SAM-REF: Introducing Image-Prompt Synergy during Interaction for Detail Enhancement in the Segment Anything Model

Supplementary Material

1. Implementation Details

1.1. Datasets

Following previous works [1, 3–5, 9, 10, 14], we perform evaluations on five different datasets to thoroughly assess our methods:

• **GrabCut**: 50 images (50 instances), each with distinct foreground and background.

• **Berkeley**: 96 images (100 instances), with some overlap with GrabCut.

• **DAVIS**: 345 specific frames from 50 videos as used in for evaluation, aligning with previous studies.

• **SBD** : 2857 validation images (6671 instances) for evaluation purposes.

• **HQSeg-44K** : 44320 images(train set) and 1537 images (validation set) [5, 10]. It is a collection of six existing image datasets, including DIS [11] (train set), ThinObject-5K [8] (train set), MSRA10K [2], FSS-1000 [7], DUT-OMRON [15], and ECSSD [13]. Each of them contains 7.4K extremely accurate image mask annotations on average.

1.2. Implementation details

We adopt the AdamW optimizer to train our proposed SAM-REF. The initial learning rate is set to $1e-6$ and raised to $1e-4$ after 1500 iteration. We then apply a polynomial decay strategy to the learning rate, setting AdamW's β_1 to 0.9 and β_2 to 0.999. We train SAM-REF at a batch size of 4 per GPU, totaling 16 samples across 4 GPUs for 80k iterations. During training, we resize the the longest side of each image to 1024 and pad it to 1024×1024. During inference, we resize test images directly to 1024×1024 without padding. All of our experiments are conducted on a server with 4 NVIDIA Tesla V100-PCIE-32GB GPUs and Intel(R) 326Xeon(R) Gold 6278C CPU.

1.3. Click Simulation

During training, we adopt InterFormer's click simulation strategy due to its simplicity [3]. We set the upper limit for simulated clicks at 20. To determine the distribution of click counts, we employ a decay coefficient γ, where the probability for a given number of clicks decreases progressively. We set the maximum simulation click at 20 and sample the number of simulations with an exponential decaying probability, where the probability of the number of clicks decreases gradually. Specifically, the probability of having i clicks is γ multiplied by the probability of having $i-1$ clicks, with the constraint that $\gamma < 1$. This method ensures a higher probability of selecting fewer clicks. It has more diverse selection clicks compared to RITM [14] and SAM [6] training methods. For the joint training of COCO and LVIS datasets, SAM-REF sets γ= 0.6. For the training of HQSeg-44k, SAM-REF use γ= 0.9, in order to more effective use the detailed annotations of HQSeg-44k.

2. Supplementary Experiments

θ	$\boldsymbol{N}_r^G$	$\boldsymbol{N}_{t-1}$	NoC90	NoC95
1.05	7.20	12.80	4.60	9.30
1.10	9.44	10.56	**4.54**	**9.10**
1.15	12.89	7.11	4.59	9.20
1.20	15.01	4.99	4.61	9.35
1.30	16.87	3.13	4.60	9.37

Table 1. **Ablation study for the threshold of θ.**

Influence of θ. Tab. 1 shows the impact of different θ. $\boldsymbol{N}_r^G$ and $\boldsymbol{N}_{t-1}$ represent the average number of pastes on $\boldsymbol{M}_r^G$ and $\boldsymbol{M}_{t-1}$ over 20 interactions. Higher θ result in more $\boldsymbol{N}_r^G$, while lower θ lead to more $\boldsymbol{N}_{t-1}$. We select θ=1.1 because it could get the best results, significantly impacting NoC95.

Method	Backbone	Latency(s)↓	5-mIoU↑	Noc90↓	Noc95↓	NoF95↓
SegNext	ViT-B	22.1	85.71	7.18	11.52	700
SAM-REF	ViT-B	5.1	89.0	6.09	9.72	596
SAM	ViT-H	4.21	88.0	6.50	10.53	653
SAM-REF	ViT-H*	5.22	89.6	5.44	9.16	566
SAM2	Hiera-L	4.12	88.26	5.90	9.87	611
SAM2-REF	Hiera-L*	5.29	89.1	5.60	8.99	565

Table 2. **Results on high-quality datasets.** All models are tested on HQSeg-44K datasets. * denotes frozen backbone.

SAM-REF vs. SegNext on Unfrozen Backbones. For a strictly fairer comparison, we provide additional contrast experiments to verify the effectiveness of SAM-REF. Tab. 2 shows the results of our models trained on COCO+LVIS datasets. As reported, with the unfrozen encoder, our method still clearly outperforms SegNext [10].

SAM2-REF vs. SAM2. Since our SAM-REF is decoupled from SAM [6], we integrate it into SAM2 [12] to validate the effectiveness and transferability of our method. As shown in Tab. 2, our method has a significant improvement over the original architecture in high-quality interactive segmentation scenarios both on SAM and SAM2. Besides, SAM2-REF outperforms SAM-REF on NoC95 and NoF95.

On 5-mIoU and Noc90, SAM2-REF underperforms SAM-REF due to the lighter encoder, but still outperforms other mainstream methods.

Method	Backbone	Params/MB↓	FPS↑	Mem/G↓
SAM [6]	ViT-H	635.6	1.70	3.7
SAM-REF	ViT-H	636.3	1.67	3.7
SAM2 [12]	Hiera-L	216.8	4.65	2.3
SAM2-REF	Hiera-L	218.0	4.34	2.3

Table 3. **Computation analysis for SAM, SAM-REF, SAM2, and SAM2-REF.**

Computation analysis. In Tab. 3, we report the comparison of model parameters (Params), inference time per image (FPS), and GPU memory (Mem). SAM2 and SAM2-REF have much fewer parameters and memory than SAM and SAM-REF, where the computing speed is also much faster. While SAM2-REF produces substantially better segmentation quality than SAM2, it adds just 1.2MB to the model parameters, with negligible increases in GPU memory use and inference time per image.

References

[1] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 1

[2] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 1

[3] You Huang, Hao Yang, Ke Sun, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, and Rongrong Ji. Interformer: Real-time interactive image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22301–22311, 2023. 1

[4] You Huang, Zongyu Lan, Liujuan Cao, Xianming Lin, Shengchuan Zhang, Guannan Jiang, and Rongrong Ji. Focsam: Delving deeply into focused objects in segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3120–3130, 2024.

[5] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[6] 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. 1, 2

[7] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020. 1

[8] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 1

[9] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 1

[10] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2024. 1

[11] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 1

[12] 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. 1, 2

[13] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4): 717–729, 2015. 1

[14] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 1

[15] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 1