Supplementary Material

Improvise, Adapt, Overcome — Telescopic Adapters for Efficient Fine-tuning of Vision Language Models in Medical Imaging

Ujjwal Mishra[1] Vinita Shukla[1] Praful Hambarde Amit Shukla

Centre for Artificial Intelligence and Robotics, Indian Institute of Technology Mandi, India
ujjwalmishra238@gmail.com, {d23097@students., praful@, amitshukla@}iitmandi.ac.in

A. Ablation Study Over Loss Functions

λ_d	λ_{BCE}	DSC(%)	IoU (%)
0	1	87.25	80.32
1	0	91.32	85.56
1.5	**1**	**92.18**	**86.12**
0.5	0.5	89.93	83.60
1	1.5	91.84	85.68

Table S1. Ablation study over the ISIC- 16 dataset on the coefficients for the composite Dice and BCE loss function. The configuration that achieved the highest performance is shown in **bold**.

We conducted an ablation study over the ISIC-16 dataset to establish the weights of the components of Dice (λ_d) and BCE (λ_{BCE}) components in our loss function (Eq. 6). The results, shown in Table S1, reveal that placing a higher weight on the Dice loss is critical for achieving optimal performance. By prioritizing the Dice component with $\lambda_d = 1.5$, we directly encourage the model to optimize spatial overlap and structural coherence, which are the primary goals of segmentation. This configuration achieved the highest scores of 92.18% DSC and 86.12% IoU, and was therefore selected for all of our experiments.

B. Alternate Adapter Placements

To examine the effect of adapter placement within the CLIPSeg architecture, we conduct an additional study using an alternative configuration shown in Fig. S1 that restricts adaptation to extracted feature representations rather than integrating adapters within transformer blocks. This simplified design isolates the contribution of decoder-level adaptation and provides insight into the importance of hierarchical versus interface-level modulation.

[1]These authors contributed equally.

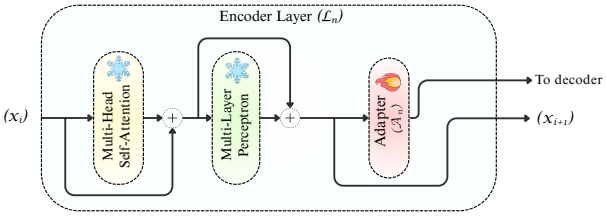

Figure S1. Alternate adapter placement strategy in an encoder layer $\mathcal{L}_n$. Unlike our main design, where adapters are inserted after the attention and MLP sublayers, this configuration integrates the adapter $\mathcal{A}_n$ after both residual connections, thereby restricting adaptation to the features passed to the decoder.

In contrast to the integration described in Section 3.1, where adapters are placed before each residual branch of the attention and MLP sublayers, the residual strategy introduces adaptation only at the interface between the frozen encoders and the decoder. This placement maintains the original transformer structure internally while enabling targeted modulation of the features passed to the segmentation decoder.

B.1. Vision Feature Adaptation

Rather than uniform adaptation, the residual strategy employs progressive scaling across the extracted vision layers similar to that described in Section 3.3 For the ith extracted layer among N total extract layers (9), the adapter dimension follows:

$$d_{\text{adapter}}^{(v,i)} = \max\left(8, \left\lfloor \frac{d_{\text{base}} \cdot i}{N} \right\rfloor\right)$$

This formulation assigns minimal capacity to early extracted features while progressively increasing adaptation strength for deeper representations that encode more task-relevant semantics. Each vision adapter follows the bot-

tleneck formulation in Eq. 2 with SiLU activation, layer normalization, and dropout regularization as described in Section 3.1.

B.2. Text Feature Adaptation

For the text encoder, only a single adapter is applied to the pooled output $\mathbf{z} \in \mathbb{R}^d$ of the text transformer. Letting $\mathbf{z}$ denote the embedding after pooling and before projection, the transformation is given by:

$$\mathcal{A}_t^{(\text{residual})}(\mathbf{z}) = \mathbf{z} + \alpha \cdot f_{\text{adapter}}(\mathbf{z})$$

where the adapter dimension is conservatively set to $d_{\text{adapter}}^{(t)} = \max(16, \lfloor d_{\text{base}}/4 \rfloor)$ to minimize interference with the pretrained text representations.

B.3. Cross-Modal Enhancement

An optional cross-modal interaction block fuses the residual-adapted text and vision embeddings in a shared projection space. The mechanism projects both modalities to a common dimension $d_{\text{shared}} = \max(32, d_{\text{base}}/2)$:

$$\mathbf{v}_{\text{proj}} = \text{LayerNorm}(\mathbf{W}_v \mathbf{v}_{\text{pooled}})$$

$$\mathbf{t}_{\text{proj}} = \text{LayerNorm}(\mathbf{W}_t \mathbf{z}_{\text{adapted}})$$

A lightweight multi-head attention layer with 4 heads enables soft alignment between modalities:

$$\mathbf{c}_{\text{cross}} = \text{MultiHeadAttn}(\mathbf{v}_{\text{proj}}, \mathbf{t}_{\text{proj}}, \mathbf{t}_{\text{proj}})$$

The cross-attended features are projected back to the conditional embedding space and added to the original conditional embeddings, providing cross-modal enhancement without replacing the hierarchical fusion in the decoder.

B.4. Conditional Adaptation and Decoder Enhancement

We retain the conditional adapter $\mathcal{A}_c$ as described in Section 3.1, applying it after projection of the text embedding with dimension $d_{\text{adapter}}^{(c)} = \max(16, \lfloor d_{\text{base}}/8 \rfloor)$.

Additionally, a lightweight decoder enhancement module refines the segmentation boundaries through a compact CNN with attention-based weighting:

$$\mathbf{A}_{\text{mask}} = \sigma(\text{Conv}_{1\times1}(\text{SiLU}(\text{BN}(\text{Conv}_{3\times3}(\mathbf{L})))))$$

$$\mathbf{L}_{\text{enhanced}} = \mathbf{L} \odot \mathbf{A}_{\text{mask}}$$

where $\mathbf{L}$ represents the decoder logits and $\odot$ denotes element-wise multiplication.

C. Comparison with LoRA at a Comparable Parameter Budget

To isolate the effect of the telescopic design from the parameter budget, we conducted a comparative experiment against a standard Low-Rank Adaptation (LoRA) [5] implementation. We configured LoRA on the CLIPSeg [7] baseline by applying it across all attention layers of both the vision and text encoders. A rank of $r = 4$ was selected, resulting in 585k trainable parameters. This configuration creates a direct comparison to our method's 613k parameters, placing both models in a nearly identical, ultra-lightweight parameter regime. All other training hyperparameters, dataset splits, and evaluation protocols were held constant, as described in the main paper.

The results in Table S2 demonstrate that our Telescopic Adapters consistently and significantly outperform the parameter-matched LoRA model across all five medical datasets. This performance discrepancy suggests that the low expressive capacity of a small, uniform rank ($r = 4$) is insufficient for this task. The model struggles to capture the complex feature shifts required for adapting from open-domain pretraining to specialized medical semantics, especially given the small size of the target datasets.

Dataset	Metric	LoRA (r=4, 585k)	Telescopic Adapters [OURS] (613k)
Kvasir-SEG [6]	DSC (%)	77.20	**89.79**
	IoU (%)	72.89	**83.50**
BKAI [8]	DSC (%)	61.80	**88.38**
	IoU (%)	58.29	**81.63**
ClinicDB [2]	DSC (%)	66.85	**91.67**
	IoU (%)	60.48	**85.19**
ISIC-16 [4]	DSC (%)	59.15	**92.18**
	IoU (%)	52.51	**86.12**
BUSI [1]	DSC (%)	43.10	**65.90**
	IoU (%)	37.32	**59.10**

Table S2. Quantitative comparison between a standard LoRA implementation with a comparable parameter budget (585k) and our Telescopic Adapters (613k) on the baseline CLIPSeg [7] model. Best results are in **bold**.

D. Additional Qualitative Analysis

Additional visual results of our proposed Telescopic Adapters are provided in Fig. S2. Across the five distinct medical datasets—$ClinicDB$ [2], $ISIC-16$ [4], $BUSI$ [1], $BKAI$ [8], and $Kvasir-SEG$ [6]. Our method, shown in column (i), consistently produces segmentation masks with exceptional fidelity to the ground truth (GT) in column (j). This stands in stark contrast to several baseline methods; for instance, zero-shot segmentation with $ClipSeg$ [7] (b) frequently fails to identify the

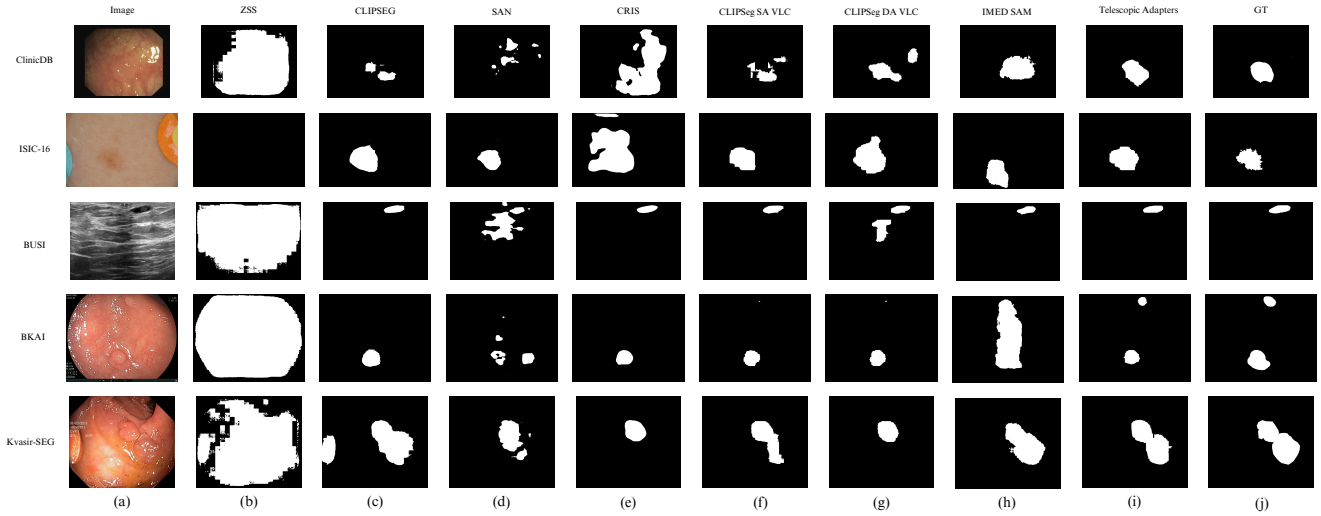

Figure S2. Additional segmentation results on samples from ClinicDB [2], ISIC-16 [4], BUSI [1], BKAI [8], and Kvasir-SEG [6] datasets (rows 1-5 respectively). Column (a) shows the original medical images, followed by segmentation masks from: (b) ZSS (zero-shot segmentation using CLIPSeg [7]), (c) CLIPSeg [7] (end-to-end fine-tuned), (d) SAN [11], (e) CRIS [9] (end-to-end fine-tuned), (f) CLIPSeg SA VLC [3], (g) CLIPSeg DA VLC [3], (h) I-MedSAM [10], (i) Telescopic Adapters, and (j) GT (ground truth).

target region, while models like SAN [11](d) often introduce significant noise and artifacts. Even when compared against strong, end-to-end fine-tuned competitors, our model demonstrates a clear advantage. Notably, on the challenging low-contrast $BUSI$ ultrasound image (row 3), our method successfully localizes the lesion, a task where most other models, including $CRIS$ [9] (e) and fine-tuned $CLIPSeg$ [7] (c), completely fail. Furthermore, the results on the $BKAI$ dataset (row 4) reveal a critical point of differentiation: while $I-MedSAM$ [10] (h) performs well on other samples, it fails catastrophically on this image. This specific case highlights not only the accuracy of our model but also its superior consistency and reliability over other state-of-the-art approaches, effectively capturing both simple and complex anatomical structures with precise boundaries and minimal error.

E. Quantitative Analysis

We conduct an ablation study on adapter placement to compare our proposed telescopic positioning where adapters are integrated within transformer encoder blocks before each residual addition, resulting in two adapter applications per layer, against the alternate configuration as discussed in Section B, as shown in Table S3.

The vision and text configuration (••) amplifies the advantages of telescopic placement when incorporating textual conditioning. With 593k parameters versus 1.7M for the alternate approach, our telescopic adapters achieve superior performance across four of five datasets. Notable improvements include Kvasir-SEG (89.67 % vs 85.5 % Dice, 83.62 % vs 77.46 % IoU) and ClinicDB (91.28 % vs 88.75 % Dice), while maintaining 2.87× parameter efficiency compared to the alternate configuration.

The complete vision, text, and conditional configuration (• • •), despite utilizing only 613k parameters compared to 1.77M for feature-level adaptation, our telescopic approach maintains superior performance on four datasets. Peak Dice coefficients on Kvasir-SEG (89.79 %), BKAI (88.38 %), and ClinicDB (91.67 %) demonstrate the efficacy of hierarchical feature modulation across encoder depths. The alternate configuration achieves competitive results solely on BUSI across all modality configurations.

This placement comparison confirms that integrating adapters within transformer blocks before residual connections enables more effective feature adaptation than feature-level modulation, achieving superior parameter efficiency and segmentation accuracy across medical datasets. The results substantiate our telescopic placement strategy over feature-level adaptation for medical VLSM fine-tuning.

Dataset	Metric	Vision only (•)		Vision & Text (••)		Vision, Text, & Conditional (•••)	
		Telescopic 498k	A. Telescopic 852k	Telescopic 593k	A. Telescopic 1.7M	Telescopic 613k	A. Telescopic 1.77M
Kvasir-SEG [6]	Dice (%)	**87.35**	85.78	**89.67**	85.50	**89.79**	84.98
	IoU (%)	**80.32**	78.67	**83.62**	77.46	**83.5**	77.28
BKAI [8]	Dice (%)	**85.53**	82.66	**87.09**	82.81	**88.38**	83.57
	IoU (%)	**77.77**	73.73	**80.00**	74.05	**81.63**	74.96
ClinicDB [2]	Dice (%)	85.39	**88.42**	**91.28**	88.75	**91.67**	88.62
	IoU (%)	78.45	**81.17**	**84.85**	81.45	**85.19**	81.42
ISIC-16 [4]	Dice (%)	**91.61**	91.19	**92.24**	91.27	**92.18**	91.35
	IoU (%)	**85.30**	84.64	**86.16**	84.73	**86.12**	84.81
BUSI [1]	Dice (%)	70.35	**71.90**	64.33	**78.58**	65.90	**70.77**
	IoU (%)	62.45	**64.57**	57.26	**64.93**	59.10	**63.42**

Table S3. Performance comparison of the proposed Telescopic Adapters and the Alternate Telescopic (A. Telescopic) Adapters across five datasets. The configurations shown are: Vision only (•), Vision and Text (••), and Vision, Text, and Conditional (•••). Best results for each modality are in **bold**.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, H.M. Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 11 2019. 2, 3, 4

[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. 2, 3, 4

[3] Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Vlsm-adapter: Finetuning vision-language segmentation efficiently with lightweight blocks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 712–722. Springer, 2024. 3

[4] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2, 3, 4

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[6] 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*, pages 451–462. Springer, 2019. 2, 3, 4

[7] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 2, 3

[8] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *Advances in visual computing: 16th international symposium, ISVC 2021, virtual event, October 4-6, 2021, proceedings, part II*, pages 15–28. Springer, 2021. 2, 3, 4

[9] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3

[10] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 90–107, Cham, 2025. Springer Nature Switzerland. 3

[11] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2945–2954, June 2023. 3