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

Kanchan Poudel*, Manish Dhakal*, Prasiddha Bhandari*, Rabin Adhikari*, Safal Thapaliya*, Bishesh Khanal

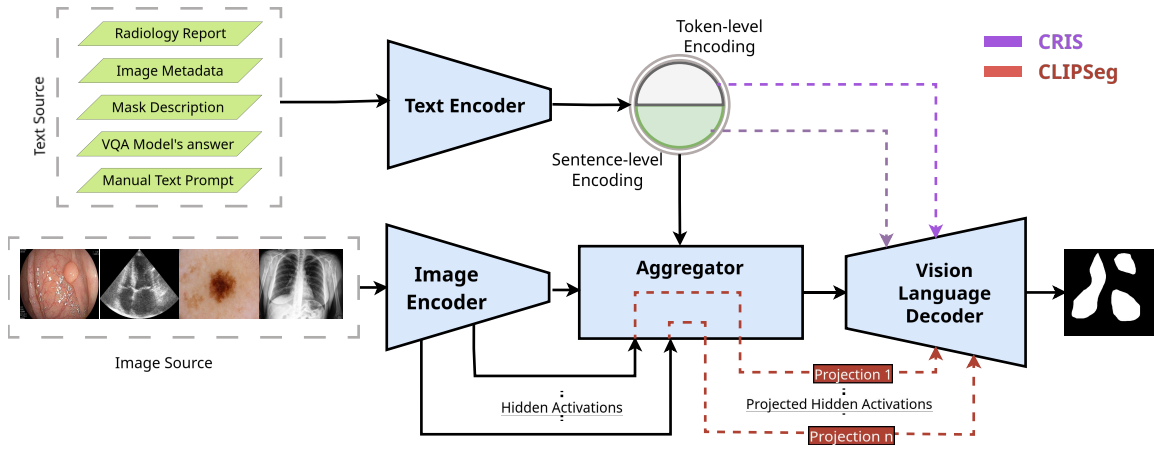
Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal
*equal contribution

MIDL Paris 2024

- Segmentation: Human interaction useful in clinics
- Language prompts capture rich semantics:
  - Shape, size, surrounding anatomies, image modality
- Vision-Language Segmentation Models (VLSMs):
  - Use language as prompts
  - Potentially more explainable outputs
  - Human-in-the-loop

Critical questions

1. Generalization: natural images —(Transfer learning)→ medical images
2. Role of language prompts and images?

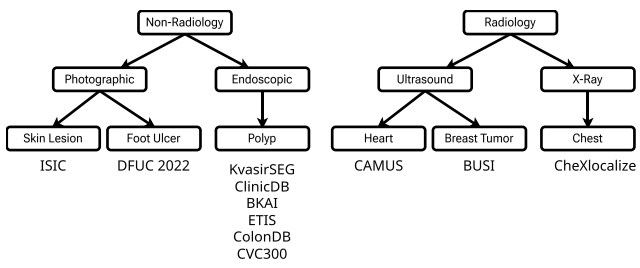


Vision-Language Segmentation Model (VLSM)

We present a systematic study on VLSM transfer learning from natural images to medical images.

Benchmarking Framework

Datasets



Models

- Pretrained on natural images → CRIS, CLIPSeg
- Pretrained on medical images → BiomedCLIPSeg, BiomedCLIPSeg-D

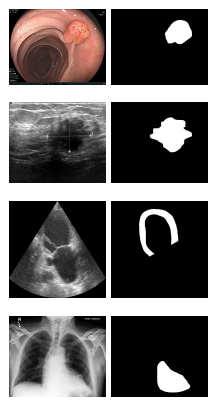
Our proposed automated prompt engineering for medical datasets

One (Number) small (Size) pink (Color) round (Shape) polyp (Target Structure) which is a projecting growth of tissue (Class Specific General Description), located in top right (Location) of the image.

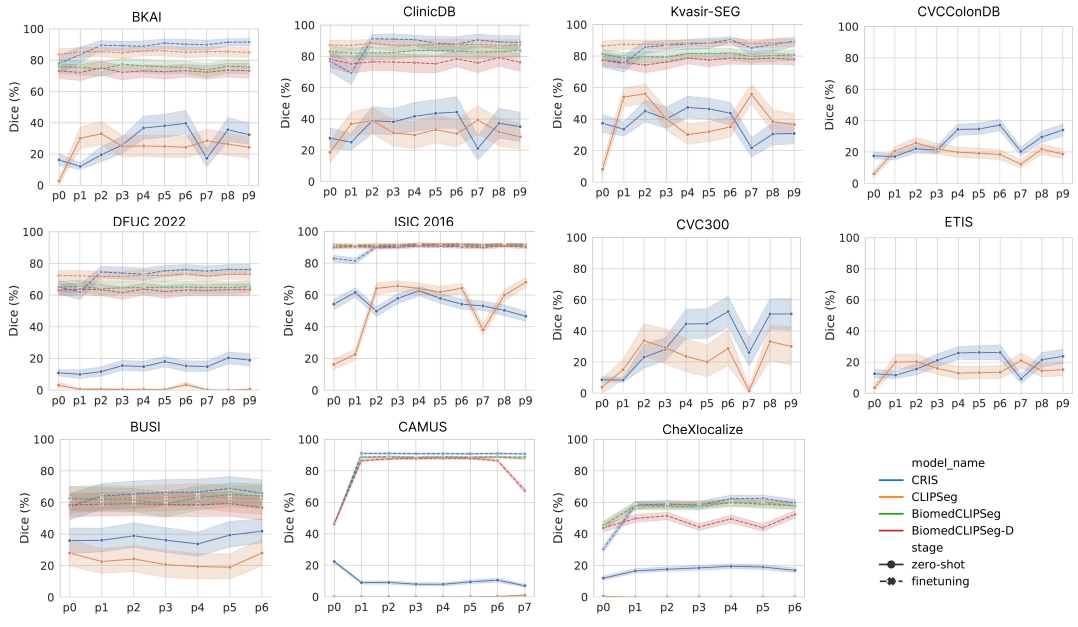
One medium circle-shaped malignant tumor at the right in the breast ultrasound image.

Myocardium of square shape in two-chamber view (View) in the cardiac ultrasound at the end of the systole cycle (Cardiac Cycle) of a seventy-one-year-old (Patient's Age) male (Patient's Sex) with medium image quality (Image Quality).

Airspace Opacity of shape rectangle, and located in bottom right of the frontal view of a Chest Xray. Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Atelectasis are present. (Pathology)



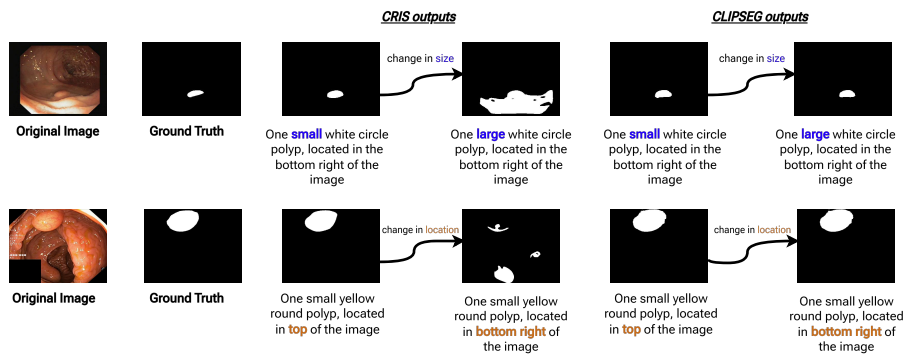
How do VLSMs perform in zero-shot and fine-tuning?



- VLSMs adapt better to non-radiology images for zero-shot segmentation.
- Some datasets perform better with image-specific attributes in prompts, while others perform better with general descriptions.
- Making prompts richer with attributes does not always improve model's performance.
- VLSMs have comparable performance to conventional segmentation models and SOTA.
- BiomedCLIP-based models perform poorly as they have not been pretrained for segmentation task.

What happens when attribute values are replaced?

- CRIS shows robust semantic learning of location and size attributes.



- VLSMs are robust to distribution shift compared to conventional segmentation models.

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

Key takeaways

- Prompt design: Include attributes familiar to models during pretraining
- Large scale dataset better than smaller-scale domain-specific dataset
- Some VLSMs leverage language semantics better than others

Future direction

- VLSMs for 3D modalities like MRI or CT scans
- Better ways to leverage language semantic to identify target anatomies
- Generate large-scale medical image-text-mask triplets

Scan to read paper