Residual-Aware Structured Pruning with Dual-Stream Alignment for Adapter-Tuned SAM

Xin Luo[1,2,3], Chunjiang Wang[1,2,3], and Shaohua Kevin Zhou[1,3,4,5,6⋆]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei, Anhui, China 230026
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, China 215123
[3] Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE) Lab, YRD-RIGHT, USTC Suzhou Institute for Advanced Research, Suzhou, Jiangsu, China 215123
[4] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu, China 215123
[5] Biomedical Basic Research Center (BBRC) of Jiangsu Province, Suzhou, Jiangsu, China 215123
[6] State Key Laboratory of Precision and Intelligent Chemistry, Hefei, Anhui, China 230026
luoxin_ustc@mail.ustc.edu.cn, chunjiang_wang@mail.ustc.edu.cn, skevinzhou@ustc.edu.cn

Abstract. Adapter-tuned Segment Anything Models are promising for medical image segmentation, yet their backbone-heavy compute and adapter overhead hinder real-time deployment. Studying SAM-Med2D as a representative adapter-tuned SAM testbed, we observe that structured masking harms Dice far more when applied to the frozen backbone than to adapters, suggesting strong asymmetry in redundancy. We propose RASP-SAM, a structured pruning framework that exploits this decoupling. It defines dependency-safe pruning units under attention, FFN, adapter, and residual couplings, ranks units with Orthogonal Residual Scoring based on the orthogonal residual of downstream gradients relative to a merged pretraining and downstream gradient subspace with cross-fitting, and restores accuracy using Dual-Stream Compensation that distills adapters and backbone in two stages under a small calibration budget. With only 20K calibration images from SA-1B and SA-Med2D-20M, RASP-SAM achieves near-lossless segmentation on SAM-Med2D while reducing encoder compute from 65.15G to 15.62G MACs and parameters from 270.1M to 63.4M at 76% sparsity, outperforming a transferred SlimSAM baseline at comparable cost. These results validate RASP-SAM in the SAM-Med2D setting and motivate broader evaluation on adapter-style SAM variants.

Keywords: Structured pruning · Segment Anything Model · Adapter.

⋆ Corresponding author.

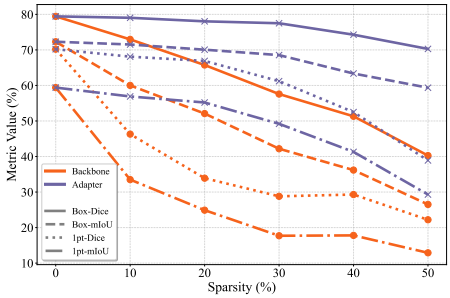

Fig. 1. Backbone vs. adapter random-mask sparsity sensitivity on SAM-Med2D.

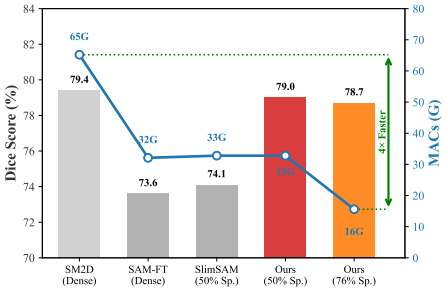

Fig. 2. RASP-SAM vs. Baselines: Dice score (left) vs. Computational cost (right).

1 Introduction

Segment Anything (SAM) [10] has become a foundation for medical segmentation, typically adapted via full fine-tuning [15] or parameter-efficient adapters [7, 8, 17, 3]. However, the heavy computational burden restricts real-time clinical deployment [3, 23, 9, 22]; in SAM-Med2D, adapter parameters can even exceed those of the base model, forcing resolution compromises [3]. Because this model exposes the frozen-backbone/trainable-adapter split and residual coupling targeted by our method, we use SAM-Med2D as the primary validation setting for adapter-aware compression.

Existing SAM acceleration strategies—spanning architecture redesign [24, 20], quantization [14, 12, 13], and pruning [2, 19, 5]—typically apply uniform compression policies, neglecting the structural distinction between the frozen backbone and trainable adapters. This oversight obscures potential efficiency gains from treating these components differentially. To investigate their distinct redundancies, we conduct a pilot study on the SAM-Med2D setting [3].

We first run a pilot study on SAM-Med2D [3] by applying random channel-wise structured masks at the same sparsity to the backbone and adapters separately. As shown in Fig. 1, pruning the backbone degrades Dice much faster than pruning the adapters. This suggests that the backbone carries fragile shared representations, while adapters exhibit higher intrinsic redundancy, motivating asymmetric compression and dual-stream recovery.

Based on this observation, we study RASP-SAM in SAM-Med2D. The framework defines dependency-safe pruning units under backbone, adapter, and residual couplings, ranks them with Orthogonal Residual Scoring (ORS), and restores accuracy with Dual-Stream Compensation. Under a 20K-image calibration budget, as shown in Fig. 2, RASP-SAM achieves large parameter/MAC reductions with near-lossless Dice on SAM-Med2D. Our contributions are:

– We analyze structural dependencies in the SAM-Med2D adapter-tuned encoder and derive practical pruning-safe units that preserve residual compatibility.

- We introduce *Orthogonal Residual Scoring*, which isolates non-substitutable gradient directions across pretraining and downstream streams to improve structured pruning decisions.
- We design *Dual-Stream Compensation*, a data-efficient recovery scheme that alternates supervision from both streams to preserve the shared backbone subspace and recover accuracy with only 20K images.

2 Related Work

Medical SAM adaptation and adapter tuning. SAM [10] has been widely adapted to medical segmentation through both full/partial fine-tuning and adapter-based parameter-efficient tuning [15, 17, 3]. SAM-Med2D is closest to our setting because it inserts trainable adapters into the image encoder and exposes a clear backbone–adapter split. More broadly, adapter-style tuning is a standard PEFT paradigm in NLP and vision [7, 18], alongside LoRA/QLoRA [8, 4]. These methods improve adaptation efficiency, but not compression of adapter-tuned foundation models. Similar SAM adaptation trends in other vision tasks [21, 1, 16] further motivate adapter-aware compression.

Efficient and compressed SAM. SAM acceleration mainly uses architecture redesign, quantization, and pruning. FastSAM [24] and MobileSAM [20] reduce cost by redesign or distillation; quantization approaches include general ViT methods [12, 13] and PQ-SAM [14]; SlimSAM [2] shows effective pruning with distillation and limited calibration. However, these methods target generic SAM compression and do not explicitly exploit the backbone–adapter decoupling in adapter-tuned models such as SAM-Med2D.

Structured pruning and dependency-aware pruning. Pruning research spans classic saliency-based methods (e.g., OBD [11], OBS [6]) to modern structured pruning for transformers. Recent ViT studies show the importance of structured saliency and pruning granularity [19], while dependency-aware pruning stresses preserving structural couplings across layers and branches [5]. This is especially relevant to adapter-tuned SAM, where backbone and adapters are decoupled in parameterization but coupled through residual pathways. Our method targets this gap under limited calibration budgets.

3 Method

Problem Setup and Overview. We focus on structured pruning of the adapter-tuned SAM-Med2D image encoder. Given an encoder with n Transformer blocks (indexed by i) and embedding dimension $d_{\text{model}} = h \cdot d_h$ (where h is the number of heads and d_h is the head dimension), our goal is to obtain a compressed model with target sparsity ρ that preserves both pretraining and medical knowledge. We utilize two small calibration streams: a pretraining stream $\mathcal{D}_{\text{pre}}$ and a downstream stream $\mathcal{D}_{\text{down}}$. The proposed RASP-SAM framework comprises three key stages: (1) Dependency-Aware Unit Definition to ensure structural consistency;

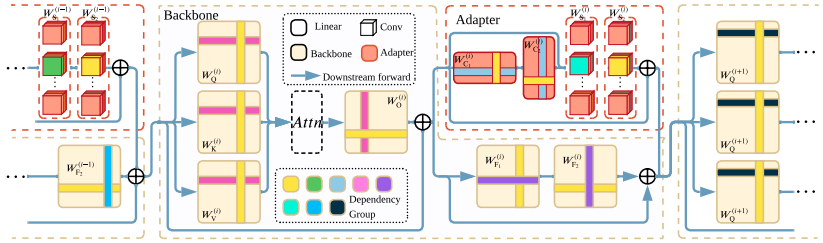

Fig. 3. Dependency schematic across encoder blocks $i-1$, i, and $i+1$. Colored rows/columns with the same color form one dependency group and must be retained or pruned jointly.

(2) Orthogonal Residual Scoring (ORS) for importance ranking; and (3) Dual-Stream Compensation for accuracy recovery. We detail each stage below.

Dependency-Aware Pruning Units. To ensure shape consistency, we classify structural couplings into local and global dependencies, prioritizing the former to maximize pruning granularity.

Local dependencies, as shown in Fig. 3 except the residual-aligned group, define the smallest independent pruning units by coupling the output dimension of one layer with the input dimension of its successor within a block. In the backbone Multi-Head Attention, preserving the inner-product requires that pruning the output channels of the query/key/value projections $\{\mathbf{W}_{\mathrm{Q}}, \mathbf{W}_{\mathrm{K}}, \mathbf{W}_{\mathrm{V}}\}$ be synchronized with pruning the input channels of the output projection $\mathbf{W}_{\mathrm{O}}$. Similarly, in FFNs and paired adapters, the output channels of the expansion/first layer are coupled with the input channels of the reduction/second layer. Adhering to these intra-block couplings allows for layer-specific sparsity patterns, which minimizes structural perturbation compared to coarser constraints.

Global dependencies arise from residual connections, which enforce a shared mask across the residual stream. As illustrated by the residual-aligned group in Fig. 3, pruning an output channel of the first block's FFN necessitates synchronized pruning of the corresponding output channel of the parallel adapter, as well as the input and output channels of all subsequent blocks. This cascading constraint couples sensitive layers with redundant ones across the entire depth, leading to overly coarse granularity. We therefore adopt a hybrid strategy: we treat blocks as locally independent by default and enforce global constraints *only* at necessary residual summation points to ensure runtime validity.

Orthogonal Residual Scoring. After adapter tuning, the backbone mainly preserves shared pretraining representations, while the adapters encode downstream corrections. We therefore score each dependency-safe unit with gradients from the pretraining and downstream calibration streams so that pruning preserves the shared backbone subspace and downstream task behavior. For a prunable module in block i, let n_u be the number of candidate units and d the flattened parameter dimension per unit. We collect unit-wise gradients averaged over calibration batches and stack them row-wise as $\mathbf{G}_{\mathrm{pre}}^{(i)} \in \mathbb{R}^{n_u \times d}$ and

$\mathbf{G}_{\text{down}}^{(i)} \in \mathbb{R}^{n_u \times d}$. During the pretraining pass, adapters are bypassed so the backbone receives only pretraining supervision, while during the downstream pass both backbone and adapters are active. Backbone units follow the dual-stream formulation below; adapters, lacking pretraining signals, default to standard Taylor scoring.

We extract dominant directions via SVD, retaining the top-r right singular vectors of the two gradient matrices:

$$\mathbf{G}_{\text{pre}}^{(i)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{U}_{\text{pre}}^{(i)} = \mathbf{V}_{[:,1:r]}, \qquad \mathbf{G}_{\text{down}}^{(i)} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top, \quad \mathbf{U}_{\text{down}}^{(i)} = \hat{\mathbf{V}}_{[:,1:r]}. \tag{1}$$

To form a joint pretraining-downstream subspace, we concatenate the two bases and orthonormalize them with a Gram construction:

$$\mathbf{C}^{(i)} = \left[\mathbf{U}_{\text{pre}}^{(i)} \;\; \mathbf{U}_{\text{down}}^{(i)}\right], \quad {\mathbf{C}^{(i)}}^\top \mathbf{C}^{(i)} = \mathbf{V}_C \mathbf{\Lambda}_C \mathbf{V}_C^\top, \quad \mathbf{U}_\cup^{(i)} = \mathbf{C}^{(i)} \mathbf{V}_{C[:,1:k]} \mathbf{\Lambda}_{C,1:k}^{-1/2}, \tag{2}$$

where $k \leq 2r$ is chosen by a fixed rank. The basis $\mathbf{U}_\cup^{(i)}$ spans directions jointly explainable by the two streams.

Directly scoring a unit with a basis estimated from the same gradients introduces optimistic bias. We therefore use two-fold cross-fitting on units: split row indices of $\mathbf{G}_{\text{down}}^{(i)}$ into disjoint sets S_1 and S_2, build the merged basis on one fold, and score units on the other. Let $\mathbf{g}_u \in \mathbb{R}^d$ be the gradient of unit u, and let $\mathbf{U}_{\cup,[1]}^{(i)}$ and $\mathbf{U}_{\cup,[2]}^{(i)}$ be the two fold-specific merged bases. The residual energy is

$$R_u^2 = \begin{cases} \|\mathbf{g}_u\|_2^2 - \|{\mathbf{U}_{\cup,[2]}^{(i)}}^\top \mathbf{g}_u\|_2^2, & u \in \mathsf{S}_1, \\ \|\mathbf{g}_u\|_2^2 - \|{\mathbf{U}_{\cup,[1]}^{(i)}}^\top \mathbf{g}_u\|_2^2, & u \in \mathsf{S}_2, \end{cases} \tag{3}$$

and the ORS score is $s_u = \sqrt{\max(R_u^2, 0)}\, \|\boldsymbol{\theta}_u\|_2$, where $\boldsymbol{\theta}_u$ is the parameter vector of unit u. Crucially, if a unit's gradient direction is reconstructible by the subspace of other units, it implies redundancy. Therefore, ORS assigns higher scores to units with large orthogonal residuals, preserving components that contribute unique directional information not spanned by the rest.

Dual-Stream Compensation. After ORS selects dependency-safe units under target sparsity ρ, pruning still perturbs the shared backbone subspace and residual-aligned branch outputs. We therefore perform a two-stage recovery with the same two calibration streams used in ORS, namely the pretraining stream $\mathcal{D}_{\text{pre}}$ and the downstream stream $\mathcal{D}_{\text{down}}$, as summarized in Fig. 4. In each stage, the teacher is the model before pruning that stage, and the student is the corresponding pruned model.

In Stage 1, we prune adapter units first and freeze the backbone. Taylor scores are computed on adapter dependency-safe units, and each adapter group is pruned to sparsity ρ. Recovery uses only $\mathcal{D}_{\text{down}}$ so that updates remain task-specific. For each encoder block i, we distill adapter outputs via MSE $(L_{\text{C}_2}^{(i)}, L_{\text{S}_2}^{(i)})$ and combine them with task loss which is a mixture of Focal, Dice, and IoU :

$$L_{\text{a}} = \lambda_{\text{a}} \sum_{i=1}^{n} \left(L_{\text{C}_2}^{(i)} + L_{\text{S}_2}^{(i)} \right) + L_{\text{task}},$$

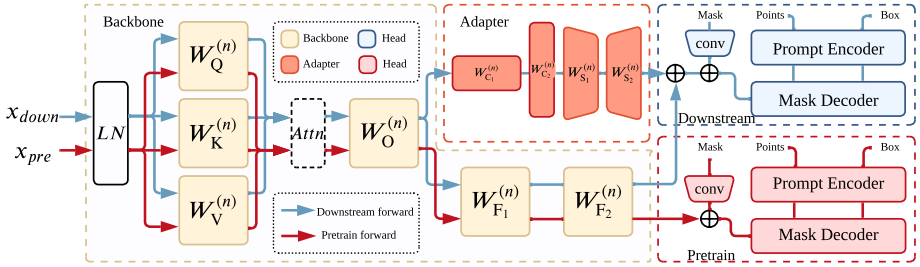

Fig. 4. Dual-Stream Compensation pipeline. Stage 1 prunes and recovers adapters with the backbone frozen using the downstream stream. Stage 2 prunes and recovers the backbone with adapters frozen using pretraining and downstream streams.

where λ_{a} balances distillation and task supervision.

In Stage 2, we prune backbone units and freeze all adapters recovered in Stage 1. ORS is applied to backbone dependency-safe units, with the shared global mask only when residual alignment requires synchronized pruning. Recovery uses both streams to preserve the shared backbone subspace while restoring downstream performance. Each iteration processes two mini-batches, $\mathcal{B}_{\mathrm{pre}} \sim \mathcal{D}_{\mathrm{pre}}$ and $\mathcal{B}_{\mathrm{down}} \sim \mathcal{D}_{\mathrm{down}}$, in two separate passes. For each block i, we distill backbone activations at $\mathbf{W}_{\mathrm{O}}^{(i)}$ and $\mathbf{W}_{\mathrm{F}_2}^{(i)}$, defining MSE distillation losses $L_{\mathrm{O}}^{(i)}$ and $L_{\mathrm{F}_2}^{(i)}$. The objective is

$$L_{\mathrm{b}} = \frac{\lambda_{\mathrm{b}}}{2} \sum_{i=1}^{n} \left[\left(L_{\mathrm{O}}^{(i)} + L_{\mathrm{F}_2}^{(i)}\right)\Big|_{\mathcal{B}_{\mathrm{pre}}} + \left(L_{\mathrm{O}}^{(i)} + L_{\mathrm{F}_2}^{(i)}\right)\Big|_{\mathcal{B}_{\mathrm{down}}} \right] + L_{\mathrm{task}}\big|_{\mathcal{B}_{\mathrm{down}}}, \tag{4}$$

where λ_{b} balances distillation and downstream supervision. The pretraining stream stabilizes backbone directions inherited from the foundation model, while the downstream stream restores task-specific behavior. This two-stage schedule matches the backbone and adapter roles used in ORS and enables recovery with a limited calibration budget.

4 Experimental Results

Experimental Settings. We build a 20K calibration set by sampling 10K images from SA-1B and 10K from SA-Med2D-20M and conduct experiments on NVIDIA H20 GPUs. To preserve foundation features during dual-stream compensation, inputs are resized to 1024×1024 for the pretraining stream and 256×256 for the downstream stream, while evaluation strictly follows the SAM-Med2D interactive protocol at 256×256. We prune the image encoder only in two stages, starting with adapters followed by the backbone, using one-shot pruning per stage with equal per-head channel pruning. ORS uses rank r=k=16. Optimization uses Adam with learning rate 1×10^{-4} and batch size 16 for 40 epochs

Table 1. Main results on SAM-Med2D. **A***xx* and **B***xx* denote adapter and backbone pruning ratios; Sp. denotes overall encoder sparsity. Params and MACs are reported for the encoder at 256×256.

Model	Strat.	Sp.	Res.	BBox	1pt	3pt	5pt	Param.	MAC
SAM-B	N/A	0.00	256^2	61.63	18.94	28.28	37.47	89.67	32.07
SAM-B	N/A	0.00	1024^2	74.49	36.88	42.76	47.57	89.67	368.86
SM2D	N/A	0.00	256^2	79.35	70.01	76.35	78.68	270.08	65.15
SAM-B	Finetune	0.00	256^2	73.56	60.11	70.95	75.51	89.67	32.07
SM2D	SlimSAM	0.50	256^2	74.10	65.23	74.64	76.93	134.43	32.75
SM2D	B50	0.15	256^2	**79.23**	**69.92**	**76.27**	**78.94**	**227.58**	**49.29**
SM2D	A50	0.33	256^2	**79.32**	**69.43**	**76.26**	**78.84**	**179.88**	**48.63**
SM2D	B75	0.24	256^2	**79.11**	**69.67**	**75.06**	**78.82**	**202.30**	**41.07**
SM2D	A75	0.50	256^2	**79.39**	**68.81**	**75.78**	**78.98**	**134.76**	**40.36**
SM2D	A87.5	0.58	256^2	**79.07**	**66.39**	**75.05**	**77.65**	**112.20**	**36.22**
SM2D	A50+B50	0.50	256^2	**79.01**	**69.43**	**76.26**	**77.62**	**134.43**	**32.75**
SM2D	A75+B75	0.76	256^2	**78.65**	**68.13**	**74.26**	**76.85**	**63.43**	**15.62**

per stage with $\lambda_a = \lambda_b = 0.5$. We denote the pruning ratio for adapters and the backbone as A*xx* and B*xx* respectively; for instance, A87.5 implies pruning adapters to 87.5% sparsity, and A75+B75 indicates 75% sparsity for both components. We report Dice under one box prompt and 1/3/5 point prompts.

Main Results. We benchmark RASP-SAM against the unpruned SAM-Med2D and a transferred SlimSAM baseline. To ensure a fair comparison, SlimSAM is calibrated using 20K downstream images to match our total budget. Table 1 shows that RASP-SAM achieves superior accuracy-efficiency trade-offs in this setting. At moderate compression, Dice remains nearly lossless. Even at high sparsity (A75+B75), where encoder MACs drop to approximately one-quarter of the original, the model retains strong performance across all prompts. Crucially, RASP-SAM outperforms the SlimSAM baseline at similar compute levels, supporting the effectiveness of our dependency-safe, dual-stream design on SAM-Med2D.

Inference Speed Analysis. To verify practical acceleration, Table 2 reports Mean latency (average ms/image) and TPS (throughput, images/s) at batch size 1. RASP-SAM (A75+B75) boosts encoder throughput by 13.7%. Crucially, this gain persists in the end-to-end pipeline with an 11.2% improvement, confirming that theoretical MAC reductions translate into effective and significant wall-clock acceleration.

Ablation on Pruning Order. Table 3 shows across both sparsity settings and all prompt types, pruning adapters before the backbone yields higher Dice than the reverse order. The gap becomes larger at higher sparsity. Pruning adapters first confines early changes to local dependencies and allows downstream-only compensation before backbone compression. Backbone pruning then uses dual-stream compensation to preserve the shared backbone subspace and reduce error accumulation. We therefore use adapter-first pruning in all other experiments.

Table 2. Inference speed. Enc. denotes encoder only; E2E denotes end to end.

	Enc.		E2E	
Strategy	**Mean**	**TPS**	**Mean**	**TPS**
Dense	22.493	44.46	28.174	35.49
A50+B50	20.979	47.67	26.629	37.55
A75+B75	19.779	50.56	25.343	39.46

Table 3. Ablation on pruning order, $A \rightarrow B$: Prune adapter first, then backbone.

Strategy	Order	BBox	1pt	3pt	5pt
A50+B50	A→B	79.01	69.43	76.26	77.62
A50+B50	B→A	78.08	67.72	75.55	76.03
A75+B75	A→B	78.65	68.13	74.26	76.85
A75+B75	B→A	77.50	65.34	72.46	75.94

Table 4. Effect of compensation data source at B75. Pre: SA-1B; Down: SA-Med2D-20M; Metrics: Dice (%).

Data	#	BBox	1pt	3pt	5pt
Pre	20K	64.95	52.47	64.91	71.74
Down	20K	68.31	64.69	73.45	78.02
Pre+Down	10K+10K	**79.23**	**69.67**	**75.06**	**78.82**

Table 5. Importance scoring ablation at B75. Metrics: Dice (%).

Method	BBox	1pt	3pt	5pt
Magnitude	77.27	67.77	73.67	75.09
Hessian	77.92	66.26	73.84	77.08
Taylor	78.63	68.22	74.26	77.34
ORS	**79.11**	**69.67**	**75.06**	**77.82**

Ablation on Compensation Data. We study compensation data sources under a fixed 20K-image budget with backbone pruning at B75. Table 4 shows, using only pretraining data preserves the foundation representation but underfits downstream prompts. Using only downstream data improves prompt adaptation but weakens preservation of the shared backbone subspace. Mixing the two streams in equal proportion gives the best Dice across all prompt types. This supports the dual-stream compensation design used in the main experiments.

Ablation on Importance Scoring. We compare Magnitude, Hessian, Taylor, and ORS under the same prune-and-compensate schedule on SAM-Med2D at B75 with 20K mixed calibration. Fig. 5 plots 1-point Dice during backbone compensation. ORS converges faster, reaches a higher plateau, and is more stable near convergence. Table 5 confirms the final Dice advantage across box and point prompts. ORS scores each unit by the orthogonal residual of its downstream gradient relative to the joint pretraining-downstream subspace, weighted by parameter scale. This suppresses substitutable directions and focuses compensation on directions that are harder to recover after pruning. Magnitude ignores direction, Taylor mixes substitutable and non-substitutable components, and Hessian is less stable under small calibration sets. These behaviors are consistent with the faster and stronger compensation observed for ORS.

5 Conclusion

We presented RASP-SAM, a structured pruning framework for the adapter-tuned SAM-Med2D setting. By leveraging the structural decoupling between the backbone and adapters, our method integrates dependency-safe pruning, Orthogonal Residual Scoring, and Dual-Stream Compensation to maximize efficiency. Experiments on SAM-Med2D demonstrate that RASP-SAM achieves a

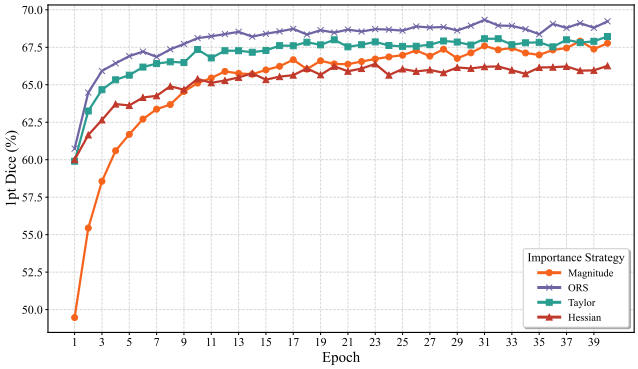

Fig. 5. 1-point Dice during backbone compensation after B75 pruning with four scoring criteria. ORS converges faster and reaches a higher plateau.

4× reduction in MACs with near-lossless performance. These results support the proposed dependency-safe and dual-stream design in a representative adapter-tuned SAM setting; extending the evaluation to other adapter-style SAM variants remains important future work.

Acknowledgments. Supported by Natural Science Foundation of China under Grant 62271465, National Key R&D Program of China under Grant 2025YFC3408300, Natural Science Foundation of Jiangsu Province (No. BK20255001), and Suzhou Basic Research Program under Grant SYG202338.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3367–3375 (2023)
2. Chen, Z., Fang, G., Ma, X., Wang, X.: Slimsam: 0.1% data makes segment anything slim. Advances in Neural Information Processing Systems **37**, 39434–39461 (2024)
3. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023), `https://arxiv.org/abs/2308.16184`
4. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)
5. Fang, G., Ma, X., Song, M., Mi, M.B., Wang, X.: Depgraph: Towards any structural pruning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16091–16101 (2023)

6. Hassibi, B., Stork, D.: Second order derivatives for network pruning: Optimal brain surgeon. Advances in neural information processing systems **5** (1992)
7. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Jiang, Y., Du, Y., Xiong, K., Huang, K., Li, T., Li, Z., Zhang, M., Gan, X., Li, Q., Liang, J., Cao, M., Sun, J., Wang, J., Duanmu, J., Li, X., Wen, Z., Jiang, Q., Yu, X., Chen, S.: Foundation model-guided multi-view semi-supervised CT segmentation of liver tumors in resource-constrained settings. npj Digital Medicine **9**(1), 31 (2025). https://doi.org/10.1038/s41746-025-02190-0
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. LeCun, Y., Denker, J., Solla, S.: Optimal brain damage. Advances in neural information processing systems **2** (1989)
12. Li, Z., Xiao, J., Yang, L., Gu, Q.: Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17227–17236 (2023)
13. Lin, Y., Zhang, T., Sun, P., Li, Z., Zhou, S.: Fq-vit: Post-training quantization for fully quantized vision transformer. arXiv preprint arXiv:2111.13824 (2021)
14. Liu, X., Ding, X., Yu, L., Xi, Y., Li, W., Tu, Z., Hu, J., Chen, H., Yin, B., Xiong, Z.: Pq-sam: Post-training quantization for segment anything model. In: European Conference on Computer Vision. pp. 420–437. Springer (2024)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.O., Huang, B.: Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. IEEE Transactions on Geoscience and Remote Sensing (2024)
17. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical Image Analysis **102**, 103547 (2025). https://doi.org/10.1016/j.media.2025.103547
18. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022)
19. Yang, H., Yin, H., Shen, M., Molchanov, P., Li, H., Kautz, J.: Global vision transformer pruning with hessian-aware saliency. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18547–18557 (2023)
20. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
21. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
22. Zhang, X., Chen, E.Z., Zhao, L., Chen, X., Liu, Y., Maihe, B., Duncan, J.S., Chen, T., Sun, S.: Adapting vision foundation models for real-time ultrasound image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2025)

23. Zhang, Y., Ye, F., Yu, X., Lian, X., Jiang, T., Yang, L., Yang, L.: Embedded framework for clinical medical image segment anything in resource limited healthcare regions. npj Digital Medicine **8**(1), 568 (2025)
24. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)