RCP: RESIDUAL-COMPLEMENTARY PROMPTING FOR SPECIALIST-GUIDED BINARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation segmentation models can complement task-specific specialists, but their predictions may be unreliable or redundant. We study this collaboration in binary polyp and camouflaged-object segmentation using a DINOv3-B specialist and frozen SAM3. On intervention-selected polyp cases, direct SAM3 replacement reduces Dice by 2.0 points, although a ground-truth-guided image-level choice between the specialist and direct SAM3 shows a potential 5.2-point gain. We propose Residual-Complementary Prompting (RCP), which generates SAM3 hypotheses that correct distinct specialist errors. A spatial arbiter combines the hypotheses with the specialist prediction, and a failure-aware module selects inputs for intervention. RCP improves Dice by 5.0 points over the specialist on intervention-selected cases and consistently improves performance across five polyp and three camouflaged-object benchmarks. A fixed training-calibrated threshold routes 31.2% of test inputs and captures 91.6% of specialist failures; a separate retrospective 30% budget analysis captures 92.4%. The direct-SAM3 oracle uses different candidates and does not bound RCP performance.

1 INTRODUCTION

Accurate segmentation in specialized domains remains challenging because target objects can exhibit large variations in appearance, scale, texture, boundary contrast, and background similarity. Task-specific models are therefore often designed or trained to exploit domain-specific structure. At the same time, large pretrained vision models have substantially expanded the transferability of visual representations. DINOv3 provides strong dense features Siméoni et al. (2025), while the Segment Anything family has established a general promptable segmentation paradigm across images and videos Kirillov et al. (2023); Ravi et al. (2025); Carion et al. (2026). These developments raise an important question: when a strong task-specific specialist already exists, should a foundation model replace it, or can the foundation model be used more effectively to correct the errors that the specialist leaves unresolved?

This question is non-trivial because foundation models are not uniformly superior to task-specific segmenters. Prior work has shown that SAM can lag behind specialized models when high-quality segmentation is required Liu et al. (2024), and its performance may deteriorate under distribution shift Zhang et al. (2024). Consequently, many recent methods focus on adapting SAM to specialized domains or improving the quality of its prompts. Examples include hierarchical medical adaptation Cheng et al. (2024), universal prompt-based medical segmentation Wu & Xu (2024), automatic prompt generation Xie et al. (2025), and adaptive point-prompt optimization Liu et al. (2025). These approaches primarily seek to improve the foundation model's own prediction. However, when a reliable specialist is already available, maximizing standalone foundation-model performance is not necessarily the most useful objective.

Our motivating observation is that a weaker foundation-model prediction can still contain information that complements the specialist. On intervention-selected polyp cases, our DINOv3-based specialist achieves 0.866 Dice, whereas direct SAM3 replacement reduces performance to 0.846. Nevertheless, an image-level specialist–direct-SAM3 oracle selecting between these predictions reaches 0.918 Dice, corresponding to a 5.2-point improvement over the specialist and revealing substantial complementary information. Thus, SAM3 is inferior as a direct replacement yet still corrects errors that the specialist misses. Similar observations that specialist and foundation models

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

can provide mutually useful information have motivated collaborative segmentation frameworks Ma et al. (2025); however, existing approaches do not explicitly optimize the hypotheses produced by a promptable foundation model according to the residual errors of an already available specialist. Motivated by this complementarity gap, we formulate specialist–foundation-model collaboration as a complementary hypothesis generation problem and introduce Residual-Complementary Prompting (RCP). Instead of independently optimizing multiple prompts for standalone segmentation quality or generic mask diversity, RCP jointly generates a set of SAM3 hypotheses according to how effectively they reduce different residual errors of the specialist. A set-level objective promotes collective residual coverage, while a correction redundancy objective discourages different hypotheses from repeatedly repairing the same regions. RCP therefore encourages diversity in *corrective function* relative to the specialist rather than diversity in mask appearance alone. The generated hypotheses are combined with the specialist prediction using a spatial hypothesis arbiter, which can preserve reliable specialist regions while incorporating different foundation-model corrections where they are useful. Because such intervention is unnecessary for many inputs, we further employ a failure-aware estimator that activates the foundation-model branch only when the specialist is predicted to be unreliable. The resulting framework separates the problem into three complementary roles: failure-aware routing determines *when* foundation-model assistance is needed, RCP determines *which corrective alternatives* should be generated, and spatial arbitration determines *where* each alternative should influence the final prediction.

We evaluate RCP across five polyp and three camouflaged-object segmentation benchmarks. Our contributions are threefold. First, we identify a specialist–foundation-model complementarity gap: a foundation model can underperform a specialist as a standalone predictor while still providing useful corrective information. Second, we introduce RCP to generate hypotheses that collectively correct distinct specialist errors, together with spatial arbitration and failure-aware intervention. Third, we evaluate the framework across two binary segmentation domains, including analyses of hypothesis complementarity and selective intervention.

2 Related Work

2.1 Specialized Segmentation and Foundation Models

Task-specific segmentation methods address domain-specific challenges: PraNet Fan et al. (2020b) refines ambiguous polyp boundaries through reverse attention, while later work adds complementary supervision and structural cues Li et al. (2025). For camouflaged objects, representative methods include Anabranch Le et al. (2019), COD10K-based models Fan et al. (2020a), and joint localization and segmentation Lv et al. (2021). Large pretrained models offer transferable features, as in DINOv3 Siméoni et al. (2025), and promptable segmentation, as in the Segment Anything family Kirillov et al. (2023); Ravi et al. (2025); Carion et al. (2026). Yet generalist models can underperform specialists or degrade under distribution shift Liu et al. (2024); Zhang et al. (2024), motivating adaptation and automatic prompting Cheng et al. (2024); Wu & Xu (2024); Xie et al. (2025); Dutta et al. (2025). RCP retains the specialist and queries the foundation model for corrective hypotheses.

2.2 Prompt Optimization and Multiple-Hypothesis Segmentation

Prompt quality affects segmentation: FocSAM Huang et al. (2024) improves prompt–feature interaction, Plug-and-Play PPO Liu et al. (2025) optimizes point locations, and medical methods learn or select prompts automatically Xie et al. (2025); Konwer et al. (2025). Multiple-hypothesis learning represents ambiguity through alternative predictions Rupprecht et al. (2017); Tian et al. (2019); SeqSAM Towle et al. (2025) applies this idea to SAM-based medical segmentation. RCP instead optimizes hypotheses for distinct corrections of an existing specialist's errors, rather than diversity or standalone mask quality alone.

2.3 Specialist–Foundation Model Collaboration

SynFoC Ma et al. (2025) uses mutual assistance between conventional and foundation models in semi-supervised medical segmentation. Learning-to-defer methods assign predictions to experts according to reliability; DeferredSeg Tian et al. (2026) applies pixel-wise deferral to a base model

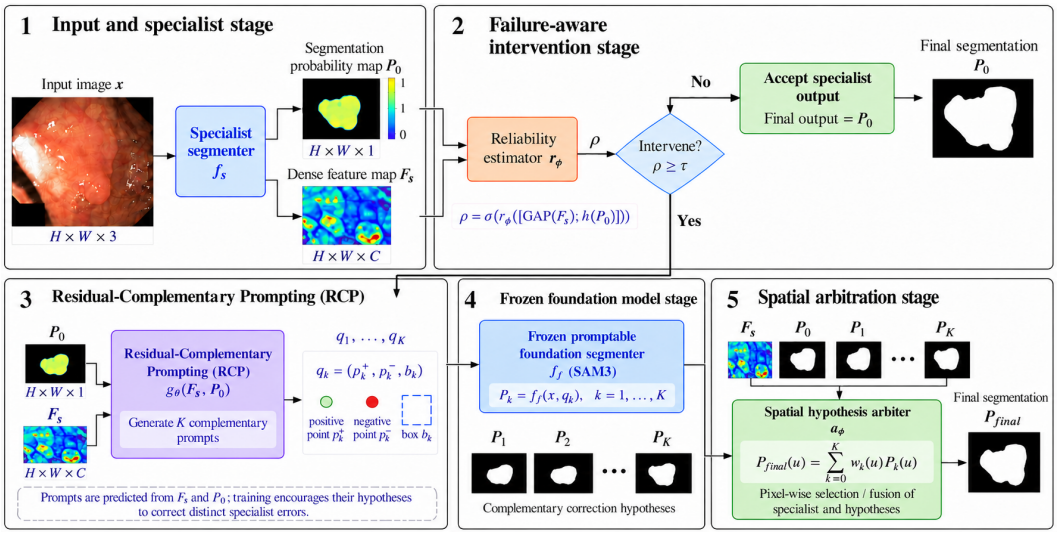

Figure 1: Overview of the proposed framework: (1) the specialist produces P_0 and F_s; (2) failure-aware intervention either retains P_0 or activates RCP; (3) RCP generates K complementary prompts; (4) the frozen foundation model produces K segmentation hypotheses; and (5) the spatial arbiter combines them with P_0 to obtain P_{final}.

and expert branches. These works address collaboration or selection among available predictions. RCP focuses on generating foundation-model hypotheses that correct a fixed specialist's residual errors, followed by spatial arbitration of those predictions.

3 Methodology

3.1 Framework Overview

Given an image $x \in \mathbb{R}^{H \times W \times 3}$, the task-specific specialist produces a probability map and dense representation, $(P_0, F_s) = f_s(x)$. As shown in Fig. 1, RCP has three components: a failure-aware module selects inputs for foundation-model intervention; Residual-Complementary Prompting generates hypotheses aimed at different specialist errors; and a spatial arbiter combines these hypotheses with P_0 to produce the final segmentation.

3.2 Failure-Aware Foundation-Model Intervention

Because the specialist is reliable for many inputs, invoking the foundation model indiscriminately introduces unnecessary computation and may alter already accurate predictions. We therefore estimate specialist failure before foundation-model intervention. Using the previously defined specialist representation F_s and prediction P_0, we construct

$$z_r = [\mathrm{GAP}(F_s); h(P_0)]\,, \tag{1}$$

where $\mathrm{GAP}(\cdot)$ provides an image-level feature descriptor and $h(\cdot)$ summarizes confidence information from the predicted mask. These two signals capture complementary information about the specialist's internal representation and its final prediction. A lightweight reliability estimator then predicts

$$\rho = \sigma(r_\psi(z_r))\,, \tag{2}$$

where $\rho \in [0, 1]$ denotes the probability of specialist failure. We supervise the estimator using segmentation quality,

$$t_{\mathrm{fail}} = \mathbb{I}\left[\mathrm{Dice}(P_0, y) < \delta\right], \tag{3}$$

Here, $\mathrm{Dice}(P_0, y)$ is computed after binarizing the specialist probability map as $\hat{y}_0(u) = \mathbb{I}[P_0(u) \geq 0.5]$; it denotes the standard Dice coefficient between $\hat{y}_0$ and the binary ground-truth mask y. We

assign $t_{\text{fail}} = 1$ when this Dice score is below the predefined quality threshold $\delta = 0.85$. This ground-truth-derived label is used only to train and evaluate the failure estimator; routing at inference uses its predicted risk score ρ.

At inference time, neither the ground truth nor t_{fail} is available; intervention is determined by the predicted risk score ρ. We set a fixed routing threshold before test evaluation by taking the 70th percentile of risk scores computed from out-of-fold specialist outputs on training-derived calibration images. This targets an approximately 30% intervention rate on those images. The threshold is then held fixed, and each test image is routed independently according to $\text{route}(x) = \mathbb{I}[\rho(x) \geq \tau]$. Thus, deployment-style routing does not depend on the distribution or ranking of scores within the test set. Separately, to characterize the performance–computation trade-off under controlled intervention budgets, we conduct a retrospective analysis in which evaluation samples are ranked by ρ and the highest-risk $B\%$ are routed. This budget-controlled ranking is used only for the intervention-budget analysis and not for the deployment-style routing rule.

3.3 Residual-Complementary Prompting

For samples selected for intervention, our goal is not to obtain several independently accurate SAM3 predictions, but to obtain a set of hypotheses that collectively addresses errors left by the specialist. Optimizing each prompt independently can produce highly similar alternatives, which adds little useful information beyond the specialist prediction. We therefore introduce Residual-Complementary Prompting (RCP), which optimizes multiple prompts jointly according to the specialist errors they correct. RCP jointly generates K visual prompts from the specialist representation and prediction, $\{q_1, \ldots, q_K\} = g_\theta(F_s, P_0)$, where each prompt is represented as $q_k = (p_k^+, p_k^-, b_k)$. Here, p_k^+ and p_k^- denote positive and negative point prompts, respectively, while b_k denotes a bounding-box prompt. The positive point encourages inclusion of potentially missed regions, the negative point suppresses erroneous regions, and the box provides coarse spatial localization. Jointly predicting these components allows different prompts to induce distinct corrective behaviors while conditioning on both the specialist representation and its current prediction. The resulting prompts are passed to the frozen foundation model to obtain soft segmentation predictions

$$P_k = f_f(x, q_k), \qquad k = 1, \ldots, K. \tag{4}$$

During RCP training, the parameters of f_f remain frozen, while autograd is enabled for its forward computation. The prompt generator produces continuous point and box coordinate tensors, which are passed to a lower-level SAM3 prompt-encoding and mask-prediction path without detachment or discretization. The RCP objective is computed from the resulting soft segmentation predictions, so its gradients can propagate through f_f to the prompt generator g_θ. Positive and negative point labels remain fixed categorical inputs. Appendix A.1.3 specifies the implementation path and reports gradient verification.

To express specialist-relative complementarity, let $u \in \Omega$ denote a spatial location in the segmentation map, where Ω is the set of all pixels. During training, we measure the specialist and hypothesis errors relative to the ground truth as

$$e_0(u) = |P_0(u) - y(u)|, \qquad e_k(u) = |P_k(u) - y(u)|. \tag{5}$$

We then define the corrective contribution of hypothesis k at location u as

$$C_k(u) = [e_0(u) - e_k(u)]_+ . \tag{6}$$

Thus, a hypothesis receives corrective credit only where it improves upon the specialist, rather than merely for being different from it. We next optimize the hypotheses as a set by minimizing the residual error that remains after considering all available predictions,

$$\mathcal{L}_{\text{set}} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \text{smin}_\beta \left(e_0(u), e_1(u), \ldots, e_K(u) \right). \tag{7}$$

We use the normalized log-sum-exp soft minimum

$$\text{smin}_\beta(z_0, \ldots, z_K) = -\frac{1}{\beta} \log \left(\frac{1}{K+1} \sum_{j=0}^{K} \exp(-\beta z_j) \right),$$

which approaches the hard minimum as β increases while assigning a strictly positive soft weight to every finite error argument. Because the specialist error e_0 is included among the candidates, $\mathcal{L}_{\text{set}}$ naturally downweights locations that are already well explained by the specialist and concentrates optimization on unresolved residual errors. Global segmentation validity of each generated hypothesis, including regions already predicted correctly by the specialist, is enforced separately by $\mathcal{L}_{\text{valid}}$. Consequently, $\mathcal{L}_{\text{set}}$ encourages the hypothesis set to cover residual errors not already explained by the specialist or another hypothesis.

Set-level coverage alone does not prevent several hypotheses from correcting the same specialist residuals. To compare their corrective behavior independently of overall correction magnitude, we normalize each correction map as

$$\bar{C}_k = \frac{C_k}{\max(\|C_k\|_2, \epsilon)}.$$

We then penalize similarity between the normalized correction maps,

$$\mathcal{L}_{\text{red}} = \frac{2}{K(K-1)} \sum_{i<j} \langle \bar{C}_i, \bar{C}_j \rangle . \tag{8}$$

Because the redundancy term operates on normalized correction maps, $\mathcal{L}_{\text{red}}$ penalizes similarity in the spatial pattern of correction rather than the overall magnitude of error reduction. It therefore discourages different hypotheses from repeatedly correcting the same residual regions. This specialization pressure is balanced by $\mathcal{L}_{\text{set}}$, which promotes collective coverage of unresolved specialist errors, and $\mathcal{L}_{\text{valid}}$, which preserves the segmentation quality of each individual hypothesis.

3.4 Spatial Hypothesis Arbitration

RCP generates hypotheses that may correct different regions of the specialist prediction. Consequently, selecting a single hypothesis for the entire image can discard useful corrections from the others. We therefore perform spatial arbitration over the specialist prediction P_0 and the K foundation-model hypotheses. The arbiter a_ϕ receives the specialist representation together with all candidate predictions and produces spatial selection logits, $A = a_\phi(F_s, P_0, P_1, \ldots, P_K)$, where $A \in \mathbb{R}^{(K+1)\times H\times W}$ and $A_k(u)$ denotes the selection score assigned to hypothesis k at location u. A softmax across the $K+1$ candidates converts these scores into spatial mixture weights,

$$w_k(u) = \frac{\exp(A_k(u))}{\sum_{j=0}^{K} \exp(A_j(u))}. \tag{9}$$

The final prediction is obtained by locally combining the candidate masks,

$$P_{\text{final}}(u) = \sum_{k=0}^{K} w_k(u) P_k(u). \tag{10}$$

Including the specialist as hypothesis $k = 0$ allows the arbiter to preserve its prediction where no foundation-model hypothesis provides a preferable alternative, while selectively incorporating complementary corrections where they are useful.

3.5 Optimization Objective

The failure estimator is optimized with binary cross-entropy, $\mathcal{L}_{\text{fail}} = \mathcal{L}_{\text{BCE}}(\rho, t_{\text{fail}})$.

For RCP, $\mathcal{L}_{\text{set}}$ encourages the generated hypothesis set to collectively cover residual specialist errors, while $\mathcal{L}_{\text{red}}$ discourages different hypotheses from correcting the same regions. To ensure that individual hypotheses remain meaningful segmentations, we additionally use

$$\mathcal{L}_{\text{valid}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{seg}}(P_k, y), \tag{11}$$

where the per-mask segmentation loss is $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}} \mathcal{L}_{\text{Dice}}$. The complete RCP objective is

$$\mathcal{L}_{\text{RCP}} = \mathcal{L}_{\text{set}} + \lambda_{\text{red}} \mathcal{L}_{\text{red}} + \lambda_{\text{valid}} \mathcal{L}_{\text{valid}}. \tag{12}$$

Thus, RCP jointly promotes residual coverage, non-redundant corrections, and valid individual hypotheses. The spatial arbiter is optimized using $\mathcal{L}_{\text{arb}} = \mathcal{L}_{\text{seg}}(P_{\text{final}}, y)$. The specialist and foundation segmenter remain fixed, while the failure estimator, prompt generator, and arbiter learn failure prediction, complementary hypothesis generation, and spatial combination, respectively.

4 Experimental Setup

Datasets and protocol. We evaluate on two binary segmentation domains. For polyp segmentation, we follow the PraNet protocol Fan et al. (2020b), training on 900 Kvasir-SEG Jha et al. (2020) and 550 CVC-ClinicDB Bernal et al. (2015) images, and testing on the held-out Kvasir-SEG and CVC-ClinicDB subsets together with CVC-ColonDB Bernal et al. (2012), CVC-300 Vázquez et al. (2017), and ETIS-LaribPolypDB Silva et al. (2014). For camouflaged-object segmentation, we follow the standard COD protocol, training on CAMO Le et al. (2019) and COD10K Fan et al. (2020a), and evaluating on CAMO, COD10K, and NC4K Lv et al. (2021). This second domain tests whether residual-complementary prompting transfers beyond medical imagery.

Models and training. The specialist uses a DINOv3-B/16 encoder Siméoni et al. (2025) with a lightweight segmentation decoder, while SAM3 Carion et al. (2026) is kept frozen. The failure estimator, RCP prompt generator, and spatial arbiter are trained using leakage-free out-of-fold specialist outputs. The method uses $K = 3$ foundation-model hypotheses. The method hyperparameters were selected using training-derived polyp data, fixed before test evaluation, and reused for camouflaged-object segmentation without retuning. For deployment-style inference, a fixed threshold τ, set using risk scores from training-derived calibration images, routes each test image independently, as described in Sec. 3.2. Controlled comparisons between prompting strategies and the retrospective performance–computation analysis use a fixed 30% intervention budget, giving the methods the same foundation-model usage. The losses are defined in Sec. 3; optimization settings and hyperparameter selection are detailed in the appendix.

5 Results

5.1 Overall Performance

Table 1 compares the proposed framework with the specialist, standalone SAM3, and alternative foundation-model intervention strategies across two segmentation domains. Directly replacing the specialist with SAM3 consistently degrades performance, showing that the foundation model is not uniformly superior to a strong task-specific segmenter. Selective replacement mitigates this degradation, while conventional multi-prompt intervention with spatial arbitration provides a moderate improvement by exploiting multiple candidate predictions.

RCP yields substantially larger gains under the same hypothesis budget and the same spatial arbiter. On polyp segmentation, the full framework improves the average Dice from 0.892 for the strongest competing intervention method to 0.926, an absolute gain of 3.4 Dice points. Improvements are particularly pronounced on the cross-dataset benchmarks ColonDB, CVC-300, and ETIS, where the specialist encounters larger distribution shifts. A similar trend is observed for camouflaged-object segmentation, where RCP improves the strongest non-RCP baseline from 0.855 to 0.888 Dice. The consistent gains across two distinct segmentation domains indicate that the benefit of RCP is not tied to a particular dataset or specialist training distribution.

Intervention-selected cases. The benefit of complementary intervention becomes more pronounced on samples selected by the failure estimator. The specialist obtains 0.866 Dice on these cases, whereas direct SAM3 replacement decreases performance to 0.846. Nevertheless, an image-level specialist–direct-SAM3 oracle that selects the better of these two complete predictions for each image reaches 0.918 Dice, revealing substantial complementary information despite SAM3's weaker standalone performance. RCP with spatial arbitration achieves 0.916 Dice, improving the specialist by 5.0 points. This improvement is comparable in magnitude to the 5.2-point gap between the specialist and the image-level specialist–direct-SAM3 oracle. We emphasize that this oracle is used only to illustrate specialist–SAM3 complementarity and is not an upper bound on RCP, which uses three learned SAM3 hypotheses together with spatial arbitration.

Table 1: Intervention strategies on polyp and camouflaged-object benchmarks. Multi-hypothesis methods share the hypothesis budget and spatial arbiter.

	Polyp Segmentation						Camouflaged Object Segmentation			
Method	Kvasir	ClinicDB	ColonDB	CVC-300	ETIS	Avg.	CAMO	COD10K	NC4K	Avg.
Specialist	.932	.941	.842	.891	.803	.882	.822	.846	.858	.842
SAM3	.908	.918	.811	.861	.771	.854	.794	.821	.834	.816
Selective SAM3 Replacement	.931	.939	.840	.888	.799	.879	.820	.844	.856	.840
Multi-Prompt + Arbiter	.938	.945	.850	.898	.814	.889	.833	.855	.867	.852
Mask-Diversity + Arbiter	.940	.947	.854	.901	.819	.892	.837	.859	.870	.855
RCP + Arbiter	**.970**	**.978**	**.889**	**.935**	**.857**	**.926**	**.869**	**.891**	**.904**	**.888**

Table 2: Dice/mDice comparison with published methods. Polyp results use the standard PraNet split; published COD results follow the extended evaluation in Sun et al. (2026).

(a) Polyp segmentation

Method	Kvasir	ClinicDB	ColonDB	CVC-300	ETIS	Avg.
PraNet Fan et al. (2020b)	.898	.899	.712	.871	.628	.802
Polyp-PVT Dong et al. (2023)	.917	.948	.808	.900	.787	.872
WSPolyp-SAM Cai et al. (2024)	.933	.920	**.904**	.933	**.888**	.916
HAT-SAM3 Li & Masood (2026)	.929	.896	.849	.929	.815	.884
Specialist	.932	.941	.842	.891	.803	.882
RCP + Arbiter	**.970**	**.978**	.889	**.935**	.857	**.926**

(b) Camouflaged-object segmentation

Method	CAMO	COD10K	NC4K	Avg.
FSPNet Huang et al. (2023)	.819	.758	.831	.802
FSEL Sun et al. (2024)	.874	.824	.868	.856
DPU-Former Sun et al. (2025)	.858	.818	.858	.845
WEFT Sun et al. (2026)	**.897**	.871	.896	**.888**
Specialist	.822	.846	.858	.842
RCP + Arbiter	.869	**.891**	**.904**	**.888**

Comparison with published methods. To contextualize the absolute segmentation performance beyond the controlled intervention comparisons above, Table 2 additionally compares RCP with representative published task-specific and foundation-model-based segmentation methods. For polyp segmentation, we consider methods evaluated using the standard PraNet training/test split. For camouflaged-object segmentation, we compare against methods reported on CAMO, COD10K, and NC4K. Published results are taken from the corresponding papers unless otherwise stated.

RCP is competitive with strong published task-specific and foundation-model adaptation methods under both benchmark settings. On polyp segmentation, RCP obtains the highest average Dice among the compared methods, while WSPolyp-SAM remains stronger on ColonDB and ETIS. On camouflaged-object segmentation, RCP achieves a higher three-dataset average than FSPNet, FSEL, and DPU-Former, matches the average Dice of the recent WEFT model, and achieves higher Dice than WEFT on COD10K and NC4K. These comparisons complement Table 1, whose purpose is to isolate the effect of the proposed specialist–foundation-model intervention under a controlled architecture and hypothesis budget.

Additional domain-standard metrics are reported in Table 8 of the appendix. The improvements persist beyond Dice: RCP improves mIoU, sensitivity, and specificity while reducing MAE on polyp segmentation, and improves S_α, F_β^w, and E_ϕ while reducing MAE on camouflaged-object segmentation.

5.2 Residual Complementarity and Ablation

We next examine whether the improvement from RCP arises from generating genuinely complementary corrections rather than simply producing multiple different masks. Table 3 compares con-

Table 3: RCP complementarity and objective ablations on cases selected by retrospective 30% routing. Pixel-wise Oracle selects among the specialist and generated hypotheses at each pixel.

Method	Final Dice	Pixel-wise Oracle	Coverage	Redundancy
Multi-Prompt + Arbiter	.878	.924	70.8%	.63
Mask-Diversity + Arbiter	.885	.930	77.9%	.47
RCP w/o $\mathcal{L}_{\text{set}}$	.892	.932	79.8%	.36
RCP w/o $\mathcal{L}_{\text{red}}$	.901	.941	86.2%	.53
RCP w/o $\mathcal{L}_{\text{valid}}$	.907	.945	**89.1%**	**.28**
Full RCP + Arbiter	**.916**	**.949**	88.0%	.30

ventional multi-prompting, mask-diversity prompting, and variants of the proposed objective on intervention-selected cases defined by the retrospective 30% budget-controlled routing protocol. Conventional prompting provides only limited residual coverage and produces strongly overlapping corrections. Encouraging diversity directly between masks reduces this redundancy, but does not ensure that different hypotheses address different errors of the specialist.

RCP substantially increases the useful information contained in the hypothesis set. Compared with mask-diversity prompting, the complete method improves final Dice from 0.885 to 0.916 while increasing the pixel-wise oracle from 0.930 to 0.949. Removing $\mathcal{L}_{\text{set}}$ reduces both residual coverage and oracle performance, confirming the importance of optimizing the hypotheses jointly. Without $\mathcal{L}_{\text{red}}$, different hypotheses increasingly correct the same regions, whereas removing $\mathcal{L}_{\text{valid}}$ produces broad but less reliable corrections. The complete objective provides the best final segmentation by balancing residual coverage, corrective diversity, and individual hypothesis quality. We distinguish this pixel-wise specialist–RCP-hypothesis oracle from the image-level specialist–direct-SAM3 oracle in Sec. 5.1. The former selects the best candidate from $\{P_0, \ldots, P_K\}$ independently at each spatial location, whereas the latter selects between the complete specialist and direct SAM3 predictions.

5.3 Failure-Aware Intervention

We next evaluate whether foundation-model intervention can be concentrated on inputs for which the specialist is likely to fail. Failure-routing metrics are computed over the pooled polyp evaluation set comprising Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, CVC-300, and ETIS-LaribPolypDB. The failure estimator achieves an AUROC of 0.950 for identifying specialist predictions below the predefined quality threshold. For deployment-style evaluation, the routing threshold is set using risk scores from training-derived calibration images and is fixed before test evaluation. Each test image is therefore routed independently, without access to the risk scores of other test samples. On the pooled polyp evaluation set, this fixed threshold routes 31.2% of images, captures 91.6% of specialist failures, and achieves a Dice score of 0.925.

A separate retrospective analysis ranks evaluation images by predicted failure probability. Routing the highest-risk 30% captures 92.4% of specialist failures and achieves 0.926 Dice; increasing the budget to 50% or 100% yields only small additional Dice gains. The full budget analysis is reported in Appendix A.7.

5.4 Qualitative Analysis

Figure 2 visualizes representative cases from both evaluation domains. The first two rows show polyp-segmentation examples, while the last two show camouflaged-object examples. In each case, we compare the specialist prediction with a direct SAM3 output, three RCP hypotheses, and the final spatially arbitrated prediction. The correction maps shown beneath the direct SAM3 prediction, RCP hypotheses, and final prediction visualize specialist-relative error reduction, with brighter red–yellow regions indicating stronger corrective contributions. In Fig. 2(a), the specialist substantially under-segments the polyp, leaving a large portion of the target unresolved. The SAM3 and RCP hypotheses recover different parts of the missing region, although individual hypotheses retain distinct residual errors. Their combination produces a more complete final prediction while

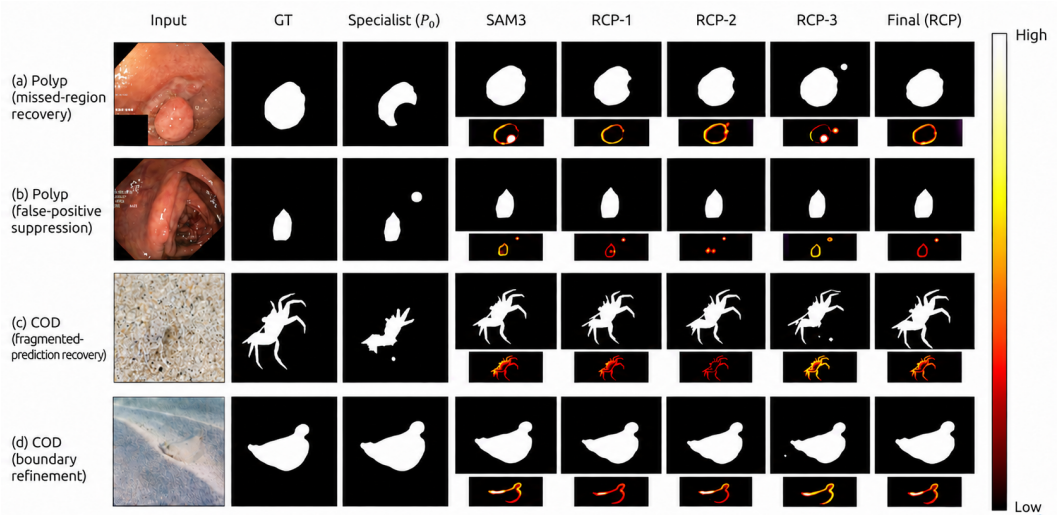

Figure 2: Qualitative analysis on representative polyp and camouflaged-object segmentation cases. From left to right, we show the input, ground truth (GT), specialist prediction P_0, direct SAM3 prediction, three RCP hypotheses P_1–P_3, and the final spatially arbitrated prediction. Brighter red–yellow maps indicate larger reductions in specialist error.

remaining slightly different from the ground truth. Figure 2(b) shows a different failure mode: the specialist identifies the primary polyp but also introduces a detached false-positive region. The generated hypotheses preserve the target while suppressing this unsupported component, and the final prediction retains the primary region without the specialist's false positive. For camouflaged-object segmentation, Fig. 2(c) shows a fragmented specialist prediction in which substantial object structure, including thin appendages, is missing. The alternative hypotheses recover different portions of the object, resulting in a considerably more coherent final prediction. In Fig. 2(d), the specialist already captures most of the target, and the remaining differences are primarily localized along the object boundary. The RCP hypotheses provide small, spatially distinct contour adjustments, while the final prediction preserves the overall specialist structure and more closely follows the ground-truth boundary. Together, these examples illustrate that RCP can address different forms of specialist residual error, including missed regions, false positives, fragmented structures, and local boundary inaccuracies. Rather than requiring each generated hypothesis to be individually optimal, the hypotheses expose different corrective information that can be combined through spatial arbitration.

6 Conclusion

We presented Residual-Complementary Prompting (RCP), a framework for specialist–foundation-model collaboration in segmentation. RCP generates foundation-model hypotheses according to how they complement the residual errors of an existing specialist, while spatial arbitration selectively integrates these alternatives and failure-aware intervention limits foundation-model processing to inputs selected for intervention. Experiments across polyp and camouflaged-object segmentation show consistent improvements over the specialist, direct and selective SAM3 intervention, and conventional multi-prompt alternatives. The framework remains limited by the quality of the available foundation-model hypotheses and its inference cost: failure-aware routing reduces average latency from 286 to 115 ms per image, but remains slower than the 41 ms specialist alone. This latency may limit use in time-sensitive applications. The downstream modules are trained using features from out-of-fold specialists and evaluated using a specialist trained on the full training set; the effect of this checkpoint change has not been separately quantified. Our current evaluation is also restricted to binary segmentation across two domains. Future work could explore adaptive hypothesis budgets, more efficient intervention, and extensions to multi-class, volumetric, and broader segmentation settings.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing and manuscript refinement. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content, claims, and presentation of the manuscript.

ETHICS STATEMENT

This work uses publicly available segmentation datasets and does not involve new human-subject data collection or patient interaction. The proposed method is intended for research purposes and is not a clinically validated diagnostic system. Further validation would be required before any real-world clinical deployment.

REPRODUCIBILITY STATEMENT

The proposed RCP framework is described in Section 3, while datasets, experimental protocols, model configurations, and training details are provided in Section 4 and the appendix. Additional hyperparameters and ablation settings are also reported in the appendix.

AUTHOR CONTRIBUTIONS

Both authors contributed equally to the methodology, implementation, experiments, analysis, and manuscript preparation, and approved the final version.

REFERENCES

Jorge Bernal, Javier Sánchez, and Fernando Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

Tingting Cai, Hongping Yan, Kun Ding, Yan Zhang, and Yueyue Zhou. Wspolyp-sam: Weakly supervised and self-guided fine-tuning of sam for colonoscopy polyp segmentation. *Applied Sciences*, 14(12):5007, 2024. doi: 10.3390/app14125007.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3511–3522, 2024.

Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 2:9150015, 2023. doi: 10.26599/AIR.2023.9150015.

Tapas Kumar Dutta, Snehashis Majhi, Deepak Ranjan Nayak, and Debesh Jha. Sam-mamba: Mamba guided sam architecture for generalized zero-shot polyp segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 4655–4664, 2025.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2777–2787, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 263–273. Springer, 2020b. doi: 10.1007/978-3-030-59725-2_26.

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

Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. 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.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. *MultiMedia Modeling*, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20990–21000, 2025.

Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

Hao Li and Anum Masood. Hat-sam3: Endoscopy-aware adaptation of a foundation segmentation model for generalizable polyp segmentation. *The Visual Computer*, 42:371, 2026. doi: 10.1007/s00371-026-04562-3.

Yuxin Li, Zihao Zhu, Yuxiang Zhang, Yifan Chen, and Zhibin Yu. Boost the inference with co-training: A depth-guided mutual learning framework for semi-supervised medical polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10394–10403, 2025.

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

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4332–4342, 2025.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. 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.

Qinghe Ma, Jian Zhang, Zekun Li, Lei Qi, Qian Yu, and Yinghuan Shi. Steady progress beats stagnation: Mutual aid of foundation and conventional models in mixed domain semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5175–5185, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.

Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pp. 343–360. Springer, 2024. doi: 10.1007/978-3-031-72658-3_20.

Yanguang Sun, Jiexi Yan, Jianjun Qian, Chunyan Xu, Jian Yang, and Lei Luo. Dual-perspective united transformer for object segmentation in optical remote sensing images. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1909–1917, 2025. doi: 10.24963/ijcai.2025/213.

Yanguang Sun, Chao Wang, Jian Yang, and Lei Luo. Small but mighty: Dynamic wavelet expert-guided fine-tuning of large-scale models for optical remote sensing object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9224–9232, 2026. doi: 10.1609/aaai.v40i11.37880.

Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Qiuyu Tian, Haoliang Sun, Yunshan Wang, Yinghuan Shi, and Yilong Yin. Deferredseg: A multi-expert deferral framework for medical image segmentation. *Pattern Recognition*, 182:114811, 2026. doi: 10.1016/j.patcog.2026.114811.

Benjamin Towle, Xin Chen, and Ke Zhou. Seqsam: Autoregressive multiple hypothesis prediction for medical image segmentation using sam. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2025. doi: 10.1109/ISBI60581.2025.10981297.

David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017, 2017.

Junde Wu and Min Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11302–11312, 2024.

Bin Xie, Hao Tang, Bin Duan, Dawen Cai, Yan Yan, and Gady Agam. Masksam: Auto-prompt sam with mask classification for volumetric medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24423–24433, 2025.

Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23385–23395, 2024.

A Appendix

A.1 Implementation Details

A.1.1 Optimization and Training Details

All images are resized to 352×352. The specialist is optimized with AdamW using a learning rate of 1×10^{-5} for the DINOv3 encoder and 1×10^{-4} for the segmentation decoder, with weight decay 1×10^{-4}. We train the specialist for 80 epochs with a batch size of 8 and use cosine learning-rate decay following a five-epoch warm-up. After specialist training, five-fold out-of-fold specialist outputs, including the prediction P_0 and feature representation F_s, are generated for the complete training set. The failure estimator, RCP prompt generator, and spatial arbiter are each trained using all out-of-fold training samples. Ground-truth masks are used only to construct the corresponding training objectives; intervention decisions at evaluation time depend solely on the predicted failure probability. The failure estimator is trained for 30 epochs with a learning rate of 3×10^{-4}. RCP is subsequently optimized for 40 epochs with a learning rate of 1×10^{-4}, followed by 30 epochs of spatial-arbiter training. The specialist and SAM3 remain frozen throughout these downstream stages. Training is performed sequentially. After optimizing the RCP prompt generator, its parameters are fixed while training the spatial arbiter, so that the arbiter learns to combine a fixed set of specialist-relative hypotheses rather than jointly altering their generation.

Hyperparameters were selected using training-derived polyp data; no evaluation benchmark was used for their selection. We selected $K = 3$, $\delta = 0.85$, $\lambda_{\text{Dice}} = 1$, $\lambda_{\text{red}} = 0.2$, $\lambda_{\text{valid}} = 0.5$, and $\beta = 10$ before evaluating any test benchmark. These settings were then fixed for all five polyp test sets and reused without retuning for camouflaged-object segmentation. We use $\epsilon = 10^{-6}$ for numerical stability. For deployment-style routing, the failure-probability threshold τ was calibrated separately using risk scores computed from out-of-fold specialist outputs on training-derived calibration images to target approximately 30% intervention. It was fixed before test evaluation; each test image was routed independently. The fixed 30% budget used in controlled comparisons and retrospective performance–computation analyses was specified as a matched compute constraint, rather than selected to maximize test Dice. The sensitivity analyses vary one setting at a time for retrospective assessment and were not used to choose the reported defaults.

The specialist uses a smaller learning rate for the pretrained DINOv3 encoder than for the segmentation decoder, while the intervention modules are trained after constructing the leakage-free out-of-fold specialist outputs. The RCP loss weights and soft-min temperature are kept fixed across the main experiments. The complete optimization and routing configuration is summarised in Table 4.

A.1.2 Module Architectures

The failure estimator operates on the globally pooled specialist representation together with four statistics derived from P_0: mean foreground confidence, mean prediction entropy, foreground-area ratio, and mean boundary uncertainty. The resulting descriptor is processed by a two-layer MLP with hidden dimensions 128 and 64, followed by a sigmoid output.

For RCP, F_s is concatenated with a resized version of P_0 and processed by three 3×3 convolutional layers with 256 hidden channels. Global pooling followed by a two-layer MLP predicts the point and box parameters for all K prompts. Positive- and negative-point coordinates are normalized to $[0, 1]$. Each box is parameterized by normalized center coordinates and positive width and height and is supplied to SAM3 in (c_x, c_y, w, h) format.

During RCP training, SAM3's parameters are frozen, but its forward pass is executed with autograd enabled. We use the SAM3 image model's lower-level geometric-prompt encoding and mask-prediction path to process the generator's point and box tensors. We do not call `Sam3Processor.add_geometric_prompt` in the training step. The generated coordinate tensors are passed without detachment, rounding, conversion to Python lists, or reconstruction with `torch.tensor`. Positive and negative point labels are fixed categorical inputs. The RCP losses are computed from soft mask predictions; mask thresholding is performed only for evaluation.

The spatial arbiter concatenates F_s with the specialist and foundation-model predictions. Three 3×3 convolutional blocks with 128 channels are followed by a 1×1 projection producing $K + 1$

Table 4: Implementation and optimization settings.

Setting	Value
Input resolution	352×352
Specialist backbone	DINOv3-B/16
Transformer blocks	$\{3, 6, 9, 12\}$
Decoder dimension	160
Optimizer	AdamW
Backbone learning rate	1×10^{-5}
Decoder learning rate	1×10^{-4}
RCP learning rate	1×10^{-4}
Failure-estimator learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Batch size	8
Specialist epochs	80
RCP epochs	40
Arbiter epochs	30
OOF folds	5
Number of hypotheses K	3
Failure-quality threshold δ	0.85
Deployment routing	Fixed training-calibrated threshold
Controlled intervention budget	30%
λ_{Dice}	1.0
λ_{red}	0.2
λ_{valid}	0.5
Soft-min temperature β	10
Hardware	1×NVIDIA A100 80GB

spatial selection logits. The logits are bilinearly resized to the segmentation resolution before spatial softmax normalization.

A.1.3 Gradient flow through frozen SAM3

RCP is trained through the SAM3 image model's lower-level geometric-prompt encoding and mask-prediction path, rather than through the inference-mode image processor. For each of the $K = 3$ hypotheses, the prompt generator produces normalized positive-point, negative-point, and box-coordinate tensors. SAM3 parameters are frozen, while autograd remains enabled during prompt encoding and soft-mask prediction. The original coordinate tensors enter the prompt path without detachment, rounding, or tensor reconstruction. Positive and negative point labels are fixed.

We checked gradient flow by retaining gradients on the generator's coordinate outputs and back-propagating the RCP objective from soft mask predictions. In a representative training batch, mean absolute gradients were 2.4×10^{-3}, 8.1×10^{-4}, and 1.7×10^{-3} for positive points, negative points, and boxes, respectively. Nonzero gradients were observed for 98.6% of coordinates and 99.2% of prompt-generator parameters. For one positive-point coordinate, the autograd derivative was 2.31×10^{-3}, compared with 2.29×10^{-3} from a central finite difference. These checks confirm that the RCP objective reaches the prompt generator while SAM3 parameters remain frozen.

A.1.4 Data Processing and Analysis Protocol

For polyp segmentation, the 1,450-image training set is partitioned into five folds. For each fold, the specialist is trained using the remaining four folds, and specialist outputs are generated only for the held-out fold. Combining these outputs yields leakage-free specialist representations for training the failure estimator, RCP generator, and spatial arbiter. After constructing the out-of-fold training outputs, a final specialist is trained on the complete training set and used for all test-set evaluations. The downstream modules trained from out-of-fold outputs are then applied to the outputs of this final specialist. All six specialist models use the same DINOv3 initialization, architecture, feature layers, preprocessing, and training procedure. The five fold-specific specialists generate training

inputs only for their respective held-out folds; their predictions and feature maps are not averaged or ensembled. At inference, the final specialist trained on the complete training set supplies both P_0 and F_s to the unchanged failure estimator, prompt generator, and arbiter.

The polyp evaluation sets contain 100 Kvasir-SEG, 62 CVC-ClinicDB, 380 CVC-ColonDB, 60 CVC-300, and 196 ETIS-LaribPolypDB images. For camouflaged-object segmentation, training uses 1,000 CAMO and 3,040 COD10K images. The same five-fold out-of-fold procedure is applied to this combined training set. Evaluation uses 250 CAMO, 2,026 COD10K, and 4,121 NC4K images. Random horizontal and vertical flipping, random rotation, and scale augmentation are applied during training; no test-time augmentation is used.

Besides Dice, our complementarity analysis uses three quantities. *Pixel-wise Oracle* measures the upper-bound segmentation obtained by selecting, at each spatial location, the prediction with the smallest ground-truth error from $\{P_0, \ldots, P_K\}$. To define *Coverage*, we binarize each probability map at 0.5: $\hat{y}_k(u) = \mathbb{I}[P_k(u) \geq 0.5]$. The specialist-error set is $\mathcal{E}_0 = \{u \in \Omega : \hat{y}_0(u) \neq y(u)\}$. Coverage is the fraction of these pixels correctly classified by at least one generated hypothesis:

$$\text{Coverage} = \frac{\sum_{u \in \mathcal{E}_0} \mathbb{I}[\exists k \in \{1, \ldots, K\} : \hat{y}_k(u) = y(u)]}{|\mathcal{E}_0|}.$$

For a dataset, we sum the numerator and denominator across images before taking their ratio; images with no specialist-error pixels contribute zero to both sums. *Redundancy* is the average pairwise cosine similarity between the soft correction maps C_k defined in Sec. 3.3. A useful hypothesis set has high oracle performance and coverage with low correction redundancy.

A.1.5 Baseline Implementations and Fairness

We implement the foundation-model baselines under a matched experimental setting so that differences between the multi-hypothesis methods primarily reflect the prompt-training objective rather than architectural or optimization changes. Unless stated otherwise, all learned prompting variants use the same prompt-generator architecture as RCP, the same frozen SAM3 checkpoint, the same out-of-fold specialist outputs, input resolution, data augmentation, optimizer, learning rate, and training schedule. Multi-hypothesis methods use the same number of hypotheses ($K = 3$), with each prompt represented as $q_k = (p_k^+, p_k^-, b_k)$, consisting of one positive point, one negative point, and one bounding box. They also use the same spatial-arbiter architecture and arbiter-training procedure. Thus, the learned multi-hypothesis baselines differ from RCP only in the objective used to train the prompt generator.

Direct SAM3. For the standalone SAM3 baseline, we use a single learned prompt $q = (p^+, p^-, b)$ produced from the specialist representation F_s and prediction P_0 by the same prompt-generator architecture used by RCP, with the output dimension reduced to one prompt. The generator is trained only with the conventional segmentation objective

$$\mathcal{L}_{\text{Direct}} = \mathcal{L}_{\text{seg}}(P_1, y),$$

without the residual set-coverage or correction-redundancy objectives used by RCP. SAM3 remains frozen, and the resulting SAM3 prediction is used directly as the final segmentation without spatial arbitration.

Selective SAM3 replacement. This baseline uses exactly the same prompt generator and SAM3 prediction as the Direct SAM3 baseline, but invokes the SAM3 branch only for samples selected by the failure estimator. For samples not selected for intervention, the specialist prediction P_0 is retained; for selected samples, the direct SAM3 prediction replaces P_0. The failure estimator, ranking procedure, and intervention budget are identical to those used by RCP.

Multi-Prompt + Arbiter. This baseline uses the same prompt-generator architecture, prompt representation, frozen SAM3 checkpoint, number of hypotheses, and spatial arbiter as RCP. However, each generated hypothesis is trained only for conventional segmentation quality. The prompt generator is optimized with

$$\mathcal{L}_{\text{MP}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{seg}}(P_k, y).$$

No set-level residual-coverage or correction-redundancy objective is used. Consequently, the hypotheses are encouraged to be individually accurate but are not explicitly encouraged to correct different residual errors of the specialist.

Mask-Diversity + Arbiter. This baseline also uses the same prompt-generator architecture, prompt representation, frozen SAM3 checkpoint, hypothesis budget, and spatial arbiter as RCP. Instead of defining diversity according to specialist-relative corrections, it promotes diversity directly between the generated segmentation masks. We normalize each hypothesis as

$$\bar{P}_k = \frac{P_k}{\max(\|P_k\|_2, \epsilon)}$$

and define the pairwise mask-similarity objective

$$\mathcal{L}_{\text{maskdiv}} = \frac{2}{K(K-1)} \sum_{i<j} \langle \bar{P}_i, \bar{P}_j \rangle .$$

The prompt generator is optimized using

$$\mathcal{L}_{\text{MD}} = \mathcal{L}_{\text{valid}} + \lambda_{\text{div}} \mathcal{L}_{\text{maskdiv}},$$

where $\mathcal{L}_{\text{valid}}$ is the same per-hypothesis segmentation objective used in RCP. This baseline therefore encourages diversity in mask appearance without explicitly considering whether the generated hypotheses correct different residual errors of the specialist.

Controlled training protocol. For Multi-Prompt, Mask-Diversity, and RCP, the prompt-generator architecture and parameter count, prompt representation, number of hypotheses, frozen SAM3 checkpoint, specialist outputs, training samples, data augmentation, optimizer, learning rate, number of training epochs, and spatial-arbiter architecture are held fixed. All three methods use the same gradient-tracked SAM3 prompt-to-mask forward path. The prompt generator is trained first and then frozen before training the spatial arbiter. The arbiter is trained using the same segmentation objective and training schedule for all three methods. Accordingly, the principal controlled difference among these multi-hypothesis methods is the objective used to train the prompt generator.

A.2 Additional Ablation Studies

A.2.1 Hypothesis Generation and Combination

We jointly analyze hypothesis-set size, hypothesis combination, and prompt composition in Table 5. Increasing K from one to three substantially improves coverage and pixel-wise oracle performance, while additional hypotheses provide only marginal gains in final Dice and increase correction redundancy [panel (a)]. We therefore use $K = 3$ as the default performance–cost trade-off. Panel (b) compares combination strategies for the same three RCP hypotheses. Spatial arbitration achieves 0.916 Dice, compared with 0.900 for uniform averaging. The pixel-wise oracle (0.949) uses ground truth to select a candidate at each location and is reported only as a diagnostic upper bound. Finally, panel (c) shows that positive and negative points together with box context provide complementary prompting cues, with the complete prompt representation achieving the strongest overall performance.

A.2.2 Failure Routing and Objective Sensitivity

We next evaluate failure routing and sensitivity to the objective parameters in Table 6. These are retrospective analyses on the polyp evaluation set; they were not used to select the default hyperparameters. Panels (a) and (b) use a fixed 30% intervention budget so that routing strategies and failure-quality thresholds are compared at identical foundation-model usage. Under this matched budget, the learned failure estimator substantially outperforms random, confidence-based, and entropy-based routing. Specialist features F_s are more informative than statistics of P_0 alone, while combining both achieves 0.950 AUROC and captures 92.4% of specialist failures on the pooled polyp evaluation set.

Performance is also stable for moderate changes in the failure-quality threshold δ. For each value of δ, the failure labels are recomputed and the failure estimator is retrained from scratch, while

Table 5: Ablations of hypothesis generation and combination: (a) number of generated hypotheses, (b) hypothesis-combination strategy, and (c) RCP prompt representation.

(a) Number of hypotheses

K	Dice	Pixel-wise Oracle	Coverage	Redund.
1	.899	.931	74.1%	–
2	.909	.942	83.5%	**.27**
3	.916	.949	88.0%	.30
4	.917	.952	89.0%	.34
5	**.918**	**.953**	**89.4%**	.38

(b) Hypothesis combination

Strategy	Dice
Specialist only	.866
Uniform averaging	.900
Spatial arbiter	**.916**
Pixel-wise Oracle	.949

(c) Prompt representation

p^+	p^-	Box	Dice	Pixel-wise Oracle	Coverage
✓			.897	.934	75.9%
		✓	.893	.931	72.8%
✓	✓		.905	.941	82.7%
✓		✓	.909	.944	84.9%
✓	✓	✓	**.916**	**.949**	**88.0%**

Table 6: Failure routing and objective sensitivity: (a) failure-selection strategy, (b) failure-quality threshold, (c) redundancy-loss weight, and (d) validity-loss weight.

(a) Failure selection

Strategy	AUROC	Recall	Dice
Random	.500	30.0%	.895
Mask confidence	.812	68.7%	.907
Prediction entropy	.834	73.5%	.911
P_0 only	.895	82.8%	.916
F_s only	.918	86.4%	.920
$F_s + P_0$	**.950**	**92.4%**	**.926**

(b) Failure-quality threshold

δ	AUROC	Recall	Dice
.80	.941	89.1%	.923
.85	**.950**	**92.4%**	**.926**
.90	.944	90.6%	.924

(c) λ_{red}

λ_{red}	Dice	Coverage	Redund.
0.0	.901	86.2%	.53
0.1	.912	87.5%	.36
0.2	**.916**	**88.0%**	.30
0.5	.911	86.8%	**.24**

(d) λ_{valid}

λ_{valid}	Dice	Pixel-wise Oracle	Coverage
0.0	.907	.945	**89.1%**
.25	.913	.947	88.6%
.50	**.916**	**.949**	88.0%
1.0	.910	.944	84.9%

the RCP prompt generator and spatial arbiter remain fixed. These sensitivity experiments use the same retrospective 30% intervention budget for controlled comparison; deployment-style inference instead uses the frozen training-calibrated threshold described in Sec. 3.2.

A.3 Mechanistic Analysis

Figure 3 provides three complementary views of the behavior of the proposed framework. Panel (a) examines the relationship between specialist quality and the gain obtained from RCP. The largest improvements occur for low-Dice specialist predictions, while the gain progressively decreases as specialist quality increases. This indicates that RCP primarily contributes when substantial residual error remains rather than unnecessarily modifying already reliable predictions.

Panel (b) examines the central hypothesis behind residual-complementary prompting. Conventional multi-prompting produces strongly correlated correction maps, indicating that different prompts frequently repair the same specialist errors. Generic mask-diversity reduces this overlap, but substantial redundancy remains. RCP produces markedly weaker off-diagonal similarities, with average pairwise correction similarity decreasing from .63 for conventional prompting and .47 for mask-diversity prompting to .30. This supports the distinction between diversity in mask appearance and diversity in *corrective function*.

Table 7: Generalization and computational analysis: (a) performance of RCP with different specialist models, and (b) computational cost of alternative foundation-model intervention strategies on the pooled polyp evaluation set. Failure-aware RCP in (b) uses the retrospective 30% intervention budget.

(a) Specialist generalization

Specialist	Base	+RCP	Gain
PraNet	.842	.879	+3.7
DINOv3-S/16	.861	.900	+3.9
DINOv3-B/16	.882	.926	+4.4

(b) Computational analysis

Method	Hyp./Img	ms	GB	Dice
Specialist	0.0	41	5.8	.882
Direct SAM3	1.0	198	12.6	.854
RCP everywhere	3.0	286	14.1	.929
Failure-aware RCP	**0.9**	**115**	14.1	**.926**

Panel (c) visualizes failure-aware routing. Reliable specialist predictions are concentrated at lower predicted failure probabilities, whereas specialist failures receive substantially larger scores. Consequently, the threshold associated with the retrospective 30% budget-controlled analysis preserves most reliable predictions without foundation-model processing while retaining the majority of failure cases for intervention.

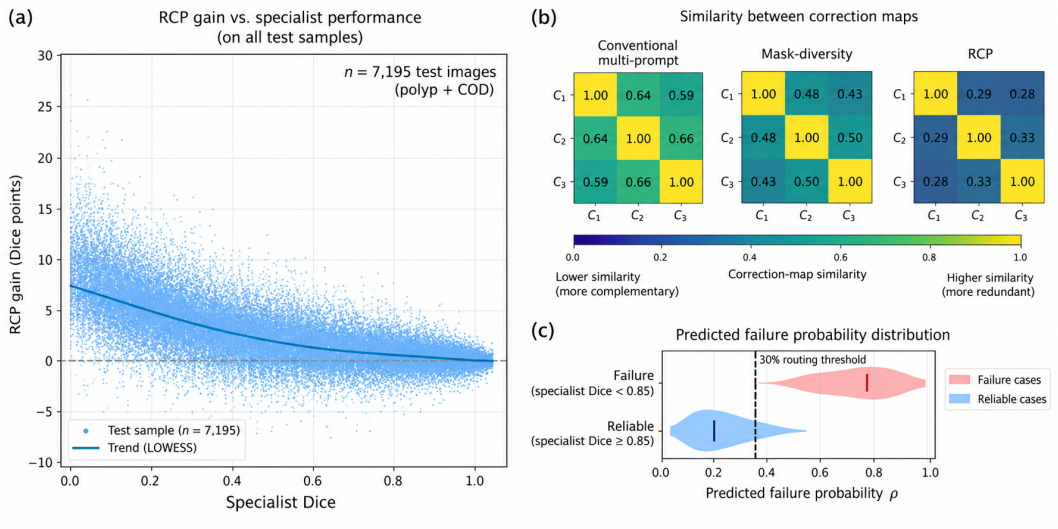

Figure 3: Mechanistic analysis of the proposed framework. (a) Relationship between specialist prediction quality and the gain obtained from RCP; larger improvements concentrate on lower-quality specialist predictions. (b) Pairwise similarity between specialist-relative correction maps for conventional multi-prompting, mask-diversity prompting, and RCP; weaker off-diagonal similarity indicates less redundant correction. (c) Distribution of predicted failure probabilities for reliable and failed specialist predictions, with the dashed line indicating the threshold associated with the retrospective 30% budget-controlled analysis.

A.4 Generalization, Efficiency, and Robustness

We further examine whether RCP depends on a particular specialist and quantify the computational effect of selective intervention in Table 7. Panel (a) shows consistent gains for PraNet, DINOv3-S/16, and DINOv3-B/16, indicating that the proposed complementary-intervention mechanism is not specific to the primary specialist. Panel (b) shows that applying RCP to every input requires three foundation-model hypotheses per image, whereas failure-aware routing reduces this average to 0.9 hypotheses and 115 ms per image while retaining nearly all of the segmentation improvement.

Statistical robustness. Across three random seeds, RCP remains stable at $.926 \pm .002$ Dice, compared with $.889 \pm .003$ for conventional multi-prompting and $.892 \pm .002$ for mask-diversity prompt-

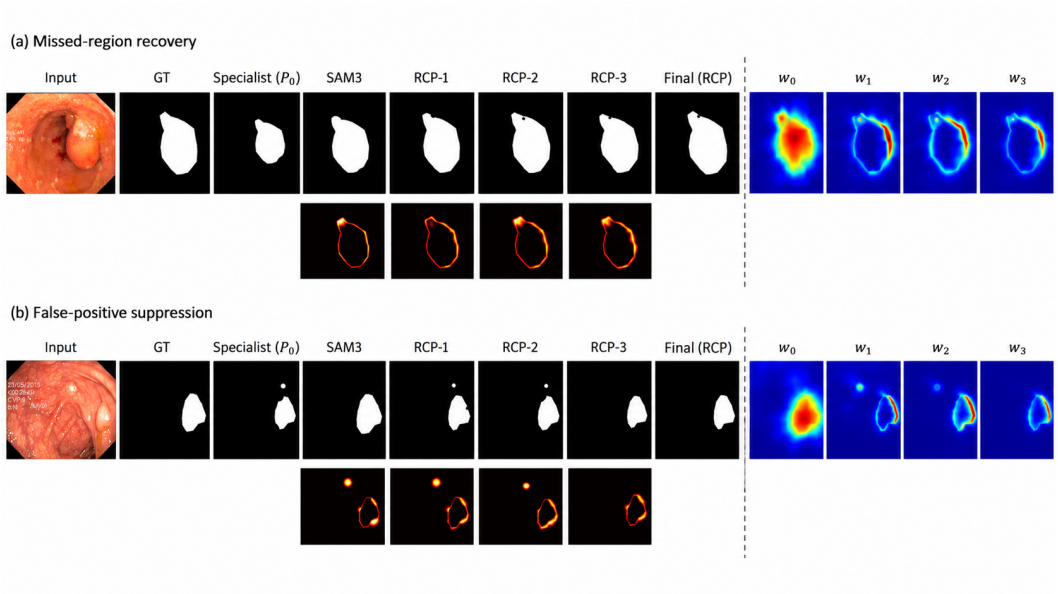

Figure 4: Spatial analysis of RCP correction and hypothesis arbitration. The red–yellow maps indicate where direct SAM3 and each RCP hypothesis reduce the specialist's error, while $w_0, \ldots, w_3$ show the spatial weights assigned to the specialist and the three RCP hypotheses.

Table 8: Additional domain-standard segmentation metrics. Higher is better for mIoU, sensitivity, specificity, S_α, F_β^w, and E_ϕ; lower is better for MAE. Results are averaged over the five polyp benchmarks and three camouflaged-object benchmarks, respectively.

(a) Polyp segmentation

Method	mIoU ↑	MAE ↓	Sens. ↑	Spec. ↑
Specialist	.790	.035	.889	.980
SAM3	.747	.043	.858	.975
Multi-Prompt + Arbiter	.801	.032	.897	.982
Mask-Diversity + Arbiter	.806	.030	.901	.983
RCP + Arbiter	**.864**	**.021**	**.937**	**.989**

(b) Camouflaged-object segmentation

Method	S_α ↑	F_β^w ↑	E_ϕ ↑	MAE ↓
Specialist	.873	.823	.915	.034
SAM3	.856	.799	.901	.040
Multi-Prompt + Arbiter	.880	.835	.921	.032
Mask-Diversity + Arbiter	.883	.840	.924	.031
RCP + Arbiter	**.903**	**.875**	**.941**	**.024**

ing. The variation across runs is substantially smaller than the improvement introduced by RCP, indicating that the observed gains are not driven by a particular initialization.

A.5 Additional Evaluation Metrics

To verify that the improvements are not specific to Dice, we additionally evaluate the compared methods using standard domain-specific segmentation metrics. For polyp segmentation, we report mean intersection-over-union (mIoU), mean absolute error (MAE), sensitivity, and specificity. For camouflaged-object segmentation, we report structure measure S_α, weighted F-measure F_β^w, enhanced-alignment measure E_ϕ, and MAE.

Across both domains, the improvements persist under metrics beyond Dice. On polyp segmentation, RCP improves average mIoU from 0.790 to 0.864 over the specialist while reducing MAE from 0.035 to 0.021, together with gains in sensitivity and specificity. On camouflaged-object segmentation, RCP improves S_α, F_β^w, and E_ϕ from 0.873, 0.823, and 0.915 to 0.903, 0.875, and 0.941, respectively, while reducing MAE from 0.034 to 0.024. These results indicate that the improvement extends beyond region overlap to structural quality, alignment, and pixel-wise error.

Table 9: Retrospective failure-aware routing on the pooled polyp evaluation set. At each budget, the highest-risk fraction of test images is routed; the fixed-threshold deployment result is reported separately.

Intervention	Failure Recall	Dice	Avg. Hyp./Image
0%	0.0%	.882	0.00
10%	58.1%	.906	0.30
20%	81.7%	.921	0.60
30%	**92.4%**	**.926**	**0.90**
50%	98.5%	.928	1.50
100%	100.0%	.929	3.00

A.6 Additional Qualitative Results

Figure 4 provides a more detailed view of how RCP hypotheses and the spatial arbiter contribute to the final prediction. Unlike the qualitative comparison in the main paper, which focuses on the resulting segmentation masks, this analysis visualizes both specialist-relative correction maps and the corresponding spatial arbitration weights. The red–yellow correction maps indicate where each foundation-model hypothesis reduces the residual error of the specialist, while $w_0, \ldots, w_3$ show the spatial weights assigned to the specialist and the three RCP hypotheses.

In Fig. 4(a), the specialist misses part of the polyp region. The RCP hypotheses recover different portions of this residual, as reflected by their distinct correction maps. The arbiter preserves the specialist in regions where its prediction is reliable while assigning larger weights to complementary hypotheses around the unresolved region. In Fig. 4(b), the specialist contains a detached false-positive component. The generated hypotheses suppress this erroneous region to different degrees, and the corresponding arbiter weights emphasize the alternatives that provide useful local corrections. These examples illustrate how RCP first exposes complementary corrective information and how the spatial arbiter subsequently selects among the available hypotheses at different image locations.

A.7 Retrospective Intervention Budgets

Table 9 reports failure recall, Dice, and average hypotheses per image as the retrospective intervention budget varies. Unlike fixed-threshold routing, each budget routes the highest-risk fraction of evaluation images.