SAMIX: Reinforcing SAM2 with Semantic Adapter and Reference Selecting Policy for Mix-Supervised Segmentation

Supplementary Material

A. Definition of Supervision Loss

In the Section 'Methodology' (Sec. 3.3) of the manuscript, we denote the supervision loss function as $\mathcal{L}_{\text{typ}}$ (typ $\in$ [mask, box, scribble, point]), which is customized for each label type, and it is used to calculate the optimization loss of the mean teacher-based semantic segmentation model (Seg-model) and define a verifiable reward, that is, the supervision reward r_γ. Let the prediction mask of the model be indicated as $\mathcal{M} \in [0,1]^{C\times H\times W}$, where the values are distributed between 0 and 1, and C, H and W represent the number of categories, mask height, and mask width, respectively. Here, we will elaborate on the specific form of $\mathcal{L}_{\text{typ}}$ when dealing with various types of labeled data:

- **Mask-labeled Data:** We directly utilize the Dice loss to Cross-entropy loss to calculate the loss between the model's predicted mask and ground-truch (GT) mask $\mathcal{M}_{\text{mask}}$:

$$\mathcal{L}_{\text{mask}} = \texttt{Dice}(\mathcal{M}, \mathcal{M}_{\text{mask}}) + \texttt{CE}(\mathcal{M}, \mathcal{M}_{\text{mask}}), \tag{1}$$

  where $\texttt{Dice}(\cdot,\cdot)$ and $\texttt{CE}(\cdot,\cdot)$ indicates the Dice loss function, and cross-entropy loss function, respectively.

- **Box-labeled Data:** We employ projection loss [9] on box-labeled data. Specifically, the GT bounding boxes are provided as coordinates, indicating the top-left and bottom-right corners of each box. Firstly, we transform the GT coordinates into a binary GT mask $\mathcal{M}_{\text{box}} \in \{0,1\}^{C\times H\times W}$ by assigning 1 to the pixels within the boxes and 0 outside them. Next, we apply 1D max-poling on both $\mathcal{M}$ and $\mathcal{M}_{box}$ horizontally and vertically, yielding four projection maps, *i.e.*, $\mathbf{o}, \in [0,1]^{C\times H\times 1}$, $\mathbf{v} \in [0,1]^{C\times 1\times W}$, and $\mathbf{o}_{\text{box}} \in [0,1]^{C\times H\times 1}$, $\mathbf{v}_{\text{box}} \in [0,1]^{C\times 1\times W}$, which align with the image width and height, respectively:

$$\begin{aligned}
\mathbf{o}[c,i,0] &= \texttt{max}(\mathcal{M}[c,i,:]),\\
\mathbf{v}[c,0,j] &= \texttt{max}(\mathcal{M}[c,:,j]),\\
\mathbf{o}_{\text{box}}[c,i,0] &= \texttt{max}(\mathcal{M}_{\text{box}}[c,i,:]),\\
\mathbf{v}_{\text{box}}[c,0,j] &= \texttt{max}(\mathcal{M}_{\text{box}}[c,:,j]),
\end{aligned} \tag{2}$$

  where $\mathcal{M}[c,i,:]$ denotes the i-th row of the c-th category and $\mathcal{M}[c,:,j]$ denotes the j-th column of the c-th category in the the predicted mask $\mathcal{M}$, and similarly for $\mathcal{M}_{\text{box}}$. Finally, we compute the Dice loss between the corresponding projection vectors:

$$\mathcal{L}_{\text{box}} = \texttt{Dice}(\mathbf{o}, \mathbf{o}_{\text{box}}) + \texttt{Dice}(\mathbf{v}, \mathbf{v}_{\text{box}}), \tag{3}$$

  where $\texttt{Dice}(\cdot,\cdot)$ indicates the Dice loss function.

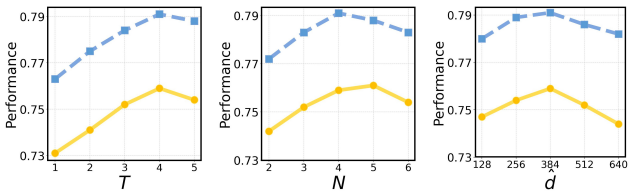

Figure A. Ablation on three hyperparameters: T, N, and $\hat{d}$. The solid yellow curve indicates performance (mIoU) on VOC *val*, while the dashed blue curve indicates performance (S_m) on COD10K (COD).

- **Scribble-labeled Data:** We employ partial cross-entropy (pCE) loss [6, 11] on the scribble-labeled data. Specifically, we denote $\mathcal{M}_{\text{scribble}} = \{0,1,-1\}^{C\times H\times W}$ as the scribble label, where 1 represents the labeled foreground pixel, 0 denotes the labeled background pixel and -1 indicates the unlabeled pixel. The partial cross-entropy loss is defined as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{scribble}} = \\
&-\sum_{c=0}^{C}\sum_{\mathcal{M}_{\text{scribble}}[i,j]\neq -1}\Big(\mathcal{M}_{\text{scribble}}[c,i,j]\log\mathcal{M}[c,i,j] \\
&+ (1-\mathcal{M}_{\text{scribble}}[c,i,j])\log(1-\mathcal{M}[c,i,j])\Big).
\end{aligned} \tag{4}$$

- **Point-labeled Data:** Similar to scribble-labeled data, we adopt partial cross-entropy loss on point-labeled data. Specifically, we denote $\mathcal{M}_{\text{point}} = \{0,1,-1\}^{C\times H\times W}$ as the point label, where 1 represents the labeled foreground pixel, 0 denotes the labeled background pixel and -1 indicates the unlabeled pixel. The partial cross-entropy loss is defined as follows:

$$\begin{aligned}
&\mathcal{L}_{\text{point}} = \\
&-\sum_{c=0}^{C}\sum_{\mathcal{M}_{\text{point}}[c,i,j]\neq -1}\Big(\mathcal{M}_{\text{point}}[c,i,j]\log\mathcal{M}[c,i,j] \\
&+ (1-\mathcal{M}_{\text{point}}[c,i,j])\log(1-\mathcal{M}[c,i,j])\Big).
\end{aligned} \tag{5}$$

- **Class-labeled and Unlabeled Data:** We do not employ $\mathcal{L}_{\text{typ}}$ on class-labeled and unlabeled data.

B. Hyperparameters Configuration.

Figure A presents the results under different hyperparameter settings, including the length of a reference set T, the number of reference sets N, and the projection dimension of the semantic adapter $\hat{d}$. As can be seen, SAMIX using

Modules	Iteration Time (ms/iter)	GPU Memory (GB)
WISH:		
Mask2Former	310	14.1
+ SAM (ViT-L)	517 (+207)	36.8 (+22.7)
SAMIX (Ours):		
Mask2Former	310	14.1
+ SAM2 (Hiera-S)	407 (+97)	24.5 (+10.4)
+ Semantic Adapter	416 (+9)	25.6 (+1.1)
+ SPNet (GRPO)	474 (+58)	31.9 (+6.3)

Table A. Comparison of iteration time and GPU memory consumption for different model configurations. Experiments were conducted on 4 NVIDIA RTX 4090 GPUs, each with 48 GB memory (custom-modified from the standard 24 GB), and a per-GPU batch size of 4.

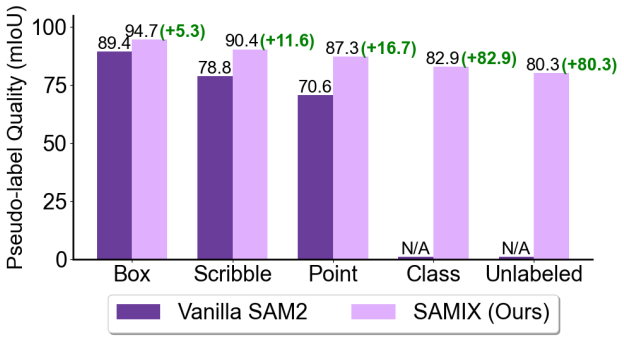

Figure B. The quality of pseudo-labels on PASCAL VOC 2012 *train* across different annotation types. Compared to vanilla SAM2, our SAMIX consistently achieves improved label quality across all annotation types.

$T = 4$, $N = 4$, and $\hat{d} = 384$ shows the best performance, respectively, and we adopt this configuration as the default setting of SAMIX.

C. Efficiency Analysis

In Table A, we compare the training cost of our SAMIX with a recent mix-supervised SOTA method WISH [5]. WISH employs SAM [4] with a ViT-L [2] backbone, and our SAMIX utilizes SAM2 [7] with a Hiera-S [8] backbone. Due to the lightweight design of the semantic adapter, as well as the compact feature input and the small number of output tokens (fixed 4 output tokens per group) in SPNet, our SAMIX achieves lower iteration time and GPU memory usage compared to WISH.

D. Analysis of Collaborative Learning across Heterogeneous Data

In Figure B, we demonstrate the quality of pseudo-labels generated by our SAMIX and vanilla SAM2 on train data (PASCAL VOC 2012 *train*) with different annotation types. Building upon SAM2, SAMIX extends an augmented dense contextual prompting mechanism that enables information interaction across data samples with different annotation types. Benefiting from the proposed Selecting Policy Network (SPNet) and Hierarchical Guidance Principle (HGP), SAMIX leverages high-quality pseudo-labels derived from strongly labeled samples to optimize the pseudo-labels of weakly annotated ones. As illustrated in Figure B, the quality of pseudo-labels across all annotation types is consistently improved, demonstrating that SAMIX effectively achieves collaborative learning cross-sample. Moreover, whereas vanilla SAM2 relies solely on visual prompts (box and point) and therefore cannot directly generate pseudo-labels for class-labeled or unlabeled data, SAMIX overcomes this limitation and further attains higher pseudo-label quality than SAM2 even on scribble-labeled data (82.9 *vs.* 78.8, 80.3 *vs.* 78.8).

E. Visualization Results on COD and IPS Datasets

In Table 1 of the manuscript, we conducted a quantitative comparison of our SAMIX with four mix-supervised state-of-the-arts (SOTAs), including MixSegNet [10], MixPolyp [3], SAM-COD [1], and WISH [5] on two specific scenarios with ambiguous boundaries, Camouflaged Object Detection (COD) and Image Polyp Segmentation (IPS) datasets. Additionally, to more clearly illustrate the characteristics of these two datasets and the performance differences among all comparison methods, we visualize the results of all methods on the COD and IPS datasets. Note that since the quantitative results in Table 1 of the manuscript show that all methods perform optimally in the unified data setting, that is, using all types of labeled data. To maintain fairness, the visualized results provided correspond to those obtained by training all methods under the same data setting.

The visualization results of all methods are presented in Figure C. From it, we can see that our SAMIX can product predicted masks that are closest to the GT mask for objects with blurred boundaries, such as polyps and camouflaged objects. This is mainly due to the fact that SAMIX can utilize dense contextual prompts provided by higher-quality pseudo-labels with precise boundaries during training, thereby promoting the entire training framework. The superior performance on these two challenging specific datasets also validates the scenario generalization capability of our method, demonstrating its effectiveness even in scenarios rarely encountered during training of visual foundation models (*e.g.*, SAM and SAM2)

Figure C. visualization results of different mix-supervised segmentation methods on image polyp segmentation benchmark (***upper***) and camouflaged object segmentation benchmark (***lower***).

References

[1] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 315–331. Springer, 2024. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[3] Yiwen Hu, Jun Wei, Yuncheng Jiang, Haoyang Li, Shuguang Cui, Zhen Li, and Song Wu. Mixpolyp: Integrating mask, box and scribble supervision for enhanced polyp segmentation. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3289–3292. IEEE, 2024. 2

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[5] Hyeokjun Kweon and Kuk-Jin Yoon. Wish: Weakly supervised instance segmentation using heterogeneous labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25377–25387, 2025. 2

[6] Xiangde Luo, Minhao Hu, Wenjun Liao, Shuwei Zhai,

Tao Song, Guotai Wang, and Shaoting Zhang. Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2022. 1

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[8] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR, 2023. 2

[9] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5443–5452, 2021. 1

[10] Ziyang Wang and Chen Yang. Mixsegnet: Fusing multiple mixed-supervisory signals with multiple views of networks for mixed-supervised medical image segmentation. *Engineering Applications of Artificial Intelligence*, 133:108059, 2024. 2

[11] Xiao Zhang, Shaoxuan Wu, Peilin Zhang, Zhuo Jin, Xiaosong Xiong, Qirong Bu, Jingkun Chen, and Jun Feng. Helpnet: Hierarchical perturbations consistency and entropy-guided ensemble for scribble supervised medical image segmentation. *Medical Image Analysis*, page 103719, 2025. 1