MEDCORE:BOUNDARY-PRESERVING MEDICAL CORE PRUNING FOR MEDSAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical segmentation foundation models such as SAM and MedSAM provide strong prompt-driven segmentation, but their image encoders are still too large for many clinical settings. Compression is also risky in medicine because a model can keep high Dice while losing boundary fidelity. We propose **MedCore**, a structured pruning framework for MedSAM. The main idea is to preserve two kinds of structures: structures that became important during SAM-to-MedSAM adaptation, and structures that have high boundary leverage. We identify the first type by a dual-intervention score comparing zeroing a group with resetting it to its original SAM weights. We identify the second type by boundary-aware Fisher estimation. We also introduce a boundary leverage principle, which relates first-order boundary displacement to boundary logit perturbation divided by the spatial gradient norm. This principle explains why boundary metrics can degrade even when Dice remains high. On polyp segmentation benchmarks, MedCore reduces parameters by 60.0% and FLOPs by 58.4% while achieving Dice 0.9549, Boundary F1 0.6388, and HD95 5.14 after post-pruning fine-tuning. It also reaches **86.6%** parameter reduction and **90.4G** FLOPs, retaining Dice and HD95 near the fine-tuned MedSAM reference, with lower BF1. Our analysis shows that MedSAM lies in a head-fragile boundary regime: the median 95th-percentile boundary leverage is 2.887 times as high for head-pruning as for MLP-pruning steps, consistent with BF1 and HD95 degradation. Our code is available at https://anonymous.4open.science/r/MedCore-8461.

1 INTRODUCTION

Segment Anything Model (SAM) changed image segmentation by turning segmentation into a prompt-driven task [17]. MedSAM adapts this idea to medical images and gives strong results across many modalities [25]. These models are useful because they can accept prompts such as bounding boxes and produce high-quality masks. Their cost is still high. The MedSAM image encoder is a ViT-based model [8], and its inference cost can be difficult to support in real-time endoscopy, point-of-care ultrasound, and other resource-limited settings.

A natural solution is model compression. Yet medical segmentation has a failure mode that ordinary compression metrics can hide. A compressed model may still identify the coarse lesion region and keep a high Dice score, while its predicted boundary becomes thick, broken, or shifted. This is a serious issue because boundary quality affects lesion measurement, treatment planning, and clinical interpretation. This paper studies structured pruning for MedSAM under a boundary-sensitive setting, directly pruning the existing adapted model rather than training a lightweight surrogate. Since MedSAM is adapted from SAM instead of trained from scratch, we exploit the transition from the SAM checkpoint $\boldsymbol{\theta}^S$ to the MedSAM checkpoint $\boldsymbol{\theta}^M$ to identify which structures remain general-purpose and which become medical-specific.

We propose **MedCore**, a boundary-preserving medical core pruning framework that keeps a compact yet medically adapted MedSAM subnetwork. For each structure group g, MedCore uses a dual-intervention score that combines its current prediction contribution under zeroing and its medical-adaptation contribution under reset to the original SAM weights, which are two counterfactual questions. A group is safe to prune only when both signals suggest low importance.

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.

We further introduce a boundary leverage principle by viewing the predicted mask boundary as a level set of the logit map. If pruning changes the logit map by $\delta_G(\boldsymbol{x})$, the first-order boundary displacement is proportional to $\delta_G(\boldsymbol{x})$ and inversely proportional to the spatial gradient norm $\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2$. This gives a direct geometric explanation of boundary collapse. Compression is dangerous when it creates a large logit perturbation on a boundary where the logit slope is small. We use this principle to motivate boundary-aware Fisher scoring and to analyze the different roles of attention heads and MLP connections.

In summary, Our three contributions are as follows:

- We formulate MedSAM pruning as medical-core extraction and introduce a dual-intervention score based on zeroing and reset-to-SAM operations.
- We propose boundary-aware Fisher and Cross-Fisher approximations to make the intervention score practical and boundary-sensitive.
- We introduce boundary leverage as a simple geometric principle for explaining Dice-boundary mismatch and head-MLP pruning trade-offs.

Our results further show that this distinction matters. Experiments on polyp endoscopy, breast ultrasound, and dermoscopy demonstrate that MedCore can remove a large fraction of MedSAM parameters while preserving strong region and boundary performance. In a 7×7 head-MLP pruning sweep, head-pruning steps exhibit $2.887\times$ larger 95th-percentile boundary leverage than MLP-pruning steps, with all 15 valid paired comparisons favoring MLP pruning. At the metric level, the BF1 and HD95 damage densities of head pruning are respectively $2.607\times$ and $2.432\times$ those of MLP pruning, indicating that MedSAM lies in a head-fragile boundary regime: attention heads are not merely low-parameter components, but can have high boundary leverage.

2 Related Work

Promptable segmentation and medical SAM. SAM formulates segmentation as a promptable task using a ViT image encoder, a prompt encoder, and a lightweight mask decoder [17, 8]. MedSAM adapts this design to medical images via large-scale image-mask pairs [25]. Subsequent work improves medical adaptation along several axes: parameter-efficient tuning (SAMed, SAM-Med2D, Medical-SAM-Adapter [44, 6, 41]; Polyp-SAM for transfer to polyp segmentation [21]), boundary-aware feature design (I-MedSAM with implicit representations [40]; LDFSAM with distilled feature prompting [46]), 3D/video extension (MedSAM2 [26]), and feature fusion with task-specific networks [47, 28]. These methods focus on prompting, fusion, or boundary localization. We study a complementary question: how to *compress* an already adapted MedSAM checkpoint while preserving the structures that became important during SAM-to-MedSAM adaptation.

Efficient SAM and medical segmentation models. The heavy ViT image encoder of SAM has stimulated extensive research on efficient segmentation, including lightweight encoder replacement, knowledge distillation, quantization, and structured pruning [45, 42, 48, 5, 24]. Medical settings also have efficient segmentation backbones, including U-Net, nnU-Net, Swin-Unet, EMCAD, MK-UNet, and quantized medical SAM variants [31, 13, 4, 30, 29, 24].Recent lightweight polyp models, such as HSSAM-Net, further show that careful multi-scale aggregation and boundary-aware enhancement can give strong accuracy with small parameter budgets [9]. These works confirm the need for efficient segmentation. Most of them redesign the architecture, train a compact surrogate, or quantize a model. MedCore instead directly prunes MedSAM and uses both the original SAM weights $\boldsymbol{\theta}^S$ and the adapted MedSAM weights $\boldsymbol{\theta}^M$.

Pruning and Transformer compression. Classic pruning methods estimate the loss increase caused by removing weights or groups. Optimal Brain Damage and Optimal Brain Surgeon use second-order information to approximate this increase [18, 12]. Deep Compression combines pruning, quantization, and coding [11]. Later methods use movement, Fisher information, Hessian-aware saliency, or group-level sensitivity to prune pretrained and Transformer models [32, 23, 43]. Attention-head pruning studies show that many heads can be redundant, but some specialized heads are important and task-dependent [27, 38]. This literature motivates structured pruning, but it usually estimates importance from one endpoint checkpoint.

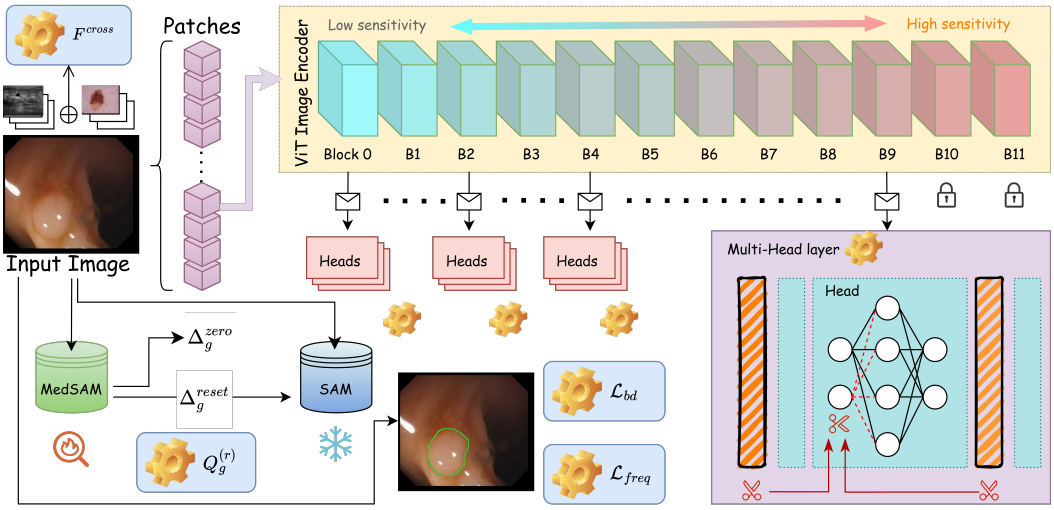

Figure 1: MedCore pruning overview. The Cross-Fisher signal F^{cross} combines SAM and MedSAM gradients to score attention heads and MLP channels via $Q_g^{(r)}$, driving block-sensitive cascade pruning of the ViT encoder (blocks B10–B11 protected) with boundary- and frequency-aware recovery. Yellow gears mark where pruning operations are applied: Cross-Fisher estimation, priority computation, per-block head pruning, and recovery-loss-driven stabilization.

Boundary preservation in medical segmentation. Medical segmentation models often optimize region losses such as Dice and cross-entropy. These losses can keep coarse foreground overlap but still miss fine contours. To address this limitation, prior work has explored boundary losses, surface- or Hausdorff-inspired objectives, topology-aware regularization, and boundary-aware architectures for more accurate contour modeling [16, 15, 34, 33, 19, 22, 36]. Different from these task-specific segmentation methods, our work brings boundary preservation into model compression by asking not only whether pruning preserves Dice, but also whether it perturbs boundary-critical structures and the boundary logit field.

3 MedCore: Boundary-Preserving Medical Core Pruning

MedCore has two design goals. The first goal is to preserve structures that became important during medical adaptation. The second goal is to preserve structures whose removal causes large boundary displacement.

3.1 Problem Formulation

A training or calibration sample is written as$\boldsymbol{z} = (I, \boldsymbol{b}, M)$, where $I \in \mathbb{R}^{3\times H\times W}$ is a medical image, $\boldsymbol{b} \in \mathbb{R}^4$ is a bounding-box prompt, and $M \in \{0,1\}^{H\times W}$ is a binary mask. MedSAM maps $(I, \boldsymbol{b})$ to a logit map $s_{\boldsymbol{\theta},\boldsymbol{\phi}}(I, \boldsymbol{b})$ and a probability mask $\widehat{M} = \sigma(s_{\boldsymbol{\theta},\boldsymbol{\phi}}(I, \boldsymbol{b}))$. The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_E, \boldsymbol{\theta}_D)$ contains the image encoder and mask decoder. The prompt encoder parameter $\boldsymbol{\phi}$ is frozen.

Given a calibration set $\mathcal{D}_{\text{cal}} = \{\boldsymbol{z}_n\}_{n=1}^N$, we define the empirical medical risk as

$$\widehat{\mathcal{R}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \ell_{\text{seg}}(\boldsymbol{z}_n; \boldsymbol{\theta}), \qquad \ell_{\text{seg}} = \ell_{\text{Dice}} + \ell_{\text{BCE}}. \tag{1}$$

We partition pruneable parameters into structured groups $\mathcal{G} = \{g_1, \ldots, g_K\}$. In our implementation, groups are attention heads and MLP connection groups in the ViT image encoder. A binary group mask $\boldsymbol{s} \in \{0,1\}^K$ gives the masked parameter $\boldsymbol{s} \odot_{\mathcal{G}} \boldsymbol{\theta}$. If c_g is the cost of group g, the ideal pruning problem is

$$\min_{\boldsymbol{\theta},\boldsymbol{s}} \widehat{\mathcal{R}}(\boldsymbol{s} \odot_{\mathcal{G}} \boldsymbol{\theta}) \quad \text{s.t.} \quad \sum_{g\in\mathcal{G}} s_g c_g \le B. \tag{2}$$

This problem is combinatorial. MedCore approximates it by scoring groups and removing low-score groups under a block-aware budget. We keep two checkpoints: $\boldsymbol{\theta}^S$ is the original SAM checkpoint, and $\boldsymbol{\theta}^M$ is the MedSAM checkpoint. Pruning starts from $\boldsymbol{\theta}^M$, while $\boldsymbol{\theta}^S$ serves as the adaptation reference.

3.2 Boundary Leverage Principle

We now state the geometric principle that motivates boundary-aware pruning. For a fixed image and prompt, we write the logit map as $s_{\boldsymbol{\theta}}(\boldsymbol{x})$, where $\boldsymbol{x}$ indexes a spatial location. The predicted boundary is the zero level set of this logit map. A compression operation G changes the model to $\boldsymbol{\theta}^{-G}$ and induces the logit perturbation $\delta_G(\boldsymbol{x}) = s_{\boldsymbol{\theta}^{-G}}(\boldsymbol{x}) - s_{\boldsymbol{\theta}}(\boldsymbol{x})$. The following assumption is standard in level-set perturbation analysis. It only says that the boundary is not locally flat in the logit field.

Assumption 3.1 (Regular boundary)**.** For a given image and prompt, the logit map $s_{\boldsymbol{\theta}}$ is continuously differentiable in a neighborhood of its zero level set $\Gamma_{\boldsymbol{\theta}} = \{\boldsymbol{x} : s_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0\}$. There exists $\kappa > 0$ such that $\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 \geq \kappa$ for every $\boldsymbol{x} \in \Gamma_{\boldsymbol{\theta}}$.

Definition 3.2 (Boundary leverage)**.** Let G be a compression operation with compression gain $\Delta C_G > 0$. Let $\mathcal{B}$ be a narrow boundary band around a ground-truth or predicted boundary. We define the boundary leverage of G as

$$\lambda_G^{\mathrm{bd}} = \frac{1}{\Delta C_G} \mathbb{E}_{\boldsymbol{x} \in \mathcal{B}} \left[\frac{|\delta_G(\boldsymbol{x})|}{\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 + \epsilon} \right]. \tag{3}$$

Here $\epsilon > 0$ is a small constant. This quantity estimates the boundary displacement caused by one unit of compression.

Theorem 3.3 (First-order boundary perturbation)**.** *Under Assumption 3.1, suppose that δ_G is sufficiently small in a neighborhood of $\Gamma_{\boldsymbol{\theta}}$. For each $\boldsymbol{x} \in \Gamma_{\boldsymbol{\theta}}$, let $u_G(\boldsymbol{x})$ be the signed normal displacement from the old boundary to the new boundary induced by G. Then*

$$u_G(\boldsymbol{x}) = -\frac{\delta_G(\boldsymbol{x})}{\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2} + O(\|\delta_G\|_{C^1}^2). \tag{4}$$

Theorem 3.3 explains why boundary metrics can fail before Dice fails. Boundary displacement depends on the logit perturbation at the boundary and on the local logit slope. A group with small parameter count can still be dangerous if it causes a large boundary logit perturbation. We prove the theorem in Appendix A.

Proposition 3.4 (Boundary-aware budget rule)**.** *Let c_H and c_M be compression budgets assigned to heads and MLPs, and let $\mathcal{E}_{\mathrm{bd}}(c_H, c_M)$ be a locally differentiable boundary error. If*

$$\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_H} > \frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_M}$$

at a feasible interior allocation, then moving a small amount of compression budget from head pruning to MLP pruning decreases boundary error to first order while keeping the total compression fixed.

Proposition 3.4 turns boundary leverage into a budget principle. It does not say that heads are always more important than MLPs. It says that the safer compression direction is the one with lower marginal boundary damage. Section 5.2 verifies that, for MedSAM, head-pruning steps have much higher boundary leverage than MLP-pruning steps.

3.3 Medical-Core Scoring via Dual Intervention

The boundary leverage principle says which structures are dangerous for boundaries. We also need to preserve structures that became important during medical adaptation. MedSAM has a useful reference checkpoint, namely the original SAM checkpoint $\boldsymbol{\theta}^S$. We therefore score each group g by two counterfactual interventions.

The zero intervention sets group g to zero while keeping all other groups unchanged. If T_g^0 denotes this operation, the exact zero cost is

$$\Delta_g^{\mathrm{zero}} = \widehat{\mathcal{R}}(\boldsymbol{s} \odot_{\mathcal{G}} T_g^0(\boldsymbol{\theta}^M)) - \widehat{\mathcal{R}}(\boldsymbol{s} \odot_{\mathcal{G}} \boldsymbol{\theta}^M). \tag{5}$$

A small value means that g contributes little to current MedSAM predictions.

The reset intervention replaces group g in MedSAM by its SAM value. If T_g^S denotes this operation, the exact reset cost is

$$\Delta_g^{\text{reset}} = \widehat{\mathcal{R}}(\boldsymbol{s} \odot_{\mathcal{G}} T_g^S(\boldsymbol{\theta}^M)) - \widehat{\mathcal{R}}(\boldsymbol{s} \odot_{\mathcal{G}} \boldsymbol{\theta}^M). \tag{6}$$

A large reset cost means that the change from $\boldsymbol{\theta}_g^S$ to $\boldsymbol{\theta}_g^M$ is functionally important for medical segmentation. This signal is different from the zero cost. It measures adaptation-specific importance rather than only endpoint importance. We combine both costs as

$$Q_g = \alpha_{b(g)} \widehat{\Delta}_g^{\text{zero}} + (1 - \alpha_{b(g)}) \widehat{\Delta}_g^{\text{reset}}, \tag{7}$$

where $b(g)$ is the Transformer block containing g. We use a block-level mixture weight because different blocks respond differently to boundary and adaptation signals. The pruning priority is

$$P_g = \frac{Q_g}{(c_g + \epsilon)^\tau}, \tag{8}$$

where $\tau \geq 0$ controls cost normalization. Groups with smaller P_g are pruned first.

3.4 Boundary-Aware Fisher and Cross-Fisher Approximation

Computing equation 5 and equation 6 exactly for every group is expensive. We use a local second-order approximation. Around $\boldsymbol{\theta}^M$, zeroing group g corresponds to the perturbation $-\boldsymbol{\theta}_g^M$, and resetting group g corresponds to the perturbation $\boldsymbol{\theta}_g^S - \boldsymbol{\theta}_g^M$. We approximate the Hessian by diagonal Fisher information.

The loss used for Fisher estimation should reflect the metric we want to preserve. We therefore use a boundary-weighted loss. Let $B(M)$ be a morphological boundary map computed by dilation minus erosion. With pixel index u, we define

$$\ell_{\text{bd}}(\boldsymbol{z}; \boldsymbol{\theta}) = \sum_u (1 + \lambda_{\text{bd}} B(M)_u)\, \ell_{\text{BCE}}(s_{\boldsymbol{\theta},\boldsymbol{\phi}}(I, \boldsymbol{b})_u, M_u) + \ell_{\text{Dice}}(\widehat{M}, M). \tag{9}$$

For a sub-distribution r, such as one dataset or modality, the MedSAM Fisher is

$$F_{r,i}^M = \frac{1}{|\mathcal{D}_{\text{cal}}^{(r)}|} \sum_{\boldsymbol{z}_n \in \mathcal{D}_{\text{cal}}^{(r)}} \left(\frac{\partial \ell_{\text{bd}}(\boldsymbol{z}_n; \boldsymbol{\theta}^M)}{\partial \theta_i} \right)^2. \tag{10}$$

This gives the zero approximation

$$\widehat{\Delta}_g^{\text{zero},(r)} = \frac{1}{2} \sum_{i \in g} F_{r,i}^M (\theta_i^M)^2.$$

For the reset cost, the perturbation is the SAM-to-MedSAM parameter change. We also compute the SAM Fisher F_i^S on the same calibration data and use a geometric mean,

$$F_{r,i}^{\text{cross}} = \sqrt{F_{r,i}^M F_i^S + \epsilon_F}. \tag{11}$$

The reset approximation becomes

$$\widehat{\Delta}_g^{\text{reset},(r)} = \frac{1}{2} \sum_{i \in g} F_{r,i}^{\text{cross}} (\theta_i^M - \theta_i^S)^2. \tag{12}$$

The geometric mean down-weights parameters that are insensitive in either SAM or MedSAM. This prevents the reset score from being dominated by large but functionally irrelevant weight shifts.

3.5 Distribution-Aware Aggregation and Budget Allocation

Medical data are heterogeneous. A structure that appears redundant on one dataset may be important on another dataset. We compute per-distribution scores $\widehat{Q}_g^{(r)}$ using equation 7 and aggregate them as

$$\widehat{Q}_g^{\text{dist}} = \sum_{r=1}^R \pi_r \widehat{Q}_g^{(r)} + \beta \operatorname{Var}_r[\widehat{Q}_g^{(r)}], \tag{13}$$

where $\pi_r \geq 0$ and $\sum_r \pi_r = 1$. The variance term makes pruning more conservative for groups that are important on only some sub-distributions.

We also allocate pruning budgets non-uniformly across Transformer blocks. Let $\mathcal{B}_\ell$ be the parameter set of block ℓ. We estimate block sensitivity by

$$S_\ell = \sum_{i \in \mathcal{B}_\ell} F_i, \qquad F_i = \sum_r \pi_r F_{r,i}^M.$$

Blocks with high S_ℓ receive smaller pruning quotas. In the main one-time pruning experiments, the deepest blocks are protected because they are closest to the mask decoder and have the highest Fisher sensitivity. In the extreme sequential setting, we apply a second conservative pruning pass to previously protected blocks after the first checkpoint has been selected.

3.6 Head-to-MLP Cascade and Lightweight Recovery

MedCore prunes in a cascade. We first prune attention heads. Then we run a short recovery phase. We then prune MLP connection groups from the head-pruned model. Order ablations show a slight advantage for head-to-MLP pruning (Appendix B), supporting this default for further mechanism analysis.

During pruning, we recover the remaining weights on the calibration set with a lightweight objective:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{bd}} + \lambda_2 \mathcal{L}_{\text{feat}} + \lambda_3 \mathcal{L}_{\text{logit}} + \lambda_4 \mathcal{L}_{\text{freq}}. \tag{14}$$

Here $\mathcal{L}_{\text{bd}}$ is boundary-weighted BCE, $\mathcal{L}_{\text{feat}}$ distills encoder features from the unpruned MedSAM teacher, $\mathcal{L}_{\text{logit}}$ distills boundary-region logits, and $\mathcal{L}_{\text{freq}}$ penalizes high-frequency mask discrepancies. Recovery is not used to define the pruning score. It lets the remaining structures re-coordinate after structural removal.

4 Experiments

4.1 Experimental Setup

We evaluate MedCore on the official MedSAM ViT-B checkpoint, using SAM ViT-B as the reference and freezing the prompt encoder. Experiments cover polyp endoscopy (CVC-ClinicDB [2], CVC-ColonDB [39], Kvasir-SEG [14]), breast ultrasound (BUSI [1]), dermoscopy (ISIC2018 [37, 7]), and multi-class cardiac MRI (ACDC Heart MR [3]). Calibration and implementation details are provided in Appendix D.

We then report Dice, Intersection over Union (IoU), Boundary F1 (BF1), 95th-percentile Hausdorff distance (HD95), parameters, and FLOPs. BF1 and HD95 measure boundary fidelity. Recovery denotes lightweight calibration-set stabilization during pruning; fine-tuning denotes post-pruning training on a target dataset. Table 1 also includes unpruned MedSAM fine-tuned with the same protocol. The one-time component ablations in Section 5.1 omit post-pruning fine-tuning to isolate component effects.

4.2 Main Results on Polyp Segmentation

Table 1 reports macro-averaged results on the three polyp benchmarks. MedCore reaches strong compression while preserving boundary quality after fine-tuning. The h50_m70 configuration removes 60.0% of parameters and reduces FLOPs by 58.4%, while achieving Dice 0.9549, BF1 0.6388, and HD95 5.14. The h70_m95 configuration removes 72.2% of parameters and still obtains BF1 0.6259 and HD95 6.78. The sequential configuration h84_m95 reaches 12.1M parameters and 90.4G FLOPs, while keeping strong region and boundary metrics.

Fine-tuned MedSAM provides the full-model reference under the same fine-tuning protocol. Relative to this reference, sequential h84_m95 has 0.0028 lower Dice, 0.0347 lower BF1, and 0.01 higher HD95. This configuration retains comparable Dice and HD95 under heavy compression, while the BF1 gap reflects the trade-off in boundary preservation.

Table 1: Macro-averaged results on CVC-ClinicDB [2], CVC-ColonDB [39], and Kvasir-SEG [14]. MedSAM + fine-tuning and MedCore use the same fine-tuning protocol. One-time uses one head-to-MLP cascade; Sequential adds a second pruning pass to h70_m95. Bold marks the best values within MedCore.

Category	Method	Params↓	FLOPs↓	Dice↑	IoU↑	BF1↑	HD95↓
Baseline	MedSAM [25]	89.7M	926.5G	0.9191	0.8648	0.5321	21.29
	MedSAM + fine-tuning	89.7M	926.5G	0.9578	0.9201	0.6809	5.11
Efficient SAM	EfficientSAM-s [42]	26.4M	188.0G	0.8765	0.8131	0.4340	29.73
	EfficientSAM-t [42]	10.2M	55.9G	0.8723	0.8103	0.4218	25.98
	SlimSAM [5]	28.0M	189.9G	0.8435	0.7758	0.3922	34.11
Medical Seg.	Swin-Unet [4]	27.2M	5.92G	0.7845	0.6933	0.4347	23.88
	nnU-Net [31, 13]	46.3M	75.1G	0.9093	0.8574	0.5382	32.15
	EMCAD [30]	26.8M	10.6G	0.9141	0.8608	0.5305	30.31
	MK-UNet [29]	0.32M	0.61G	0.8658	0.7986	0.4635	46.21
Medical SAM	SAMed [44]	91.2M	206.7G	0.7419	0.6640	0.3283	97.57
	QMedSAM [24]	9.79M	75.7G	0.8089	0.7108	0.4118	14.55
MedCore One-time	h40_m30	61.3M	620.7G	0.9547	0.9173	**0.6508**	7.23
	h50_m70	35.8M	385.2G	0.9549	0.9169	0.6388	5.14
	h70_m70	30.1M	287.5G	0.9524	0.9130	0.6342	6.23
	h70_m95	24.9M	245.0G	0.9522	0.9124	0.6259	6.78
MedCore Sequential	h60_m70	15.6M	132.6G	0.9547	**0.9176**	0.6506	5.91
	h72_m84	13.7M	112.4G	0.9536	0.9153	0.6337	5.47
	h84_m95	**12.1M**	**90.4G**	**0.9550**	0.9174	0.6462	**5.12**

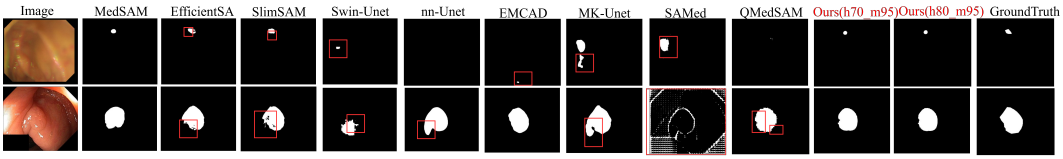

Figure 2: Qualitative segmentation comparison between our two high-pruning configurations and other models. Incorrect segmentation regions and boundary artifacts are highlighted with red boxes.

4.3 Cross-Modality Generalization

Table 2 evaluates the same pruning pipeline on CVC-ClinicDB, BUSI, ISIC2018, and ACDC Heart MR. On ClinicDB and ISIC2018, MedCore improves boundary metrics after fine-tuning. On BUSI, moderate compression preserves performance closely, while aggressive compression begins to degrade BF1 and HD95 (Appendix Table 6). This pattern is useful because it shows both the strength and the limit of direct cross-modality transfer. MedCore is not a replacement for target-domain validation. It is a compression framework that can preserve strong performance when the calibration and fine-tuning data are representative.

5 Analysis and Ablation

5.1 Component Ablation: What Drives Boundary Preservation?

Table 3 ablates the main components. The one-time columns report pruning results before post-pruning fine-tuning to isolate the structural effect of pruning decisions. The sequential column uses the same post-pruning fine-tuning protocol for all variants.

Boundary-aware Fisher provides the most direct contour protection. Removing it reduces BF1 across the reported settings, with the largest reduction at the most aggressive one-time setting h70_m95. Reset-to-SAM scoring contributes less at moderate compression and becomes important as the pruning budget becomes severe. This pattern supports its role as a safeguard for medical adaptation

Table 2: Cross-modality results with post-pruning fine-tuning. Params/FLOPs: 89.7M/926.5G for MedSAM and 24.9M/240.9G for MedCore h70_m95.

Method	Metric	CVC-ClinicDB [2]	BUSI [1]	ISIC2018 [37, 7]	ACDC Heart MR [3]			
					Myocardium	LV	RV	Average
MedSAM	Dice↑	0.9264	0.9334	0.9314	0.7934	0.9235	0.9577	0.8915
	BF1↑	0.5477	0.4989	0.2401	0.7685	0.8232	0.8805	0.8240
	HD95↓	13.77	7.30	7.47	7.20	2.58	0.87	3.55
MedCore h70_m95	Dice↑	0.9523	0.9298	0.9530	0.8173	0.9069	0.9463	0.8902
	BF1↑	0.6194	0.4703	0.4753	0.7121	0.7450	0.7949	0.7507
	HD95↓	6.30	9.43	5.86	5.87	4.65	1.51	4.01

Table 3: Component ablation on macro-average BF1 across the three polyp benchmarks. One-time columns are reported without post-pruning fine-tuning. The sequential column uses the same post-pruning fine-tuning for all variants. Δ is the difference from the full method at the same compression level.

	h40_m30		h50_m70		h70_m70		h70_m95		Seq. h84_m95	
Method	BF1	Δ	BF1	Δ	BF1	Δ	BF1	Δ	BF1	Δ
Full MedCore	0.5162	–	0.4078	–	0.3144	–	0.4050	–	0.6462	–
w/o Boundary Fisher	0.5035	-0.013	0.3724	-0.035	0.3096	-0.005	0.2781	-0.127	0.6304	-0.016
w/o reset	0.5149	-0.001	0.3967	-0.011	0.3628	+0.048	0.3227	-0.082	0.6279	-0.018
w/o variance	0.5221	+0.006	0.3997	-0.008	0.3812	+0.067	0.3249	-0.080	0.6462	0.000

under aggressive compression. The distribution variance term similarly protects structures whose importance varies across datasets. Its contribution becomes more useful when fewer structures remain to support predictions across the different datasets.

These results support a hierarchy of component roles within MedCore. Boundary Fisher provides the most direct protection for boundary-core structures, while reset-to-SAM scoring preserves medical-core structures created by adaptation. Variance aggregation complements these signals by retaining groups with uneven importance across datasets. Together, the components couple contour protection with safeguards for medical adaptation under aggressive compression.

5.2 Boundary Leverage Reveals a Head-Fragile Regime

We test the empirical prediction in Definition 3.2 using a 7×7 sweep over $h \in \{0.3, \ldots, 0.9\}$ and $m \in \{0.3, 0.5, 0.7, 0.8, 0.85, 0.9, 0.95\}$, with 384 calibration images (128 each from Kvasir, ClinicDB, ColonDB). For each adjacent head step $G_H(i,j)$:$(h_i, m_j) \to (h_{i+1}, m_j)$ and MLP step $G_M(i,j)$:$(h_i, m_j) \to (h_i, m_{j+1})$, we compute the 95th-percentile version of Eq. equation 3 on the boundary band, skipping steps with zero parameter reduction.

Table 4 reports median $\lambda_{95}^{\mathrm{bd}}$ values of 3.961 for head steps and 1.372 for MLP steps, giving a head/MLP ratio of 2.887. Head steps have higher leverage in all 15 valid paired comparisons, supporting the empirical law $\mathrm{Median}(\lambda_G^{\mathrm{bd}} \mid G \in \mathcal{H}) > \mathrm{Median}(\lambda_G^{\mathrm{bd}} \mid G \in \mathcal{M})$ for MedSAM.

The BF1 and HD95 rows confirm the connection between logit-level boundary leverage and final-mask degradation: head pruning incurs 2.4–2.6× greater BF1 and HD95 damage per one percent parameter reduction. Head pruning also has a high full-image perturbation ratio, indicating perturbations across both the boundary band and the wider image. These combined observations identify MedSAM heads as high-boundary-leverage components, making aggressive head pruning risky for boundary-sensitive segmentation.

5.3 Head-MLP Compression Landscape

The boundary leverage result explains how the head/MLP allocation shapes the observed compression landscape. At approximately 60% parameter reduction, h30_m80 yields BF1 0.376 while h50_m70 yields 0.408 before post-pruning fine-tuning, showing that boundary quality depends on structural

Table 4: Step-level boundary leverage analysis on the head-MLP pruning sweep. The logit-level boundary leverage result agrees with metric-level BF1 and HD95 damage densities.

Quantity	Head steps	MLP steps	Head/MLP ratio
Median $\lambda_{95}^{\mathrm{bd}}$	3.961	1.372	2.887
BF1 damage density	0.007692	0.002951	2.607
HD95 damage density	0.025406	0.010445	2.432
Valid steps	35	21	–
Paired win rate		15/15 = 1.000	

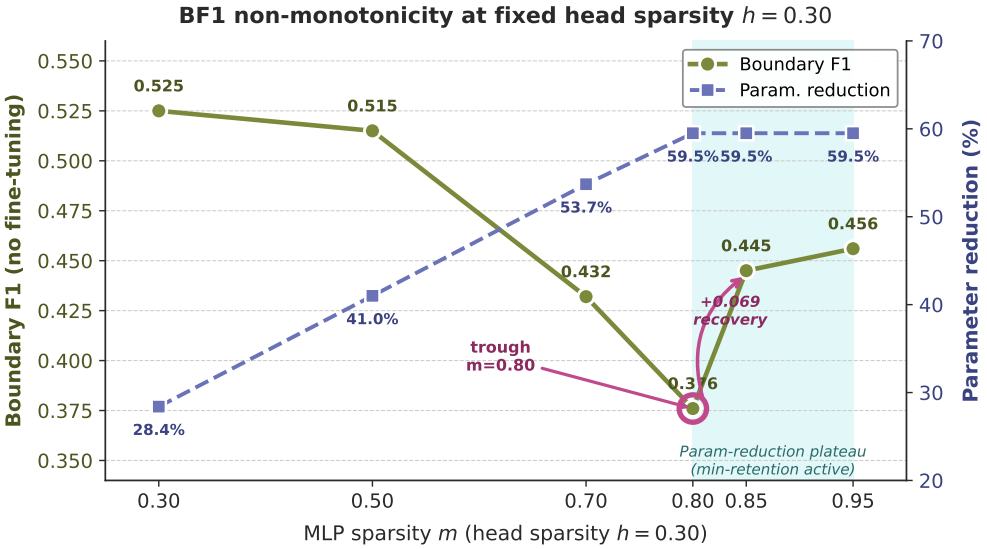

Figure 3: BF1 non-monotonicity at fixed head sparsity $h = 0.30$ (no post-pruning fine-tuning). BF1 drops to a trough of 0.376 at $m = 0.80$, then recovers to 0.456 at $m = 0.95$ along the 59.5% parameter-reduction plateau where the min-retention constraint is active.

allocation even at similar model sizes. Figure 3 examines a slice at fixed $h = 0.30$: BF1 initially decreases as m increases, then rises once the effective parameter count saturates at the min-retention plateau. This non-monotonic behavior shows that both block allocation and structure type determine the relationship between sparsity and boundary performance.

Boundary-aware allocation can therefore matter as much as nominal sparsity in choosing a compression configuration. The benefit of further sparsity depends on the resulting allocation of surviving structures. Together with Table 4, these results support conservative head pruning and assigning most compression to lower-leverage MLP structures, with calibration analysis guiding the head/MLP balance for each target setting.

6 Conclusion

We introduced MedCore, a structured pruning framework that treats MedSAM pruning as medical-core extraction. A reset-to-SAM intervention preserves structures that became important during SAM-to-MedSAM adaptation, while boundary-aware Fisher estimation protects contour-critical structures. Boundary leverage provides a geometric view of compression-induced boundary damage, explaining how high Dice coexists with degraded boundaries and why attention heads require conservative pruning in our setting. Experiments demonstrate substantial parameter and FLOPs reductions while retaining strong region and boundary performance across medical segmentation datasets. MedCore inherits MedSAM's underlying capabilities and biases, so clinical deployment requires task-specific validation and prospective safety evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, interpreting results, and generating synthetic data sets, cleaning and reformating dataset are not applicable to this work. Additionally, we used generative AI tools for organizing software code, brainstorming, sourcing/searching for information, and editing a research paper to improve readability. We have reviewed all AI-assisted work. We manually checked LLM-assisted proofs for correctness and consistency with the paper's core arguments. LLM-generated code was also verified and tested for correctness by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The research conducted in the paper conform, in every respect, with the Code of Ethics. The work uses public medical segmentation datasets and public model checkpoints for model compression research, and it does not involve deceptive data collection, unsafe data release, or human-subject interaction. We also discuss clinical validation and deployment risks in the impact statement. We have also checked and confirmed that this article does not involve potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics).

REPRODUCIBILITY STATEMENT

The paper describes the model structure, datasets, calibration sample sizes, pruning groups, pruning schedule, recovery setting, metrics, and algorithms in Sections 5.2- 5.1 and Appendices B- E. These details are sufficient to reproduce the main experimental pipeline. The experiments use public datasets and public SAM/MedSAM checkpoints, and the current submission provide an anonymized code release. However, we do not provide complete reproduction scripts or private checkpoints. We plan to release code and pruned checkpoints after acceptance, together with data preparation and evaluation instructions.

Besides, for details, Section 4.1 and Appendix E specify the datasets, fixed train/test split protocol, calibration size, prompt setup, optimizer, learning rates, batch sizes, recovery steps, post-pruning fine-tuning, and evaluation metrics.

REFERENCES

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2019.104863. URL https://www.sciencedirect.com/science/article/pii/S2352340919312181.

[2] 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. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2015.02.007. URL https://www.sciencedirect.com/science/article/pii/S0895611115000567.

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc

Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

[5] Zigeng Chen, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Slimsam: 0.1% data makes segment anything slim. *Advances in Neural Information Processing Systems*, 37:39434–39461, 2024.

[6] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023. URL https://arxiv.org/abs/2308.16184.

[7] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. URL https://arxiv.org/abs/1902.03368.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[9] Qing Feng, Shahzad Ahmed, Yueming Zhang, Lan He, and Muhammad Yaqub. HSSAM-Net: Hyper-scale shifted aggregation network for precise colorectal polyp segmentation in endoscopic images. *Scientific Reports*, 15:38146, 2025. doi: 10.1038/s41598-025-21954-y. URL https://www.nature.com/articles/s41598-025-21954-y.

[10] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks, 2015. URL https://arxiv.org/abs/1506.02626.

[11] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

[12] Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

[13] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[14] 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. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

[15] Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39 (2):499–513, 2020.

[16] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pp. 285–296. PMLR, 2019.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

[18] Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, 1989.

[19] Qin Li, Tianchi Zhang, Parvej Md Mosharaf, and Jing Zhang. BMFA-Net: Boundary constraint multi-level feature aggregation framework for precise polyp segmentation. *Applied Sciences*, 14(10):4063, 2024. doi: 10.3390/app14104063. URL https://www.mdpi.com/2076-3417/14/10/4063.

[20] Yawei Li, Kamil Adamczewski, Wen Li, Shuhang Gu, Radu Timofte, and Luc Van Gool. Revisiting random channel pruning for neural network compression, 2022. URL https://arxiv.org/abs/2205.05676.

[21] Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. Polyp-SAM: Transfer SAM for polyp segmentation. In *Medical imaging 2024: computer-aided diagnosis*, volume 12927, pp. 749–754. SPIE, 2024.

[22] Dingzhou Liu, Hongmin Deng, Zhengwei Huang, and Jinghao Fu. FCA-Net: Fully context-aware feature aggregation network for medical segmentation. *Biomedical Signal Processing and Control*, 91:106004, 2024. doi: 10.1016/j.bspc.2024.106004. URL https://www.sciencedirect.com/science/article/pii/S1746809424000624.

[23] Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group fisher pruning for practical network compression. In *International conference on machine learning*, pp. 7021–7032. PMLR, 2021.

[24] Haisheng Lu, Yujie Fu, Fan Zhang, and Le Zhang. Efficient quantization-aware training on segment anything model in medical images and its deployment. In *Medical Image Segmentation Foundation Models. CVPR 2024 Challenge: Segment Anything in Medical Images on Laptop*, pp. 137–150. Springer Nature Switzerland, 2025.

[25] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[26] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3d medical images and videos, 2025. URL https://arxiv.org/abs/2504.03600.

[27] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

[28] Zhaoting Mu, Yu Ning, Yueqi Zhao, Rong Zhang, Fuad Md Mubtasim, and Hailong Ning. SAM2-PolypNet: SAM2 with adaptive context enhancement model for polyp segmentation. *Biomedical Signal Processing and Control*, 2026. URL https://www.sciencedirect.com/science/article/pii/S1746809426009353. In press.

[29] Md Mostafijur Rahman and Radu Marculescu. Mk-unet: Multi-kernel lightweight cnn for medical image segmentation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1053–1062. IEEE, 2025.

[30] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11769–11779. IEEE, 2024.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[32] Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in neural information processing systems*, 33:20378–20389, 2020.

[33] Hao Shao, Yang Zhang, and Qibin Hou. Polyper: Boundary sensitive polyp segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4731–4739, 2024. doi: 10.1609/aaai.v38i5.28274. URL https://ojs.aaai.org/index.php/AAAI/article/view/28274.

[34] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. cldice-a novel topology-preserving loss function for tubular structure segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16555–16564. IEEE, 2021.

[35] Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and fisher pruning, 2018. URL https://arxiv.org/abs/1801.05787.

[36] Tao Tong, Wen Zhang, and Wanni Zu. BEGA-UNet: Boundary-explicit guided attention U-Net with multi-scale feature aggregation for colonoscopic polyp segmentation. medRxiv preprint, 2026. URL https://www.medrxiv.org/content/10.64898/2026.03.04.26347608v2.

[37] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL http://dx.doi.org/10.1038/sdata.2018.161.

[38] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

[39] 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(1):4037190, 2017. doi: https://doi.org/10.1155/2017/4037190. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/4037190.

[40] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-MedSAM: Implicit medical image segmentation with segment anything. In *European Conference on Computer Vision*, pp. 90–107. Springer, 2024. doi: 10.1007/978-3-031-72684-2_6. URL https://arxiv.org/abs/2311.17081.

[41] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[42] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16111–16121. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01525.

[43] Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18547–18557. IEEE, 2023.

[44] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023. URL https://arxiv.org/abs/2304.13785.

[45] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. URL https://arxiv.org/abs/2306.12156.

[46] Xuanbo Zhao, Cheng Wang, Huaxing Xu, Hong Zhou, Zekuan Yu, Tao Chen, Xiaoling Wei, and Rongjun Zhang. LDFSAM: Localization distillation-enhanced feature prompting sam for medical image segmentation. *Journal of Imaging*, 12(2):74, 2026. doi: 10.3390/jimaging12020074. URL https://www.mdpi.com/2313-433X/12/2/74.

[47] Han Zhong, Jiatian Zhang, and Lingxiao Zhao. MedSAM/MedSAM2 feature fusion: Enhancing nnUNet for 2d TOF-MRA brain vessel segmentation. *Journal of Imaging*, 11(6):202, 2025. doi: 10.3390/jimaging11060202. URL https://www.mdpi.com/2313-433X/11/6/202.

[48] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for sam. *International Journal of Computer Vision*, 133(12):8452–8468, 2025.

APPENDIX

The appendix contains proofs, implementation details, and additional information about the analysis. We keep the main text focused on the method and key empirical findings.

A PROOFS

A.1 PROOF OF THEOREM 3.3

Proof. Fix a point $\boldsymbol{x} \in \Gamma_{\boldsymbol{\theta}}$. By definition of the old boundary, $s_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$. Since Assumption 3.1 gives $\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 > 0$, the unit normal direction of the level set at $\boldsymbol{x}$ is

$$\boldsymbol{n}(\boldsymbol{x}) = \frac{\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})}{\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2}.$$

The new logit map after compression is $\widetilde{s}(\boldsymbol{x}) = s_{\boldsymbol{\theta}}(\boldsymbol{x}) + \delta_G(\boldsymbol{x})$. We seek the new boundary point on the normal line through $\boldsymbol{x}$. We write it as

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x} + u_G(\boldsymbol{x})\boldsymbol{n}(\boldsymbol{x}),$$

where $u_G(\boldsymbol{x})$ is the signed normal displacement. Because $\widetilde{\boldsymbol{x}}$ lies on the new boundary, it satisfies

$$0 = \widetilde{s}(\widetilde{\boldsymbol{x}}) = s_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{x}}) + \delta_G(\widetilde{\boldsymbol{x}}).$$

We expand both terms around $\boldsymbol{x}$. The first-order Taylor expansion gives

$$s_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{x}}) = s_{\boldsymbol{\theta}}(\boldsymbol{x}) + u_G(\boldsymbol{x})\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})^\top \boldsymbol{n}(\boldsymbol{x}) + O(u_G(\boldsymbol{x})^2).$$

Since $s_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$ and $\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})^\top \boldsymbol{n}(\boldsymbol{x}) = \|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2$, this becomes

$$s_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{x}}) = u_G(\boldsymbol{x})\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 + O(u_G(\boldsymbol{x})^2).$$

The perturbation term satisfies

$$\delta_G(\widetilde{\boldsymbol{x}}) = \delta_G(\boldsymbol{x}) + O(|u_G(\boldsymbol{x})|\|\delta_G\|_{C^1}).$$

Substituting the two expansions into the boundary equation gives

$$0 = u_G(\boldsymbol{x})\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 + \delta_G(\boldsymbol{x}) + O(u_G(\boldsymbol{x})^2) + O(|u_G(\boldsymbol{x})|\|\delta_G\|_{C^1}).$$

For small perturbations, the implicit function theorem guarantees that $u_G(\boldsymbol{x})$ is also small and has the same order as $\delta_G(\boldsymbol{x})$. Therefore the two remainder terms are second order in $\|\delta_G\|_{C^1}$. Solving the first-order part gives

$$u_G(\boldsymbol{x}) = -\frac{\delta_G(\boldsymbol{x})}{\|\nabla s_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2} + O(\|\delta_G\|_{C^1}^2).$$

This is Eq. equation 4. □

A.2 PROOF OF PROPOSITION 3.4

Proof. Let the current allocation be (c_H, c_M) and keep the total compression fixed at $C = c_H + c_M$. We move a small amount $\eta > 0$ of compression budget from heads to MLPs. The new allocation is

$$(c_H - \eta, c_M + \eta).$$

The first-order Taylor expansion of $\mathcal{E}_{\mathrm{bd}}$ at (c_H, c_M) gives

$$\mathcal{E}_{\mathrm{bd}}(c_H - \eta, c_M + \eta) = \mathcal{E}_{\mathrm{bd}}(c_H, c_M) - \eta\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_H} + \eta\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_M} + o(\eta).$$

The first-order change in boundary error is therefore

$$\Delta\mathcal{E}_{\mathrm{bd}} = \eta\left(\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_M} - \frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_H}\right) + o(\eta).$$

If $\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_H}$ is larger than $\frac{\partial \mathcal{E}_{\mathrm{bd}}}{\partial c_M}$, then the coefficient of η is negative. For sufficiently small positive η, the first-order term dominates the $o(\eta)$ remainder, so $\Delta\mathcal{E}_{\mathrm{bd}} < 0$. The boundary error decreases while the total compression remains C. □

Table 5: Pruning-order ablation on CVC-ClinicDB. Each cell reports light/full results, without/with post-pruning fine-tuning. Best results in each setting are in bold.

Order	Dice↑	IoU↑	BF1↑	HD95↓
Head→MLP (default)	**0.8646 / 0.9080**	**0.7788 / 0.8461**	**0.2861 / 0.4991**	**28.19 / 17.39**
MLP→Head (reversed)	0.8629 / 0.9078	0.7770 / 0.8457	0.2757 / 0.4912	28.70 / 17.71
Simultaneous	0.8608 / 0.9069	0.7733 / 0.8452	0.2792 / 0.4934	28.83 / 18.14

B Head–MLP Pruning Order and Interactions

We compare head-to-MLP, reversed MLP-to-head, and simultaneous pruning on CVC-ClinicDB [2]. The light setting uses lightweight recovery on the calibration set during pruning, without post-pruning fine-tuning. The full setting additionally includes post-pruning fine-tuning.

Table 5 shows that head-to-MLP pruning is numerically best on all four metrics in both settings. Its advantage over the reversed order is modest, while its advantage over simultaneous pruning is consistent. These results support head-to-MLP as a practical default for mechanism analysis and show limited sensitivity to order reversal in this experiment. They do not establish theoretical optimality or statistical significance.

The cascade also accounts for directional head–MLP interactions. After head pruning and the fixed lightweight recovery, we recompute the MLP Fisher and dual-intervention scores on the resulting checkpoint. MLP groups are therefore not scored independently on the intact model: their ranking is conditioned on the selected head mask and the updated surviving parameters. This captures conditional head-to-MLP interactions rather than exhaustively searching all joint head–MLP masks. The simultaneous variant omits this conditional re-scoring and performs slightly worse than the default on every reported metric.

C Algorithmic Summary

Algorithm 1 summarizes MedCore. The algorithm separates the scoring stage from the recovery stage. The reset-to-SAM operation is used only to compute importance. The final compressed model removes groups or keeps them active. It does not reset kept groups to SAM weights.

D Additional Experimental and Boundary Leverage Details

D.1 Additional Experimental Settings

Training and test splits use a fixed random seed. Polyp Fisher calibration uses 128 images in total: 64, 38, and 26 from Kvasir-SEG, CVC-ColonDB, and CVC-ClinicDB, respectively, with $\pi_r = (0.5, 0.3, 0.2)$ and batch size 1. We set $\lambda_{\mathrm{bd}} = 3.0$, $\beta = 0.3$, $\tau = 0$, and $\varepsilon = 10^{-8}$, keeping these hyperparameters fixed for BUSI and ISIC2018.

Unless stated otherwise, MedCore uses 100-step lightweight recovery on the calibration set during pruning at a learning rate of 10^{-5}, with feature, boundary, logit, and frequency loss weights of 0.5, 1.0, 2.0, and 0.5, respectively. This recovery is distinct from post-pruning fine-tuning on a target dataset. ACDC results include per-structure metrics and their macro-average.

D.2 Boundary Leverage Evaluation

We compute boundary leverage on 384 calibration images, with 128 images from each polyp dataset. The sweep has seven head sparsity levels and seven MLP sparsity levels. We compute adjacent head steps and adjacent MLP steps. Steps with zero additional parameter reduction are undefined and are omitted.

For a head step, we fix the MLP sparsity and increase head sparsity. For an MLP step, we fix head sparsity and increase MLP sparsity. We use the same parameter-reduction unit for both families.

Algorithm 1 MedCore pruning

Require: MedSAM checkpoint $\boldsymbol{\theta}^M$, SAM checkpoint $\boldsymbol{\theta}^S$, frozen prompt encoder ϕ, calibration sets $\{\mathcal{D}_{\text{cal}}^{(r)}\}_{r=1}^{R}$, structured groups $\mathcal{G}$, target budget B
Ensure: Pruned and recovered MedSAM checkpoint
1: Initialize all group masks as active.
2: **for** each sub-distribution r **do**
3: Compute boundary-aware MedSAM Fisher $F_{r,i}^{M}$ using Eq. equation 10.
4: **end for**
5: Compute SAM Fisher F_i^S on the same calibration data.
6: **for** each group $g \in \mathcal{G}$ **do**
7: Estimate $\widehat{\Delta}_g^{\text{zero},(r)}$ using $F_{r,i}^{M}$.
8: Estimate $\widehat{\Delta}_g^{\text{reset},(r)}$ using Eq. equation 12.
9: Compute the fused score $\widehat{Q}_g^{(r)}$ and the distribution-aware score $\widehat{Q}_g^{\text{dist}}$.
10: Compute the cost-normalized priority P_g using Eq. equation 8.
11: **end for**
12: Allocate head-pruning budgets across blocks with block sensitivity.
13: Prune attention heads with the lowest priorities under the block budgets.
14: Run a short recovery fine-tuning stage.
15: Recompute or reuse priorities for MLP connection groups in the head-pruned model.
16: Prune MLP connection groups until the target budget is reached.
17: Run post-pruning recovery fine-tuning with Eq. equation 14.
18: Physically remove pruned structures and return the compact checkpoint.

This normalization is important because a 10% increase in head sparsity and a 10% increase in MLP sparsity do not remove the same number of parameters.

The boundary band width is tested at 3 and 5 pixels. Both choices give the same qualitative conclusion in our experiments. The main text reports the 95th-percentile boundary leverage because HD95 is also a high-percentile boundary metric. The mean boundary leverage gives the same direction but is less aligned with HD95.

The summary values are: $\text{BLR}_{95} = 2.887$, $\text{RLR} = 3.064$, $\text{BSR} = 0.942$, paired win rate $= 1.000$, and median paired difference $= 2.603$. BSR is the ratio between boundary leverage ratio and region leverage ratio. Since BSR is slightly below one, we do not claim that head pruning is boundary-specific. We claim that head pruning is high-boundary-leverage. This is the statement needed to explain boundary fragility.

E Implementation Details

We use the official MedSAM ViT-B checkpoint as $\boldsymbol{\theta}^M$ and the SAM ViT-B checkpoint as $\boldsymbol{\theta}^S$. The prompt encoder remains frozen. Fisher estimation uses calibration images and the same bounding-box prompt protocol as MedSAM. We compute boundary maps by morphological dilation minus erosion of the binary mask. We normalize losses so that the boundary-weighted BCE term does not dominate the Dice term at the beginning of recovery.

For attention-head pruning, each group contains the corresponding Q, K, V, and output-projection parameters associated with one head. For MLP pruning, each group contains a connection group in the feed-forward block. We use physical removal after mask selection so that parameter count and FLOPs reflect the actual compact model.

The main one-time configurations protect the deepest encoder blocks because they have high Fisher sensitivity and connect directly to the mask decoder. The sequential extreme setting is different. It starts from the h70_m95 checkpoint and then performs a conservative second pruning pass on the previously protected deepest blocks. This setting is used only to test the limit of compression.

Table 6: Full cross-modality results with post-pruning fine-tuning.

Dataset	Method	Params↓	FLOPs↓	Dice↑	BF1↑	HD95↓
CVC-ClinicDB [2]	MedSAM	89.7M	926.5G	0.9264	0.5477	13.77
	MedCore h50_m70	35.8M	377.0G	0.9601	0.6603	2.64
	MedCore h70_m70	30.1M	283.4G	0.9564	0.6409	3.85
	MedCore h70_m95	24.9M	240.9G	0.9523	0.6194	6.30
	MedCore h80_m95	23.2M	224.6G	0.9507	0.5951	4.46
BUSI [1]	MedSAM	89.7M	926.5G	0.9334	0.4989	7.30
	MedCore h50_m70	35.8M	377.0G	0.9357	0.5048	7.35
	MedCore h70_m70	30.1M	283.4G	0.9336	0.4913	7.82
	MedCore h70_m95	24.9M	240.9G	0.9298	0.4703	9.43
	MedCore h80_m95	23.2M	224.6G	0.9316	0.4833	8.58
ISIC2018 [37, 7]	MedSAM	89.7M	926.5G	0.9314	0.2401	7.47
	MedCore h50_m70	35.8M	377.0G	0.9525	0.4806	6.12
	MedCore h70_m70	30.1M	283.4G	0.9525	0.4732	5.57
	MedCore h70_m95	24.9M	240.9G	0.9530	0.4753	5.86
	MedCore h80_m95	23.2M	224.6G	0.9526	0.4708	5.94

F Full Cross-Modality Results

Table 6 reports results for all available pruning configurations on CVC-ClinicDB, BUSI, and ISIC2018, complementing the representative h70_m95 comparison in Table 2. All MedCore configurations include post-pruning fine-tuning.

On CVC-ClinicDB and ISIC2018, all reported MedCore configurations improve Dice, BF1, and HD95 relative to the MedSAM baseline. On BUSI, h50_m70 preserves performance closely, while the more aggressive configurations have lower BF1 and higher HD95 than the baseline. These results show that the compression–performance trade-off depends on the target dataset and support the need for target-domain validation.

G Comparison with Structured Pruning Baselines

The main paper compares MedCore against external efficient SAM variants and medical segmentation backbones (Table 1), whose architectures and training pipelines differ. This appendix provides a controlled comparison of the *scoring criterion*. All matched baselines operate on the same MedSAM ViT-B encoder, share identical structured groups (attention heads and MLP connection groups), calibration data, and evaluation metrics, and use the same target sparsity levels as MedCore. All methods receive the same light 100-step inter-stage stabilization, but **terminal post-pruning fine-tuning is disabled**, including for MedCore. The results therefore compare importance scores under matched stabilization rather than after full recovery.

Baselines. We compare against eight structured pruning criteria: (i) **Random Pruning** [20], which selects groups uniformly at random within each block; (ii) **Magnitude Pruning** [10, 11], which ranks groups by their parameter ℓ_2 norm; (iii) **Zero-only Fisher** [27], which uses only Δ^{zero} from Eq. (5) and discards the reset-to-SAM signal; (iv) **Vanilla Fisher** [35], which uses the standard Dice+BCE loss without boundary weighting, Cross-Fisher, or distribution-aware aggregation; (v) **Original Medical Core**, which retains dual zero/reset intervention but omits boundary-aware Fisher, Cross-Fisher, multi-dataset aggregation, and block-sensitivity allocation; (vi) **Movement Pruning** [32], which uses first-order movement scores adapted to the shared structured groups; (vii) **Group Fisher** [23], which estimates group importance through Fisher information; and (viii) **Hessian-aware ViT** [43], which uses Hessian-aware structural saliency.

Table 7: Comparison with structured pruning baselines on macro-averaged polyp segmentation metrics. All methods use MedSAM ViT-B with the same 100-step inter-stage stabilization and no terminal post-pruning fine-tuning. Parameter counts and FLOPs are identical across methods within each compression level. Best results within each compression level are in bold.

Compression	Method	Dice↑	IoU↑	BF1↑	HD95↓
Low: h40_m30	Random Pruning	0.9076	0.8436	0.4598	23.12
	Magnitude Pruning	0.8278	0.7228	0.2081	41.63
	Zero-only Fisher	0.9137	0.8541	0.5035	22.13
	Vanilla Fisher	0.9113	0.8512	0.4972	23.55
	Original Medical Core	0.9123	0.8519	0.4988	23.50
	Movement Pruning	0.8751	0.7943	0.3343	30.77
	Group Fisher	0.9141	0.8547	0.5044	22.38
	Hessian-aware ViT	0.9144	0.8550	0.5041	22.14
	MedCore (ours)	**0.9191**	**0.8615**	**0.5162**	**20.30**
Mid: h50_m70	Random Pruning	0.8394	0.7391	0.2272	40.16
	Magnitude Pruning	0.8237	0.7135	0.1665	43.60
	Zero-only Fisher	0.8881	0.8124	0.3911	29.10
	Vanilla Fisher	0.8854	0.8080	0.3859	29.44
	Original Medical Core	0.8870	0.8106	0.3879	28.69
	Movement Pruning	0.8450	0.7475	0.2442	37.18
	Group Fisher	0.8866	0.8101	0.3888	29.48
	Hessian-aware ViT	0.8868	0.8103	0.3890	29.46
	MedCore (ours)	**0.8908**	**0.8179**	**0.4078**	**27.07**
High: h70_m70	Random Pruning	0.8371	0.7344	0.2121	41.59
	Magnitude Pruning	0.8200	0.7101	0.1761	44.74
	Zero-only Fisher	0.8555	0.7630	0.2784	39.26
	Vanilla Fisher	0.8555	0.7634	0.2803	38.95
	Original Medical Core	0.8558	0.7629	0.2799	36.25
	Movement Pruning	0.8273	0.7191	0.1837	41.09
	Group Fisher	0.8556	0.7637	0.2794	38.90
	Hessian-aware ViT	0.8555	0.7635	0.2794	38.92
	MedCore (ours)	**0.8738**	**0.7899**	**0.3144**	**31.35**
Extreme: h84_m95	Random Pruning	0.8180	0.7046	0.1468	43.95
	Magnitude Pruning	0.8161	0.7033	0.1643	45.93
	Zero-only Fisher	0.8312	0.7258	0.2003	42.18
	Vanilla Fisher	0.8338	0.7298	0.2048	42.39
	Original Medical Core	0.8346	0.7312	0.2046	42.46
	Movement Pruning	0.8343	0.7302	0.1946	42.51
	Group Fisher	0.8348	0.7314	0.2072	42.38
	Hessian-aware ViT	0.8349	0.7314	0.2054	**41.84**
	MedCore (ours)	**0.8492**	**0.7542**	**0.2539**	41.92

The **head-importance heuristic** [27] is reported separately because it prunes attention heads only and defines no MLP-neuron importance. We retain this head-only setting without introducing an auxiliary MLP criterion. Its smaller actual parameter reductions do not constitute an equal-total-budget comparison with the joint head–MLP methods.

Results. Table 7 reports macro-averaged metrics on the three polyp benchmarks across four compression levels. Table 8 reports the head-only results with their actual parameter reductions.

Table 8: Head-importance heuristic on macro-averaged polyp segmentation metrics. Only attention heads are pruned; actual parameter reductions are relative to the unpruned MedSAM.

Head budget	Param↓	Dice↑	IoU↑	BF1↑	HD95↓
h40	12.7%	0.9130	0.8534	0.5031	23.31
h50	15.8%	0.9068	0.8430	0.4752	24.72
h70	22.2%	0.8797	0.8005	0.3640	30.49
h84	24.1%	0.8594	0.7700	0.2930	34.95

Discussion. Three observations follow from Table 7, consistent with the conceptual claims in the main paper.

First, Magnitude Pruning severely degrades boundary metrics across all compression levels: BF1 falls to 0.2081 even at h40_m30, and HD95 exceeds 41. This is consistent with Theorem 3.3, which relates boundary displacement to the boundary logit perturbation and local logit slope rather than weight magnitude alone. A small-magnitude attention head can carry substantial boundary leverage that a norm-based criterion does not directly capture.

Second, Zero-only Fisher and Vanilla Fisher improve Dice and BF1 over Random and Magnitude Pruning at every level, supporting data-dependent sensitivity estimation. Group Fisher and Hessian-aware ViT provide additional competitive comparisons, but MedCore achieves the best Dice, IoU, and BF1 at all four levels and the best HD95 at three, leading in 15 of 16 metric–budget cells. At the sole exception, h84_m95, Hessian-aware ViT obtains HD95 41.84 versus 41.92 for MedCore, while MedCore improves BF1 by 0.0485. These results are consistent with the component ablation in Section 5.1, where boundary-aware Fisher, reset-to-SAM scoring, and distribution-aware aggregation contribute to boundary preservation.

Third, MedCore's absolute BF1 margin over the strongest competing baseline increases monotonically from 0.0118 at h40_m30 to 0.0467 at h84_m95, corresponding to relative gains of 2.3% and 22.5%. This widening supports the boundary leverage principle as a design lens. As compression increases, identifying boundary-critical structures becomes more consequential, and generic sensitivity scores leave a larger gap in contour preservation. The advantage therefore extends beyond the internal Fisher variants to the additional structured-pruning criteria.

These comparisons complement Table 1, which evaluates architecturally distinct efficient and medical segmentation models. They address a narrower question: whether boundary-aware, adaptation-aware scoring improves pruning decisions on the same MedSAM encoder under matched calibration and stabilization. Table 7 supports this conclusion, with the largest BF1 advantage at the highest compression level.

H LIMITATIONS

MedCore uses calibration data to estimate Fisher information and boundary leverage. If this calibration set is small or biased, the score may under-protect rare structures. The reset-to-SAM score also assumes that the SAM and MedSAM checkpoints are aligned parameter by parameter. This is true for the standard MedSAM adaptation path, but it may not hold for methods that change the architecture.

The boundary leverage theorem is local. It explains first-order boundary displacement under small logit perturbations. Very aggressive pruning can create non-local effects, such as disconnected masks or missing components, that are not fully captured by the first-order formula. We therefore view boundary leverage as a useful diagnostic and design principle, not as a complete replacement for final BF1 and HD95 evaluation.

Finally, our main experiments focus on 2D prompt-driven segmentation with bounding boxes. Extension to 3D medical segmentation and other prompt types requires additional validation. The same concepts can be applied, but the boundary band, distance metric, and computational cost must be adjusted for 3D volumes.