Electronic Supplementary Material (Online Resource 1) for *MC-SAM: A Stability-Constrained Coupled Adaptation Framework for SAM in Camouflaged Scene Segmentation*

Zihua Liu[1], Yonghao Wu[1,*], Jiajie Murong[1], Chang Liu[2]

The Visual Computer

[1]School of Electronic Science and Engineering (School of Microelectronics), South China Normal University, Foshan, China

[2]St. Petersburg School of Shipbuilding and Ocean Technology, Guangdong Ocean University, Zhanjiang, China

*Corresponding author: 2023025071@m.scnu.edu.cn

This document accompanies the manuscript. It records the implementation and evaluation protocol, gives the scope of the mixer-level argument, and provides explicit parameter accounting. Main-text table and figure numbers are not reused here.

1 Training, model selection, and test protocol

The source training pool comprises 3,040 COD10K training images and 1,000 CAMO training images (4,040 in total). A fixed split with seed 42 reserves 404 images for validation and leaves 3,636 images for parameter updates. The official COD10K and CAMO test splits, the complete NC4K benchmark, and the 76 CHAMELEON images are used only for final evaluation. No image from an official test set is used to update parameters or select a checkpoint.

Training runs for a maximum of 20 epochs with batch size 1. After each epoch, RankDice-RMA is enabled for validation inference; the validation score is the arithmetic mean of S_α, adaptive E-measure, and F_β^w computed from those predictions. Fixed-budget validation-score maximization selects the highest-scoring checkpoint among the 20 epochs. The same selected checkpoint is evaluated without RankDice-RMA for the main benchmark table and probability-map ablation rows, and with RankDice-RMA only for the explicitly marked optional-inference comparisons. RankDice-RMA is not used in the training objective. Official test sets play no role in checkpoint selection. The reported hardware is one NVIDIA RTX 5090 with 32 GB of memory.

Images are converted to RGB, resized to 1024×1024, and normalized by ImageNet channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. Masks use nearest-neighbor resizing. No random image augmentation is applied. AdamW uses an initial learning rate of 5×10^{-5}, weight decay 0.01, and cosine scheduling.

2 Component configuration

The original ViT-H SAM image encoder is frozen. The SAM prompt encoder is frozen, while its mask decoder is trainable. The MM-SAM-supplied BLIP captioner (https://huggingface.co/Salesforce/blip-image-captioning-large) and Mamba text encoder (https://huggingface.co/state-spaces/mamba-130m-hf) remain frozen and are outside the integrated segmentation-model parameter total. Captions are generated using the BLIP implementation's default generation configuration.

CSPG forms a Gram affinity $A = UU^\top$ from row-normalized text and visual embeddings and applies an elementwise exponential with temperature $\epsilon = 1.0$. Five alternating row–column normalization pairs approximate a doubly stochastic pre-mixing matrix. A learned post-mapping is column-normalized before producing the two sparse prompts. MCA uses four streams in a 128-dimensional bottleneck, with 20 row–column normalization pairs. One MCA parameter set is reused at SAM blocks 8, 16, and 24. The MCA gates are initialized to 0.01 before softmax. BLIP visual-feature alignment has rank 48.

The dense-prompt head contains two depthwise-separable convolution stages with 256 input and output channels, each followed by GELU. HyperCond maps the scalar threshold through widths $1 \to 64 \to 128$ and $128 \to 256 \to 128$, then projects to 256 channels and applies a 256-channel depthwise-separable spatial projection. The resulting condition enters the dense prompt. The boundary-loss weight λ_b is not a HyperCond input.

3 Loss and inference-only mask selection

For prediction logits z and ground truth G, the training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}} + \lambda_b \mathcal{L}_{\text{boundary}}. \tag{1}$$

The base Dice term uses squared predictions. A soft boundary map B is obtained by 3×3 morphological dilation minus erosion of G, followed by 3×3 average pooling and normalization by its maximum. When $\sum_i B_i > 10$,

$$\mathcal{L}_{\text{boundary}} = 0.05(\mathcal{L}_{\text{Dice},B} + \mathcal{L}_{\text{BCE},B}), \tag{2}$$

and otherwise it is zero. The boundary Dice term compares $\text{sigmoid}(z)B$ with GB; boundary BCE applies binary cross-entropy with logits to zB and GB and averages over the full map. The outer λ_b is applied once. Independent $\text{Beta}(2, 2)$ samples are rescaled to $t \in [0.4, 0.6]$ and $\lambda_b \in [0.3, 1.0]$ during training. At inference $t = 0.5$ by default, and λ_b is absent.

RankDice-RMA is an optional inference-only selection step operating on the output of the already-trained model. It adds no training loss and no trainable parameters. MC-SAM denotes Base+HC+CSPG+MCA without this selector; MC-SAM+RankDice-RMA uses the same trained weights with the selector enabled. Its selected threshold τ^* is distinct from HyperCond's supplied condition t.

The implementation returns the continuous sigmoid map $\mathbf{S}$ when the selector is disabled, and the binary mask $M = \mathbb{I}(\mathbf{S} \geq \tau^*)$ when it is enabled. Standard COD scores on these two output objects describe different representations. They should not be interpreted as a change in the decoder probabilities. For the same checkpoint, image preprocessing, and HyperCond condition, $\mathbf{S}$ is unchanged. Any ECE, threshold-sweep AUC, or best-threshold statistic computed from this same map is therefore shared by the two selection modes. A comparison of binary decision rules requires evaluating both a fixed-threshold mask and the selected mask from the same probability map.

4 Scope of the stability argument

Let H be an *exact* doubly stochastic mixing matrix: $H_{ij} \geq 0$, $H\mathbf{1} = \mathbf{1}$, and $H^\top \mathbf{1} = \mathbf{1}$. The Birkhoff–von Neumann theorem expresses H as a convex combination of permutation matrices, $H = \sum_k \theta_k P_k$, with $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. Since every permutation matrix has spectral norm one, convexity of the operator norm yields $\|H\|_2 \leq 1$. Thus, for a *fixed* H, $\|HZ\|_F \leq \|Z\|_F$ and $\|H(Z_1 - Z_2)\|_F \leq \|Z_1 - Z_2\|_F$.

This statement concerns only the exact mixing operator. A finite number of Sinkhorn iterations approximates double stochasticity. The learned H also depends on its input, so

$$H(Z_1)Z_1 - H(Z_2)Z_2 = H(Z_1)(Z_1 - Z_2) + \big(H(Z_1) - H(Z_2)\big)Z_2. \tag{3}$$

The fixed-operator bound controls the first term only. It does not guarantee non-expansiveness of the complete residual block, preservation of neighborhood distances, or improved segmentation. Those properties are assessed empirically in the main manuscript.

5 Parameter accounting

Table S1 counts registered parameters in the integrated ViT-H segmentation model. A parameter is trainable if it has gradient updates enabled. The shared MCA is counted once even though it is invoked at three blocks. Frozen BLIP and Mamba cue extractors are excluded from the denominator, and their inference costs must not be inferred from this parameter count.

Table S1: Module-level parameters in the reported integrated ViT-H model.

Module	Trainable	Frozen
Original SAM image encoder	0	637,026,048
SAM prompt encoder	0	6,220
Shared bottleneck MCA	580,276	0
BLIP feature alignment	99,584	0
CSPG	100,868	0
HyperCond	75,264	0
Conditional channel projection	32,768	0
Conditional spatial projection	68,096	0
Dense-prompt head	136,192	0
SAM mask decoder	4,058,340	0
Total	5,151,388	637,032,268

6 Computational cost and inference throughput

Tables S2 and S3 report the measured comparison with the locally run MM-SAM configuration using the same ViT-H backbone and BLIP–Mamba text-generation pathway. RankDice-RMA introduces no trainable parameters. Its FLOP entries indicate that the main network is unchanged; they do not claim that post-processing itself has zero computational cost.

Table S2: Measured latency, throughput, and peak memory. Both configurations use BLIP and Mamba for text generation.

Model	Latency (ms/img)	Throughput (img/s)	Peak memory (GB)
MM-SAM (ViT-H)	563	1.8	10.5
MC-SAM w/o RankDice-RMA	472	2.1	9.0
MC-SAM w/ RankDice-RMA	485	2.06	9.1

Table S3: Parameters and FLOPs per image. T denotes 10^{12} operations. RankDice-RMA FLOPs refer to the unchanged main network and exclude selector post-processing.

Model	Trainable	Integrated total	Full (T)	SAM+adapters (T)
MM-SAM (ViT-H)	9,164,932	646,190,980	3.45	3.02
MC-SAM w/o RankDice-RMA	5,151,388	642,183,656	3.24	2.77
MC-SAM w/ RankDice-RMA	5,151,388	642,183,656	3.24*	2.77*

*Main-network FLOPs only.

Compared with local MM-SAM, MC-SAM without RankDice-RMA reduces latency by 91 ms/img and peak memory by 1.5 GB. The optional selector adds 13 ms/img and 0.1 GB. These are implementation-specific measurements, not hardware-independent guarantees.