

Method	1/64(316)	1/32(632)	1/16(1263)	1/8(2526)
Confidence=0.7	20.19	20.57	20.73	21.54
SEEM+CP	**20.72**	**20.82**	**21.19**	**21.59**

Table 9. Experimental results with constant confidence thresholds on the original PASCAL VOC and ADE20K datasets.

In other words, the quantile map in ConformalSAM is a constant in every pixel. Specifically, we set the confidence threshold to 0.7 and experimented on the original PASCAL VOC and ADE20K datasets. The result as shown in Tab. 9. The class-conditional label selection plays a crucial role in PASCAL VOC. On ADE20K, the CP-based filtering is always better than the fixed confidence filtering.

E. Overview of ConformalSAM(AllSpark). We show the model framework for applying our approach to AllSpark in Fig. 6

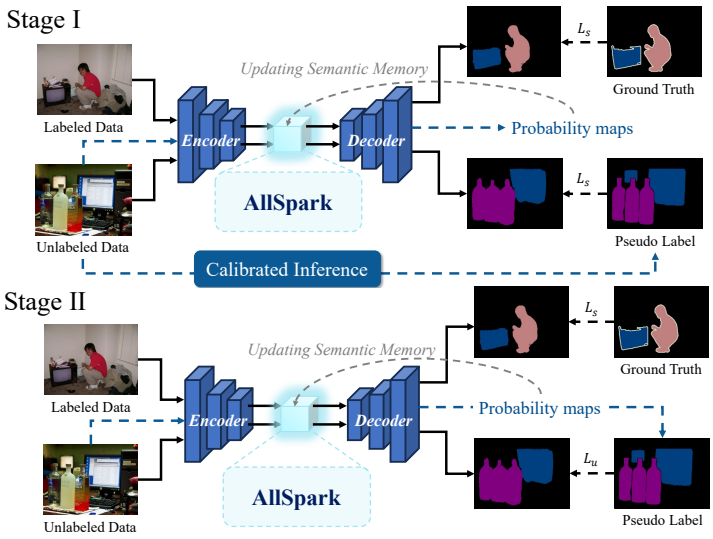

Figure 6. Overview of ConformalSAM(AllSpark). We use the framework proposed by AllSpark [34] in ConformalSAM.

Cityscapes	1/6(186)	1/8(372)	1/4(744)	1/2(1488)
AllSpark(reproduced)	74.95	**77.88**	79.05	79.43
ConformalSAM(AllSpark)	**76.07**	77.17	78.87	**80.08**
ConformalSAM	74.67	77.30	**79.06**	79.64

Table 10. ConformalSAM experimental results on cityscapes.

F. More results on Cityscapes datasets. We report the experimental results on cityscapes [7] in the Tab 10, it shows competitive performance of ConformalSAM. According to the authors' explanation, the code open-sourced in github accidentally deleted the temperature coefficient for calculating similarities with semantic memory, and the code has not been updated so far. As a result, the results we reproduced on cityscapes differ from those reported in the paper.

G. Quantitative measure of SEEM filtering. We report mIoU of CP-filtered masks vs. SEEM-generated masks in the Tab. 11. After the CP calibration, the correctness of the masks is substantially improved.

PASCAL VOC original	1/16(92)	1/8(183)	1/4(366)	1/2(732)
SEEM	30.29	30.26	30.32	30.34
SEEM+CP	**73.25**	**73.33**	**73.27**	**73.59**

Table 11. Results of mIoU evaluation of SEEM and SEEM+CP on original PASCAL VOC 2012 dataset.

H. Impact of the unsupervised loss weight λ. We experimented with different initial values of λ, with results shown in the Tab. 12. The small initial λ is generally better because the SEEM-generated masks provide high-quality supervision signals for the initial training stage.

I. Performance on GLAMM. We agree that it is important to show the performance of ConformalSAM using a different foundational segmentation model. Thus, we conducted additional experiments using GLAMM [25]. We obtain

PASCAL VOC *original*	λ	1/16(92)	1/2(732)
ConformalSAM	0.01	81.21	83.52
	0.25	79.85	83.15
	0.5	78.91	82.55

Table 12. Ablation experiments with different initial values of unsupervised loss weights

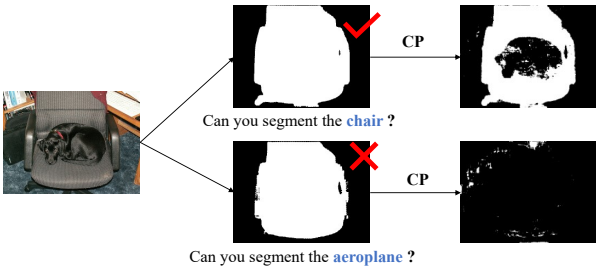

Figure 7. A mask of GLAMM and GLAMM+CP

segmentation results by performing referring-expression segmentation for each category in PASCAL VOC 2012 and then use ConformalSAM to calibrate the output. As shown in Fig. 7, CP successfully calibrates the segmentation results by filtering out some distracting categories generated by GLAMM. We conducted experiments on PASCAL VOC 2012, and the results are presented in the Tab. 13.

PASCAL VOC *original*	1/8(183)	1/2(732)
GLAMM	76.22	78.85
GLAMM+CP	78.64	81.03

Table 13. Results of using GLAMM on the *original* VOC dataset.

J. More Visualizations. We show the segmentation masks from AllSpark, ConformalSAM and ground-truth labels in Fig. 8 & 9 to give a qualitative comparison.

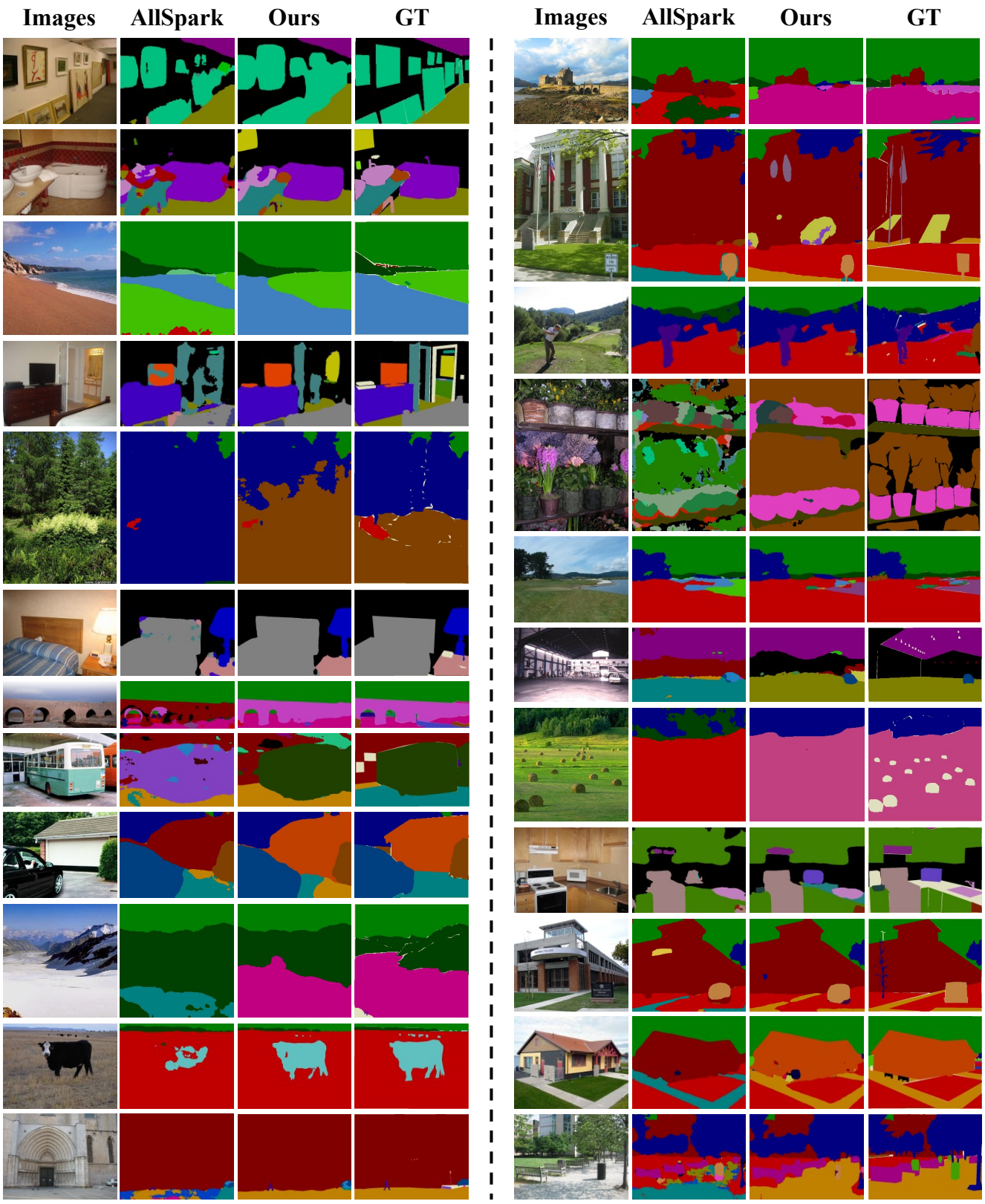

Figure 8. Visualization of the segmentation results on ADE20K validation set, compared with AllSpark [34]

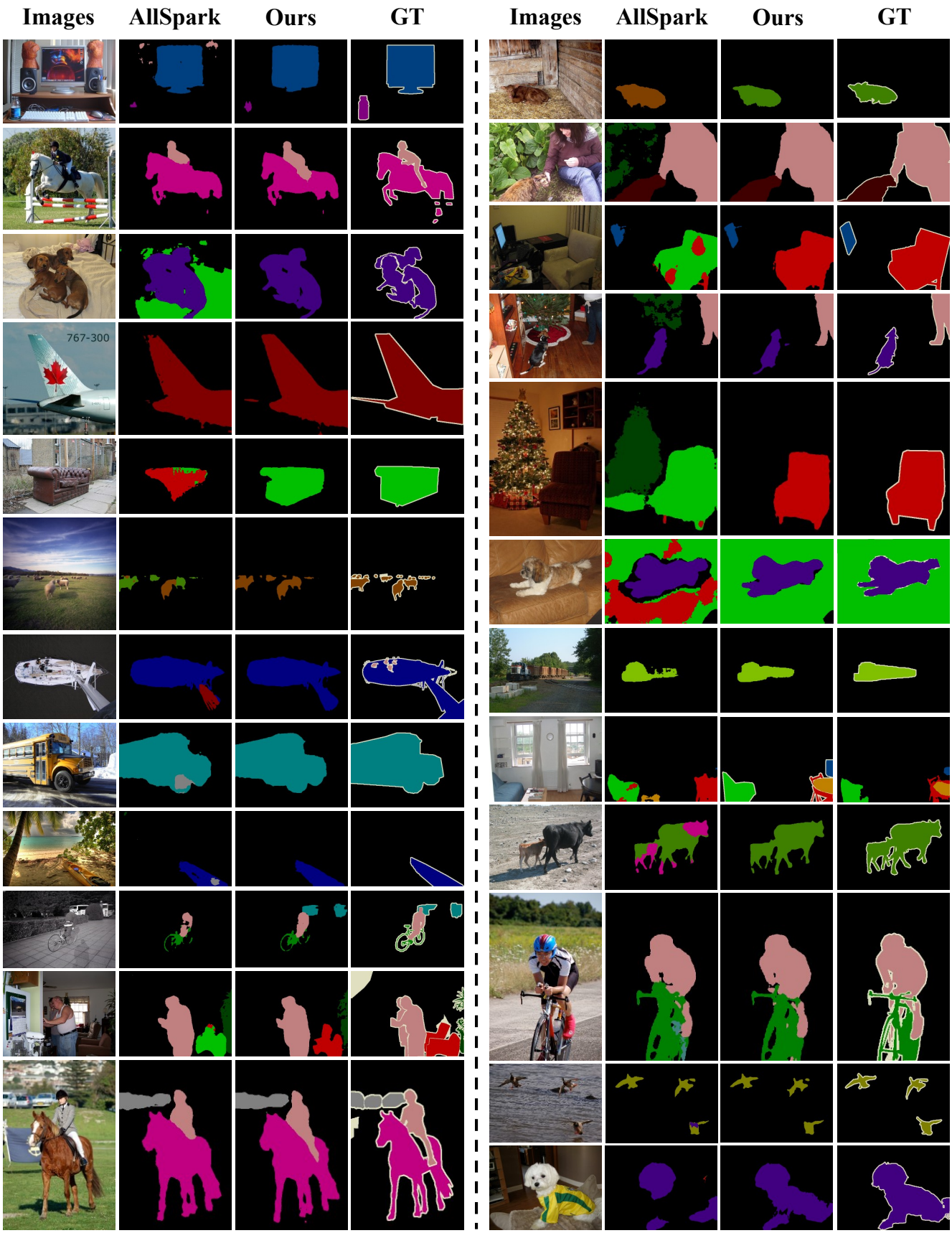

Figure 9. Visualization of the segmentation results on Pascal validation set, compared with AllSpark [34]