IP-SAM: Rethinking Prompt-Conditioned Segmentation for Prompt-Absent Deployment

Huiyao Zhang[1,2], Jin Bai[1,2], Rui Guo[1,2], Jianwen Tan[1,2], Hongfei Wang[1,2], and Ye Li[1]✉

[1] Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
zhanghuiyao25@csu.ac.cn, baijin25@mails.ucas.ac.cn, liye@csu.ac.cn

Abstract. Prompt-conditioned foundation segmenters rely on explicit spatial prompts, such as points, boxes, or masks, to guide mask decoding. In prompt-absent deployment, however, no external/user prompts or detector-generated prompts are available at test time, creating a mismatch between the expected prompt-conditioned interface and automatic inference. We propose IP-SAM, a prompt-space adaptation framework that restores prompt-conditioned decoding without external prompts. Instead of bypassing SAM2's prompt interface through direct feature-space adaptation, IP-SAM synthesizes complementary intrinsic foreground/background prompts and routes their dense logits through SAM2's frozen prompt encoder, translating task-specific spatial cues into the native prompt embedding space. To reduce background leakage under severe camouflage, Prompt-Space Gating uses the intrinsic background prompt as an asymmetric suppressive constraint before decoding. Under a deterministic no-external-prompt protocol, IP-SAM achieves state-of-the-art performance across four COD benchmarks, including MAE 0.017 on COD10K, with only 21.26M trainable parameters from SPG, PSG, a SAM2-initialized downstream mask decoder, and image-encoder LoRA, while keeping the prompt encoder frozen. Medical polyp experiments in the main paper and supplementary SOD results further show that the same prompt-space adaptation route transfers within foreground–background segmentation.

Keywords: Prompt-Absent Segmentation · Prompt-Space Conditioning · SAM2 Adaptation · Camouflaged Object Detection

1 Introduction

Prompt-conditioned segmentation has emerged as a new paradigm for visual foundation models. Instead of directly predicting masks, these models rely on explicit spatial prompts—such as points, boxes, or masks—to guide the decoding process. Representative systems such as the Segment Anything Model (SAM) and its successor SAM2 [16, 28] demonstrate remarkable flexibility under interactive settings, enabling users to steer segmentation through simple spatial cues.

✉ Corresponding author.

However, this paradigm reveals a fundamental mismatch in fully automatic deployment. In many real-world scenarios, segmentation must operate without human interaction, leaving the model strictly prompt-absent at inference time. This creates a structural paradox: the decoder is designed to rely on prompt-conditioned signals, yet no prompts are available during deployment.

This issue becomes particularly evident in Camouflaged Object Detection (COD), where targets blend into surrounding textures [6]. Without disambiguating prompts, models frequently activate visually similar background regions, leading to unstable masks and severe background leakage. Although we evaluate on COD as a challenging stress-test benchmark, the prompt-absent deployment problem arises broadly across automatic segmentation tasks.

To address this challenge, most existing adaptations modify intermediate features by introducing adapters, lateral fusion modules, or task-specific decoders [4, 7]. While these feature-space modifications improve task alignment, they inadvertently bypass the native prompt interface that the architecture was originally designed to exploit. As a result, the prompt-conditioned decoding pathway remains weakly activated, leading to suboptimal adaptation. We argue that bypassing the prompt interface weakens the design assumption of prompt-conditioned segmenters and limits prompt-conditioned decoding.

To resolve this mismatch, we propose Intrinsic Prompting SAM (IP-SAM), which revisits adaptation from a prompt-space perspective through explicit prompt-space conditioning. Instead of injecting task information into intermediate features, IP-SAM synthesizes intrinsic prompts that activate the model through its native prompt interface. Concretely, a Self-Prompt Generator (SPG) distills aliased image contexts into complementary prompt candidates that serve as coarse regional anchors. These cues are projected through SAM2's frozen prompt encoder, translating task-specific signals into the model's native prompt manifold and restoring prompt-conditioned decoding.

Furthermore, camouflaged scenes exhibit severe foreground–background ambiguity. Traditional feature fusion tends to symmetrically mix foreground and background cues, often amplifying deceptive textures. To address this, we introduce Prompt-Space Gating (PSG), which uses the intrinsic background prompt as an asymmetric suppressive constraint. By filtering deceptive activations prior to decoding, PSG suppresses background-induced false positives through prompt-space interaction.

We evaluate IP-SAM under a deterministic no-external-prompt protocol across four COD benchmarks. With only 21.26M tunable parameters—optimizing SPG, PSG, and a task-specific mask decoder alongside image-encoder LoRA while keeping the prompt encoder strictly frozen—IP-SAM consistently outperforms heavy specialist COD models and recent SAM adaptation baselines. In essence, IP-SAM restores the missing prompt signal by synthesizing intrinsic prompts and conditioning SAM2 through its native prompt interface.

Contributions. (i) We formulate prompt-absent deployment as an interface mismatch in prompt-conditioned segmentation and introduce prompt-space conditioning, which restores SAM2's native prompt pathway without external prompts.

(ii) We develop IP-SAM with complementary intrinsic foreground/background prompts and Prompt-Space Gating (PSG), using background-derived prompt embeddings as asymmetric suppressive constraints before decoding. (iii) Under a deterministic no-external-prompt protocol, IP-SAM achieves state-of-the-art COD performance with 21.26M trainable parameters, and further demonstrates transferability within foreground–background segmentation through main-paper medical polyp experiments and supplementary SOD evaluations.

2 Related Work

2.1 Camouflaged Object Detection

Camouflaged Object Detection (COD) aims to segment objects intentionally concealed within their surroundings, a task plagued by extreme foreground-background aliasing [6, 25]. While early approaches relied on heavy task-specific designs (e.g., explicit boundary modeling, multi-scale aggregation) [6, 18, 37], recent Transformer-based models [15, 26, 35] have improved global structural modeling. Nevertheless, specialist COD models trained from scratch on limited datasets can struggle with domain generalization and diverse camouflage topologies. This motivates exploring large-scale pretrained foundation segmenters, which provide richer semantic representations and geometric priors.

2.2 Adapting SAM for Downstream Tasks

The Segment Anything Model (SAM) [16, 28] has inspired extensive domain adaptation efforts [4, 24]. To handle prompt-absent scenarios, prevalent methods rely heavily on feature-space adaptation, injecting structural patches—such as adapters or heavy decoding heads—into the backbone or decoder [3, 4, 7]. Recent SAM-based COD models largely follow this feature-centric paradigm [1, 7, 20, 36]. However, by predominantly modulating intermediate features, these methods often underutilize SAM's defining characteristic: its prompt-conditioned decoding interface. This raises a critical question: rather than relying primarily on intermediate feature modulation, how can we translate task requirements into SAM's native prompt manifold while preserving prompt-conditioned decoding?

2.3 Automatic Prompting and Parameter-Efficient Tuning

To automate SAM, recent works train auxiliary networks to generate explicit geometric cues (points, boxes) [2, 29, 34], explore prompt-free deployment via self-prompting and test-time calibration [5, 38], or leverage class activation maps and domain-adaptive prototypes for medical segmentation [33]. In COD, however, automatically generated explicit prompts can be brittle. Due to extreme visual aliasing, predicted boxes may misregister on deceptive distractors, injecting background-induced false positives into the decoder. Furthermore, while some prior work generates dense conceptual prompts [21], they typically rely on symmetric fusion without actively isolating background interference.

Our approach avoids this fragility. Rather than relying on discrete coordinate predictions, heavy feature-space addition, or test-time optimization, we translate complementary dense spatial priors into SAM's frozen prompt manifold. Crucially, we distinguish our method from prior self-prompting approaches by introducing Prompt-Space Gating (PSG) to preemptively filter noise via asymmetric suppression. Resolving semantic conflicts before decoding maximizes the effectiveness of Parameter-Efficient Fine-Tuning (PEFT) like LoRA [10]. Without reliable prompt conditions, PEFT can be forced to absorb much of the feature realignment burden. By reactivating the prompt manifold, IP-SAM provides a more direct adaptation route with efficient updates.

3 Method

3.1 Overview

To resolve the structural paradox of prompt-absent deployment, IP-SAM revisits adaptation by conditioning SAM2 through its native prompt space. Specifically, IP-SAM synthesizes task-aware prompts and feeds them through SAM2's native prompt interface. As illustrated in Fig. 1, a Self-Prompt Generator (SPG) synthesizes dense intrinsic prompts. These cues are projected through SAM2's frozen prompt encoder, structurally aligning the task with the prompt-conditioned decoding process used by our task-specific decoder. To reduce false positives, Prompt-Space Gating (PSG) (Fig. 2) imposes background-derived asymmetric suppressive constraints strictly prior to decoding.

3.2 Intrinsic Prompt Generation and Frozen Encoding

Image embedding. Given an input image I, the SAM2 image encoder $\mathcal{F}$ extracts high-dimensional visual semantics to produce an image embedding $E = \mathcal{F}(I) \in \mathbb{R}^{C\times H\times W}$.

Self-Prompt Generator (SPG). Instead of relying on heuristic pseudo-prompts (boxes/points) that are brittle under extreme camouflage, we learn dense mask-like priors as intrinsic prompts. We design a dual-branch Self-Prompt Generator $\mathcal{G}$ driven by learnable task queries $Q \in \mathbb{R}^{N\times d}$. Specifically, $E \in \mathbb{R}^{256\times 32\times 32}$ denotes the highest-semantic-level feature map produced by the FPN neck, at $\frac{1}{16}$ of the input resolution. We flatten E into a sequence of length $(HW)\times C$ before cross-attention. $\mathcal{G}$ employs a 2-layer TwoWayTransformer decoder (structurally identical to SAM's default transformer block, with $d = 256$, 8 heads, MLP dim 2048). Here, Q (comprising standard sparse prompt slots K and $K_m = 4$ learnable mask tokens) extracts complementary foreground and background semantics from the flattened E via cross-attention (*i.e.*, $\mathrm{Attn}(Q, E, E)$). Through a dedicated projection head (which upsamples the feature map by a factor of 4 using bilinear upsampling and a 3×3 convolution), SPG distills this aggregated context into complementary dense prompt logits ($P^+, P^- \in \mathbb{R}^{1\times 128\times 128}$) alongside continuous sparse prompt tokens (S^+, S^-). Notably, the positive branch

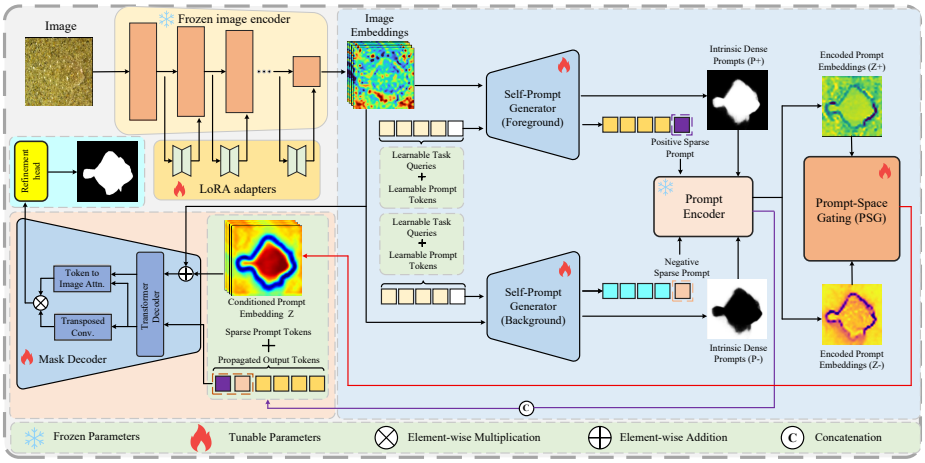

Fig. 1: Overall architecture of IP-SAM. Shifting the paradigm to prompt-space conditioning, IP-SAM utilizes a Self-Prompt Generator (SPG) to synthesize complementary intrinsic prompts ($P^{\pm}$). These are projected into SAM2's native manifold via the frozen prompt encoder. Prior to decoding, Prompt-Space Gating (PSG) leverages the negative embedding (Z^{-}) to asymmetrically suppress deceptive cues in the positive embedding (Z^{+}). The purified condition (Z) explicitly steers the mask decoder. For visual clarity, the optional auxiliary lateral inhibition module (Lat., ablated in Sec. 4.3) is omitted.

additionally yields a set of specialized propagated output tokens $\mathcal{T}_{\text{prop}}$ (depicted as the purple block in Fig. 1) to serve as strong semantic anchors:

$$\{P^{+}, S^{+}, \mathcal{T}_{\text{prop}}\}, \{P^{-}, S^{-}\} = \mathcal{G}(E, Q), \tag{1}$$

where $\mathcal{T}_{\text{prop}} \in \mathbb{R}^{K_m \times C}$ strictly matches the number ($K_m = 4$) and dimensionality ($C = 256$) of SAM2's default mask tokens. Given the extreme foreground-background aliasing in COD, these synthesized prompts inherently inherit positional ambiguity and structural noise from the camouflaged scene. They provide essential high-level spatial intents, while the mask decoder performs fine-grained disambiguation.

Manifold projection and sparse alignment. We feed the real-valued dense logits $P^{\pm}$ directly into SAM2's frozen native prompt encoder $\mathcal{E}$, yielding the encoded dense prompt embeddings $Z^{\pm} = \mathcal{E}(P^{\pm}) \in \mathbb{R}^{C \times H_e \times W_e}$. In practice, although native masks are binary, we intentionally feed continuous logits into the frozen prompt encoder without explicit binarization. Empirically, we found that binarizing the masks or applying temperature scaling consistently degrades performance, suggesting that continuous logits provide richer spatial uncertainty for prompt conditioning while remaining compatible with the encoder's embedding space. For sparse prompts, SPG bypasses the discrete coordinate-to-embedding step by directly outputting continuous token embeddings $S^{\pm}$. We use the same number of sparse prompt slots as SAM2 (denoted K) and add the corresponding pre-trained (positive/negative) point-type embeddings to form the encoded

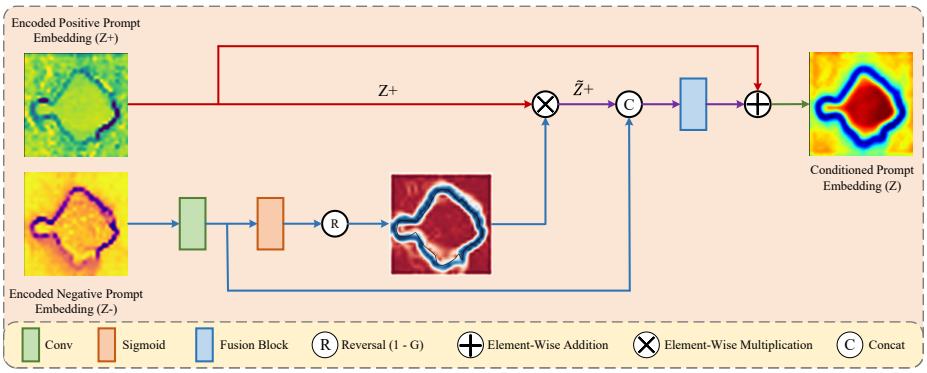

Fig. 2: Detailed schematic of Prompt-Space Gating (PSG). PSG suppresses background-induced false positives before they enter the decoder. It formulates a localized gate decision map from the negative prompt embedding (Z^-) to suppress deceptive cues in the positive counterpart (Z^+). The gated feature ($\widetilde{Z}^+$) is then fused with unactivated background cues and compensated via an anchored residual connection to yield the conditioned prompt embedding (Z).

sparse conditions $U^{\pm} \in \mathbb{R}^{K\times C}$. Keeping $\mathcal{E}$ strictly frozen translates task-specific cues into SAM2's native prompt embedding space used for decoding, formalizing our prompt-space conditioning.

3.3 Prompt-Space Gating (PSG)

In camouflaged scenes, standard symmetric feature fusion often amplifies deceptive textures. Although the prompt encoder can successfully embed the negative dense prior P^- into the native prompt space, directly providing the resulting dense background embedding to the decoder as a symmetric condition tends to amplify background activations under camouflage, leading to more false positives. Therefore, we use the negative dense branch exclusively to form a suppressive gate prior to decoder interaction.

First, the negative dense prompt embedding Z^- is transformed into an unactivated background cue $B = \phi_{\text{feat}}(Z^-)$ and a localized gate map G:

$$B = \phi_{\text{feat}}(Z^-), \quad G = \sigma(\phi_{\text{gate}}(B)), \quad \widetilde{Z}^+ = Z^+ \odot (1 - G). \tag{2}$$

Here, $(1-G)$ suppresses background-correlated activations in the positive dense prompt embedding Z^+, yielding the gated feature $\widetilde{Z}^+$.

To restore structural integrity while retaining useful background cues, we fuse $[\widetilde{Z}^+, B]$ through a fusion block ψ and apply an anchored residual correction:

$$Z = \psi([\widetilde{Z}^+, B]) + Z^+. \tag{3}$$

The fusion block maps the concatenated $2C$-channel feature back to C channels using a 3×3 conv, LayerNorm, GELU, and a 1×1 conv. The residual is anchored to the original Z^+ rather than $\widetilde{Z}^+$, since the suppressive gate may attenuate fine-grained foreground signals; anchoring to Z^+ preserves the positive prompt structure while allowing ψ to compensate gated prompt-space noise.

3.4 Decoding and Optional Enhancements

Prompt-conditioned decoding. To maintain compatibility with SAM2's prompt interface, we provide our task-specific mask decoder with a comprehensive sparse condition. As depicted in Fig. 1, we concatenate the continuous sparse prompt tokens from both branches to form a unified sparse condition $[U^+; U^-]$. Furthermore, the SAM2 decoder inherently relies on a set of standard learnable output tokens (an IoU token and multiple mask tokens). To fully leverage our SPG, we replace the default mask tokens with the propagated output tokens $\mathcal{T}_{\text{prop}}$ from the positive SPG branch, while retaining the standard IoU token (denoted as T_{iou}). This maintains the expected token sequence length while injecting robust semantic anchors. The task-specific mask decoder $\mathcal{D}$ is thus driven by the image embedding E, the condition-purified dense prompt embedding Z, the concatenated sparse tokens $[U^+; U^-]$, and the customized output tokens $[T_{\text{iou}}; \mathcal{T}_{\text{prop}}]$, yielding a robust coarse mask:

$$M_c = \mathcal{D}\big(E, Z, [U^+; U^-], [T_{\text{iou}}; \mathcal{T}_{\text{prop}}]\big). \tag{4}$$

This achieves robust decoding under camouflage while preserving the native conditioning interface.

Optional enhancements. Optionally, we add two lightweight components (ablated in Sec. 4.3): (i) an auxiliary lateral inhibition module (Lat.) using the detached coarse mask M_c to form a soft gate (*e.g.*, $\sigma(M_c)$) suppressing multi-scale FPN features; and (ii) a pixel refinement head $\mathcal{R}$ (fusing the coarse mask, the conditioned prompt embedding, and 256-channel image features via a 64-channel bottleneck) to predict a residual correction $M = M_c + \mathcal{R}(\cdot)$ via a zero-initialized 1×1 conv for fine boundary alignment.

3.5 Training Objective and Efficient Adaptation

Losses & Adaptation. Let $Y \in \{0,1\}^{1\times H_0 \times W_0}$ be the ground truth. We supervise SPG with complementary targets: $\mathcal{L}_{\text{SPG}} = \text{BCE}(\sigma(P^+), Y) + \text{BCE}(\sigma(P^-), 1 - Y)$. Although SPG is supervised to learn mask-like priors, the final prediction is produced by a task-specific mask decoder conditioned on their prompt-space embeddings, rather than directly from $P^{\pm}$. The dense mask loss $\mathcal{L}_{\text{mask}}$ combines standard BCE, IoU, and L1 components. The overall objective is $\mathcal{L} = \lambda_{\text{spg}}\mathcal{L}_{\text{SPG}} + \lambda_{\text{c}}\mathcal{L}_{\text{mask}}(M_c, Y) + \lambda_{\text{r}}\mathcal{L}_{\text{mask}}(M, Y)$, where $\lambda_{\text{r}} = 0$ when refinement is disabled. We utilize LoRA on the qkv and output projections of the SAM2 image encoder across all Transformer blocks to enable lightweight domain adaptation. Importantly, the prompt encoder $\mathcal{E}$ remains strictly frozen to preserve the pre-trained prompt embedding manifold. A lightweight task-specific mask decoder follows the architectural design of SAM2's mask decoder and is initialized from shape-compatible SAM2 mask-decoder weights when available. We additionally evaluate a true-scratch decoder variant to isolate the effect of decoder initialization. Overall, this results in only 21.26M trainable parameters.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We evaluate on four standard COD benchmarks: CAMO [17], CHAMELEON [30], COD10K [6], and NC4K [23]. Following [8,9,11–13], we train on the COD10K and CAMO training splits, and test on their respective test sets plus the full CHAMELEON and NC4K datasets. Performance is assessed using Mean Absolute Error (MAE) $\downarrow$, weighted F-measure (F^{ω}) $\uparrow$, Structure-measure (S_m) $\uparrow$, and Enhanced-alignment measure (E_{ϕ}) $\uparrow$.

Implementation & evaluation. IP-SAM is built upon the SAM2-L (Hiera-Large) image encoder. Images are resized to 512×512. We train end-to-end for 100 epochs using AdamW [22] (LR 1×10^{-4}, batch size 1) on a single RTX 3090. Loss weights are fixed at $\lambda_{\text{spg}} = \lambda_{\text{c}} = \lambda_{\text{r}} = 1.0$. For fair comparison, all trainable specialist and SAM-based adaptation baselines are re-trained under identical data and resolution settings, while zero-shot diagnostic baselines such as SAM/SAM2, Oracle, and SAM2-AMG are evaluated under the same resolution and deterministic protocol. During evaluation, we strictly adhere to a deterministic prompt-absent protocol: standard SAM/SAM2 baselines use empty tokens and zero-mask inputs, whereas IP-SAM synthesizes intrinsic conditions ($U^{\pm}, Z^{\pm}$). For SAM2-AMG, we select the top-1 mask via its native `predicted_iou`, since top-k selection requires ground-truth knowledge.

Parameter-efficient tuning. To isolate the effect of our prompt-space conditioning, we keep the prompt encoder strictly frozen. Param and T-Param denote the total and trainable parameters of the inference model, respectively, reported in millions (M). Tunable parameters include only the lightweight SPG/PSG modules, a task-specific mask decoder, and LoRA (rank $r = 32$) applied to the attention projections of the SAM2 image encoder.

4.2 Comparisons with State-of-the-Art Methods

Quantitative superiority. Table 1 shows IP-SAM sets a new state-of-the-art in the prompt-absent setting. On the heavily cluttered CAMO dataset, IP-SAM achieves MAE 0.032 and S_m 0.912, outperforming both the strongest specialist baseline (ZoomNeXt, MAE 0.041, S_m 0.892) and the strongest feature-adapted SAM2-L baseline (MAE 0.037, S_m 0.900). IP-SAM also substantially outperforms SAM2-AMG. On CHAMELEON, SAM2-AMG trails the naive null-prompt baseline (MAE 0.500 vs. 0.182); lacking confidence on concealed targets, AMG's top-1 proposal often gravitates to well-defined background regions, inflating errors.

Discussion & Generalization. Compared to specialist models, IP-SAM consistently reduces errors without human prompts. To probe the zero-shot prompting capacity of SAM-style models, we include an Oracle baseline explicitly prompted with a ground-truth centroid. Although this Oracle is not prompt-absent, it provides a diagnostic reference for single-point prompting under severe camouflage. IP-SAM outperforms the zero-shot SAM2-Large Oracle on COD10K

Table 1: Quantitative comparison (no-external-prompt). The default IP-SAM configuration is highlighted in gray.

Method	Param (M)	T-Param (M)	CAMO				CHAMELEON				COD10K				NC4K			
			MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑	MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑	MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑	MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑
Specialist COD Methods (CNN & Transformer-based)																		
FSPNet [14]	274.2	274.2	.060	.761	.864	.880	.036	.766	.887	.889	.034	.671	.843	.858	.043	.772	.878	.891
ZoomNet [26]	33.4	33.4	.070	.731	.809	.856	.030	.812	.881	.928	.031	.719	.833	.882	.045	.774	.847	.887
DGNet [15]	8.30	8.30	.057	.769	.839	.901	.029	.816	.890	.934	.033	.693	.822	.877	.042	.784	.857	.907
CamoFormer [35]	71.3	71.3	.048	.824	.868	.923	.022	.860	.901	.942	.023	.777	.868	.928	.031	.841	.888	.935
ZoomNeXt [27]	28.46	28.46	.041	.857	.892	.935	.020	.865	.903	.941	.018	.834	.898	.939	.030	.859	.902	.932
SAM-based Adaptation Methods																		
SAM (Oracle, 1-pt GT)*	641.08	0	.201	.460	.580	.574	.221	.484	.588	.582	.083	.583	.710	.714	.135	.579	.678	.683
SAM2 (Oracle, 1-pt GT)*	224.51	0	.086	.714	.771	.802	.041	.814	.848	.888	.030	.794	.850	.888	.046	.811	.849	.883
SAM [16]	641.08	0	.251	.105	.413	.329	.196	.077	.433	.326	.157	.056	.449	.344	.239	.096	.426	.340
SAM2 [28]	224.51	0	.198	.348	.658	.506	.182	.311	.660	.512	.132	.287	.669	.531	.159	.393	.722	.572
YOLOv8+SAM2(Auto)	268.1	0	.219	.134	.450	.372	.464	.154	.422	.323	.318	.112	.463	.400	.387	.175	.461	.375
SAM2-AMG [28]	224.51	0	.329	.115	.358	.443	.500	.113	.283	.327	.262	.151	.424	.522	.302	.207	.423	.507
SAM-Adapter [4]	641.27	4.25	.072	.763	.836	.858	.034	.801	.878	.891	.025	.797	.876	.904	.043	.810	.875	.891
MedSAM [24]	93.74	93.73	.069	.757	.847	.873	.036	.800	.895	.918	.037	.718	.854	.892	.047	.791	.878	.905
MDSAM [7]	101.06	14.57	.061	.782	.839	.885	.027	.838	.891	.942	.030	.752	.846	.904	.042	.811	.866	.908
SAM2-Adapter [3]	224.51	3.94	.048	.817	.884	.915	.025	.847	.909	.945	.020	.814	.891	.936	.030	.851	.906	.934
SAM2(SAM2-L, r80) [3]	234.79	18.96	.037	.862	.900	.935	.022	.857	.908	.942	.020	.817	.890	.933	.029	.860	.904	.936
IP-SAM Variants (Ours)																		
IP-SAM (SAM2-L, FT)	231.86	231.86	.040	.862	.900	.933	**.017**	**.894**	**.929**	**.969**	.018	.834	.901	.942	.030	.858	.901	.935
IP-SAM (SAM2-T, r32)	45.53	16.92	.148	.415	.632	.649	.097	.477	.705	.728	.080	.407	.674	.711	.105	.520	.711	.729
IP-SAM (SAM2-B, r32)	89.07	18.07	.093	.644	.770	.795	.049	.708	.835	.876	.045	.626	.790	.839	.060	.721	.826	.858
IP-SAM (SAM2-L, r16)	234.47	18.65	.040	.851	.896	.929	.022	.850	.907	.944	.020	.821	.894	.939	.029	.863	.907	.941
IP-SAM (SAM2-L, r32)	237.08	21.26	**.032**	**.878**	**.912**	**.946**	.018	.874	.920	.956	**.017**	**.838**	**.906**	**.947**	**.026**	**.873**	**.911**	**.943**
IP-SAM (SAM2-L, r64)	242.30	26.48	.037	.867	.904	.938	.019	.874	.919	.956	.017	.841	.904	.947	.027	.871	.912	.943

Notes. Param/T-Param in millions (M). *Oracle uses a ground-truth centroid point to show theoretical capacity. IP-SAM metrics are averaged over 3 seeds. **FT**: Full fine-tuning. **T/B/L**: Hiera variants. **AMG**: Auto Mask Generator (top-1 via predicted IoU).

(MAE 0.017 vs. 0.030), suggesting that a single centroid point can be insufficient when foreground and background are heavily aliased. By synthesizing dense intrinsic prompts with PSG, IP-SAM provides richer prompt-space conditions for resolving such ambiguity.

As an additional auto-prompting diagnostic, YOLOv8+SAM2(Auto) frequently collapses under camouflage, with 35.4% full-image fallbacks on COD10K and a high MAE of 0.318. Compared with controlled feature-space adaptations, IP-SAM shows larger gains on structure-sensitive metrics (S_m, E_ϕ), supporting the benefit of suppressing deceptive cues before decoding. IP-SAM also maintains strong cross-dataset performance on CHAMELEON and NC4K. Efficiency-wise, it runs at 29 FPS compared with 35 FPS for base SAM2, introducing a moderate overhead for prompt-space conditioning. Finally, the null-prompt baseline in Table 1 and the controlled diagnostics in Table 1 indicate that the SOTA gains are primarily associated with prompt-space conditioning rather than decoder redesign alone. Across three random seeds, IP-SAM achieves a stable COD10K MAE of 0.017 ± 0.001.

4.3 Ablation Study

Setup and core components. We systematically evaluate each module at 512×512. Our baseline (Tab. 2, row 1) trains only the image-encoder LoRA ($r = 32$) and our task-specific mask decoder (T-Param=9.15M in total), utilizing a frozen prompt encoder with null prompts (empty sparse tokens and a zero-mask input) for determinism. Tab. 2 supports our architectural evolution: (i) integrating SPG improves the null-prompt baseline, suggesting that dense self-prompts reactivate the prompt-conditioned decoding pathway; (ii) introducing

PSG yields the largest stepwise error reduction on CAMO, reducing MAE from 0.041 to 0.034, indicating that asymmetric suppression is crucial under severe camouflage; and (iii) Lat. and Ref. provide complementary refinements, where Lat. further reduces CAMO MAE to 0.032 and the refinement head mainly improves structure-sensitive metrics and boundary fidelity.

Controlled alternatives. Table 2 verifies the incremental contribution of SPG, PSG, Lat., and Ref. within IP-SAM. We further ask whether the observed gains can be reduced to stronger image-feature decoding, latent prompt bypassing, explicit learned auto-prompting, or decoder initialization. To this end, Table 3 reports controlled diagnostics under matched data, resolution, loss, and evaluation settings. These controls remain below IP-SAM, indicating that the gain is primarily associated with routing continuous intrinsic logits through SAM2's frozen prompt encoder, followed by asymmetric prompt-space suppression, rather than with any single alternative factor alone.

Key design choices. Tab. 4 supports our core design philosophy. **(1) Conditioning paradigm & SPG capacity:** To isolate the contribution of Prompt-Space Conditioning (PSC) from the SPG's parameter capacity, we introduce a Minimal-SPG baseline by replacing the TwoWayTransformer with a lightweight 3-layer CNN. When used directly as a segmentation head (Minimal-SPG (Direct)), the resulting coarse logits yield a severely degraded MAE of 0.051 on COD10K. Passing these logits through the prompt-space conditioning pipeline (Minimal-SPG + PromptSpace) reduces the MAE to 0.040, indicating that reac-

Table 2: Ablation on core components. Incrementally adding SPG, PSG, Lat., and Ref. to the baseline.

SPG	PSG	Lat.	Ref.	T-Param (M)	COD10K				CAMO			
					MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑	MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑
✗	✗	✗	✗	9.15	0.024	0.788	0.857	0.902	0.043	0.818	0.859	0.893
✓	✗	✗	✗	17.00	0.022	0.802	0.872	0.916	0.041	0.826	0.875	0.915
✓	✓	✗	✗	18.98	0.019	0.826	0.893	0.938	0.034	0.868	0.902	0.938
✓	✓	✓	✗	19.13	0.018	0.835	0.899	0.941	0.032	0.871	0.907	0.942
✓	✓	✓	✓	**21.26**	**0.017**	**0.838**	**0.906**	**0.947**	**0.032**	**0.878**	**0.912**	**0.946**

Notes. Baseline (row 1) trains only the image-encoder LoRA and our task-specific mask decoder with null prompts. ✓ indicates the module is enabled.

Table 3: Controlled diagnostics against alternative explanations. All variants use matched training and evaluation settings.

Method	T-Param	Control	COD10K	CAMO
Direct-Head++	19.46M	Strong image-feature head	.019/.824	.039/.857
Latent-Prompt++	22.68M	Latent prompt bypass	.019/.831	.037/.867
AutoPrompt++ (B/P/M)	18.20M	Learned explicit prompts	.018/.831	.036/.866
True-scratch dec.	21.26M	No SAM2 decoder init.	.019/.824	.036/.863
Core w/o Lat./Ref.	18.98M	Core SPG+PSG only	.019/.826	.034/.868
IP-SAM	**21.26M**	**Full model**	**.017/.838**	**.032/.878**

Notes. Metrics are MAE↓/F^ω ↑. Direct-Head++ replaces prompt-space conditioning with a stronger image-feature-only head. Latent-Prompt++ bypasses the frozen prompt encoder with a higher-capacity latent embedding mapper. AutoPrompt++ predicts box/point/mask prompts and feeds them into SAM2 under the same prompt-absent protocol. Core w/o Lat./Ref. removes the optional lateral inhibition and refinement modules.

tivating the native prompt-conditioned decoding pathway itself provides a clear performance benefit. Scaling the generator back to the full SPG further clarifies the division of labor: using the full-capacity SPG directly as a predictor (Full-SPG (Direct)) achieves MAE 0.025, while enabling prompt-space conditioning with the prompt-conditioned decoder yields our best result of 0.017. These results suggest a complementary relationship where SPG extracts coarse regional anchors, while fine-grained disambiguation is ultimately performed by the prompt-conditioned decoder operating in the frozen prompt embedding space. Conditioning strictly through the native prompt manifold (PromptSpace) also consistently outperforms conventional feature-space injection (FeatAdd), and unfreezing the prompt encoder degrades performance, confirming that keeping it strictly frozen helps preserve pre-trained geometric priors against camouflage noise.

(2) PSG operator and residual anchoring: We compare our asymmetric gating mechanism with several standard feature interaction strategies. Asymmetric gating consistently outperforms symmetric fusion methods such as concatenation and subtraction, which tend to amplify both foreground and background signals in highly deceptive scenes. We further analyze residual anchoring: attaching the residual to the purified $\widetilde{Z}^+$ degrades performance compared with anchoring to the original Z^+ (AsymGate). This supports our design choice that residual learning in prompt space should primarily function as a noise cancel-

Table 4: Ablation on design choices. (I) Conditioning paradigm and SPG capacity, (II) PSG operator and residual anchoring, (III) PSG and token replacement, and (IV) LoRA injection strategy.

Setting	T-Param (M)	COD10K				CAMO			
		MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑	MAE↓	F^ω ↑	S_m ↑	E_ϕ ↑
(I) Conditioning paradigm & SPG Capacity									
Minimal-SPG (Direct)	13.85	0.051	0.802	0.887	0.896	0.058	0.727	0.873	0.885
Minimal-SPG + PromptSpace	13.85	0.040	0.851	0.894	0.928	0.049	0.818	0.893	0.936
Full-SPG (Direct)	21.26	0.025	0.765	0.883	0.899	0.043	0.829	0.897	0.913
FeatAdd	19.02	0.020	0.825	0.896	0.940	0.039	0.864	0.904	0.936
Trainable Prompt Encoder	21.26	0.019	0.820	0.891	0.938	0.039	0.854	0.895	0.932
Full-SPG + PromptSpace (Ours)	**21.26**	**0.017**	**0.838**	**0.906**	**0.947**	**0.032**	**0.878**	**0.912**	**0.946**
(II) PSG operator and residual anchoring									
None (no interaction)	19.27	0.020	0.818	0.877	0.918	0.040	0.830	0.881	0.910
Subtraction ($Z^+ - Z^-$)	19.93	0.019	0.829	0.898	0.941	0.038	0.863	0.902	0.936
Concatenation	20.37	0.019	0.823	0.893	0.934	0.038	0.851	0.895	0.933
CrossAttn Fusion	19.53	0.019	0.828	0.899	0.942	0.037	0.867	0.904	0.938
Anchor to $\widetilde{Z}^+$	21.26	0.019	0.820	0.891	0.935	0.039	0.852	0.894	0.927
AsymGate (Anchor Z^+)	**21.26**	**0.017**	**0.838**	**0.906**	**0.947**	**0.032**	**0.878**	**0.912**	**0.946**
(III) PSG and token replacement									
w/o PSG (Refine kept)	19.11	0.021	0.816	0.887	0.931	0.040	0.846	0.891	0.922
w/ PSG + default mask tokens	21.26	0.019	0.828	0.898	0.940	0.039	0.859	0.900	0.933
Full IP-SAM	**21.26**	**0.017**	**0.838**	**0.906**	**0.947**	**0.032**	**0.878**	**0.912**	**0.946**
(IV) LoRA injection strategy									
qkv only	19.51	0.019	0.821	0.895	0.937	0.038	0.859	0.900	0.935
Deep stages	20.87	0.034	0.689	0.822	0.873	0.066	0.734	0.825	0.853
Shallow stages	16.42	0.024	0.772	0.869	0.915	0.057	0.785	0.857	0.886
Sparse (50%)	18.63	0.019	0.820	0.894	0.937	0.039	0.859	0.899	0.932
Full (ours)	**21.26**	**0.017**	**0.838**	**0.906**	**0.947**	**0.032**	**0.878**	**0.912**	**0.946**

Notes. *Direct* denotes directly using the SPG positive logit as the final mask, bypassing the prompt-conditioned decoding pipeline (i.e., prompt encoder, PSG, and mask decoder). `FeatAdd` adds dense prompts to image embeddings. `None` uses only Z^+, discarding Z^-. Detailed settings follow Sec. 4.3.

lation signal. By anchoring to the uncorrupted Z^+, the fusion block ψ focuses on suppressing isolated noise rather than reconstructing damaged foreground features, simplifying optimization while preserving fine structural details.

(3) PSG vs. token replacement: Removing PSG (while retaining the refinement module) noticeably degrades performance, indicating that PSG provides essential preemptive disambiguation that cannot be replicated by post-hoc refinement. To isolate the role of dynamic token replacement, we retain PSG but revert to the default mask tokens of the decoder. Although propagated tokens $\mathcal{T}_{\text{prop}}$ provide additional improvements, the model without them still substantially outperforms the baselines, confirming that PSG remains the primary contributor.

(4) LoRA injection strategy: Finally, we evaluate several LoRA injection strategies. Distributing LoRA across all attention blocks (Full) yields the best balance between parameter efficiency and adaptation capacity compared with partial injection schemes such as shallow-only, deep-only, or sparse insertion.

4.4 Qualitative Results

Robust Ambiguity Resolution. As illustrated in Fig. 3, the dominant failure mode in prompt-absent COD is background-induced false positives with eroded boundaries. Without explicit cues, baseline adaptations often fail to separate targets from homologous clutter. By contrast, IP-SAM yields substantially cleaner predictions. By using intrinsic background prompts as asymmetric suppressive constraints, our PSG mechanism suppresses deceptive cues before decoder interaction. Concurrently, strict alignment with SAM2's native prompt embedding space allows IP-SAM to fully exploit SAM2's pre-trained geometric priors, better preserving delicate, low-contrast topologies (*e.g.*, thin insect antennas) even under extreme visual ambiguity.

Visualizing the Prompt-Space Gating Mechanism. To visualize the internal behavior of prompt-space conditioning, we show feature evolution in the PSG module in Fig. 4. Without explicit prompts, the baseline model (Fig. 4, Col 2) suffers severe feature spillover, activating homologous background regions that closely resemble the target appearance. With the Self-Prompt Generator (SPG), IP-SAM disentangles the ambiguous context into two complementary representations: the foreground prompt (P^+, Col 5) anchoring the primary target structure and the background prompt (P^-, Col 6) capturing deceptive distractors. The Gate Decision Map (Col 7) illustrates asymmetric gating: red regions remain unaffected, while dark blue regions denote suppressive constraints derived from the encoded background prompt Z^-. Rather than symmetrically fusing P^+ and P^-, the network uses Z^- to form a localized suppressive boundary (the dark blue "isolation ring" in Col 7) that encloses the camouflaged target and isolates it from surrounding clutter. This boundary suppresses deceptive leakage before decoding, corroborated by the feature energy change map (Col 8, where warmer colors indicate larger adjustments). Consequently, IP-SAM reduces background-induced false positives (Col 4), allowing target-specific prompt embeddings to guide the final segmentation (Fig. 4, Col 3).

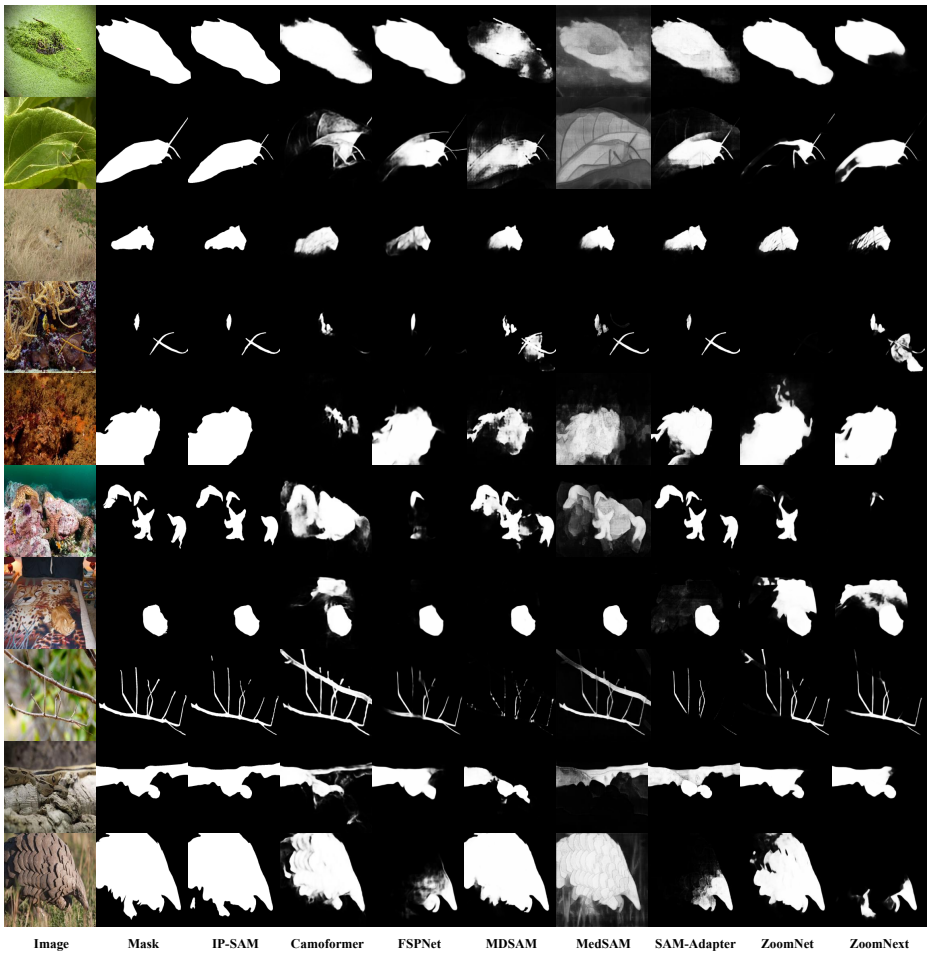

Fig. 3: Qualitative comparison on challenging camouflaged scenarios. Visual predictions of our IP-SAM against representative specialist COD models and SAM-based adaptations under the prompt-absent setting.

4.5 Cross-Dataset Generalization

Cross-Dataset Medical Segmentation. To demonstrate the generality of prompt-space conditioning beyond camouflaged object detection, we evaluate IP-SAM on medical polyp segmentation under the same prompt-absent protocol. The model is trained solely on the Kvasir-SEG training split using the same hyperparameter configuration as in the COD experiments, without introducing domain-specific augmentation or task-specific heuristics.

As shown in Table 5, IP-SAM achieves the best in-domain performance on Kvasir-SEG and matches or outperforms representative SAM-based adaptations such as MDSAM and SAM2-Adapter under the same training protocol. On CVC-ClinicDB, IP-SAM achieves the highest mDice (0.840), mIoU (0.778), and S_m

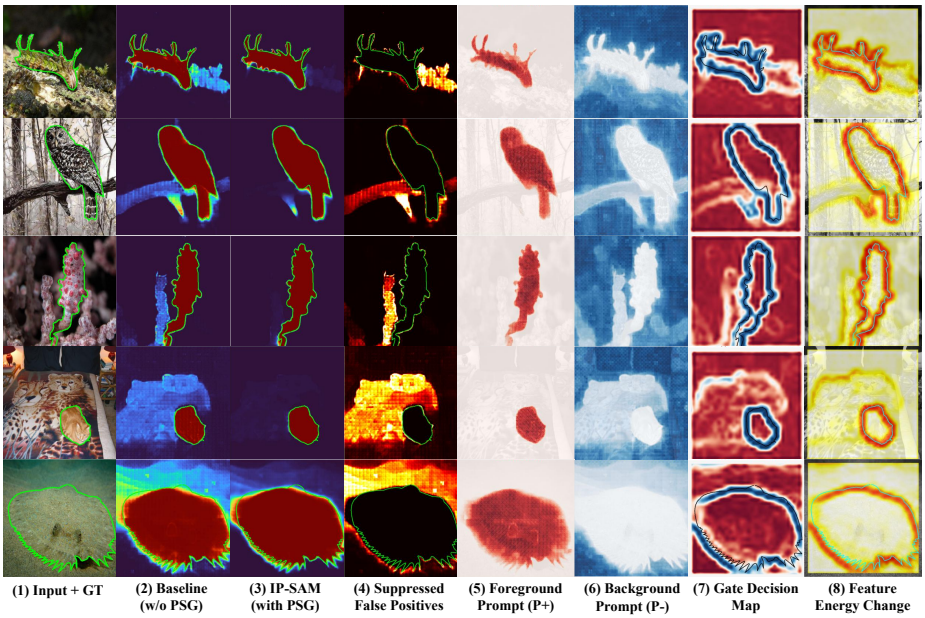

Fig. 4: Internal feature visualization of the Prompt-Space Gating (PSG) mechanism. Feature evolution from the baseline's leakage (Col 2) to IP-SAM's clean prediction (Col 3). Intermediate columns depict suppressed false positives (Col 4), complementary intrinsic prompts (Cols 5-6), gate decision map (Col 7), and feature energy adjustments (Col 8).

(0.897), surpassing representative specialist models such as LACFormer on overlap and structure metrics while remaining competitive in MAE. On the more challenging ETIS dataset, which exhibits notable domain shifts due to varying endoscopic equipment and smaller polyp scales, IP-SAM achieves the best SAM-based performance on mDice, mIoU, and S_m, with a tied best SAM-based MAE.

Table 5: Cross-dataset medical polyp segmentation. Models are trained on Kvasir-SEG and evaluated on CVC-ClinicDB and ETIS. Best overall results are underlined and best SAM-based results are boldfaced; ties share the same mark.

Method	**Kvasir-SEG** (In-domain)				**CVC-ClinicDB**				**ETIS**			
	mDice↑	mIoU↑	S_m ↑	MAE↓	mDice↑	mIoU↑	S_m ↑	MAE↓	mDice↑	mIoU↑	S_m ↑	MAE↓
SEPNet [32]	.896	.839	.917	.028	.815	.744	.874	.028	.741	.663	.847	.018
LACFormer [31]	.906	.851	.922	.029	.826	.758	.889	.026	.798	.722	.884	.016
ASPS [19]	.788	.704	.844	.063	.670	.580	.786	.069	.362	.313	.660	.039
ZoomNeXt [27]	.892	.831	.909	.031	.805	.731	.863	.031	.693	.625	.841	.023
SAM-Adapter [4]	.624	.514	.724	.117	.450	.366	.671	.093	.223	.189	.596	.070
MDSAM [7]	.906	.851	.922	.028	.796	.724	.863	.034	.753	.692	.858	**.021**
SAM2-Adapter [3]	.899	.846	.918	.031	.801	.738	.873	.030	.733	.663	.847	.026
IP-SAM (Ours)	**.913**	**.861**	**.929**	**.025**	**.840**	**.778**	**.897**	**.029**	**.764**	**.696**	**.868**	**.021**

Notes. All SAM-based adaptations and specialist models were re-trained by us on the Kvasir-SEG training split and evaluated under the identical strictly prompt-absent protocol.

These results suggest that conditioning through SAM2's frozen prompt interface provides a stable adaptation pathway for prompt-absent deployment, enabling the model to better exploit the foundation model's pre-trained geometric priors without relying solely on task-specific feature modifications.

5 Conclusion and Limitations

We address the structural mismatch between prompt-conditioned segmenters and prompt-absent deployment. Rather than treating automatic segmentation as a feature-space adaptation problem, IP-SAM shows that task-specific spatial cues can be translated back into SAM2's native prompt manifold. By synthesizing intrinsic foreground/background prompts, encoding their dense logits with the frozen prompt encoder, and applying PSG before decoding, IP-SAM restores prompt-conditioned inference without external prompts. Under a deterministic no-external-prompt protocol, IP-SAM establishes state-of-the-art COD performance with only 21.26M trainable parameters, while medical polyp experiments in the main paper and SOD experiments in the Supplementary Material indicate that the same prompt-space adaptation route transfers within foreground–background segmentation.

The current evidence is intentionally centered on prompt-absent foreground–background segmentation. Extending this paradigm to COCO-style multi-instance segmentation, semantic segmentation, and open-vocabulary instance parsing remains an important direction. IP-SAM also depends on foundation-model capacity, as smaller SAM2-T/B variants weaken prompt separability. In extremely ambiguous cases, P^- may encroach on weak target regions and cause PSG oversuppression. These operating boundaries point to future work on uncertainty-aware prompt generation and instance-aware prompt-space conditioning.

Acknowledgments

This study was supported in part by National Natural Science Foundation of China under Grants 61971404 and 61901454; in part by the Project of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences under Grant T503471; in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant 2019168; and in part by the Project of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences under Grant CSU-JJKT-2023-4.

References

1. Chen, H., Wei, P., Guo, G., Gao, S.: Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection. IEEE Transactions on Circuits and Systems for Video Technology **35**(5), 4635–4647 (2024)
2. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024)
3. Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2408.04579 (2024)
4. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3367–3375 (2023)
5. Chen, Y., Son, M., Hua, C., Kim, J.Y.: Aop-sam: Automation of prompts for efficient segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2284–2292 (2025)
6. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)
7. Gao, S., Zhang, P., Yan, T., Lu, H.: Multi-scale and detail-enhanced segment anything model for salient object detection. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 9894–9903 (2024)
8. He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. Advances in Neural Information Processing Systems **36**, 30726–30737 (2023)
9. He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 781–789 (2023)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
11. Hu, J., Cheng, Z., Gong, S.: Int: Instance-specific negative mining for task-generic promptable segmentation. arXiv preprint arXiv:2501.18753 (2025)
12. Hu, J., Lin, J., Gong, S., Cai, W.: Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12511–12518 (2024)
13. Hu, J., Lin, J., Yan, J., Gong, S.: Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. Advances in Neural Information Processing Systems **37**, 107171–107197 (2024)
14. Huang, Z., Dai, H., Xiang, T.Z., Wang, S., Chen, H.X., Qin, J., Xiong, H.: Feature shrinkage pyramid for camouflaged object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5557–5566 (2023)
15. Ji, G.P., Fan, D.P., Chou, Y.C., Dai, D., Liniger, A., Van Gool, L.: Deep gradient learning for efficient camouflaged object detection. Machine Intelligence Research **20**(1), 92–108 (2023)

16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
17. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer vision and image understanding **184**, 45–56 (2019)
18. Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., Dai, Y.: Uncertainty-aware joint salient object and camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10071–10081 (2021)
19. Li, H., Zhang, D., Yao, J., Han, L., Li, Z., Han, J.: Asps: Augmented segment anything model for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 118–128. Springer (2024)
20. Liu, J., Kong, L., Chen, G.: Improving sam for camouflaged object detection via dual stream adapters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21906–21916 (2025)
21. Liu, L., Chang, J., Yu, B.X., Lin, L., Tian, Q., Chen, C.W.: Prompt-matched semantic segmentation. arXiv preprint arXiv:2208.10159 (2022)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11591–11601 (2021)
24. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
25. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8772–8781 (2021)
26. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 2160–2170 (2022)
27. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoomnext: A unified collaborative pyramid network for camouflaged object detection. IEEE transactions on pattern analysis and machine intelligence **46**(12), 9205–9220 (2024)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
29. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
30. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished manuscript **2**(6), 7 (2018)
31. Van Nguyen, Q., Nguyen, M., Nguyen, T.T., Quang, H.T., Van, T.P., Bao, L.D.: Lacformer: Toward accurate and efficient polyp segmentation. In: BMVC. pp. 411–416 (2023)
32. Wang, T., Qi, X., Yang, G.: Polyp segmentation via semantic enhanced perceptual network. IEEE Transactions on Circuits and Systems for Video Technology **34**(12), 12594–12607 (2024)

33. Wei, Z., Dong, W., Zhou, P., Gu, Y., Zhao, Z., Xu, Y.: Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15008, pp. 533–543. Springer Nature Switzerland (2024)
34. Xie, Z., Guan, B., Jiang, W., Yi, M., Ding, Y., Lu, H., Zhang, L.: Pa-sam: Prompt adapter sam for high-quality image segmentation. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
35. Yin, B., Zhang, X., Fan, D.P., Jiao, S., Cheng, M.M., Van Gool, L., Hou, Q.: Camoformer: Masked separable attention for camouflaged object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10362–10374 (2024)
36. Yuan, C., Liu, L., Li, Y., Li, J.: Sam2-dfbcnet: A camouflaged object detection network based on the heira architecture of sam2. Sensors **25**(14), 4509 (2025)
37. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8582–8591 (2020)
38. Zhang, M., Dai, X., Sun, Y., Wang, J., Yao, Y., Gong, X., Cong, F., Wang, F., Lv, Y.: Hierarchical self-prompting sam: A prompt-free medical image segmentation framework. arXiv preprint arXiv:2506.02854 (2025)