Uncertainty-Aware Vision-Language Segmentation for Medical Imaging

Supplementary Material

Table 1. Comparison of Monomodal and Multimodal methods on polyp segmentation across four datasets: ClinicDB, ColonDB, CVC-300, and ETIS. Metrics include Dice score (%) and mean Intersection over Union (mIoU, %). **Black**: best, **Green**: second best, **Blue**: third best values.

Modality	Method	ClinicDB		ColonDB		CVC-300		ETIS	
		Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
MonoModels	UNet [1]	82.49	75.61	52.33	45.58	71.35	62.84	39.82	33.52
	UNet++ [2]	79.49	72.95	48.35	41.07	70.74	62.45	40.15	34.47
	TransUNet [3]	85.12	78.23	58.94	51.47	76.88	67.91	46.33	38.72
	Swin-Umamba [4]	83.67	76.89	56.21	49.15	74.45	65.83	44.28	37.09
	U-Mamba [5]	84.35	77.48	57.62	50.31	75.91	67.12	45.71	38.15
Multimodal	BiomedClip [6]	83.61	76.51	67.40	50.83	72.89	63.25	52.46	40.89
	LViT [7]	89.20	82.54	73.12	57.62	84.53	71.88	63.24	53.80
	Ariadne [8]	**93.87**	**88.45**	**77.82**	**63.70**	**90.31**	**82.34**	**82.36**	**70.01**
	M-Adapter [9]	**95.46**	**91.32**	**78.85**	**65.80**	**91.94**	**85.09**	**82.85**	**70.72**
	Our Model	**96.74**	**92.87**	**81.43**	**68.91**	**94.17**	**87.38**	**85.92**	**74.16**

1. Additional Experiments

For the additional experiments, we evaluate our methods on four benchmark polyp segmentation datasets.

Datasets: **ClinicDB** [10] includes 612 colonoscopy images with manual region annotations for polyps, selected to capture variability in polyp shape, size, and background complexity. For vision-language model training, each image is paired with a synthetic language prompt detailing polyp attributes, uniformly curated under the MedVLSM [11]. **ColonDB** [12] contains 380 frames sampled from colonoscopy videos, representing diverse polyp presentations, illumination conditions, and camera angles, with region-level masks and MedVLSM-aligned descriptive prompts for clinically relevant attributes. **CVC-300** [13] consists of 60 carefully selected colonoscopy images annotated with polyp region masks, complemented with synthetic captions curated via MedVLSM to ensure annotation consistency. Finally, **ETIS-Larib** [14] comprises 196 challenging images featuring subtle and small polyps in complex mucosal backgrounds, each with expert region annotations and MedVLSM-curated textual prompts to facilitate multimodal learning.

Experimental evaluation across four benchmark datasets demonstrates multimodal approaches superiority over monomodal segmentation for polyp detection. On **ClinicDB**, while UNet and TransUNet achieve 82.49% and 85.12% Dice respectively, our multimodal model reaches 96.74% Dice (+1.28% over M-Adapter's 95.46%). The challenging **ColonDB** dataset emphasizes monomodal limitations, with TransUNet achieving only 58.94% Dice versus our model's 81.43% (+2.58% over M-Adapter's 78.85%). On **CVC-300**, monomodal methods plateau below 80% Dice, while our approach achieves 94.17% (+2.23% over M-Adapter). Most compelling results emerge from **ETIS**, where small, low-contrast polyps limit UNet to 39.82% Dice. Our model achieves 85.92% Dice (+3.07% over M-Adapter's 82.85%), representing clinically significant improvements for challenging segmentation tasks.

Performance trends on **CVC-300** confirm the superiority of cross-modal fusion strategies, with monomodal architectures plateauing below 80% Dice despite advanced designs like U-Mamba (75.91% Dice) and TransUNet (76.88% Dice). Our multimodal approach achieves 94.17% Dice and 87.38% mIoU, outperforming M-Adapter by +2.23% Dice and +2.29% mIoU, demonstrating robust segmentation capabilities across diverse polyp morphologies and imaging conditions. The most compelling evidence emerges from the highly challenging **ETIS** dataset, where small, low-contrast polyps embedded in complex backgrounds severely limit monomodal performance, with UNet achieving only 39.82% Dice and TransUNet reaching 46.33% Dice. While advanced multimodal methods like Ariadne (82.36% Dice) and M-Adapter (82.85% Dice) show substantial improvements, our model further advances the State-of-the-Art to 85.92% Dice and 74.16% mIoU, representing significant gains of +3.07% Dice and +3.44% mIoU that could meaningfully impact polyp detection rates in challenging scenarios.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241. 1

[2] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*. Springer, 2018, pp. 3–11. 1

[3] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524002056 1

[4] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang, H. Zheng, and S. Wang, " Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15009. Springer Nature Switzerland, October 2024. 1

[5] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2401.04722 1

[6] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Large-scale domain-specific pretraining for biomedical vision-language processing," *arXiv preprint arXiv:2303.00915*, vol. 2, no. 3, p. 6, 2023. 1

[7] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: Language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 96–107, 2024. 1

[8] Y. Zhong, M. Xu, K. Liang, K. Chen, and M. Wu, "Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 724–733. 1

[9] X. Zhang, B. Ni, Y. Yang, and L. Zhang, " MAdapter: A Better Interaction between Image and Language for Medical Image Segmentation ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15009. Springer Nature Switzerland, October 2024. 1

[10] J. Bernal, J. M. Sanchez, J. Vila, A. Iglesias, C. Rodriguez, G. Fernández-Esparrach, P. Radeva, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," in *Computerized Medical Imaging and Graphics*, vol. 43, 2015, pp. 99–111, clinicDB dataset. 1

[11] K. Poudel, M. Dhakal, P. Bhandari, R. Adhikari, S. Thapaliya, and B. Khanal, "Exploring transfer learning in medical image segmentation using vision-language models," *arXiv preprint arXiv:2308.07706*, 2023. 1

[12] J. Bernal, N. Tajkbaksh, J. M. Sanchez, R. Mazo, Q. Angermann, P. Hammel, P. Radeva, C. Mérie, X. Garcia, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," in *Pattern Recognition*, vol. 48, no. 11, 2015, pp. 3707–3718, colonDB dataset. 1

[13] ——, "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *Medical Image Analysis*, vol. 35, pp. 327–340, 2017, cVC-300 dataset. 1

[14] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward automatic polyp detection with a polyp appearance model," in *Pattern Recognition*, vol. 48, no. 11, 2014, pp. 3707–3718, eTIS-Larib dataset. 1