Exploring Transfer Learning in Medical Image Segmentation using Vision-Language Models

Kanchan Poudel* KANCHAN.POUDEL@NAAMII.ORG.NP
Manish Dhakal* MANISH.DHAKAL@NAAMII.ORG.NP
Prasiddha Bhandari* PRASIDDHA.BHANDARI@NAAMII.ORG.NP
Rabin Adhikari* RABIN.ADHIKARI@NAAMII.ORG.NP
Safal Thapaliya* SAFAL.THAPALIYA@NAAMII.ORG.NP
Bishesh Khanal BISHESH.KHANAL@NAAMII.ORG.NP
Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal

Editors: Under Review for MIDL 2024

Abstract

Medical image segmentation allows quantifying target structure size and shape, aiding in disease diagnosis, prognosis, surgery planning, and comprehension. Building upon recent advancements in foundation Vision-Language Models (VLMs) from natural image-text pairs, several studies have proposed adapting them to Vision-Language Segmentation Models (VLSMs) that allow using language text as an additional input to segmentation models. Introducing auxiliary information via text with human-in-the-loop prompting during inference opens up unique opportunities, such as open vocabulary segmentation and potentially more robust segmentation models against out-of-distribution data.

Although transfer learning from natural to medical images has been explored for image-only segmentation models, the joint representation of vision-language in segmentation problems remains underexplored. This study introduces the first systematic study on transferring VLSMs to 2D medical images, using carefully curated 11 datasets encompassing diverse modalities and insightful language prompts and experiments. Our findings demonstrate that although VLSMs show competitive performance compared to image-only models for segmentation after finetuning in limited medical image datasets, not all VLSMs utilize the additional information from language prompts, with image features playing a dominant role. While VLSMs exhibit enhanced performance in handling pooled datasets with diverse modalities and show potential robustness to domain shifts compared to conventional segmentation models, our results suggest that novel approaches are required to enable VLSMs to leverage the various auxiliary information available through language prompts. The code and datasets are available at https://github.com/naamiinepal/medvlsm.

1. Introduction

Medical image segmentation is crucial for various clinical applications such as diagnosis, prognosis, and surgery planning. The latest supervised segmentation models exhibit promising outcomes across diverse imaging modalities, anatomies, and diseases (Milletari et al., 2016; Havaei et al., 2017; Zhou et al., 2018; Chen et al., 2021; Isensee et al., 2021; Hatamizadeh et al., 2022; Oktay et al., 2022; Wazir and Fraz, 2022). Despite their success, these models are constrained to predefined foreground classes on specific modalities and

* Contributed equally. The order is in the ascending order of the authors' first names.

© 2024 CC-BY 4.0, K. Poudel, M. Dhakal, P. Bhandari, R. Adhikari, S. Thapaliya & B. Khanal.

anatomies, lacking adaptability to auxiliary information and hindering their application outside extensive population-based studies.

The integration of VLMs (Huang et al., 2020; Jia et al., 2021; Li et al., 2021; Radford et al., 2021; Fürst et al., 2022; Singh et al., 2022; Zhai et al., 2022) into VLSMs (Lüddecke and Ecker, 2022; Rao et al., 2022; Wang et al., 2022) presents a paradigm shift in medical image segmentation. Models like CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2023a), capable of joint text-image representation, allow for auxiliary information incorporation through language prompts during segmentation. This approach can enhance interpretability and robustness against domain shift and out-of-distribution data.

While transfer learning from natural to medical images for image-only representation learning has been extensively explored (Ghafoorian et al., 2017; Cheplygina et al., 2019; Amin et al., 2019), only a few such studies have been done for joint vision-language representation (Qin et al., 2022). Yet, two critical questions persist (**i**) the generalizability of this approach across multiple VLSMs for segmentation tasks, and (**ii**) the nuanced role of language prompts vs. images during finetuning and the VLSMs' capacity to handle pooled dataset training and out-of-distribution data.

This work presents the first systematic study on VLSM transfer learning to the medical images, using four models based on the two most popular contrastive VLMs: CLIP pretrained on natural image-text pairs and BiomedCLIP pretrained in the medical domain.

Key contributions include meticulous dataset selection (11 datasets) across four 2D medical image modalities, diverse anatomical structures, and pathology. We also enrich existing datasets with diverse language prompts generated through automated methods utilizing image metadata, VQA models, and segmentation masks. Our extensive experiments with four VLSMs, diverse datasets, and carefully designed prompts explore intricate relationships between language and image during joint representation adaptation for medical images. We evaluate robustness against domain shift and the ability to handle pooled datasets with diverse modalities, attributes, and targets. Finally, we open-source our framework, source code, and prompts, promoting transparency and reproducibility in the scientific community.

2. Method

2.1. CLIP- and BiomedCLIP-based Medical VLSMs

We create four medical VLSMs using CLIP and BiomedCLIP: (**i**) Finetuning CLIP-based VLSMs, **CLIPSeg** (Lüddecke and Ecker, 2022) and **CRIS**[1] (Wang et al., 2022), pretrained on natural image-text pairs, and (**ii**) Building two new VLSMs for the medical domain by adding a decoder to BiomedCLIP, pretrained on medical image-text pairs. The proposed new models are **BiomedCLIPSeg-D** (with a pretrained CLIPSeg decoder) and **BiomedCLIPSeg** (with a randomly initialized decoder of CLIPSeg). A sample from the datasets in our experiments is a triplet of a medical image, a segmentation mask, and a text prompt. Figure 1 displays the overall VLSM architecture.

CLIPSeg accommodates both CNN and ViT (Dosovitskiy et al., 2020) backbones, whereas CRIS only supports a CNN-based CLIP backbone. BiomedCLIPSeg-based models include transformer-based backbones for both the encoders. We study CLIPSeg and CRIS

1. We used unofficial weights from a GitHub issue since the authors haven't released the model weights yet.

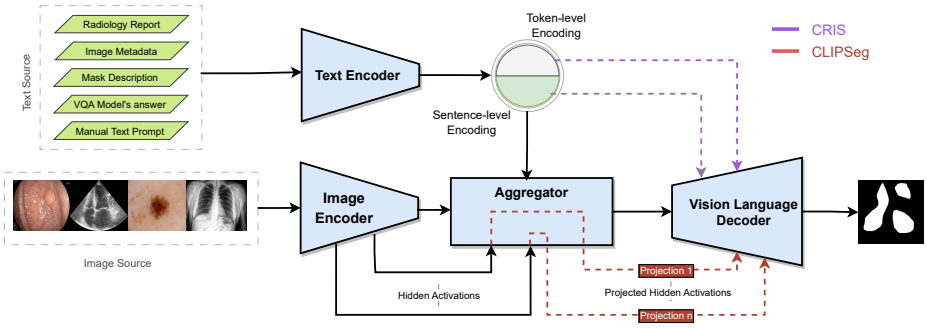


Figure 1: CRIS and CLIPSeg-variants include Text and Image encoders, an Aggregator, and a Vision-Language Decoder.

in both zero-shot and finetuning while only finetuning for BiomedCLIPSeg-based models as they lack an end-to-end pretrained encoder-decoder.

2.2. Datasets

We collected 11 2D medical imaging datasets of diverse modalities, organs, and pathologies covering both radiology and non-radiology images for binary and multi-class segmentation tasks (see Table 1). All the datasets are used for finetuning separately or combined (as a single pooled dataset) except the last three endoscopy datasets (ETIS, ColonDB, and CVC300), which are used only as the test split to study domain shift robustness.

Table 1: Datasets overview for single and multi-class segmentation tasks.

Category	Modality	Organ	Name	Foreground Class(es)	# train/val/test
Non-Radiology	Endoscopy	Colon	Kvasir-SEG	Polyp	800/100/100
			ClinicDB		490/61/61
			BKAI		800/100/100
			ETIS		0/0/196
			ColonDB		0/0/380
			CVC300		0/0/60
	Photography	Skin	ISIC 2016	Skin Lesion	810/90/379
		Foot	DFU 2022	Foot Ulcer	1600/200/200
Radiology	Ultrasound	Heart	CAMUS	Myocardium, Left ventricular, and Left atrium cavity	4800/600/600
		Breast	BUSI	Benign and Malignant Tumors	624/78/78
	X-Ray	Chest	CheXlocalize	Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices	1279/446/452

2.3. Generating Language Prompts

Although language prompts enable injecting rich information into VLSMs, manually crafting individual image-specific prompts becomes impractical for large-scale evaluations. Thus, we implement an automated prompt generation system for extensive assessments of medical VLSMs. This involves incorporating semantic concepts such as size, position, color, and specific medical attributes like gender, age, and pathology.

In addition to automated prompts, we introduce manual prompts that provide general class-level information applicable to all samples within a given dataset. The generated language prompts encapsulate a comprehensive set of attributes and information, comprising: (**i**) Inspired by Tomar et al. (2022), *number*, *size*, and *relative location* are derived through image processing on segmentation masks. (**ii**) Motivated by Qin et al. (2022), we use *shape* and *color* information from VQA queries. (**iii**) *General class information*, extracted for photographic images from online medical journals, provides overarching details applicable across different datasets. Notably, Qin et al. (2022) used PubMedBERT (Gu et al., 2021) for this purpose; however, our experiments revealed its unreliability, leading us to manually gather this information from online medical journals (see Table 5). (**iv**) Attributes like *age*, *gender of patients*, *image quality*, *cardiac cycle*, and *tumor type* are extracted whenever available, contributing valuable context to the language prompts. There are 14 such attributes, (**a1** to **a14**), which we combined in various ways to build nine distinct prompt types (**P1** to **P9**) for each dataset (Table 9; Appendix G). Each prompt type caters to specific attribute combinations, prioritizing the class name as the foundational attribute and enhancing the versatility of the generated prompts.

2.4. Implementation Details

We finetuned VLSMs with minimal hyperparameter changes from the original pretraining settings. AdamW (Loshchilov and Hutter, 2017) optimizer with weight decay of 10^{-3}, and initial learning rates of 2×10^{-3} (CLIPSeg) and 2×10^{-5} (CRIS) were utilized. Dice loss was used alongside Binary Cross Entropy loss scaled by 0.2. The learning rate was reduced by 10 times if validation loss did not decrease for 5 consecutive epochs. Batch sizes of 128 and 32 were used for CLIPSeg and CRIS, respectively, due to the difference in model sizes[2].

3. Results

VLSMs adapt better to non-radiology images in Zero-Shot Setting (ZSS). Both CRIS and CLIPSeg barely work in ZSS for radiology images except for CRIS in BUSI dataset but get a Dice score in the range of $20\% - 70\%$ for non-radiology datasets, with 67.98% being the highest Dice score for ISIC (Figure 2). Adding more attributes to the prompt generally improved performance, but the gain is inconsistent across prompts and datasets.

Image-specific-attributes or general descriptions? In the ZSS, CRIS performs better on endoscopy datasets when prompts contain image-specific attributes (*size*, *number*, and *location*; **P4**, **P5**, and **P6**; Figure 2), but degrades with non-image-specific attributes added (**P7**, **P8**, **P9**). Interestingly, prompts with general descriptions (**P8** and **P9**) achieve the highest performance on the DFU 2022 dataset, possibly due to pretrained models' familiarity with feet and skin compared to the colon. This highlights the complex relationship between pretraining data, VLSM architecture, and the medical segmentation task.

Making prompts richer does not always help during finetuning. Figure 2 shows that the DSC variation across prompt type is minimal in the finetuned setting for all the models. Prompt with only *class name* (**P1**) improves segmentation performance in

2. Further details are in Appendix C.

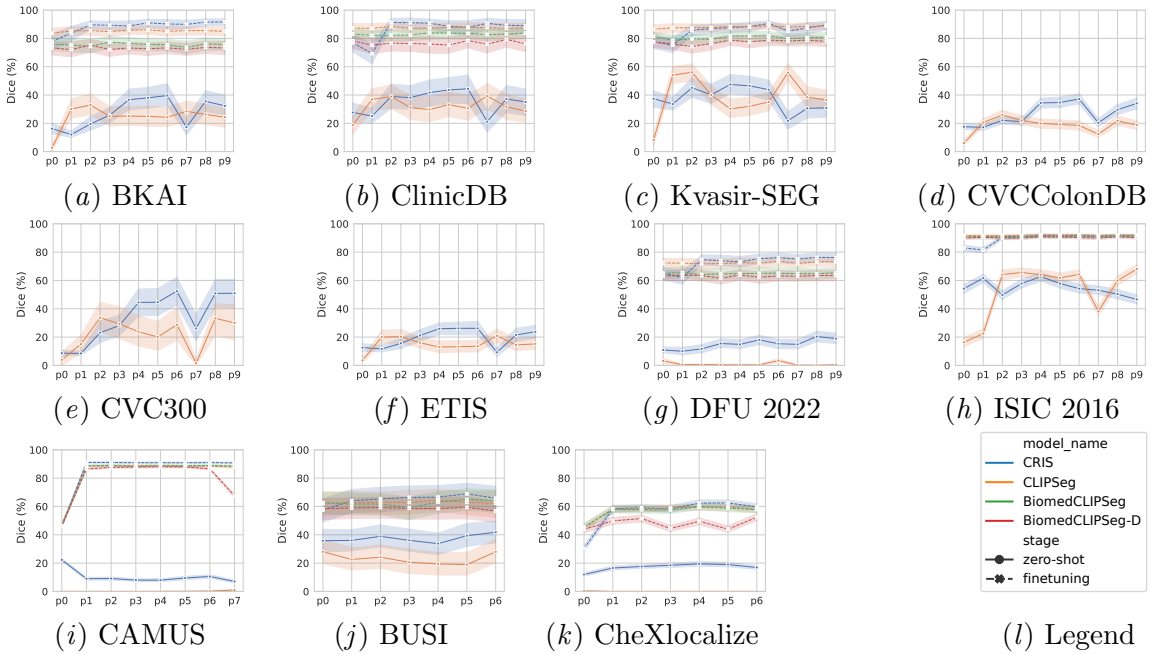


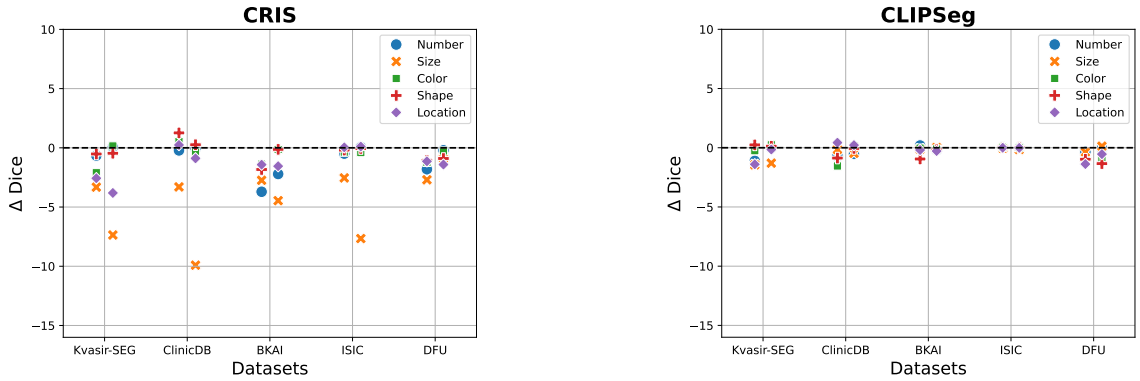
Figure 2: Zero-shot and finetuning performance of CRIS, CLIPSeg, BiomedCLIPSeg, and BiomedCLIPSeg-D model on non-radiology (first two rows) and radiology datasets (last row). Finetuning using the prompts improves performance compared to the empty prompt, particularly in multi-class settings.

radiology datasets for all four VLSMs. While CRIS' performance almost saturates after adding the *class name* and *mask shape* (**P2**), the rest of the models have similar performance for all the prompts except **P0** with multi-class segmentation (CAMUS and CheXlocalize).

BiomedCLIPSeg and BiomedCLIPSeg-D, despite being based on a VLM pretrained on medical data, consistently perform poorly across all prompts compared to CLIP and CLIPSeg. This is likely because it has not been further pretrained for segmentation tasks on a large-scale dataset. Subsequent experiments use better performing CLIPSeg and CRIS to study the impact of individual attributes and robustness of VLSMs[3].

When finetuned, CRIS captures some language semantics better than CLIPSeg. We replaced attribute values of the input prompts during inference with random uncommon English words and semantically wrong or opposite values to assess whether VLSMs leverage the language semantics. Figure 3 shows that altering attributes minimally impacts CLIPSeg's performance but notably deteriorates CRIS's. To further investigate CLIPSeg's indifference to attribute values, we provided only the *class name*(**P1**) as input during inference to the model trained on rich prompts **P6**; the results were very similar to providing the rich prompts, reinforcing the minimal impact of attributes in CLIPSeg.

3. Additionally, we have also trained both the models, keeping their encoders frozen whose results are shown in Appendix F.1.

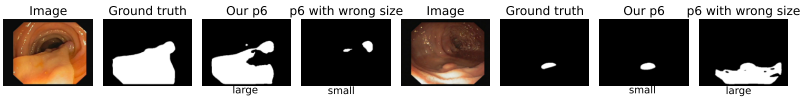


(*a*) CRIS (*b*) CLIPSeg

Figure 3: Relative change in percentage dice score on replacing attribute values by a random uncommon English word (left of vertical lines) or semantically opposite value such as replacing 'large' with 'small' (right of vertical lines) in prompt *P6*.

CRIS's performance decreases notably for attributes like size and location. The decline is more significant when providing semantically opposite values than random uncommon English words, indicating robust semantic learning. A qualitative examination of predicted segmentation masks confirms this trend (Figure 4).

(*a*) Kvasir-SEG for attribute *location*



(*b*) Clinicdb for attribute *size*

Figure 4: Examples of images with the highest drops in dice score for two datasets when values for sensitive attributes are replaced with another value within the value set of the attributes in the dataset in *P6*.

Finetuned VLSMs comparable to SOTA segmentation models. Table 2 compares VLSMs vs. traditional CNN-based models (Ronneberger et al., 2015; Chen et al., 2018; Zhou et al., 2018) on their ability to learn in two scenarios: when trained on (**i**) individual specialized datasets or (**ii**) a pooled dataset that combines diverse datasets into a single training set. While the segmentation models (CNNs and VLSMs) achieve better on pooled endoscopy datasets than individual endoscopy datasets, performance mainly drops when

Table 2: Performance of VLSMs (Dice (%)) and CNN models when finetuning in different combinations of datasets. For each column, **Bold** and **Bold with underline** represent the best result among all models for the specific dataset combination and all combinations, respectively.

Finetuned Dataset	Model \ Tested Dataset →	Kvasir-SEG	ClinicDB	BKAI	CVC-300	CVC-ColonDB	ETIS	ISIC	DFU	CAMUS	BUSI	CheXlocalize
Individual	CRIS	**91.39**	**91.69**	**92.40**	-	-	-	91.94	**76.13**	**91.09**	69.31	**62.57**
	CLIPSeg	89.51	88.74	86.47	-	-	-	**92.12**	73.24	88.85	64.32	59.56
	UNet	84.77	85.65	83.79	-	-	-	90.40	67.87	90.19	**75.21**	50.29
	UNet++	84.70	84.16	84.61	-	-	-	90.12	69.95	89.95	72.55	49.53
	DeepLabv3+	84.11	89.11	84.95	-	-	-	90.66	67.89	90.43	70.57	49.95
	*SOTA**	95.02	95.73	90.23	-	-	-	92.00	72.87	94.10	89.80	-
Pooled	CRIS	**90.23**	**91.88**	**90.21**	**88.99**	**78.07**	**75.93**	**91.99**	**75.55**	**91.00**	67.89	**61.01**
	CLIPSeg	87.25	87.49	87.30	87.24	71.32	69.64	91.34	71.94	88.76	66.02	56.60
	UNet	36.60	26.10	37.70	4.94	8.55	12.00	64.90	38.60	76.82	44.60	38.00
	UNet++	80.52	78.21	77.87	87.80	51.92	48.16	88.41	65.78	89.99	75.59	53.88
	DeepLabv3+	82.40	82.70	77.60	84.40	59.30	48.30	89.60	67.70	90.17	**77.80**	54.56
Endoscopy Pooled	CRIS	**91.25**	**92.94**	**92.35**	**90.42**	**81.00**	**79.67**	-	-	-	-	-
	CLIPSeg	89.62	88.96	86.98	88.98	75.23	71.18	-	-	-	-	-
	UNet	85.45	88.17	84.70	90.27	67.87	61.84	-	-	-	-	-
	UNet++	83.99	85.44	82.27	89.4	66.61	55.62	-	-	-	-	-
	DeepLabv3+	87.87	87.60	84.38	87.54	69.95	65.24	-	-	-	-	-
	**SOTA Sources*	Dumitru et al. (2023)	Fitzgerald and Matuszewski (2023)	Tomar et al. (2022)	-	-	-	Hasan et al. (2022)	Liao et al. (2022)	Ling et al. (2022)	Zhang et al. (2023b)	-

training on a pooled set comprising all the datasets. VLSMs outperform image-only off-the-shelf CNN-based methods in most cases. We have compared with the best method reported in the literature for each dataset as well.[4] The state-of-the-art results[5] are better, although VLSMs seem to have competitive performance.

VLSMs adapt better to distribution shifts. To assess the ability of the segmentation models to transfer knowledge learned from one dataset to another similar one, we train the models on each large endoscopy dataset (Kvasir-SEG, ClinicDB, and BKAI) and evaluate them on all endoscopy datasets. Table 3, shows that VLSMs perform better in all the cases than the conventional models for endoscopic datasets. VLSMs show smaller performance drops than conventional models when trained on a different distribution from the test set.

4. Discussion, Limitations, and Conclusion

VLSMs pretrained on natural images show suboptimal zero-shot accuracy with medical images for practical use but provide a foundation for joint text-image representation. Our study provides intriguing insights into prompt design, attributes' roles, and models' performance when finetuning across diverse datasets. The zero-shot segmentation performance showed improvement across all non-radiology datasets when compared to the radiology datasets. This could be attributed to the non-radiology medical imaging modalities being closer to open-domain images, as well as the potential familiarity with organs such as skin and feet (for ISIC and DFU datasets) during pretraining. The best-performing prompts vary with datasets but often include attributes familiar to models during pretraining. For

4. To ensure a thorough comparison across datasets with diverse modalities and state-of-the-art methods, we report the state-of-the-art for each dataset from literature, apart from implementing a few commonly used CNN baselines.
5. Except for CAMUS and ISIC, may have different training, validation, and test splits due to the unavailability of the standard splits in literature.

Table 3: Segmentation performance (Dice (%)) on out-of-distribution endoscopy datasets. For each column, **Bold** and **Bold with underline** show the best result across the model concerning the tested dataset for each finetuning dataset and across the finetuning datasets, respectively. The shaded results correspond to results in test sets of the same distribution, while the rest are on out-of-distribution test sets.

Tested on → Finetuned on ↓	Model ↓	Kvasir-SEG	ClinicDB	BKAI	CVC-300	CVC-ColonDB	ETIS
Kvasir-SEG	CRIS	**91.39**	**82.99**	**83.26**	86.15	**76.87**	**62.99**
	CLIPSeg	89.51	80.21	77.89	**86.49**	70.46	62.83
	UNet	84.77	64.84	66.22	77.16	50.81	34.98
	UNet++	84.70	68.15	61.76	79.35	52.3	32.81
	DeepLabv3+	84.11	68.0	63.57	76.93	58.41	33.81
ClinicDB	CRIS	82.66	**91.69**	**76.21**	**87.47**	**76.14**	**64.62**
	CLIPSeg	**84.02**	88.74	72.04	87.07	67.91	60.09
	UNet	65.80	85.65	35.26	73.91	55.01	29.66
	UNet++	61.93	84.16	38.81	71.15	55.05	23.16
	DeepLabv3+	66.63	89.11	40.89	82.05	61.79	39.53
BKAI	CRIS	**83.74**	**78.18**	**92.40**	79.48	**65.30**	66.72
	CLIPSeg	83.70	76.07	86.47	**86.06**	63.59	**66.97**
	UNet	68.42	62.20	83.79	60.13	44.52	42.91
	UNet++	70.64	62.66	84.61	82.44	55.60	46.84
	DeepLabv3+	69.02	61.99	84.95	77.47	53.15	49.61

instance, CRIS trained on RefCOCO (Kazemzadeh et al., 2014) for referring image segmentation captures size, location, and number well.

The ability of CRIS to leverage better language semantics than CLIPSeg might be due to (**i**) CRIS's architecture that focuses on token-level intervention instead of CLIPSeg's sentence-level embedding, and (**ii**) end-to-end VLSM training of CRIS compared to CLIPSeg's training for segmentation task with frozen CLIP encoder. Interestingly, models based on CLIP performing better than those based on BiomedCLIP (pretrained with image-text pairs of 400 million natural domain versus 15 million medical domain) shows that large-scale dataset has the benefit that is hard to achieve with smaller-scale domain-specific data.

Our study aims to build insights into how well VLSMs leverage textual information and perform transfer learning in the medical domain. It proposes pragmatic prompt settings and systematic experiments instead of implementing an exhaustive list of VLSMs and only grossly comparing their performance. The four CLIP-based VLSMs cover significant variations in architecture to capture global vs. token level information in prompts, training approach with end-to-end for referring image segmentation vs. finetuning only decoder for segmentation, and based on VLM pretrained on natural vs. medical domain, etc. We focus only on 2D medical images, excluding 3D modalities like MRI or CT scans, as most existing VLSMs are suitable only for 2D images, requiring further research in building 3D VLSMs.

While the VLSMs' performance seems on par with image-only architectures, and some of the VLSMs use information injected via text prompts, our results show that further research is needed to develop novel approaches that can better leverage the rich information provided via prompts. Moreover, interesting future directions can explore how these prompts could help build more robust and explainable models against out-of-distribution data. Our work serves as an essential first step in this direction, offering a valuable evaluation framework, datasets enriched with prompts, and fascinating insights for future investigation.

Acknowledgments

We thank Kathmandu University for its invaluable support in granting us access to its supercomputer infrastructure. This enabled the successful execution of the experiments crucial to this paper.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Javeria Amin, Muhammad Sharif, Mussarat Yasmin, Tanzila Saba, Muhammad Almas Anjum, and Steven Lawrence Fernandes. A new approach for brain tumor segmentation and classification based on score level fusion using transfer learning. *Journal of medical systems*, 43:1–16, 2019.

Nguyen S An, Phan N Lan, Dao V Hang, Dao V Long, Tran Q Trung, Nguyen T Thuy, and Dinh V Sang. Blazeneo: Blazing fast polyp segmentation and neoplasm detection. *IEEE Access*, 10:43669–43684, 2022.

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.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Li-Ching Chen, Po-Chih Kuo, Ryan Wang, Judy Gichoya, and Leo Anthony Celi. Chest x-ray segmentation images based on mimic-cxr. 2022.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using duck-net for polyp image segmentation. *Scientific Reports*, 13(1):9803, 2023.

Kerr Fitzgerald and Bogdan Matuszewski. Fcb-swinv2 transformer for polyp segmentation. *arXiv preprint arXiv:2302.01027*, 2023.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35:20450–20468, 2022.

Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 516–524. Springer, 2017.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

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.

Md Kamrul Hasan, Md Toufick E Elahi, Md Ashraful Alam, Md Tasnim Jawad, and Robert Martí. Dermoexpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation. *Informatics in Medicine Unlocked*, 28:100819, 2022.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

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.

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 *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019a.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

Connah Kendrick, Bill Cassidy, Joseph M Pappachan, Claire O'Shea, Cornelious J Fernandez, Elias Chacko, Koshy Jacob, Neil D Reeves, and Moi Hoon Yap. Translating clinical delineation of diabetic foot ulcers into machine interpretable segmentation. *arXiv preprint arXiv:2204.11618*, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2021.

Ting-Yu Liao, Ching-Hui Yang, Yu-Wen Lo, Kuan-Ying Lai, Po-Huai Shen, and Youn-Long Lin. Hardnet-dfus: Enhancing backbone and decoder of hardnet-mseg for diabetic foot ulcer image segmentation. In *Diabetic Foot Ulcers Grand Challenge*, pages 21–30. Springer, 2022.

Hang Jung Ling, Damien Garcia, and Olivier Bernard. Reaching intra-observer variability in 2-d echocardiographic image segmentation with a simple u-net architecture. In *IEEE International Ultrasonics Symposium (IUS)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

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.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2022.

Ziyuan Qin, Hua Hui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. In *The Eleventh International Conference on Learning Representations*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: text-guided attention for improved polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 151–160. Springer, 2022.

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

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.

Saad Wazir and Muhammad Moazam Fraz. Histoseg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images. In *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023a.

Shuai Zhang, Mei Liao, Jing Wang, Yongyi Zhu, Yanling Zhang, Jian Zhang, Rongqin Zheng, Linyang Lv, Dejiang Zhu, Hao Chen, et al. Fully automatic tumor segmentation of breast ultrasound images with deep learning. *Journal of Applied Clinical Medical Physics*, 24(1):e13863, 2023b.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.

Appendix A. Impact on Society

The incorporation of language prompts in medical image segmentation has the potential to impact society significantly, particularly in clinical settings. By enabling radiologists to quickly and accurately segment complex shapes using just a few words, language prompts offer a more interpretable and explainable approach compared to traditional visual prompts such as points or boxes.

One significant advantage of language prompts is their ability to convey detailed information about normal and abnormal structures' texture, shape, and spatial relationships. This allows for a more comprehensive understanding of medical images, facilitating more accurate segmentation results. Additionally, language prompts can be easily adapted to new classes, making them highly versatile and adaptable in various medical scenarios.

Using language prompts in medical image segmentation can improve the efficiency and effectiveness of radiologists' work, potentially leading to faster diagnoses and treatment decisions. Moreover, the interpretability of language prompts can aid in building trust and confidence among healthcare professionals and patients as the reasoning behind the segmentation process becomes more transparent.

Overall, the integration of language prompts in medical image segmentation has the potential to revolutionize clinical practices, providing radiologists with a powerful tool to enhance their segmentation capabilities and ultimately improve patient care outcomes.

We strongly encourage and invite other researchers to contribute to this field of study. This research paper has no negative impact on society or further research in medical imaging, as we have adhered to ethical considerations in medical imaging and have not expressed disapproval of any previous studies.

Appendix B. Dataset and Code Access

The GitHub repository[6] contains the source code with detailed documentation, the generated prompts for all the datasets, and thorough instructions along with the relevant links to access the individual image-mask pair datasets used in this work.

Appendix C. Experiments

C.1. VLSM Finetuning Experiments

CLIPSeg and CRIS internally resize the three-channeled input images to 352×352 and 416×416, respectively. The dice scores mentioned in the paper are calculated after resizing the output of the models back to the original size (before respective resizing). We normalize the resized images with means and standard deviations provided by the respective models and haven't performed other preprocessing and post-processing to access the models' raw performance.

For the five non-radiology datasets (Kvasir-SEG, ClinicDB, BKAI, ISIC, and DFU), we finetune VLSMs with ten prompts for an individual dataset, resulting in 50 experiments for each VLSM. Similarly, in the case of radiology datasets (CAMUS, BUSI, and CheXlocalize), we have a total of 22 finetuning experiments for each VLSM. We also finetune CRIS and

6. https://github.com/naamiinepal/medvlsm

Optimizers	{Adam, AdamW}
Learning Rates (LRs)	[1e-5,1e-2]
LR Schedulers	{CosineAnealingLR, ConstantLR, ReduceLROnPlateau}
Batch sizes	{16, 32, 64, 128 }

Table 4: Different settings of hyperparameters that have been experimented with to select the optimal one.

CLIPSeg with the pooled datasets comprising only endoscopic and all datasets. Thus, including all varieties with the VLSMs and the different prompting mechanisms, we have 442 finetuning experiments.

The average time to fine-tune CRIS for a dataset on a prompt is approximately 60 minutes in our training setup, running 45 epochs on average. For CLIPSeg, the average training time is 40 minutes, running for 90 epochs on average. BiomedCLIPSeg's and BiomedCLIPSeg-D's average training times are 20 minutes and 30 minutes, running for 80 epochs and 50 epochs, respectively. We monitored the segmentation metric on the held-out validation sets for early stopping, with patience of 50 epochs for CLIPSeg variants and 10 epochs for CRIS.

C.2. Hyperparameters Search

We experiment with multiple sets of hyperparameters including learning rates, optimizers, batch sizes, and schedulers. We select the optimal setting of hyperparameters (as mentioned in the main paper) that showed optimal performance in most datasets (Table 4).

C.3. CNN-based Experiments

For comparative analysis, we consider three of the conventional CNN-based segmentation models: UNet (Ronneberger et al., 2015), UNet++(Zhou et al., 2018), and DeepLabV3+ (Chen et al., 2018). For all of the models, we use pretrained ResNet-50 (He et al., 2016) as the backbone, and default parameters given by the framework *Segmentation Models PyTorch*[7] are chosen as the model hyperparameters. We use Dice loss for error propagation within the models with Adam optimizer (Kingma and Ba, 2014) of learning rate 10^{-3} and zero weight decay.

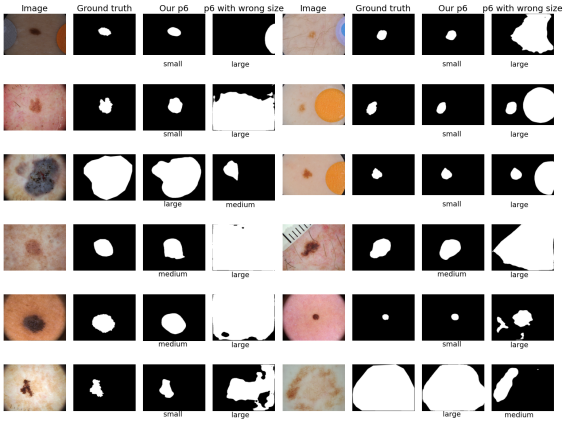
Appendix D. PubMedBERT's failure to give reliable output

Table 5 contains the predictions of PubMedBERT for the masked language modeling in different datasets.

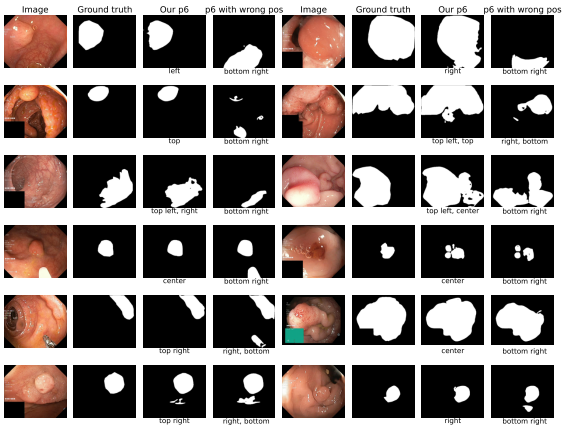
Appendix E. Some visualizations and qualitative analysis

Some visualizations and qualitative analysis are shown in Figures 5 and 6.

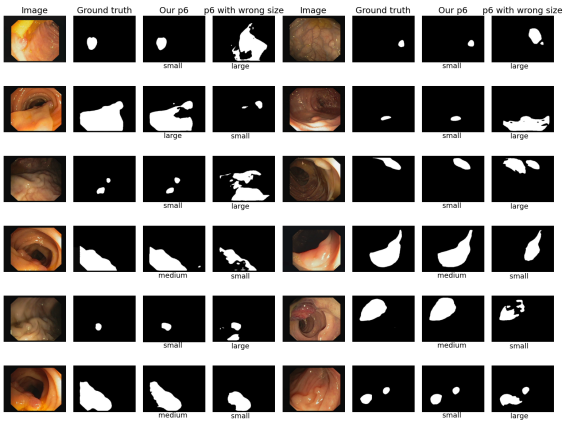
7. https://github.com/qubvel/segmentation_models.pytorch



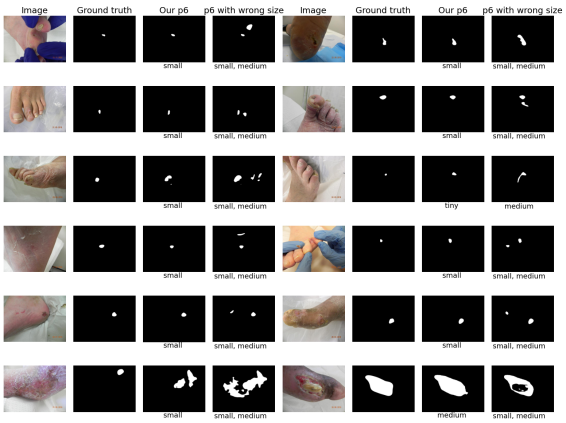
(*a*) ISIC for attribute *size*



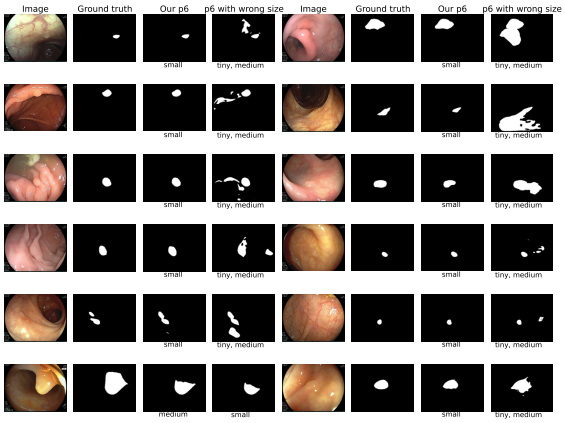
(*b*) Kvasir-SEG for attribute *location*



(*c*) ClinicDB for attribute *size*



(*d*) DFU for attribute *size*



(*e*) BKAI for attribute *size*

Figure 5: Visualization of CRIS's performance when prompt attributes are changed using a wrong attribute value. For each medical image, three corresponding masks are displayed: ground truth mask, output mask for the corresponding prompt, and output mask after altering an attribute value of the prompts.

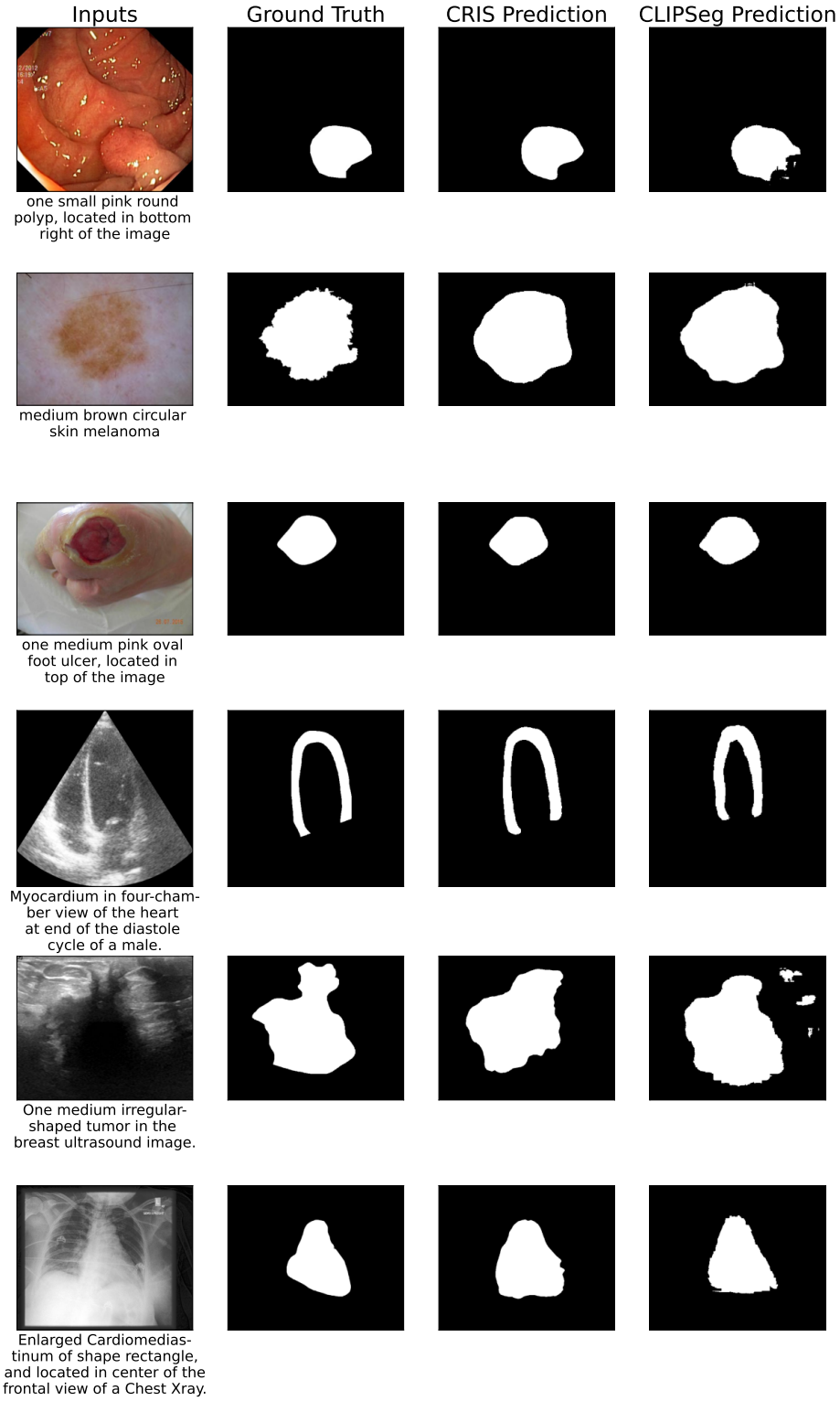


Figure 6: Sample input, ground truth, and models' predictions

Table 5: PubMedBERT's top five predictions for the masked language modeling inference. The predictions are ordered in the descending order of the probability generated by the model. The model has high uncertainty as the maximum probability is about 0.1. The predictions are almost the same and uninformative, which is more prominent in the radiology datasets.

Dataset	Masked sentence	Top-5 Predictions
All Endoscopy*	The location of the polyp is [MASK].	[variable, unknown, varied, unpredictable, uncertain]
	Polyp is located at [MASK].	[bifurcation, apex, rectum, midline, right]
	The shape of polyp is [MASK].	[irregular, variable, oval, round, different]
	Polyp is [MASK] in shape.	[oval, irregular, round, spherical, cylindrical]
	The color of the polyp is [MASK].	[yellow, red, blue, brown, pink]
	Polyp is [MASK] in color.	[yellow, white, red, black, green]
ISIC	The location of skin melanoma is [MASK].	[unknown, variable, unusual, unpredictable, rare]
	The color of skin melanoma is [MASK].	[red, yellow, brown, black, blue]
	Skin melanoma is [MASK] in texture.	[heterogeneous, variable, soft, irregular, fibrous]
	Skin cancer is located at [MASK].	[extremities, birth, puberty, adolescence, skin]
	Skin cancer is [MASK] in texture.	[heterogeneous, unique, variable, diverse, distinctive]
DFU	The location of a diabetic foot ulcer is at [MASK].	[first, rest, ankle, home, foot]
	Diabetic foot ulcer is located at [MASK].	[ankle, heel, foot, extremities, feet]
	The location of the foot ulcer is [MASK].	[ankle, knee, first, heel, night]
	Foot ulcer is located at [MASK].	[ankle, heel, foot, knee, night]
CAMUS	The left ventricular cavity is [MASK] in shape.	[spherical, triangular, normal, oval, round]
	The myocardium is [MASK] in shape.	[spherical, cylindrical, circular, round, triangular]
	The left atrium cavity is [MASK] in shape.	[oval, round, triangular, spherical, irregular]
	The left ventricular cavity is located at [MASK].	[diastole, apex, rest, 90°, 45°]
	The myocardium is located at [MASK].	[rest, apex, risk, diastole, birth]
	The left atrium cavity is located at [MASK].	[diastole, right, left, 90°, apex]
BUSI	The malignant breast tumor is [MASK] in shape.	[round, irregular, oval, solid, spherical]
	The benign breast tumor is [MASK] in shape.	[oval, round, irregular, solid, spherical]
CheXlocalize	Airspace Opacity is [MASK] in shape.	[irregular, oval, round, triangular, globular]
	Enlarged Cardiomediastinum is [MASK] in shape.	[oval, triangular, irregular, round, rounded]
	Cardiomegaly is [MASK] in shape.	[irregular, triangular, normal, oval, round]
	Lung Opacity is [MASK] in shape.	[irregular, round, oval, nodular, reticular]
	Consolidation is [MASK] in shape.	[spherical, circular, triangular, irregular, round]
	Atelectasis is [MASK] in shape.	[irregular, oval, triangular, spherical, round]
	Pleural Effusion is [MASK] in shape.	[irregular, round, oval, spherical, solid]

*This includes six datasets of endoscopy: Kvasir-SEG, ClinicDB, BKAI, CVC-300, CVC-ColonDB, ETIS

Appendix F. Results

F.1. Finetuning only the Decoders for CLIP-based VLSMs

Tables 6 and 7 show the results of VLSMs with finetuned the decoder while keeping the encoders frozen.

F.2. Using radiology reports for lung segmentation

To examine the usage of free-text radiology reports of chest x-rays for segmentation, we utilize 1,141 frontal-view CXRs randomly selected from the MIMIC-CXR database (Johnson et al., 2019a,b; Chen et al., 2022). This dataset contains the segmentation of lungs, which has been verified manually. We use the free-text radiology reports provided in the MIMIC-CXR Database (Johnson et al., 2019a) as the only prompt (P1), and the results are reported in Table 8.

Table 6: Finetuned segmentation dice score (%) of CRIS on different datasets on different sets of prompts with frozen CLIP.

Prompt → / Dataset ↓	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
Kvasir-SEG	75.49±27.22	76.03±26.35	82.18±22.40	81.89±21.78	84.26±20.39	**86.39±17.01**	85.37±17.29	82.43±22.11	85.06±18.89	85.02±19.10
ClinicDB	49.48±33.67	46.98±34.30	81.07±24.37	82.72±23.96	84.88±24.00	85.01±21.84	83.31±22.74	81.66±26.10	**87.13±21.38**	84.65±22.25
BKAI	77.98±28.73	75.01±29.97	81.93±24.66	82.49±24.7	82.39±23.65	84.65±21.75	85.75±21.48	84.91±23.06	**86.40±20.49**	85.07±22.01
ISIC	87.64±14.37	85.77±18.29	90.25±10.37	90.32±10.93	91.28±7.45	91.23±8.56	**91.29±8.10**	90.46±10.90	91.29±8.11	91.28±7.65
DFU	66.30±29.57	66.14±29.81	**70.28±27.11**	67.24±30.22	69.19±28.98	68.55±29.56	68.93±29.41	69.35±28.75	68.36±29.62	70.15±28.59
CAMUS	46.15±9.69	88.87±8.49	**89.18±6.79**	88.94±7.05	88.92±6.69	88.02±7.37	88.96±6.84	89.04±6.85	N/A	N/A
BUSI	47.11±39.12	61.49±36.03	63.18±36.89	62.87±37.60	65.10±36.60	66.69±35.68	**66.76±35.77**	N/A	N/A	N/A
CheXlocalize	41.03±24.96	54.18±25.77	54.57±25.06	53.30±25.16	**56.17±24.73**	56.03±24.49	52.48±25.89	N/A	N/A	N/A

Table 7: Finetuned segmentation dice score (%) of CLIPSeg on different datasets on different sets of prompts with frozen CLIP.

Prompt → / Dataset ↓	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
Kvasir-SEG	86.38±17.8	87.50±15.35	87.49±14.29	87.68±14.60	88.33±10.95	88.25±12.11	**88.98±11.98**	87.97±13.93	88.39±14.72	88.71±11.4
ClinicDB	87.23±14.93	87.07±14.43	**88.41±11.01**	87.17±14.73	87.25±15.09	87.73±13.52	87.76±13.56	87.57±13.98	87.05±14.79	87.46±14.39
BKAI	83.64±18.59	85.26±15.40	85.47±15.15	84.7±16.94	85.93±14.66	**86.01±14.84**	85.02±17.23	85.45±14.76	85.50±15.68	84.99±17.11
ISIC	91.71±8.68	91.45±8.47	91.66±8.29	91.85±8.36	**92.11±6.87**	92.02±6.88	92.09±7.00	91.77±7.73	91.89±7.70	91.90±7.21
DFU	72.35±25.04	72.19±25.69	71.79±25.05	71.88±24.83	72.5±24.43	72.31±25.27	**73.53±23.68**	72.1±25.48	73.11±23.98	73.31±23.81
CAMUS	46.48±9.07	88.67±6.25	88.70±5.93	**88.81±6.15**	88.77±6.22	88.47±6.55	88.53±6.29	87.82±7.01	N/A	N/A
BUSI	62.03±38.3	62.79±37.55	62.97±37.27	62.85±36.66	**64.47±37.54**	62.83±38.19	62.33±38.68	N/A	N/A	N/A
CheXlocalize	45.35±25.18	58.10±25.03	58.37±24.50	58.95±24.48	59.49±25.11	**59.56±24.70**	58.06±25.34	N/A	N/A	N/A

Appendix G. Prompt Composition

The prompts used during the training for various datasets are shown below. If there are multiple templates for the same prompts for a dataset, one is randomly chosen during the training to increase the regularization for the models.

G.1. Non-radiology images

G.1.1. Endoscopy Datasets

A total of six endoscopy datasets (polyp segmentation image-mask pairs) have been used for finetuning and evaluating our proposed models: Kvasir-SEG (Jha et al., 2020), ClinicDB (Bernal et al., 2015), BKAI (Ngoc Lan et al., 2021; An et al., 2022), CVC-300 (Vázquez et al., 2017), CVC-ColonDB (Tajbakhsh et al., 2015), and ETIS (Silva et al., 2014). The last three datasets have a small number of image-masks pairs, so they are used only for testing and evaluating the trained models.

1. **P0**: “” (No prompt)
2. **P1**: “*class name*”
   - *polyp*
3. **P2**: “*shape class name*”

Table 8: Zero-shot and finetuning dice scores (%) of the CRIS and CLIPSeg Manually labeled Chest X-ray Segmentation Dataset. We have used the actual radiology reports as **P1**. P0 indicates an empty prompt.

Models ↓	Prompt → / Experiment ↓	P0	P1
CRIS	**Zero-shot**	$44.8_{\pm 18.97}$	$40.73_{\pm 18.95}$
	Finetuning	$81.66_{\pm 5.65}$	$90.99_{\pm 1.41}$
CLIPSeg	**Zero-shot**	$0.26_{\pm 2.35}$	$0.09_{\pm 0.88}$
	Finetuning	$91.39_{\pm 1.09}$	$91.22_{\pm 1.26}$

Table 9: Different prompts are formed for each dataset using combinations of 14 potential attributes. Although some attributes, like *Pathology*, are specific to some particular datasets, others, like *Class Keywords*, are common to all the datasets.

Attributes → **a1:** Class Keyword; **a2:** Shape; **a3:** Color; **a4:** Size; **a5:** Number; **a6:** Location; **a7:** General Class Info; **a8:** View; **a9:** Pathology; **10:** Cardiac Cycle; **a11:** Gender; **a12:** Age; **a13:** Image Quality; **a14:** Tumor Type

Prompts → / Datasets ↓	P1	P2	P3	P4	P5	P6	P7	P8	P9
Non-Radiology	a1	a1a2	a1a2a3	a1a2a3a4	a1a2a3a4a5	a1a2a3a4a6	a1a7	a1a2a3a4a5a7	a1a2a3a4a5a6a7
Example Prompt	**P9** → **one small pink round polyp** which is **often a bumpy flesh in rectum** located in **center** of the image								
CheXlocalize	a1	a1a8	a1a2a8	a1a2a6a8	a1a2a6a8a9	a1a9	N/A	N/A	N/A
Example Prompt	**P5** → **Airspace Opacity** of shape **rectangle**, and located in **right** of the **frontal** view of a Chest Xray. **Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Consolidation, Atelectasis, Pleural Effusion** are present.								
CAMUS	a1	a1a8	a1a8a10	a1a8a10a11	a1a8a10a11a12	a1a8a10a11a12a13	a1a8a10a11a12a13a2	N/A	N/A
Example Prompt	**P7** → **Left ventricular cavity** of **triangular shape** in **two-chamber view** in the cardiac ultrasound at the end of the **diastole cycle** of a **40-year-old female** with **poor image quality**.								
BUSI	a1	a1a14	a1a14a5	a1a14a5a4	a1a14a5a4a6	a1a14a5a4a6a2	N/A	N/A	N/A
Example Prompt	**P6** → **Two medium square-shaped benign tumors** at the **center, left** in the breast ultrasound image.								

- *round polyp*

4. **P3**: "*color shape class name*"
   - *pink round polyp*
5. **P4**: "*size color shape class name*"
   - *medium pink round polyp*
6. **P5**: "*number size color shape class name*"
   - *one medium pink round polyp*
7. **P6**: "*number size color shape class name*, located in the *location* of the image"
   - *one medium pink round polyp*, located in the *top left* of the image
8. **P7**: "*class name*, which is a *general description of the class*"

- *polyp*, which is a *small lump in the lining of colon*

9. **P8**: "*number size color shape class name*, which is a *general description of the class*"
   - *one medium pink round polyp*, which is a *small lump in the lining of colon*
10. **P9**: "*number size color shape class name*, which is a *general description of the class* located in the *location* of the image "
   - *one medium pink round polyp*, which is a *small lump in the lining of colon* located in the *top left* of the image

For *General Description of the class*, prompts were built using information about the subject on the internet. Five such descriptions were designed for each dataset, and one random sample was selected each time as the *general description of the class* attribute whenever the prompts **p7**, **p8**, and **p9** were used.

G.1.2. ISIC and DFU-2022

The templates of prompts for the DFU-2022 (Kendrick et al., 2022) and ISIC (Gutman et al., 2016) datasets used were the same as the above examples for endoscopy images, with *class name* and *general description of the class* being different. We used class names **skin melanoma** and **foot ulcer** for the two datasets, respectively.

The five *General Description of the class* for each of the three types of photographic datasets used is listed in the table below.

Table 10: General Descriptions selected for each of the photographic datasets.

Endoscopy Datasets	**ISIC**	**DFU-2022**
→ a projecting growth of tissue	→ a spot with dark speckles	→ a wound in foot and toes
→ often a bumpy flesh in rectum	→ a spot with irregular texture	→ a sore in foot and toes
→ a small lump in the lining of colon	→ a dark sore with irregular texture	→ a sore in skin of foot and toe
→ a tissue growth that often resemble mushroom-like stalks	→ an irregular sore with speckles	→ an abnormality in foot and toes
→ an abnormal growth of tissues projecting from a mucous membrane	→ a rough wound on skin	→ an open sore or lesion in foot and toes

G.2. Radiology Images

G.2.1. CheXlocalize

The prompts for the CheXlocalize (Saporta et al., 2022) dataset are listed below.

1. **P0**: "" (No prompt)

2. **P1**: "*labels* in a chest Xray."
    - *Airspace Opacity* in a chest Xray.
3. **P2**: "*labels* in the *xray_view* view of a Chest Xray."
    - Airspace Opacity in the *frontal* view of a Chest Xray.
4. **P3**: "*labels* of shape *shape* in the *xray_view* view of a Chest Xray."
    - Airspace Opacity of shape *rectangle* in the frontal view of a Chest Xray.
5. **P4**: "*labels* of shape *shape*, and located in *location* of the *xray_view* view of a Chest Xray."
    - Airspace Opacity of shape rectangle, and located in *right* of the frontal view of a Chest Xray.
6. **P5**: "*labels* of shape *shape*, and located in *location* of the *xray_view* view of a Chest Xray. *pathology* are present."
    - Airspace Opacity of shape rectangle, and located in right of the frontal view of a Chest Xray. *Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Consolidation, Atelectasis, Pleural Effusion* are present.
7. **P6**: "*labels* in a Chest Xray. *pathology* are present."
    - Airspace Opacity in a Chest Xray. Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Consolidation, Atelectasis, Pleural Effusion are present.

G.2.2. CAMUS

The prompts for the CAMUS (Leclerc et al., 2019) dataset are listed below.

1. Class of Current Image
    - *Left ventricular cavity*, *Myocardium*, or *Left atrium cavity* of the heart
    - [*class*] in the cardiac ultrasound
2. Include the chamber information
    - Left ventricular cavity in *two-chamber view* of the heart.
    - Left ventricular cavity in *two-chamber view* in the cardiac ultrasound.
3. Include the cycle
    - Left ventricular cavity in two-chamber view of the heart at the *end of the diastole cycle*.
    - Left ventricular cavity in two-chamber view in the cardiac ultrasound at the *end of the diastole cycle*.

4. Include the gender
    - Left ventricular cavity in two-chamber view of the heart at the end of the diastole cycle of *a female*.
    - Left ventricular cavity in two-chamber view in the cardiac ultrasound at the end of the diastole cycle of *a female*.
5. Include the age
    - Left ventricular cavity in two-chamber view of the heart at the end of the diastole cycle of a *forty-six-year-old* female.
    - Left ventricular cavity in two-chamber view in the cardiac ultrasound at the end of the diastole cycle of a *forty-six-year-old* female.
6. Include the image quality
    - Left ventricular cavity in two-chamber view of the heart at the end of the diastole cycle of a 40-year-old female with *poor image quality*.
    - Left ventricular cavity in two-chamber view in the cardiac ultrasound at the end of the diastole cycle of a 40-year-old female with *poor image quality*.
7. Include the mask shape
    - Left ventricular cavity of *triangular shape* in two-chamber view of the heart at the end of the diastole cycle of a 40-year-old female with *poor image quality*.
    - Left ventricular cavity of *triangular shape* in two-chamber view in the cardiac ultrasound at the end of the diastole cycle of a 40-year-old female with *poor image quality*.

G.2.3. Breast Ultrasound Images Dataset

The prompts for the Breast Ultrasound Images (BUSI) (Al-Dhabyani et al., 2020) dataset are listed below.

1. Presence of tumor
    - *[No] tumor* in the breast ultrasound image
2. Tumor Type
    - *Benign* tumor in the breast ultrasound image
    - *Regular-shaped* tumor in the breast ultrasound image
3. Tumor Number
    - *Two* benign tumors in the breast ultrasound image
    - *Two* regular-shaped tumors in the breast ultrasound image

4. Tumor Coverage
   - Two *medium* benign tumors in the breast ultrasound image
   - Two *medium* regular-shaped tumors in the breast ultrasound image
5. Tumor Location
   - Two medium benign tumors *at the center, left* in the breast ultrasound image
   - Two medium regular-shaped tumors *at the center, left* in the breast ultrasound image
6. Tumor Shape
   - Two medium *square-shaped* benign tumors at the center, left in the breast ultrasound image
   - Two medium *square-shaped* regular tumors at the center, left in the breast ultrasound image