A Novel SAM Coupling Mechanism for SAR Image Segmentation

Yang Liu*, **Miao Fu**, **Yingqi Gao**, **Lin Lin**, **Jiarui Li** and **Rui Liu**
School of Computer Science and Artificial Intelligence, Liaoning Normal University, Dalian, China
yangliu.0816@lnnu.edu.cn

Abstract

The Segment Anything Model **(SAM)** provides a new paradigm for visual perception tasks through large-scale pre-training and interactive prompts. However, the quality of mask generation is limited by the difficulty of aligning the optimization direction between the backbone and the prompt, which is especially significant in non-natural domain images with low contrast and indistinct semantic edge structure features, especially in SAR images. Therefore, this paper proposes a coupling optimization mechanism that integrates the originally independent prompt generation process with the backbone into a closed loop through the Point Impact Decomposition **(PID)** module to guide the iterative optimization of the prompt. Specifically, this paper introduces the concept of PID in the decoder to quantify the contribution of prompt points to the masks, constructing an interpretable reward variable to drive the reinforcement learning prompt optimizer. This optimizer designs a reward function based on hypothesis testing ideas, achieving iterative updates of prompt points, and providing feedback on the effects of mask generation based on prompt points. Experiments show that this mechanism can provide a prompt that fits image features; its performance exceeds the existing baseline, which effectively improves the generalization ability of SAM on **SAR** images. Code is available at https://github.com/Fm336/PID-SAM.

1 Introduction

The Segment Anything Model (SAM) [Kirillov *et al.*, 2023; Ravi *et al.*, 2024; Carion *et al.*, 2025] achieves zero-shot object segmentation based on the SA-1B dataset; however, due to its training data predominantly comprising natural images, its performance is constrained in non-natural domain images, especially in Synthetic Aperture Radar (SAR) images characterized by weak boundary contrast and low signal-to-noise ratios, as well as in fine-grained object segmentation tasks. Interactive prompts tend to confuse foreground objects with background noise. Furthermore, SAM employs the Vision Transformer **(ViT)** architecture [Liu *et al.*, 2021], which relies on global feature matching for prompt-based inference, thus lacking effective constraints on pixel-level boundary delineation.

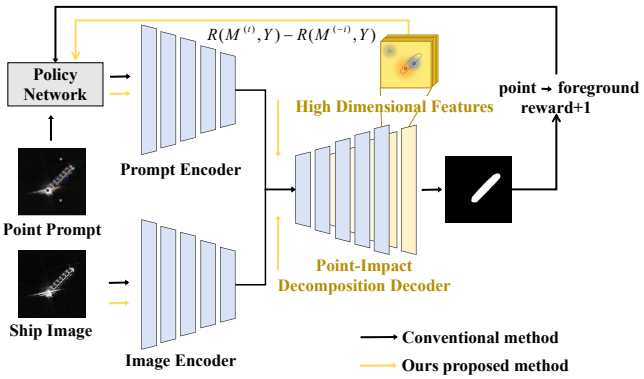

Figure 1: Comparison of method flow. 1) The **black arrow** represents the existing separation paradigm of cues and the main structure. 2) The **yellow arrow** represents the closed-loop coupling mechanism proposed in ours method.

To enhance the generalization capability of the SAM in non-natural domains, existing research primarily focuses on improvements in structural adaptation and prompt mechanisms. On the one hand, approaches such as intermediate representations, lightweight adaptation [Chen *et al.*, 2024a; Gong *et al.*, 2024; Yang *et al.*, 2023], internal structure enhancement [Chen *et al.*, 2023; Pu *et al.*, 2025; Zhang *et al.*, 2025a], multi-scale and local feature modeling [Chen *et al.*, 2024c], and hybrid architecture designs [Xiong *et al.*, 2026] are employed to mitigate issues such as low contrast, complex textures, and fine-grained edge ambiguity. On the other hand, strategies involving semantic or visual language prompts [Kweon and Yoon, 2024; Ren *et al.*, 2024], prompt self-correction [Radman and Laaksonen, 2025], and multi-prompt fusion are adopted to strengthen target representation and improve segmentation performance in complex scenarios [Hu *et al.*, 2024]. However, current methods often treat backbone modeling and prompt generation as disjoint processes, lacking effective collaboration, which restricts cross-scene adaptability.

*Corresponding author.

We categorize the synergistic relationship between the backbone network and prompt generation into two types: strong coupling and weak coupling. Strong coupling training frequently results in gradient conflicts and unstable training processes due to inconsistent optimization objectives among networks [Chen *et al.*, 2018]. Consequently, most approaches adopt a weakly coupled multi-stage strategy, introducing adaptive or hierarchical instructions during the learning of backbone characteristics to achieve joint optimization [Wang *et al.*, 2025; Li *et al.*, 2024]. A representative method, AlignSAM [Huang *et al.*, 2024], optimizes prompt configurations in the final binary mask layer through reinforcement learning, as illustrated by the black arrow in Figure 1. However, this kind of coupling mechanism operates in the low-dimensional output layer, imposing limited constraints on the prompt generation process and struggling to handle complex structures and weak boundary scenarios.

In this paper, we train a reinforcement learning network based on Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] as a point prompt optimizer and design a hypothesis testing-based reward function inspired by statistical principles to provide an interpretable credit assignment for prompt generation. The optimizer determines subsequent prompt points based on the current state and the reward signal from the backbone, thus achieving a high-dimensional closed-loop coupling between the backbone network and the prompt mechanisms, as illustrated by the yellow arrow in Figure1. The framework relies solely on the standard SAM interface, ensuring compatibility with various versions. Experimental results validate its effectiveness, particularly in cross-domain applications such as SAR image segmentation. Our main contributions are summarized as follows:

- A novel higher-dimensional coupling mechanism for SAM has been proposed, which enables synergistic and iterative optimization between prompt generation and the primary architecture through reward signal feedback prior to binary masking.
- A point impact decomposition module was designed that constructs point-level contributions from high-dimensional features before masking and employs contribution regularization to achieve a fine-grained representation of the mask boundaries.
- A PPO-based reinforcement learning prompt optimizer is introduced, coupled with the establishment of a reward function grounded in point-wise contribution hypothesis testing, thereby enabling stable point-level prompt strategy learning.
- The methodology has been validated on the **FSS-seg** and **HRSID** datasets, demonstrating performance that significantly outperforms existing baseline approaches.

2 Related Work

2.1 Non-natural Domain Image Adaptation of SAM

Early applications of SAM in non-natural domain imagery have primarily focused on medical images, which are characterized by low boundary contrast and poor signal-to-noise ratios. To address its limitations in feature representation and spatial modeling, research efforts have been made to refine SAM through intermediate representations [Yang *et al.*, 2023], prompt modeling, and lightweight adaptation architectures [Gong *et al.*, 2024; Chen *et al.*, 2024a]. Subsequently, SAM has been extended to remote sensing and radar domains, where performance in SAR imaging, multi-resolution scenarios, and noisy or blurry environments [Pu *et al.*, 2025; Chen *et al.*, 2024c; Zhang *et al.*, 2025a] is enhanced through parameter fine-tuning, category-level adaptation, cross-resolution feature alignment, and image priors.

2.2 Prompting Methods in Interactive Segmentation

Early interactive segmentation methodologies predominantly used energy minimization and graph-cut frameworks [Rother *et al.*, 2004], leveraging user-supplied points or boundary constraints to achieve global optimization [Boykov and Jolly, 2001]. With the advent of deep learning, prompts such as clicks, extreme points, or coarse bounding boxes have been encoded into end-to-end neural networks, enabling more efficient segmentation [Maninis *et al.*, 2018; Chen *et al.*, 2022]. Reinforcement learning approaches model the segmentation process as sequential decision-making, autonomously generating or refining prompts to minimize manual intervention [Liao *et al.*, 2020; Wang *et al.*, 2024b; Chen *et al.*, 2025b; Huang *et al.*, 2025]. Following the introduction of SAM, research has explored prompt enhancement [Dai *et al.*, 2023], sequential optimization, as well as the integration of reinforcement and adversarial learning [Wang *et al.*, 2025; Liu *et al.*, 2025b], to achieve automated prompt generation and calibration [Chen *et al.*, 2025c; Li *et al.*, 2024], thus enhancing the robustness and performance of segmentation.

2.3 Weak Coupling Between the Trunk and Prompts in SAM

Current collaborative research based on SAM predominantly employs weakly-coupled strategies involving frozen backbone modules [Chen *et al.*, 2025c; Li *et al.*, 2024; Zhang *et al.*, 2025b]. GenSAM [Hu *et al.*, 2024] maps text prompts to visual prompts, but struggles with cross-modal alignment and provides limited generalization enhancement for SAM through textual information. DAPSAM [Wei *et al.*, 2024] addresses the key connection between general-purpose models and non-natural domains via domain adaptation, but faces challenges in balancing extreme domain adaptation with lightweight efficacy. The representative AlignSAM [Huang *et al.*, 2024] Binary mask feedback reinforcement learning optimization. However, its reliance on low-dimensional mask-level prompts restricts the localization accuracy for fine-grained structures and low-contrast boundaries.

3 Methodology

The framework of the proposed method is shown in Figure 2. Based on the SAM segmentation backbone , we design modules and strategies such as PID and Hypothetical Testing

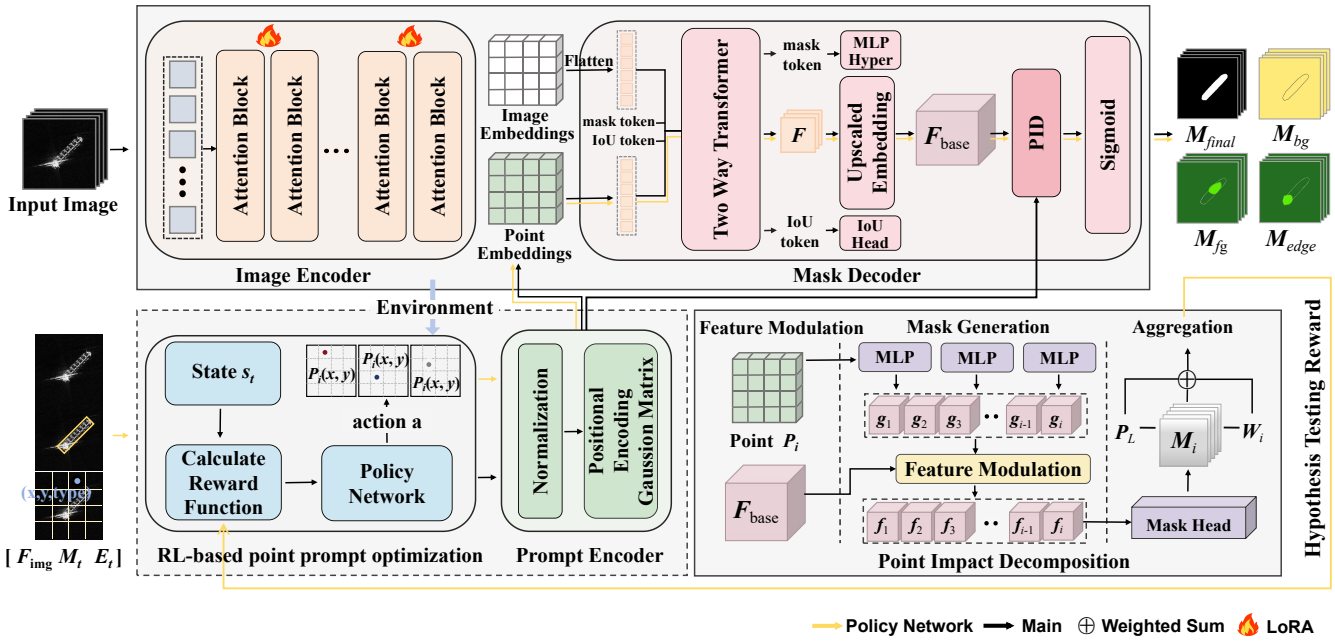

Figure 2: Overview of the coupling optimization mechanism. This mechanism quantifies the contribution of prompt points to mask generation through the point impact decomposition module, and thereby constructs a reward function based on hypothesis testing; it then drives the reinforcement learning strategy network to achieve iterative optimization of prompt points, ultimately forming a closed-loop coupling between the segmentation backbone and prompt generation.

Reward, and provide a unified feature and decoding interface for the SAM series of models.

In SAM, the input image is first processed by the image encoder to extract multi-scale features, which are flattened into a sequence F_{flat}. The point prompts are embedded in the feature vectors by the prompt encoder and concatenated with the learnable Mask tokens T_{mask}, the IoU token T_{iou} and F_{flat} to form the Transformer input sequence. This sequence is then processed by a bidirectional Transformer to fuse image features with prompt information.

To improve SAM's transfer efficiency in cross-domain tasks, we incorporate Low-Rank Adaptation (LoRA) [Zhong *et al.*, 2024] into the multi-head self-attention layers of the SAM image encoder. For each projection matrix W, a trainable low-rank side path is added to the Query and Value projections. During fine-tuning, only the side-path parameters A and B are updated, while the original parameters W_0 are kept fixed.

3.1 Point Impact Decomposition

Within the SAM mask decoder, we incorporate a point impact decomposition mechanism that disassembles the intricate effects of point prompts on mask prediction into relatively independent feature representations, thereby rendering the point impact observable and steerable. This framework comprises three components: point-wise feature modulation, point-wise mask generation, and linear mask reconstruction, which collectively enable explicit modeling and aggregation of the per-point contributions from each prompt.

Conditional Feature Modulation at the Point Level. To embed the information of the prompt points into the backbone feature maps and obtain the corresponding point modulation vectors, each point embedding p_i is processed by a lightweight MLP:

$$g_i = \phi(p_i) = W_2 \cdot \mathrm{ReLU}(W_1 \cdot p_i + b_1) + b_2 \quad (1)$$

Here, $\phi(\cdot)$ represents a fully connected two-layer network, where W_1 and W_2 are learnable weight matrices, and b_1 and b_2 are bias terms. The modulation vector g_i encodes the relative importance of the point i for each channel of features. Subsequently, a channel-wise multiplication operation $\odot$ is applied to transform the global feature F_{base} into the corresponding point-specific feature map f_i. This enables each point to derive a semantically distinct feature representation from the shared feature F_{base}, guided by its own modulation vector:

$$f_i = g_i \odot F_{\text{base}} \quad (2)$$

Point-level Mask Generation. Each point-specific feature map f_i is decoded into a spatial contribution mask through a shared lightweight convolutional head:

$$M_i = h(f_i) = \mathrm{Conv}_{1\times1}\Big(\mathrm{ReLU}\big(\mathrm{LayerNorm}\big(\mathrm{Conv}_{3\times3}(f_i)\big)\big)\Big) \quad (3)$$

The convolutional head $h(\cdot)$ consists of a 3×3 convolution, layer normalization, ReLU activation, and a 1×1 convolution. Generates the logarithmic map of the original contribution M_i for each point. The shared convolutional head design ensures that all points follow the same decoding rules, making the contribution masks of different points comparable.

Linear Mask Reconstruction. The ultimate segmentation mask is generated through a linear aggregation of pixel-level contributions to reconstruct the final prediction:

$$M_{\text{final}} = \sigma\left(\sum_{i=1}^{N} w_i \cdot M_i + b\right),$$
$$w_i = \begin{cases} w_{jg}, & P_L[i] = 1, \\ w_{bg}, & P_L[i] = 0, \\ w_{edge}, & P_L[i] = 2. \end{cases} \quad (4)$$

The weights w_i and biases b are learnable parameters, w_{fg} and w_{bg} are globally learnable scalars, P_L denotes point-type labels, and the final probability mask is obtained through a Sigmoid function per-pixel $\sigma(\cdot)$. This framework transforms the global segmentation task into a linear combination of points-level contributions, where each M_i explicitly represents the spatial impact of the corresponding point, allowing visualization and interpretable analysis of the segmentation process and providing reward variables for subsequent point-prompter optimization.

PID Regularization. To ensure that the point-level contribution mask M_i output of the PID decoder is interpretable, distinguishable, and locally concentrated, this paper introduces two important regularization expressions:

$$L_{PID} = \lambda_{local} L_{local} + \lambda_{overlap} L_{overlap} \quad (5)$$

The local consistency in pointwise L_{local} serves as a basis for the interpretability of PID. It constrains the contribution mask of each point prompt to exhibit behavioral patterns within its local neighborhood that are consistent with the semantic meaning of the point type. For foreground points, higher activation values are encouraged at the point location p_i compared to other positions within a radius neighborhood r; on the contrary, suppression is applied for background points. Mathematically, this is expressed as the summation of difference penalties computed separately for different point types based on activation variances within their respective neighborhoods:

$$L_{local} = \frac{1}{B \cdot N} \sum_{m=1}^{B} \sum_{i=1}^{N} \frac{1}{|N_r|} \sum_{q \in N_r} \max \left(0, \text{sign}_{PL_i} \cdot \left(M_i^{(m)}(q) - M_i^{(m)}(p_i)\right)\right) \quad (6)$$

Here, B denotes the batch size, N represents the number of prompt points per sample, m is the batch index, i is the point index within the sample, and N_r refers to the set of neighborhood points centered at the point p_i with radius r, where $|N_r|$ indicates the number of neighboring points and is used to average the loss within the neighborhood. $M_i^{(m)}(p_i)$ represents the contribution mask value of the i-th point in the m-th sample at its own position, while $M_i^{(m)}(q)$ denotes the contribution mask value of the same point at the position of a neighborhood point q. signPL_i is a sign function determined based on the point label. The term $\max(0, \cdot)$ denotes the Hinge function, which imposes a penalty only when the local relational constraint is violated.

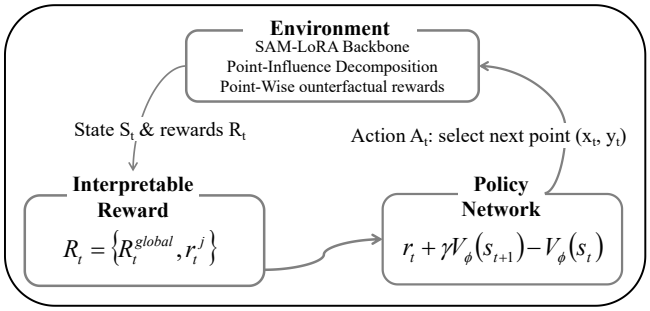

Figure 3: Environment for cooperative point-prompt tuning

The loss of overlap L_{overlap} is designed to improve discriminability between the contribution masks of different points, thereby preventing multiple points from generating redundant and ambiguous contributions to the same region. It penalizes the sum of the element-wise spatial products of the contribution masks for any pair of points:

$$L_{overlap} = \frac{1}{B \cdot \binom{N}{2}} \sum_{m=1}^{B} \sum_{i<j} \left\| M_i^{(m)} \odot M_j^{(m)} \right\|_1 \quad (7)$$

The normalization factor $B \cdot \binom{N}{2}$ is introduced to eliminate the impact of batch size B and the number of points per sample N on the magnitude of the loss. For each sample m, the summation is performed on all pairs (i, j) satisfying $i < j$ to avoid redundant calculations. Here, $M_i^{(m)}$ and $M_j^{(m)}$ represent the contribution masks of the points i-th and j-th in the sample m-th, respectively. The L1 norm of the product matrix yields a quantitative measure of the degree of overlap. Minimizing L_{overlap} encourages non-overlapping contribution masks across distinct points, thereby enhancing the discriminability and complementarity of pointwise contributions.

3.2 RL-based Point Prompt Optimization

Building upon the reinforcement learning-based point prompt optimizer, a new **Hypothesis Test Reward Function** based on the reward variable provided by PID has been designed., enabling precise credit assignment at the point level for each action generated by the policy network. The iterative procedure is elaborated in Figure 3.

State Representation and Action Space. At each iterative step t of the reinforcement learning process, the state s_t of the policy network is constructed by concatenating three component elements of the characteristic $[F_{img}; M_{t-1}; E_t]$, which correspond to the deep features of the image, the prediction from the previous step and the embedding of the current prompt, respectively. In this way, the policy network integrates visual semantics, segmentation progression, and interaction history together. Its action space consists of selecting the next coordinate point on the image plane as the prompt.

Reward Function. The global reward metric quantifies the final segmentation quality by evaluating the predicted mask obtained in the current interaction round using the Dice coefficient. Specifically, upon completion of the t-th interaction

round, the global reward is defined as:

$$R_{global}^{(t)} = Dice\left(M_{final}^{(t)}, Y\right) = \frac{2 \cdot \left|M_{final}^{(t)} \cap Y\right| + \varepsilon}{\left|M_{final}^{(t)}\right| + |Y| + \varepsilon} \quad (8)$$

Here, $M_{\text{final}}^{(t)}$ denotes the predicted mask obtained under the current prompt set $P^{(t)}$, Y represents the ground truth mask, and ε is a numerical stability constant.

To resolve credit assignment in interactive point-based prompt learning, we introduce a hypothesis testing-based reward that quantifies the marginal contribution of each prompt point, allowing backbone-guided prompt optimization. At the t-th decoding step, the hypothetical testing reward for the j-th action point is defined as:

$$\begin{aligned} r_j^{(t)} = R_{global}&\left(\sigma\left(\sum_{k=1}^{N_t} w_k M_k + b_3\right), Y\right) \\ &- R_{global}\left(\sigma\left(\sum_{k=1}^{N_{t-1}} w_k M_k + b_3\right), Y\right) \end{aligned} \quad (9)$$

Where N_t denotes the number of points triggered in the decoding process in the interaction t-th, M_k represents the point-level mask generated by the PID decoder for the k-th point prompt, w_k is a type-dependent weight learnable and b_3 is the bias term. The first term corresponds to the global segmentation quality under the actual prediction where point j is active, while the second term corresponds to the hypothetical scenario where point j is inactive, constructed by removing its corresponding contribution from the decoded logits. The difference between these two terms quantifies the performance gain attributable to point j in the final outcome of the segmentation.

Policy Network. We employ a proximal policy optimization algorithm based on the Actor–Critic architecture[Konda and Tsitsiklis, 1999] to train the policy network. The Critic network, denoted as V_ϕ, functions as a state-value estimator to compute advantage functions to reduce the variance in policy gradients. For advantage estimation, we adopt the generalized advantage estimation method:

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \quad (10)$$

Here, r_t represents the immediate reward obtained after executing an action in the state s_t. γ denotes the discount factor, while $V_\phi(\cdot)$ signifies the state-value function parameterized by the Critic network. δ_t refers to the temporal-difference error and A^t corresponds to the generalized advantage estimate (GAE), where λ serves as the decay coefficient of GAE to balance bias and variance.

3.3 Loss Function

The segmentation backbone loss is composed of binary cross-entropy loss, Dice loss, and point-guided loss:

$$L_{seg} = w_{BCE} \cdot L_{BCE} + w_{Dice} \cdot L_{Dice} + \alpha \cdot L_{point} \quad (11)$$

The binary cross-entropy loss quantifies the pixel-wise agreement between the predicted probabilities and the ground truth labels. Given the predicted logit map M, the loss function is defined as follows:

$$\begin{aligned} L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\Big[&Y_i \log \sigma(M_i) \\ &+ (1 - Y_i)\log(1 - \sigma(M_i))\Big] \end{aligned} \quad (12)$$

The Dice loss function optimizes the overlap between the predicted foreground region and the ground truth foreground region, demonstrating strong robustness against class imbalance issues. Given the predicted mask as $\hat{Y}$, the loss is defined as:

$$L_{Dice} = 1 - \frac{2 \cdot \left|\hat{Y} \cap Y\right| + \varepsilon}{\left|\hat{Y}\right| + |Y| + \varepsilon} \quad (13)$$

Point-guided loss encourages the alignment of predicted mask boundaries with the user-provided point prompt locations to enhance the model's comprehension of interactive intent:

$$L_{point} = \frac{1}{B \cdot N}\sum_{m=1}^{B}\sum_{i=1}^{N} \max\left(0, dist\left(p_i, \partial\hat{Y}_m\right) - \delta\right) \quad (14)$$

For each point prompt p_i, the distance from its location to the boundary of the predicted mask $\partial\hat{Y}_m$ is calculated, where ∂ denotes the boundary and $\hat{Y}_m$ represents the predicted mask for the m-th sample. δ is a non-negative threshold hyperparameter defining the acceptable alignment tolerance along the boundary. The discrepancy between the computed distance and the threshold δ yields a non-positive value when the distance is less than or equal to δ, and a positive value when the distance exceeds δ. Minimizing this loss term encourages the predicted mask boundary to align as closely as possible with the point prompt location, thereby enhancing the accuracy of the model in boundary localization.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. Based on FuSARShip1.0 [Hou *et al.*, 2020], a pixel-level semantic segmentation dataset, termed FSS-Seg, is constructed. Through annotation of the original data, the resulting data set contains 11,857 ship images that span seven semantic categories. Following the standard evaluation protocol, the dataset is divided into 9,470 images for training and 2,387 images for testing. Additionally, we also evaluate our method on the public HRSID dataset, which contains 5,604 SAR ship images and 16,951 ship targets. Detailed dataset statistics and specifications are reported in Appendix.

Implementation Details. All models were implemented using NVIDIA GeForce RTX 3090 24GB GPUs and PyTorch 2.7.1. The models were trained with AdamW optimizer, starting with an initial learning rate of $1e{-}3$, a batch size of 8, and ran for 100 epochs. The input images were uniformly resized to $224{\times}224$ pixels, and the random seed was fixed to 1234.

Method	Fss-seg				HRSID (Offshore)				HRSID (Inshore)			
	mDice	mIoU	Prec.	HD95	mDice	mIoU	Prec.	HD95	mDice	mIoU	Prec.	HD95
SAM (ViT-L) [Kirillov *et al.*, 2023]	77.04	63.39	80.98	10.33	57.90	41.03	65.11	180.27	59.67	52.10	60.14	176.33
SAM-Adapter [Chen *et al.*, 2023]	78.29	64.98	81.92	11.02	61.56	51.48	63.46	136.01	63.62	63.65	66.23	147.65
CAM-SAM [Kweon and Yoon, 2024]	81.29	69.23	78.81	36.14	68.27	59.16	67.39	42.20	61.93	67.73	68.64	40.90
Grounded SAM [Ren *et al.*, 2024]	81.01	69.08	77.26	36.23	77.19	68.05	79.28	32.31	70.59	71.52	75.31	63.77
PPO [Liu *et al.*, 2025b]	81.72	70.05	79.24	30.49	83.22	69.16	88.01	31.62	81.56	76.84	86.38	40.52
Ours (SAM1+PIDD+RL)	**87.18**	**77.89**	**89.23**	**7.29**	**86.93**	**76.50**	**91.30**	**14.03**	**83.64**	**79.22**	**90.99**	**17.98**
SAM2 (ViT-L) [Ravi *et al.*, 2024]	54.36	67.15	70.75	32.40	58.31	43.87	67.35	174.64	59.95	50.53	62.76	289.73
SAM2-Adapter [Chen *et al.*, 2024b]	83.99	73.79	84.38	7.55	65.04	68.00	72.47	121.41	68.20	61.96	70.95	166.72
Det-SAM2 [Wang *et al.*, 2024b]	78.65	66.96	84.51	11.15	73.71	60.88	74.66	108.16	77.07	65.58	81.37	178.04
SAM-OCTA2 [Chen *et al.*, 2025b]	84.05	73.00	85.02	6.62	86.80	75.69	81.30	30.09	83.49	72.08	85.61	96.00
CA-SAM2 [Huang *et al.*, 2025]	84.42	73.54	84.82	8.36	83.50	**78.48**	83.19	21.18	82.54	74.04	88.74	56.69
Ours (SAM2+PIDD+RL)	**89.02**	**82.87**	**90.64**	**6.30**	**86.84**	75.19	**91.44**	**16.12**	**87.91**	**77.24**	**90.50**	**26.92**
SAM3 (ViT-L) [Carion *et al.*, 2025]	77.32	63.57	73.16	43.86	67.87	62.13	69.32	235.37	63.25	69.92	66.09	203.00
SAM3-Adapter [Chen *et al.*, 2025a]	83.18	**71.29**	85.00	27.85	73.29	68.67	85.00	54.81	71.86	70.53	83.86	65.01
Ours (SAM3+PIDD+RL)	**85.13**	69.60	**86.33**	**10.19**	**83.33**	**72.14**	**93.73**	**33.54**	**86.79**	**72.45**	**91.53**	**25.83**
CWSAM [Pu *et al.*, 2025]	82.51	70.23	90.03	23.63	77.35	74.52	83.87	51.26	75.03	71.64	82.54	53.58
PolSAM [Wang *et al.*, 2024a]	76.81	63.85	78.44	10.13	73.40	61.20	77.40	54.18	69.11	60.80	71.26	67.21
PointSAM [Liu *et al.*, 2025a]	78.69	66.10	85.36	52.12	75.37	65.71	81.07	70.20	73.47	68.15	79.98	66.39

Table 1: Quantitative results on Fss-seg, HRSID (Offshore), and HRSID (Inshore) datasets. Metrics include Mean Dice (%), Mean IoU (%), Mean Precision (%), and Mean HD95. Bold indicates the results of our proposed methods.

	K	Mean Global↑	% r>0↑	% r<0↓
Random	8	0.5961	54.17%	45.83%
Heuristic	8	0.7457	63.10%	36.90%
PPO	8	**0.8173**	**87.50%**	**12.50%**

Table 2: Global and point-wise rewards under different prompt strategies on the FSS-Seg dataset. Comparison of Random, Heuristic, and PPO under $K = 8$.

Upon completion of the segmentation backbone training, the learning rate for the policy network was set at $2e{-}4$, with a discount factor $\gamma = 0.98$ and a GAE parameter $\lambda = 0.97$. Each interactive episode was limited to a maximum of 8 steps, starting with an initial token count of 1.

4.2 Results and Analysis

The experiments cover three generations of SAM base models and their representative enhanced variants, as well as SAR-oriented specialized adaptation methods. Quantitative evaluations were conducted using Dice coefficient, Intersection over Union (IoU), Precision, and 95th percentile Hausdorff Distance (HD95). The experimental results are summarized in Table 1. The following analysis focuses on the performance on FSS-Seg.

SAM1. We compared the proposed mechanism with the original SAM1[Kirillov *et al.*, 2023], SAM-Adapter[Chen *et al.*, 2023], CAM-SAM[Kweon and Yoon, 2024], Grounded SAM[Ren *et al.*, 2024], and PPO[Liu *et al.*, 2025b]. The presented method demonstrates substantial improvements over the original SAM and its enhanced variants in key accuracy metrics. Compared to the reinforcement learning-based PPO approach, our method achieves relative gains of 6.68% in Dice and 11.19% in IoU.

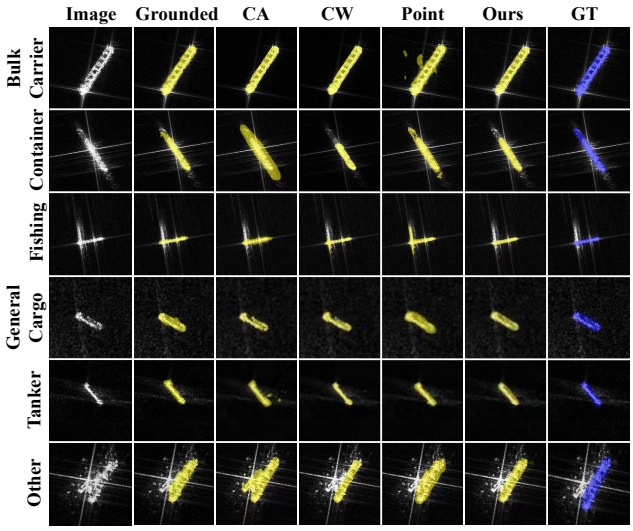

Figure 4: Qualitative comparison of segmentation results on the FSS-Seg dataset.

SAM2. We conducted a comparative analysis of the proposed mechanism against the original SAM2[Ravi *et al.*, 2024], SAM2-Adapter[Chen *et al.*, 2024b], Det-SAM2[Wang *et al.*, 2024b], SAM-OCTA2[Chen *et al.*, 2025b] and CA-SAM2[Huang *et al.*, 2025]. Compared to SAM2, the proposed approach achieved improvements of 63.75% and 23.41% in Dice and IoU, respectively. Furthermore, despite the coexistence of multiple strong adaptation

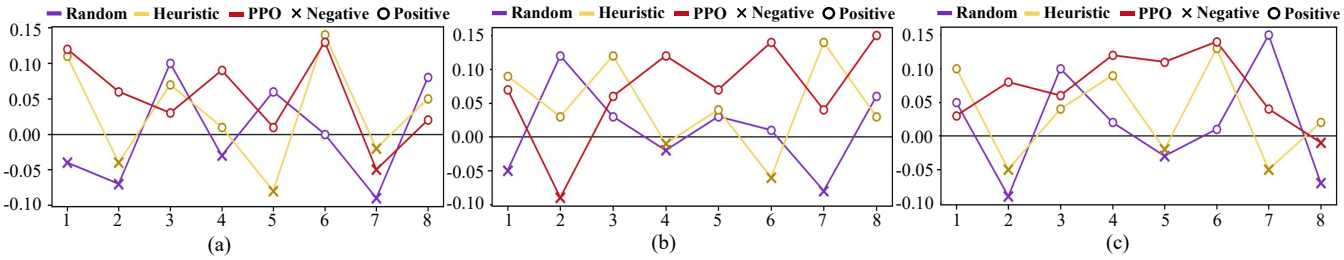

Figure 5: Dots and crosses denote the corresponding measurements on the FSS-Seg dataset. The three experiments correspond to: (a) Bulk Carrier sample, (b) Container sample, (c) General Cargo sample. The horizontal axis represents the point index, and the vertical axis represents the hypothesis testing reward score r. Positive and negative rewards are indicated by $r > 0$ and $r < 0$, respectively.

and heuristic prompt methods, our method still attained the best performance in three evaluation metrics.

SAM3. Our approach still improves overall accuracy, although with a smaller improvement compared to SAM1 and SAM2. However, the results validate the cross-generational transferability of this mechanism and confirm its capacity to sustain stable performance gains on more advanced models.

SAR-specific methods. We selected CWSAM[Pu *et al.*, 2025], pointSAM[Liu *et al.*, 2025a], and PolSAM[Wang *et al.*, 2024a] for a comparative evaluation. The CWSAM approach achieved a comparatively high accuracy of 90.03%, but exhibited notable boundary discrepancies, with HD95 reaching 23.6307 and IoU maintained at 70.23%. In contrast, the proposed method achieves a more favorable balance between overlap quality and boundary precision.

Visualization. For a comprehensive comparative analysis, Figure4 presents qualitative results on the FSS-Seg dataset, comparing our model with three other approaches including Grounded SAM[Ren *et al.*, 2024] and CASAM[Kweon and Yoon, 2024]. It can be observed that, under challenging scenarios, such as strong scattering points, streak artifacts, tailing interference, or elongated target structures, certain methods are prone to generating false background detections or exhibiting fragmented object contours. In contrast, the proposed method demonstrates superior capability in preserving the structural continuity and contour integrity of vessel regions while effectively suppressing clutter interference.

4.3 Module Comparison

In three distinct experiments, we conducted distributional measurements of rewards r for three types of policy, namely random policy[Dai *et al.*, 2023], heuristic policy[Dai *et al.*, 2023], and proximal policy optimization[Schulman *et al.*, 2017], at designated points $K = 8$. As indicated in Table2, the PPO policy exhibits the optimal global performance, substantially outperforming the other two methods. Furthermore, Figure5 illustrates the distribution of rewards between the eight measurement points. Compared with alternative approaches, the PPO policy maintains more consistent positive rewards in the majority of points, with fewer occurrences of negative rewards, and the reward distribution is more concentrated.

Method's Variants			FSS-Seg			
PIDD	**RL**	**Interop.**	**Dice**	**IoU**	**Precision**	**HD95**
–	–	–	77.04	63.39	80.98	10.34
–	√	–	85.87	75.23	87.03	8.06
√	–	–	84.48	73.78	84.11	7.60
√	√	–	84.45	75.96	85.18	20.10
√	√	√	**87.18**	**77.89**	**89.23**	**7.29**

Table 3: Ablation study of different variants on the FSS-Seg dataset. Constantly add key components to verify their actual effects.

4.4 Ablation Study

Table 3 presents the significant performance improvements achieved by each component and their synergy effects. The PPO strategy [Schulman *et al.*, 2017] improves the Dice coefficient, achieving a balance between filling in omissions and suppressing false detections. The PID module improves the Dice coefficient and the IoU by optimizing boundary delineation, but its effect in reducing false detections remains limited. When they operate independently without synergy, inconsistent optimization directions may lead to boundary jittering. However, the collaborative integration proposed in this paper resolves such conflicts, achieving the best performance and the lowest HD95 value in all three metrics, thus simultaneously optimizing regional characteristics and boundary features.

5 Conclusions

This study proposes a high-dimensional coupled optimization mechanism for SAM to address the joint optimization of prompt generation and segmentation. Through pointwise impact decomposition, the backbone decoder quantifies per-pixel contributions and constrains contribution maps via regularization. A hypothesis testing-based reward mechanism is further introduced for the reinforcement learning policy network to guide autonomous prompt optimization with fine-grained weight allocation, enabling weakly coupled optimization between the backbone and prompt modules. Experimental results on SAR semantic segmentation datasets demonstrate significant improvements in segmentation performance and cross-domain adaptability, providing an interpretable and learnable framework for automated prompt optimization.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62002146, and in part by the Youth Project of the Liaoning Provincial Department of Education under Grant LJ212410165083.

References

[Boykov and Jolly, 2001] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001.

[Carion *et al.*, 2025] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[Chen *et al.*, 2018] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[Chen *et al.*, 2022] 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.

[Chen *et al.*, 2023] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2304.09148*, 2023.

[Chen *et al.*, 2024a] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024.

[Chen *et al.*, 2024b] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024.

[Chen *et al.*, 2024c] Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4091, 2024.

[Chen *et al.*, 2025a] Tianrun Chen, Runlong Cao, Xinda Yu, Lanyun Zhu, Chaotao Ding, Deyi Ji, Cheng Chen, Qi Zhu, Chunyan Xu, Papa Mao, et al. Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation. *arXiv preprint arXiv:2511.19425*, 2025.

[Chen *et al.*, 2025b] Xinrun Chen, Chengliang Wang, Haojian Ning, Mengzhan Zhang, Mei Shen, and Shiying Li. Sam-octa2: Layer sequence octa segmentation with fine-tuned segment anything model 2. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Chen *et al.*, 2025c] Yi Chen, Muyoung Son, Chuanbo Hua, and Joo-Young Kim. Aop-sam: Automation of prompts for efficient segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2284–2292, 2025.

[Dai *et al.*, 2023] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023.

[Gong *et al.*, 2024] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024.

[Hou *et al.*, 2020] Xiyue Hou, Wei Ao, Qian Song, Jian Lai, Haipeng Wang, and Feng Xu. Fusar-ship: Building a high-resolution sar-ais matchup dataset of gaofen-3 for ship detection and recognition. *Science China Information Sciences*, 63(4):140303, 2020.

[Hu *et al.*, 2024] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12511–12518, 2024.

[Huang *et al.*, 2024] Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3205–3215, 2024.

[Huang *et al.*, 2025] Hanbin Huang, Hongliang He, Liying Xu, Xudong Zhu, Siwei Feng, and Guohong Fu. Ca-sam2: Sam2-based context-aware network with auto-prompting for nuclei instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 86–95. Springer, 2025.

[Kirillov *et al.*, 2023] 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.

[Konda and Tsitsiklis, 1999] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[Kweon and Yoon, 2024] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment any-

thing model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024.

[Li *et al.*, 2024] Yuchen Li, Li Zhang, Youwei Liang, and Pengtao Xie. Am-sam: Automated prompting and mask calibration for segment anything model. *arXiv preprint arXiv:2410.09714*, 2024.

[Liao *et al.*, 2020] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9394–9402, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2025a] Nanqing Liu, Xun Xu, Yongyi Su, Haojie Zhang, and Heng-Chao Li. Pointsam: Pointly-supervised segment anything model for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[Liu *et al.*, 2025b] Xueyu Liu, Rui Wang, Yexin Lai, Guangze Shi, Feixue Shao, Fang Hao, Jianan Zhang, Jia Shen, Yongfei Wu, and Wen Zheng. Plug-and-play ppo: An adaptive point prompt optimizer making sam greater. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4332–4342, 2025.

[Maninis *et al.*, 2018] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–625, 2018.

[Pu *et al.*, 2025] Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[Radman and Laaksonen, 2025] Abduljalil Radman and Jorma Laaksonen. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23947–23956, 2025.

[Ravi *et al.*, 2024] 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.

[Ren *et al.*, 2024] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[Rother *et al.*, 2004] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Wang *et al.*, 2024a] Yuqing Wang, Zhongling Huang, Shuxin Yang, Hao Tang, Xiaolan Qiu, Junwei Han, and Dingwen Zhang. Polsam: Polarimetric scattering mechanism informed segment anything model. *arXiv preprint arXiv:2412.12737*, 2024.

[Wang *et al.*, 2024b] Zhiting Wang, Qiangong Zhou, and Zongyang Liu. Det-sam2: Technical report on the self-prompting segmentation framework based on segment anything model 2. *arXiv preprint arXiv:2411.18977*, 2024.

[Wang *et al.*, 2025] Yuli Wang, Victoria Shi, Wen-Chi Hsu, Yuwei Dai, Sophie Yao, Zhusi Zhong, Zishu Zhang, Jing Wu, Aaron Maxwell, Scott Collins, et al. Optimizing prompt strategies for sam: advancing lesion segmentation across diverse medical imaging modalities. *Physics in Medicine & Biology*, 70(17):175014, 2025.

[Wei *et al.*, 2024] Zhikai Wei, Wenhui Dong, Peilin Zhou, Yuliang Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–543. Springer, 2024.

[Xiong *et al.*, 2026] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4(1):2, 2026.

[Yang *et al.*, 2023] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023.

[Zhang *et al.*, 2025a] Jie Zhang, Yunxin Li, Xubing Yang, Rui Jiang, and Li Zhang. Rsam-seg: A sam-based model with prior knowledge integration for remote sensing image semantic segmentation. *Remote Sensing*, 17(4):590, 2025.

[Zhang *et al.*, 2025b] Mengmeng Zhang, Xingyuan Dai, Yicheng Sun, Jing Wang, Yueyang Yao, Xiaoyan Gong, Fuze Cong, Feiyue Wang, and Yisheng Lv. Hierarchical self-prompting sam: A prompt-free medical image segmentation framework. *arXiv preprint arXiv:2506.02854*, 2025.

[Zhong *et al.*, 2024] Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. *arXiv preprint arXiv:2401.17868*, 2024.