MIDL
Paris 2024

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

Medical Imaging in Deep Learning
Paris, France
5th July 2024

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

NAAMII, Nepal

**equal contribution*

Outline

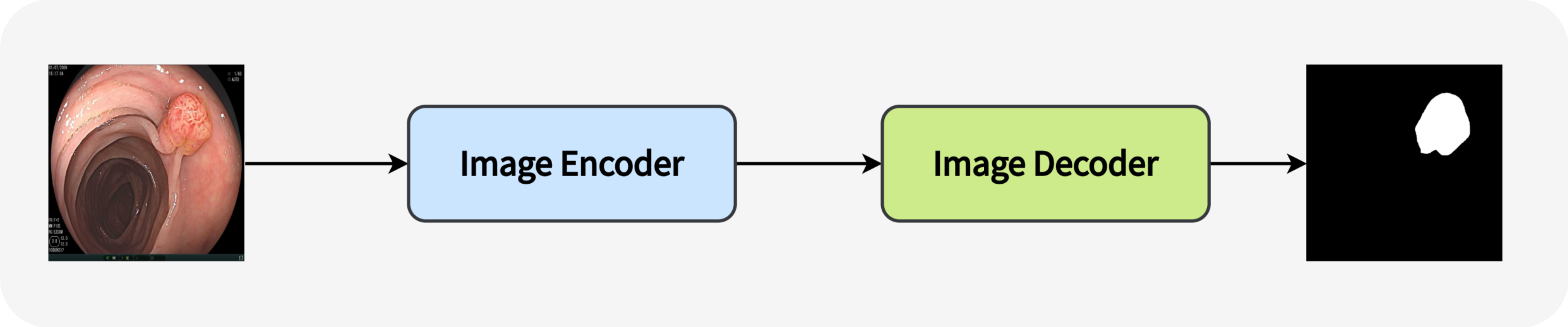
- Human Interactive Image Segmentation
- Vision Language Segmentation Models (VLSMs)
- Benchmarking Framework
- Prompt Generation
- Results

Outline

- **Human Interactive Image Segmentation**
- Vision Language Segmentation Models (VLSMs)
- Benchmarking Framework
- Prompt Generation
- Results

Problems with unimodal segmentation models

Typical Image Segmentation Models:



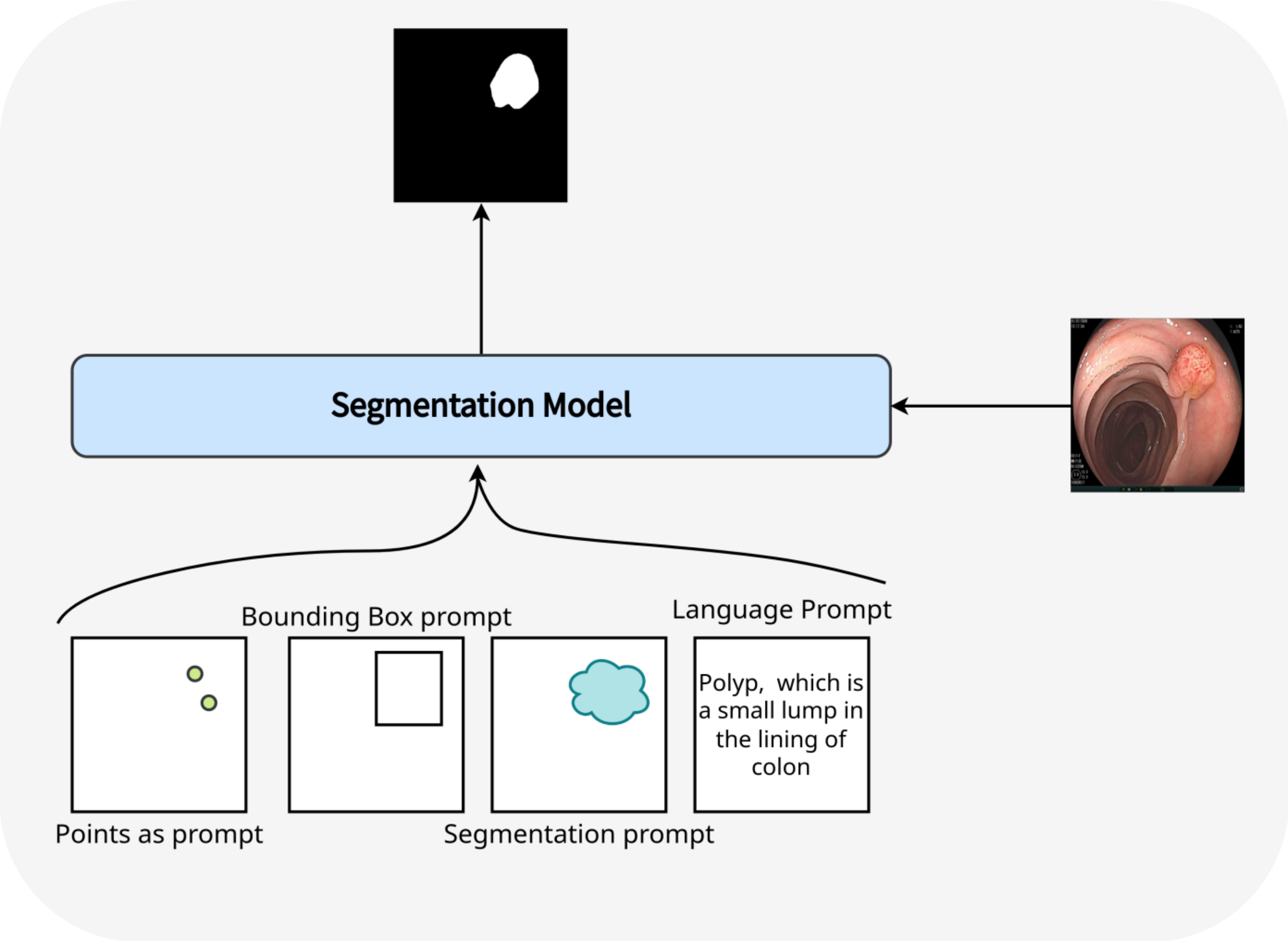
Constrained to predefined foreground classes

Requires retraining when new classes are introduced

Minimal human interaction

Lack explainability

Multimodal Image Segmentation with prompts



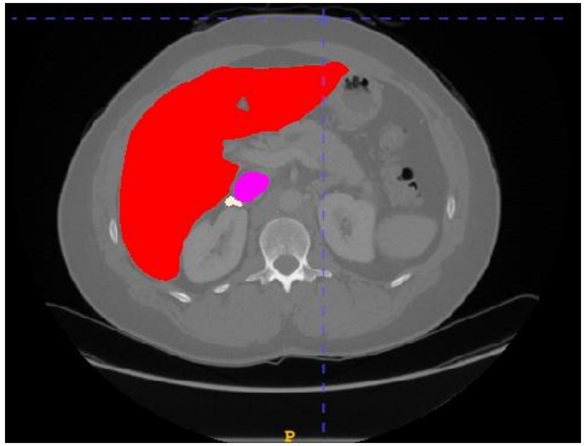
Enables human interaction ☺

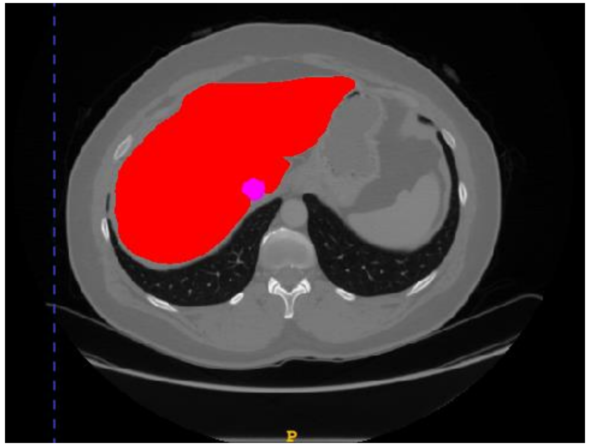
Language prompts possibilities:

- more explainable during inference
- open vocabulary segmentation
- robustness to out of distribution data

Multimodal Image Segmentation with prompts

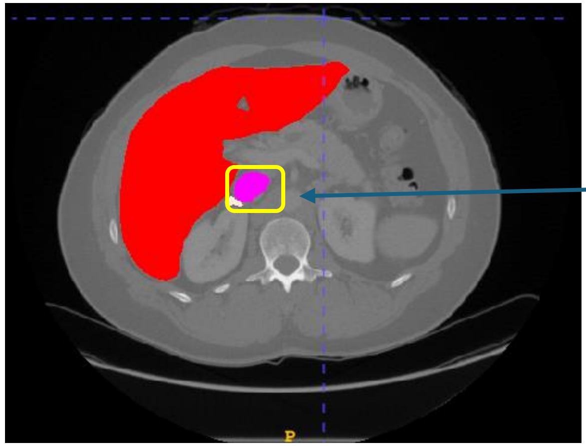
Two different slices of an abdominal CT Scan with liver (red) and inferior vena cava (pink)



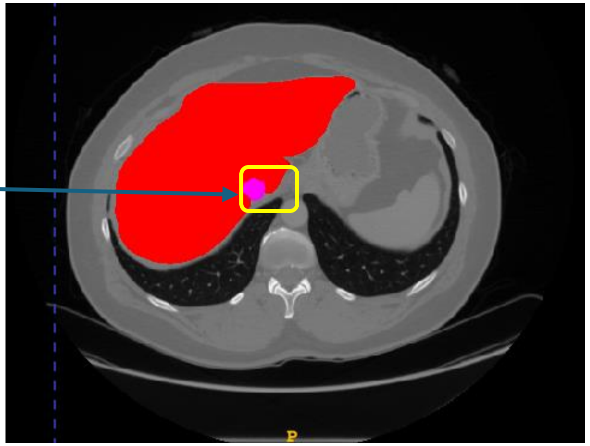


Multimodal Image Segmentation with prompts

Two different slices of an abdominal CT Scan with liver (red) and inferior vena cava (pink)



Example Prompts

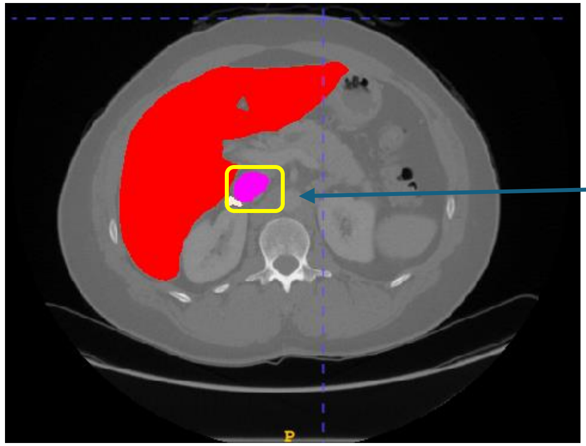


Inferior Vena Cava in the center of the image, oval in shape, close to liver

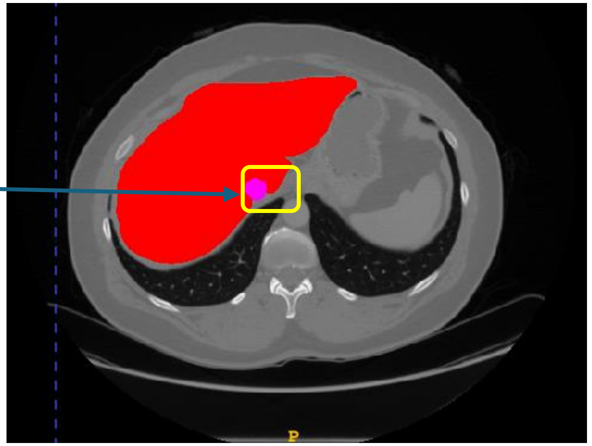
Inferior Vena Cava in the center of the image, circular in shape, enclosed by liver

Multimodal Image Segmentation with prompts

Two different slices of an abdominal CT Scan with liver (red) and inferior vena cava (pink)



Example Prompts



Inferior Vena Cava in the center of the image, oval in shape, close to liver

Inferior Vena Cava in the center of the image, circular in shape, enclosed by liver

Language prompts are more expressive and powerful than others

Recent Advancements in Leveraging Text Prompts

Foundation Vision Language Models (VLMs) from large natural image and human text pairs

Vision Language Segmentation Models (VLSMs) built on top of VLMs in natural image domain

Some VLMs finetuned further in medical data or trained from scratch

Recent Advancements in Leveraging Text Prompts

Foundation Vision Language Models (VLMs) from large natural image and human text pairs

Vision Language Segmentation Models (VLSMs) built on top of VLMs in natural image domain

Some VLMs finetuned further in medical data or trained from scratch

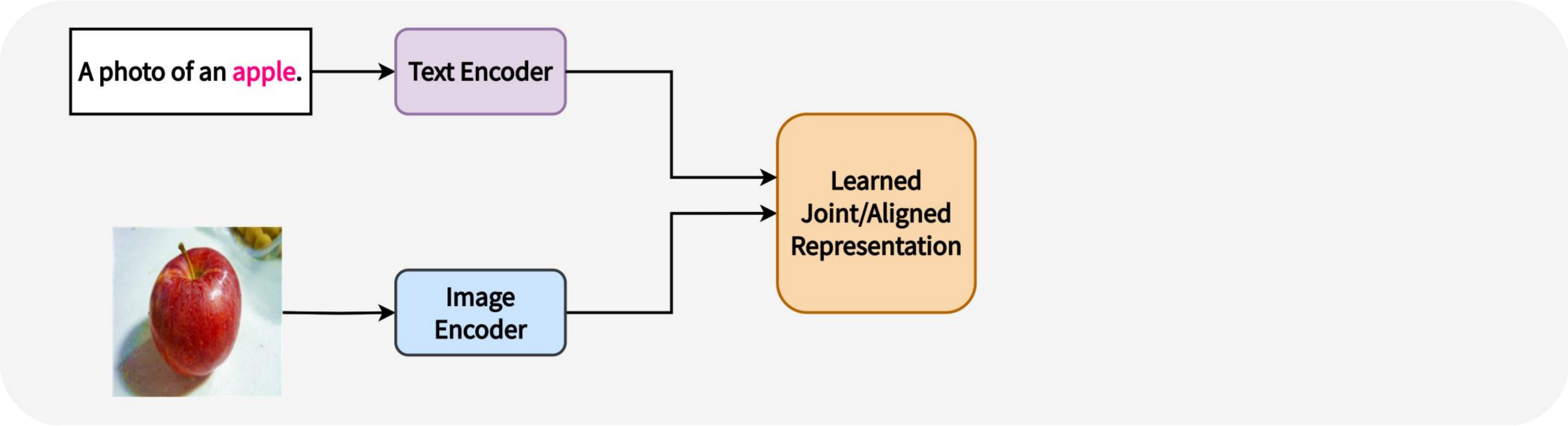
How does transfer learning of VLMs/VLSMs to limited medical image data look like?

Are VLSMs really leveraging the rich semantics that the text prompts can provide?

Outline

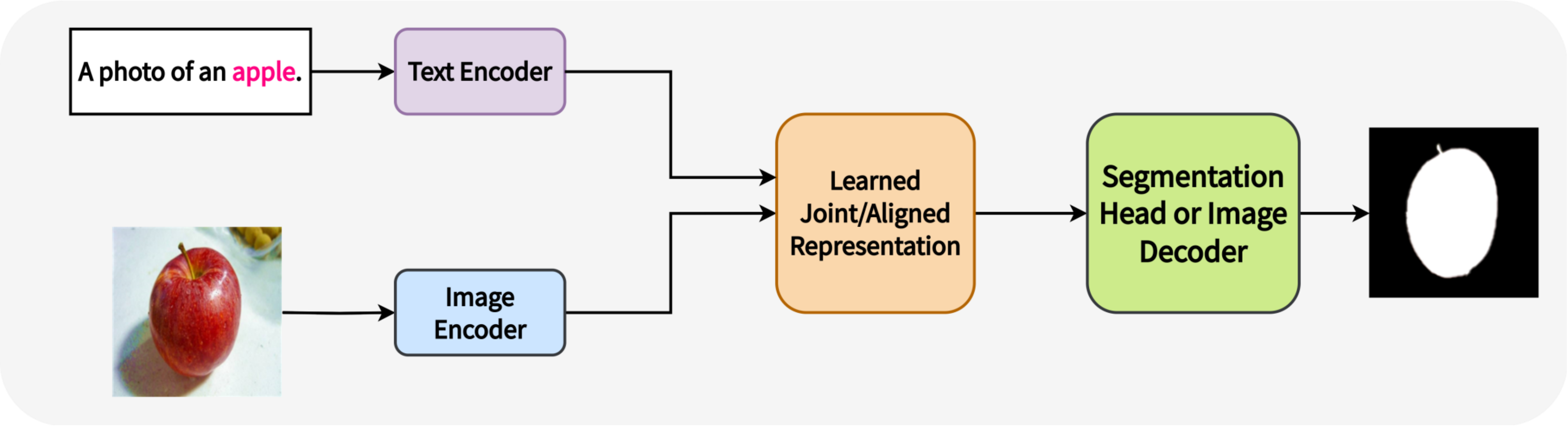
- Human Interactive Image Segmentation
- **Vision Language Segmentation Models (VLSMs)**
- Benchmarking Framework
- Prompt Generation
- Results

Foundation Vision Language Models (VLMs)



- Large scale pretraining to align text and image representations
- Millions of image-text pairs

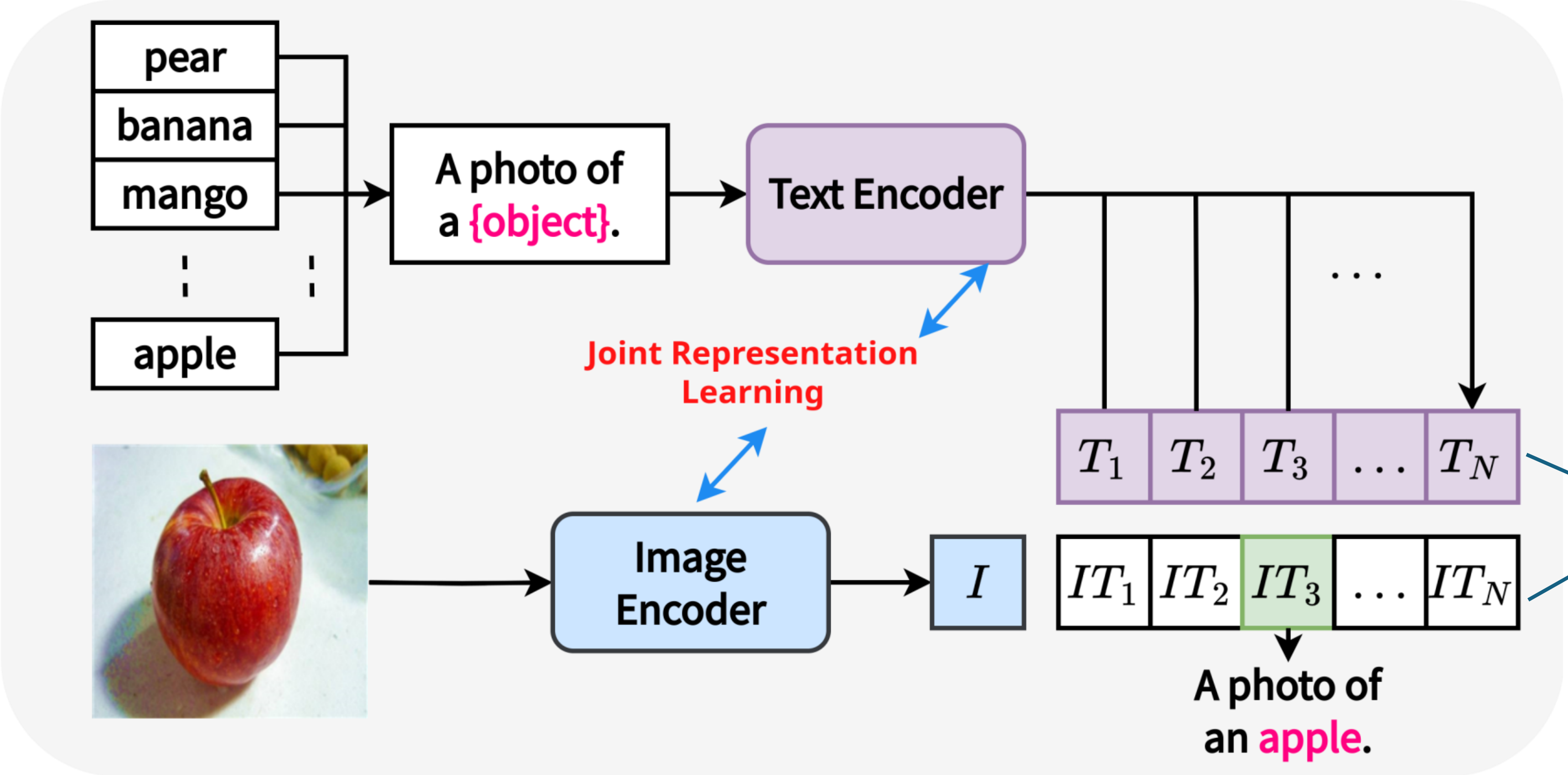
VLSMs using Foundation Models



- Large scale pretraining to align text and image representations
- Millions of image-text pairs
- VLSMs by adding a segmentation decoder

Vision Language Foundation Model: CLIP

The most popular vision language model trained on 400 million image-text pairs



Contrastive Loss

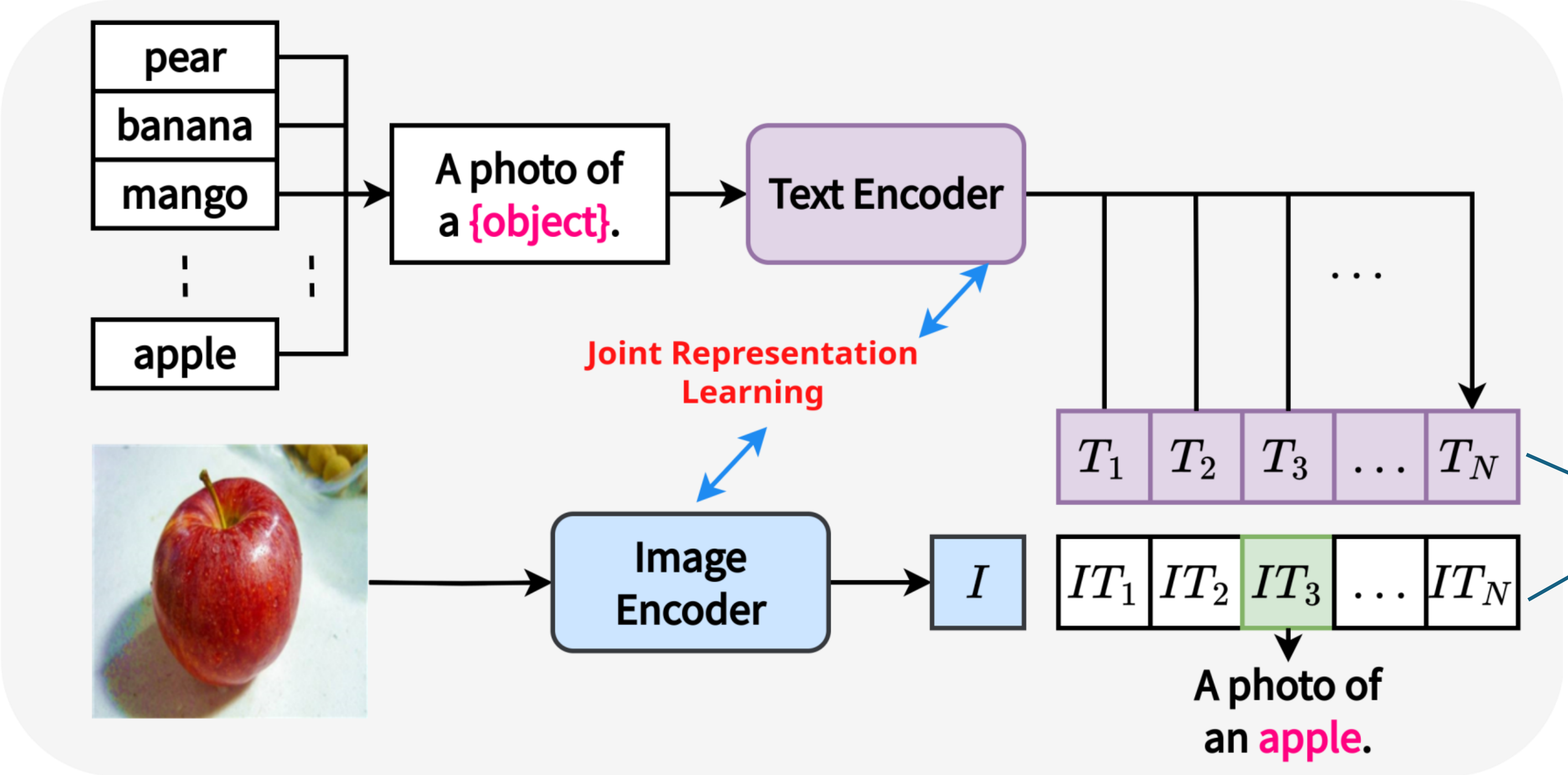
Pull similar pair closer

Dissimilar Apart

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Vision Language Foundation Model: CLIP

The most popular vision language model trained on 400 million image-text pairs



Contrastive Loss

Pull similar pair closer

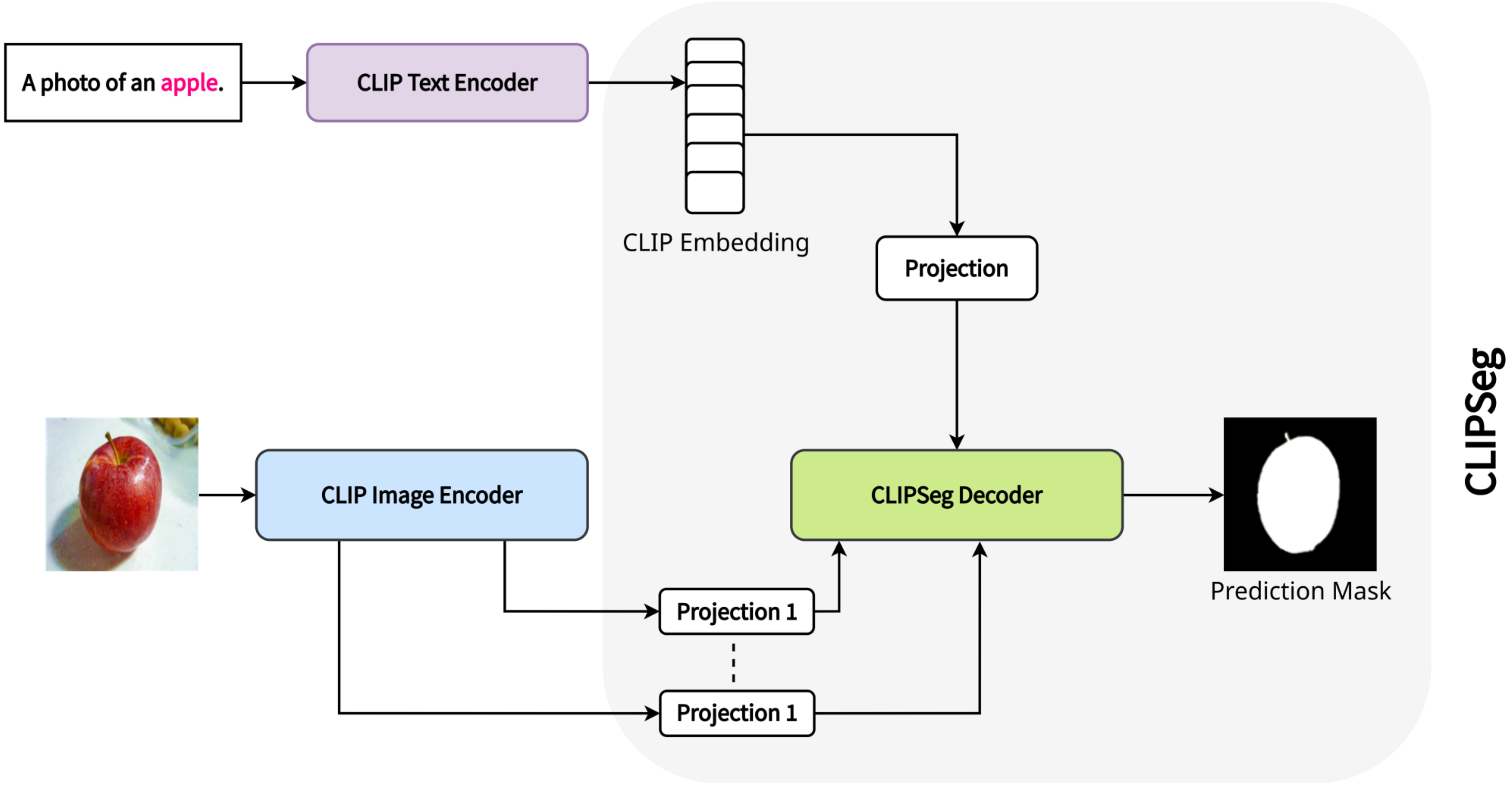
Dissimilar Apart

Reusing the encoders that have learnt powerful representations for building VLSMs

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

CLIPSeg

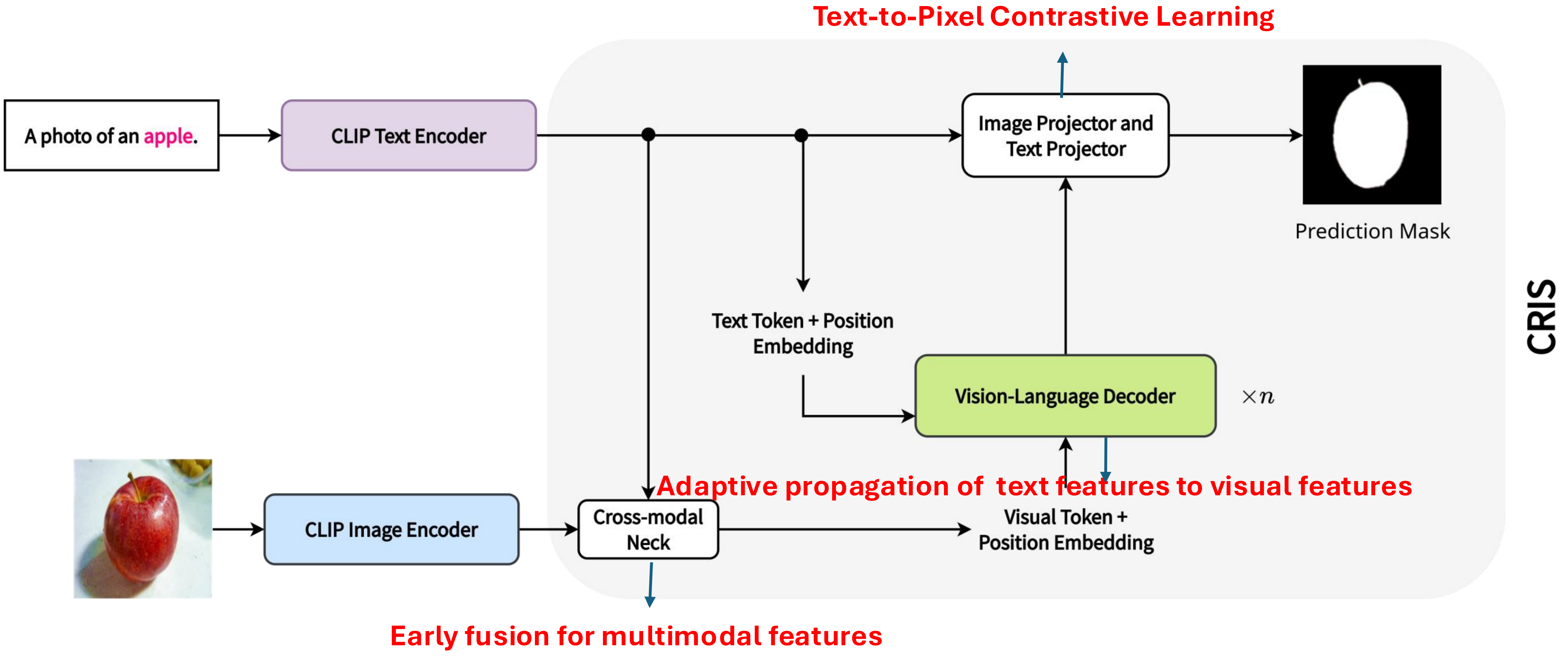
Trained on PhraseCut Dataset with 340,000 image-text paira



1. Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7086-7096).

CRIS

Trained on RefCOCO with 142,210 image-text pairs



Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., & Liu, T. (2022). Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11686-11695).

CLIP-based VLSMs for Medical Domain

BiomedCLIP: A VLM based on CLIP architecture

Trained from scratch on 15 million biomedical image-text pairs

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., ... & Poon, H. (2023). BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*

CLIP-based VLSMs for Medical Domain

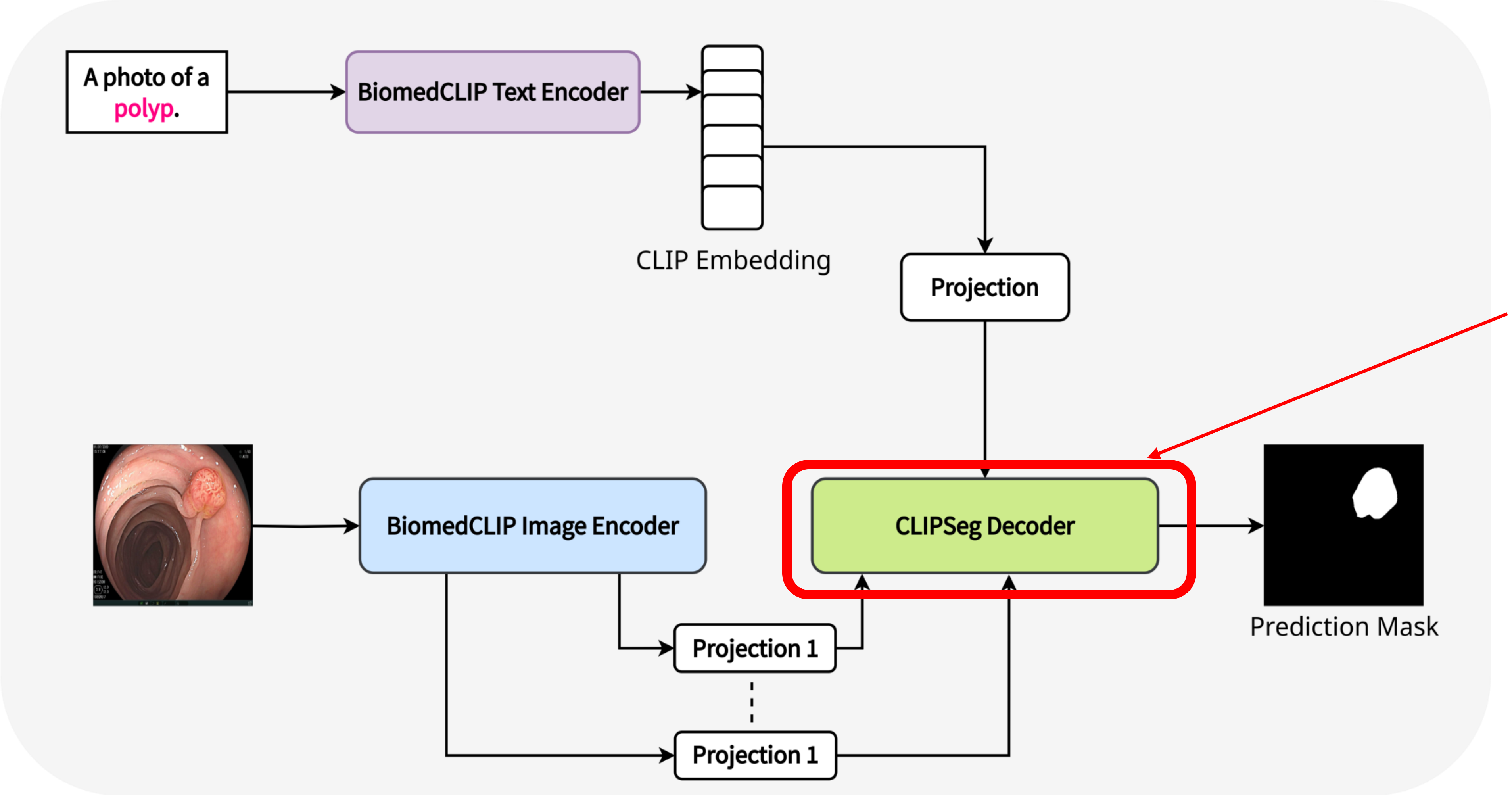
BiomedCLIP: A VLM based on CLIP architecture

Trained from scratch on 15 million biomedical image-text pairs

But no VLSMs built on top of BiomedCLIP!

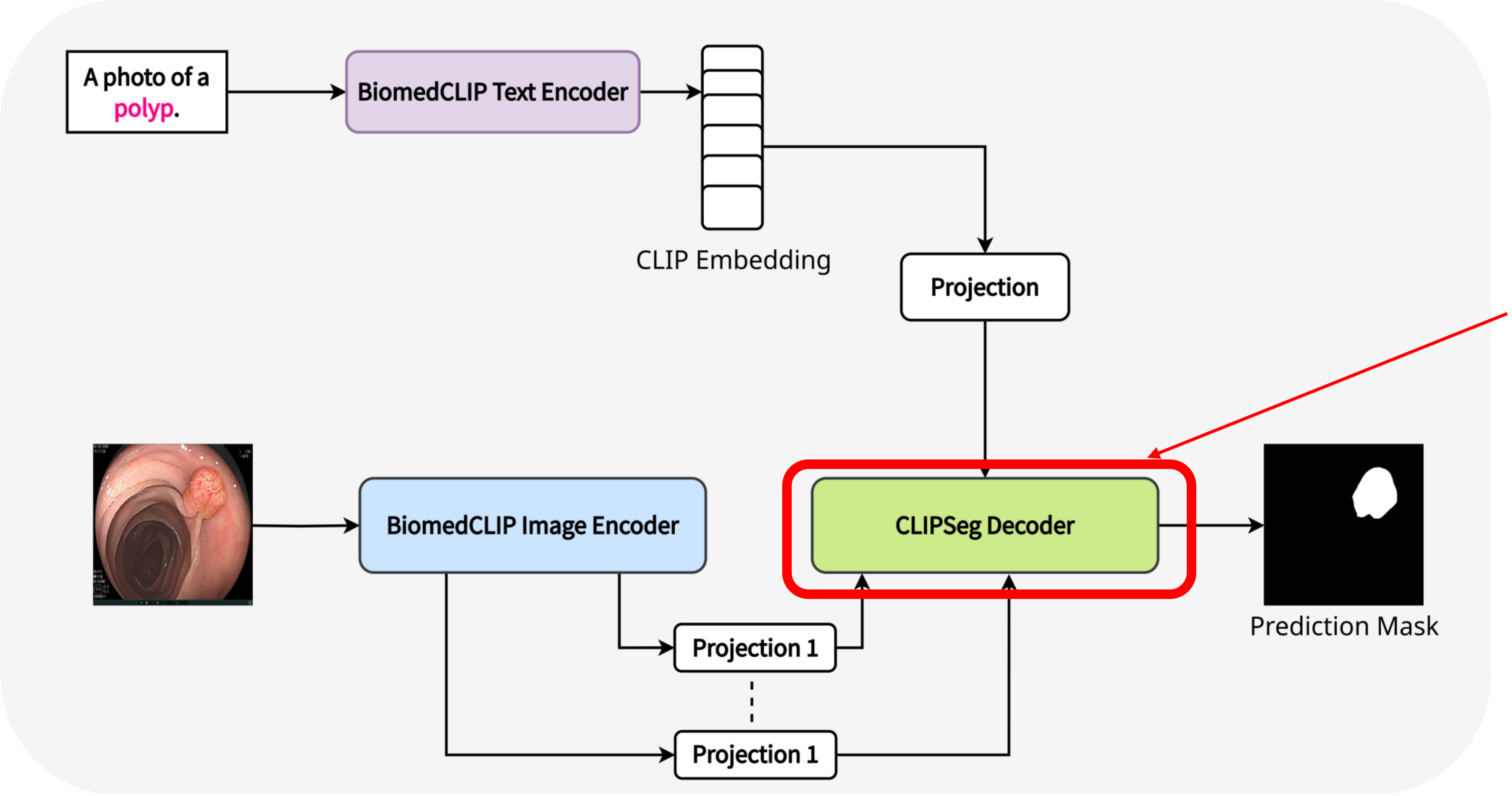
Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., ... & Poon, H. (2023). BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*

Potential Design: BiomedCLIPSeg



Randomly Initialized

Potential Design: BiomedCLIPSeg-D



Pretrained in CLIPSeg

Outline

- Human Interactive Image Segmentation
- Vision Language Segmentation Models (VLSMs)
- **Benchmarking Framework**
- Prompt Generation
- Results

Our Benchmarking Framework

Models	Pretrained on natural image-text pairs: **CRIS, CLIPSeg** Pretrained on medical image-text pairs: **BiomedCLIPSeg, BiomedCLIPSeg-D**

Our Benchmarking Framework

Models

Pretrained on natural image-text pairs: **CRIS, CLIPSeg**
Pretrained on medical image-text pairs: **BiomedCLIPSeg, BiomedCLIPSeg-D**

Key Questions

- Adaptation from natural domain to medical domain?
- Roles of text prompts and images during transfer learning?
- Pretrained on natural data vs pretrained on medical data?

Our Benchmarking Framework

Models

Pretrained on natural image-text pairs: **CRIS, CLIPSeg**
Pretrained on medical image-text pairs: **BiomedCLIPSeg, BiomedCLIPSeg-D**

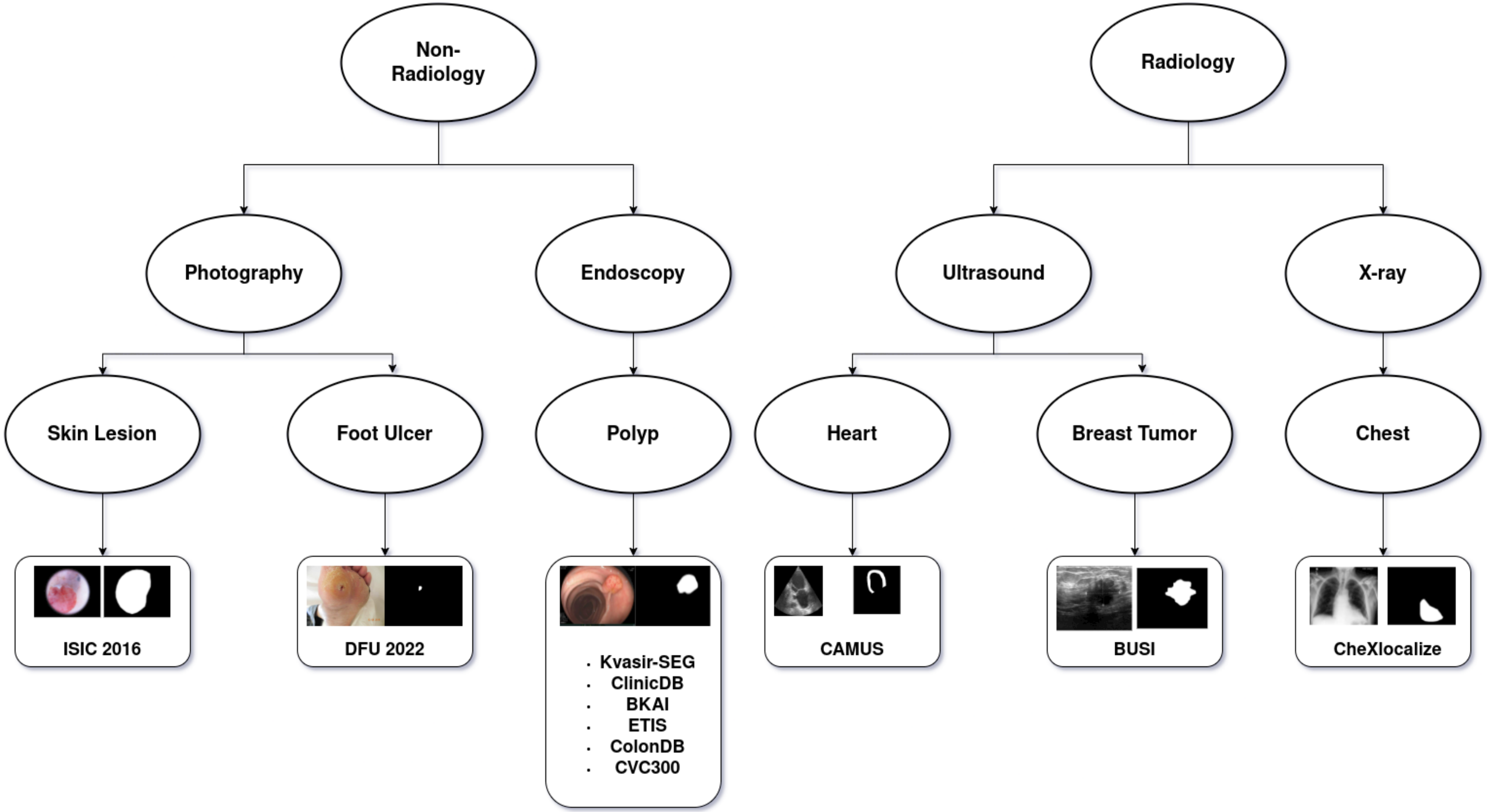
Key Questions

- Adaptation from natural domain to medical domain?
- Roles of text prompts and images during transfer learning?
- Pretrained on natural data vs pretrained on medical data?

Datasets

11 diverse medical imaging datasets with upto 9 different language prompts

2D Medical Imaging Datasets

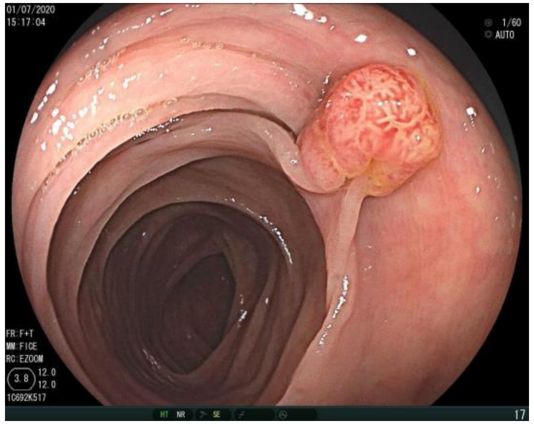


Outline

- Human Interactive Image Segmentation
- Vision Language Segmentation Models (VLSMs)
- Benchmarking Framework
- **Prompt Generation**
- Results

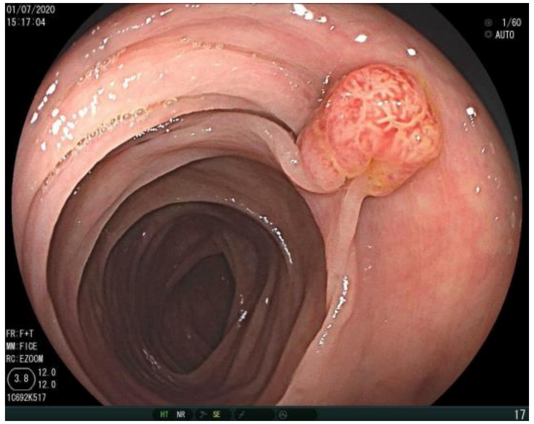
In a real clinical setting

This endoscopic image contains a polyp in the top right side, which is roughly circular in shape and has some reddish texture.



In a real clinical setting

This endoscopic image contains a polyp in the top right side, which is roughly circular in shape and has some reddish texture.

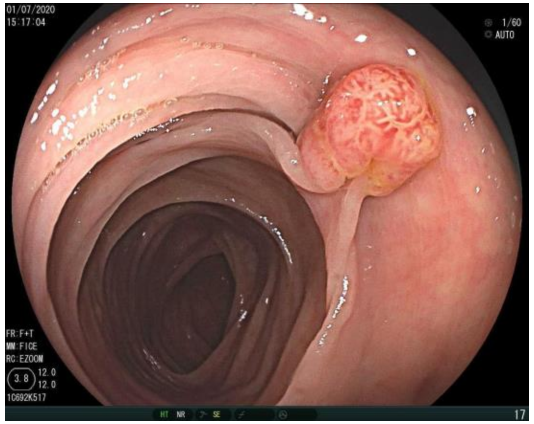


Training with rich prompts

No ground truth for prompts in our datasets

In a real clinical setting

This endoscopic image contains a polyp in the top right side, which is roughly circular in shape and has some reddish texture.



Training with rich prompts

No ground truth for prompts in our datasets

Automated Prompt Generation from existing datasets

Automatic Prompt generation

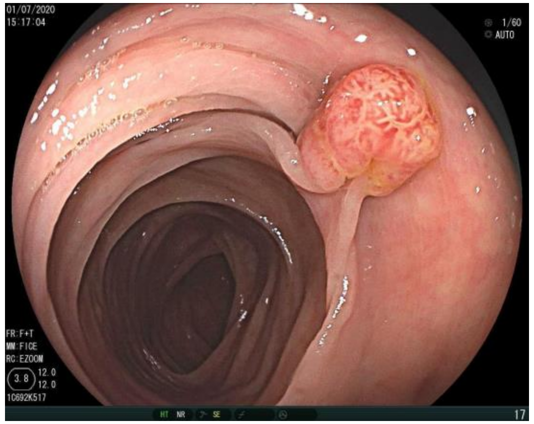
Prompt attributes generated using:

- Image processing on masks
- VQA models
- Class information from online medical journals
- Metadata and radiology reports in datasets when available

Automatic Prompt generation

Prompt attributes generated using:

- **Image processing on masks**
- VQA models
- Class information from online medical journals
- Metadata and radiology reports in datasets when available



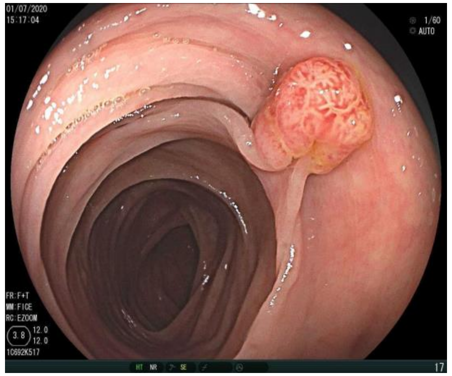
→ **Image Processing** →

Number: **one**
Size: **small**
Location: **top right**

Automatic Prompt generation

Prompt attributes generated using:

- Image processing on masks
- **VQA models**
- Class information from online medical journals
- Metadata and radiology reports in datasets when available



VQA model

Color: **pink**
Shape: **round**

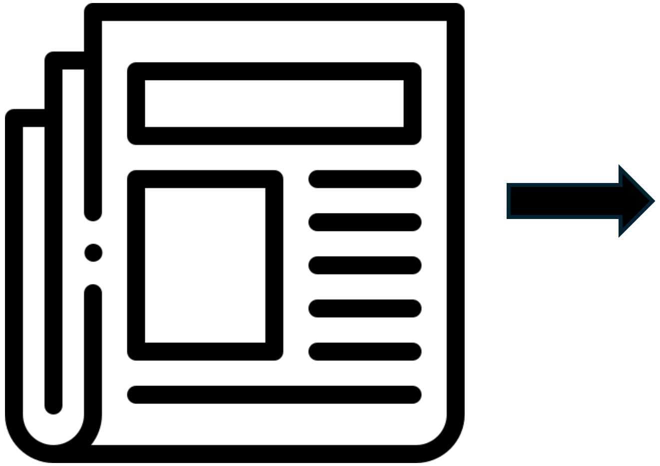
What is the color of the target object in the image?

What is the shape of the target object in the image?

Automatic Prompt generation

Prompt attributes generated using:

- Image processing on masks
- VQA models
- **Class information from online medical journals**
- Metadata and radiology reports in datasets when available



"polyp, which is a projecting growth of tissue"

"skin melanoma, which is a rough wound on skin"

"foot ulcer, which is an open sore or lesion in foot and toes"

Automatic Prompt generation

Prompt attributes generated using:

- Image processing on masks
- VQA models
- Class information from online medical journals
- **Metadata and radiology reports in datasets when available**

Metadata in datasets →

- View
- Pathology
- Cardiac Cycle
- Gender
- Age
- Image Quality

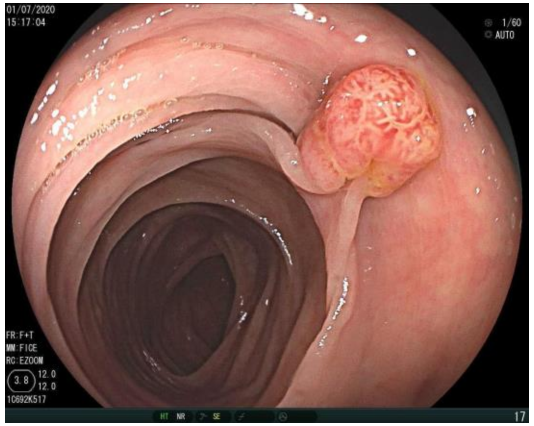
Automatic Prompt generation

Prompt attributes generated using:

- Image processing on masks
- VQA models
- Class information from online medical journals
- Metadata and radiology reports in datasets when available

Upto 9 different prompts with growing complexity

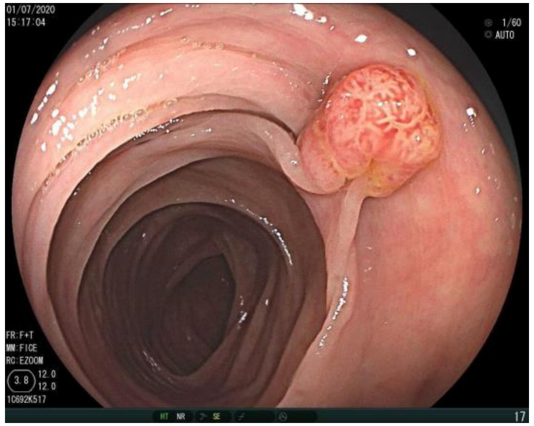
9 different prompts with growing complexity



Target
Structure
polyp '

Prompt for Endoscopy Datasets: P1

9 different prompts with growing complexity

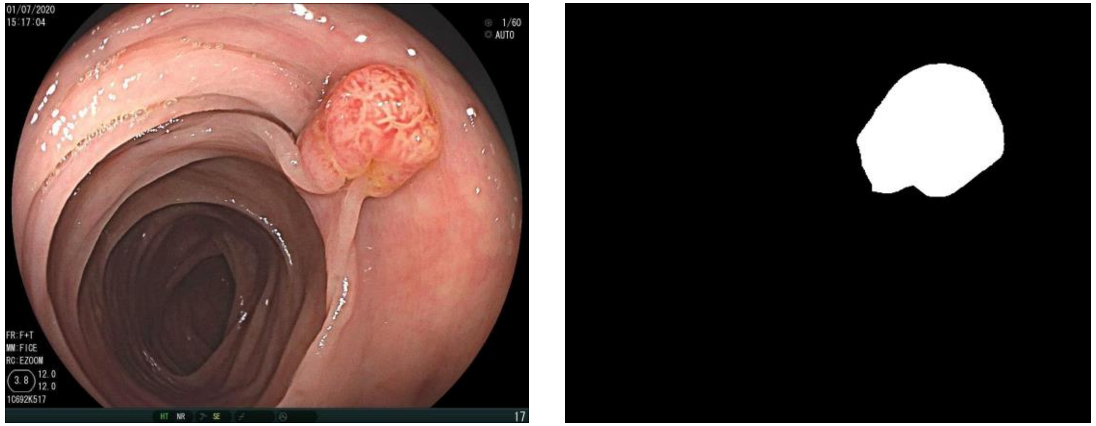


Shape — round
Target Structure — polyp

round polyp

Prompt for Endoscopy Datasets: P2

9 different prompts with growing complexity

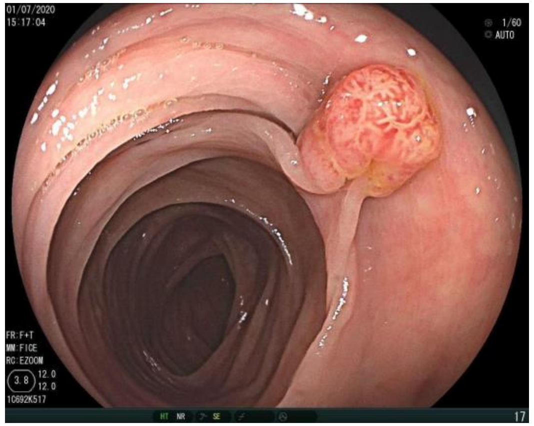


Color | Shape | Target Structure

pink round polyp

Prompt for Endoscopy Datasets: P3

9 different prompts with growing complexity

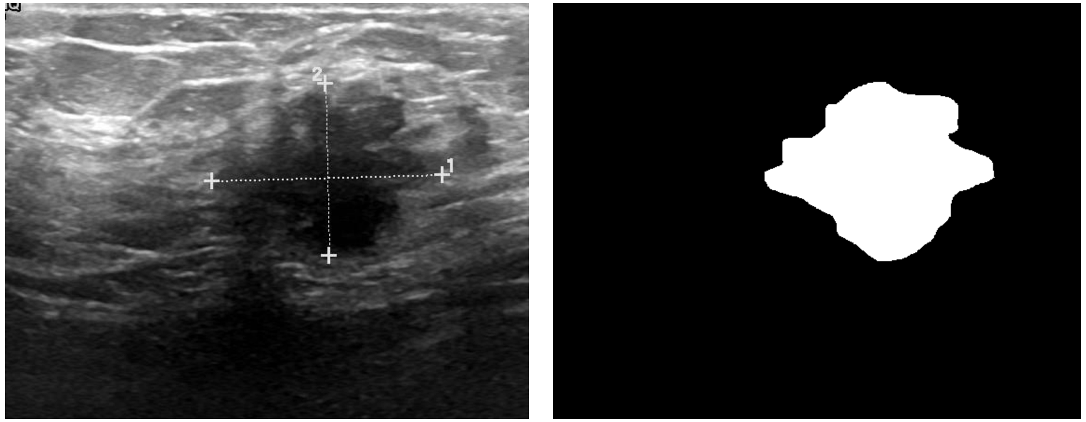


Number: One | Size: small | Color: pink | Shape: round | Target Structure: polyp | Class Specific General Description: a projecting growth of tissue | Position: top right

One small pink round polyp which is a projecting growth of tissue, located in top right of the image.

Prompt for Endoscopy Datasets: P9

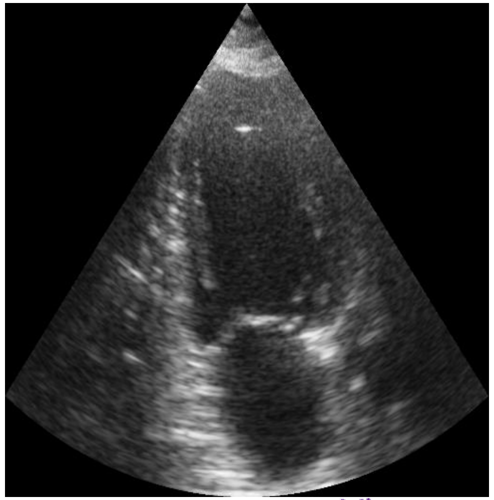
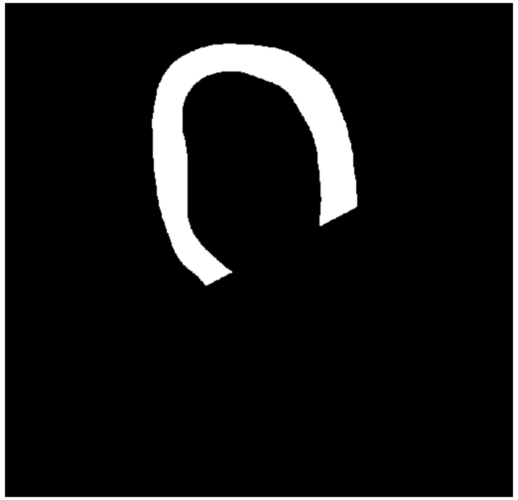
9 different prompts with growing complexity



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

Prompt for BUSI:P6

9 different prompts with growing complexity

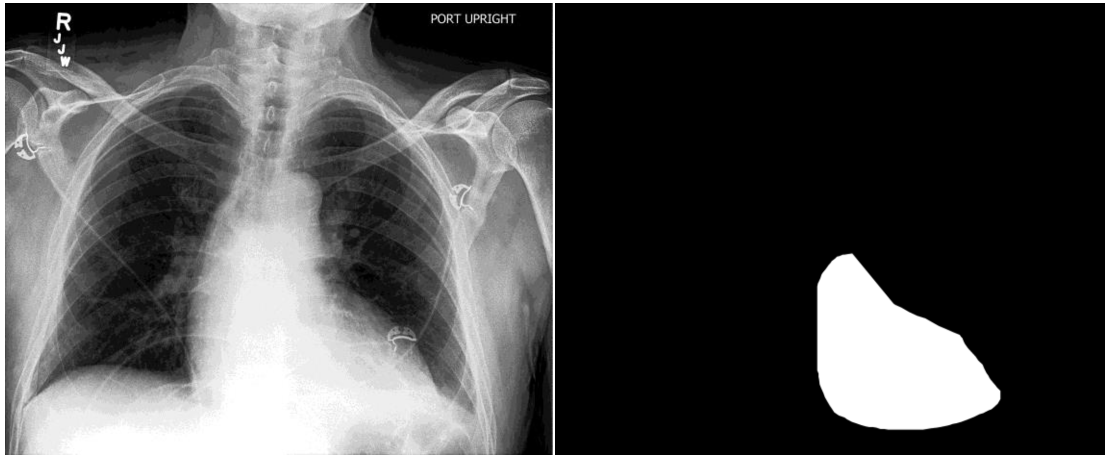
View

Myocardium of square shape in two-chamber view in the cardiac ultrasound at the end of the systole cycle of a seventy-one-year-old male with medium image quality.

Cardiac Cycle Patient's Age Patient's Sex Image Quality

Prompt for Camus: P7

9 different prompts with growing complexity



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

Prompt for CheXlocalize: P5

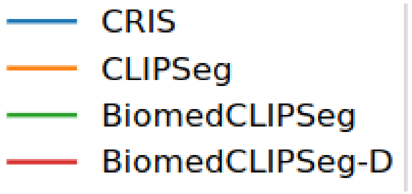
Outline

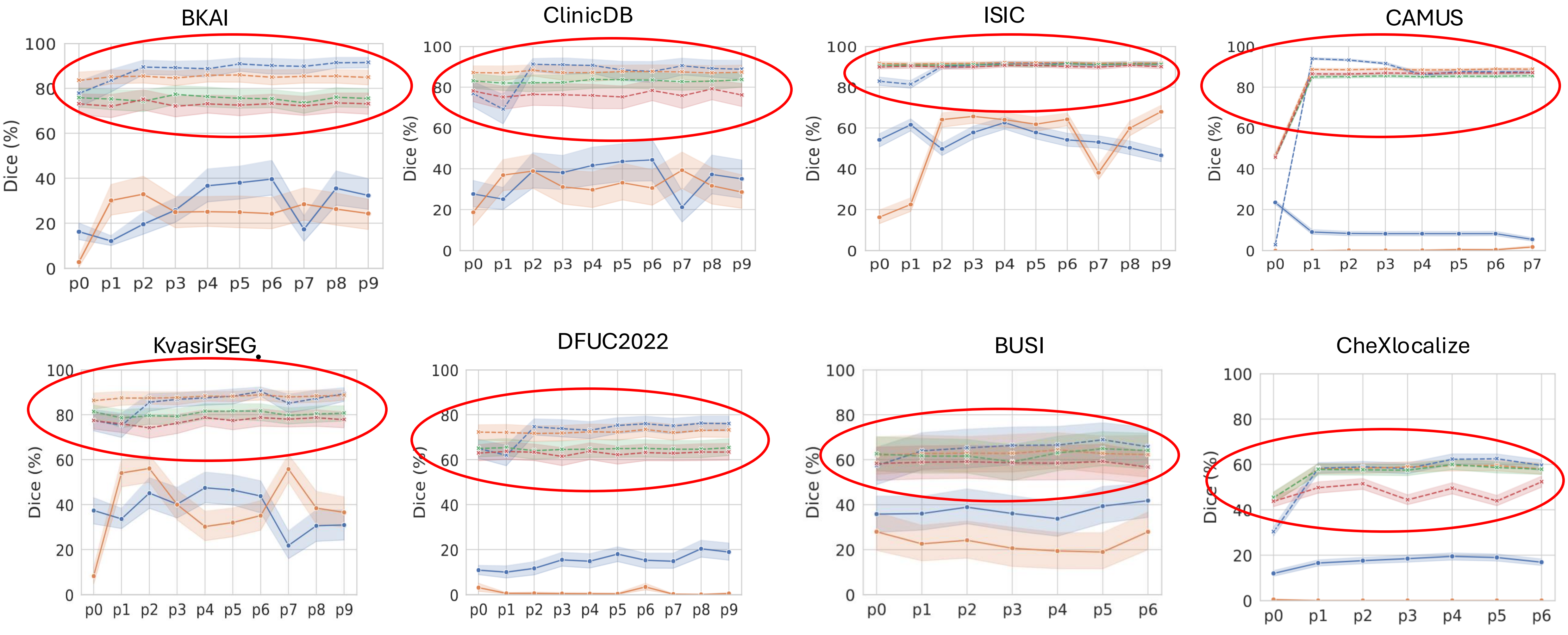
- Human Interactive Image Segmentation
- Vision Language Segmentation Models (VLSMs)
- Benchmarking Framework
- Prompt Generation
- **Results**

Does making prompt richer improve finetuning performance?

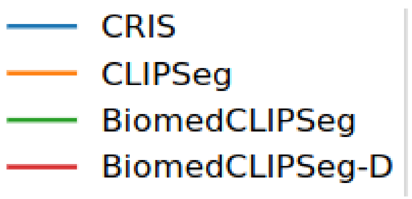
Does making prompt richer improve finetuning performance?

- Well....not much!

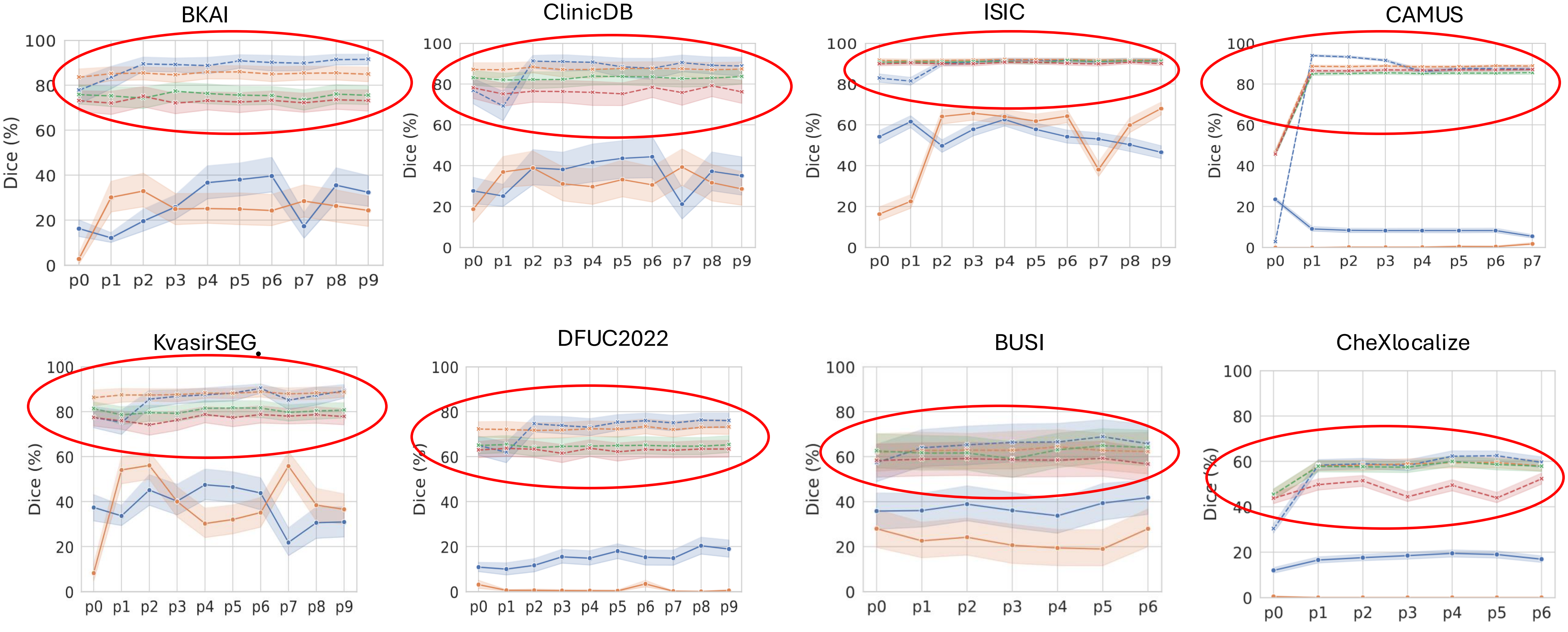




Does making prompt richer improve finetuning performance?

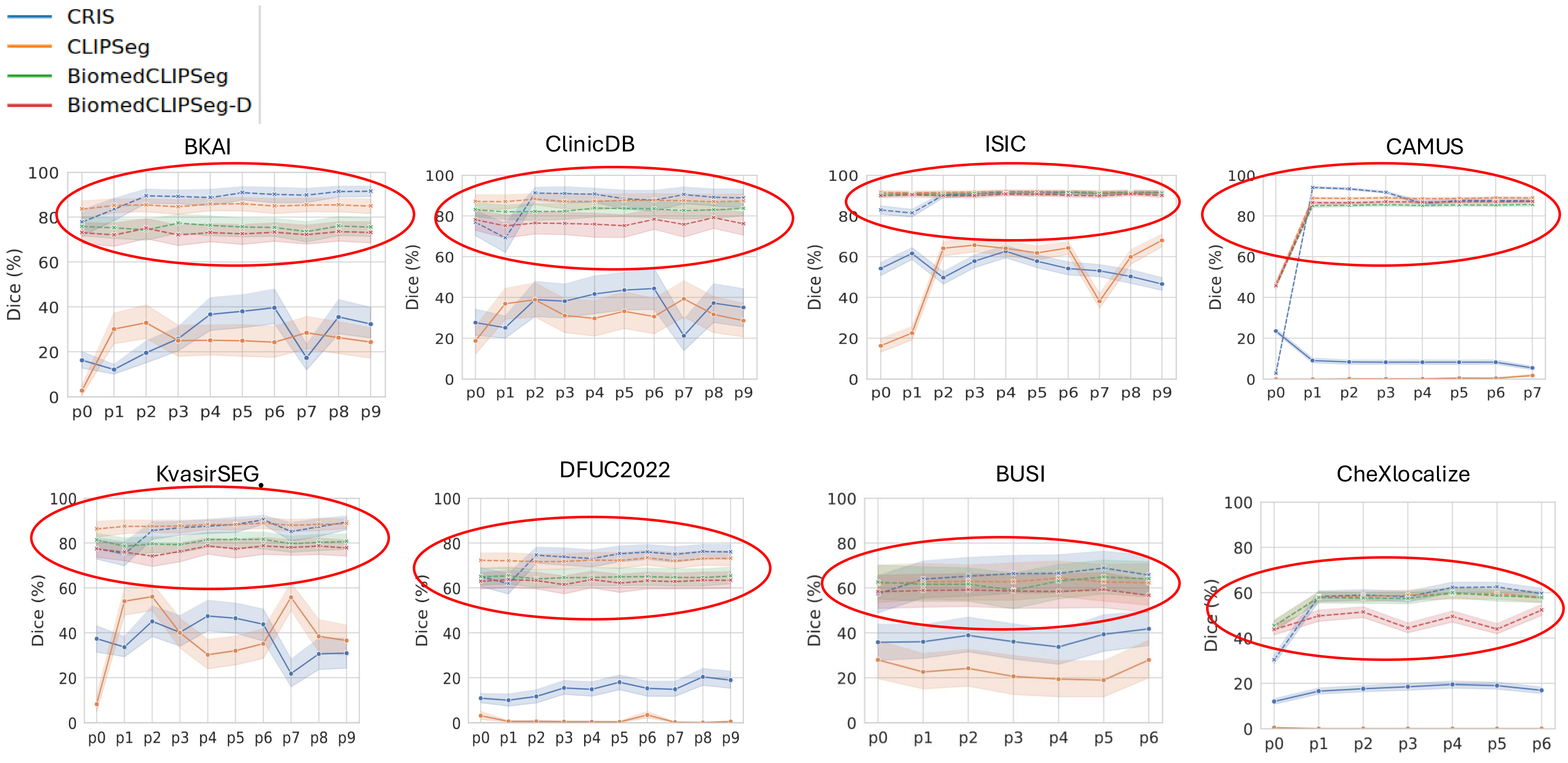


- Well....not much!
- Minimal DSC variation across all models and datasets
- Performance mostly saturates after adding only the class name (P1)



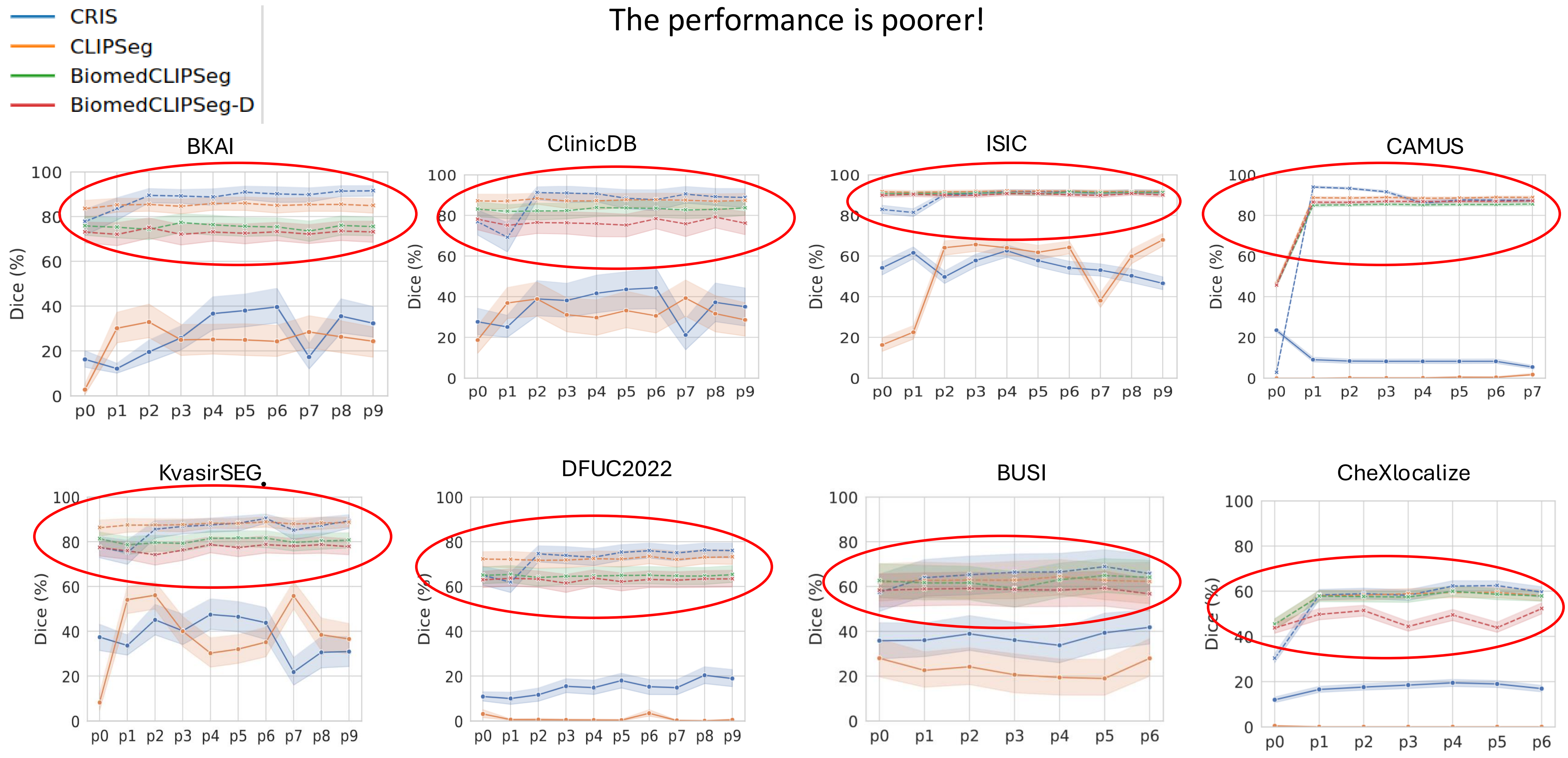
How do models pretrained on medical image-text pair perform?

How do models pretrained on medical image-text pairs perform?



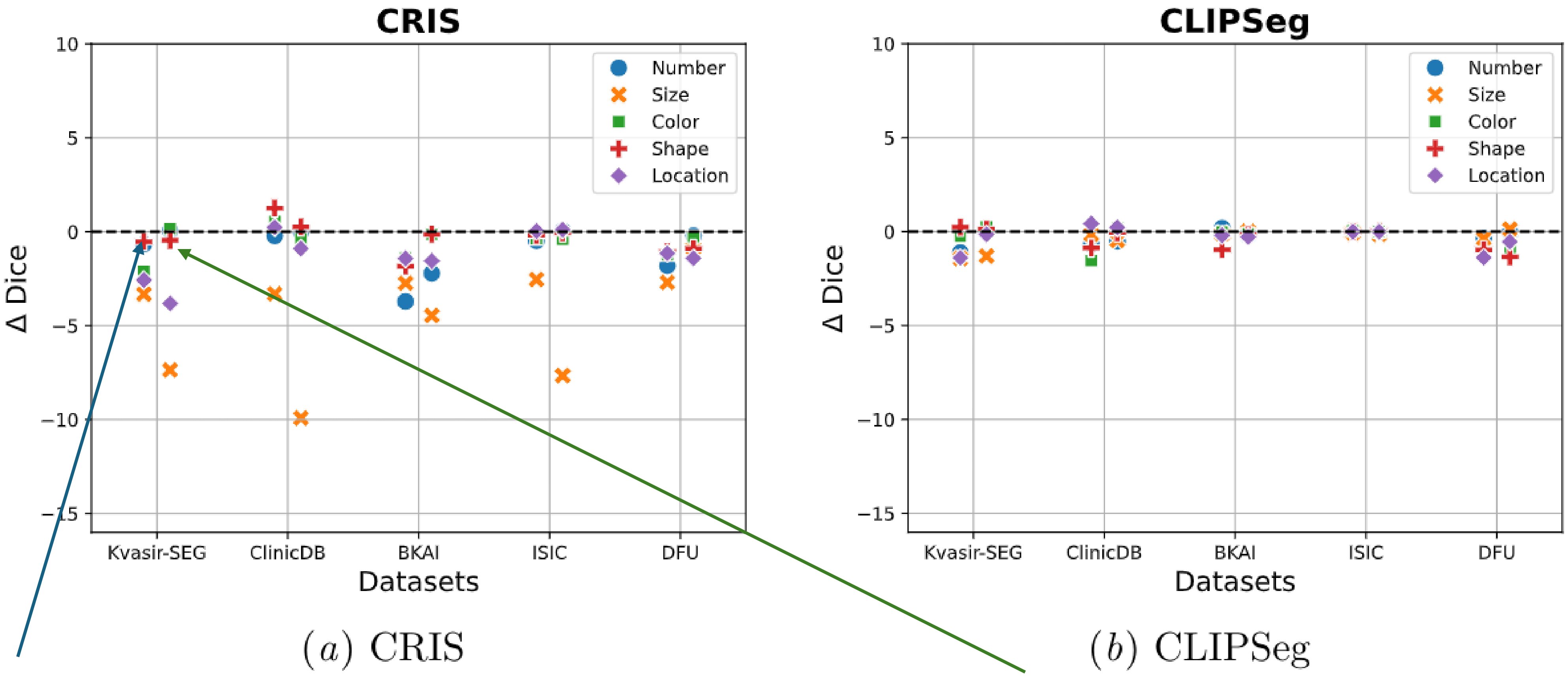
How do models pretrained on medical image-text pairs perform?

The performance is poorer!



Do models capture language semantics well?

Do models capture language semantics well?



(a) CRIS

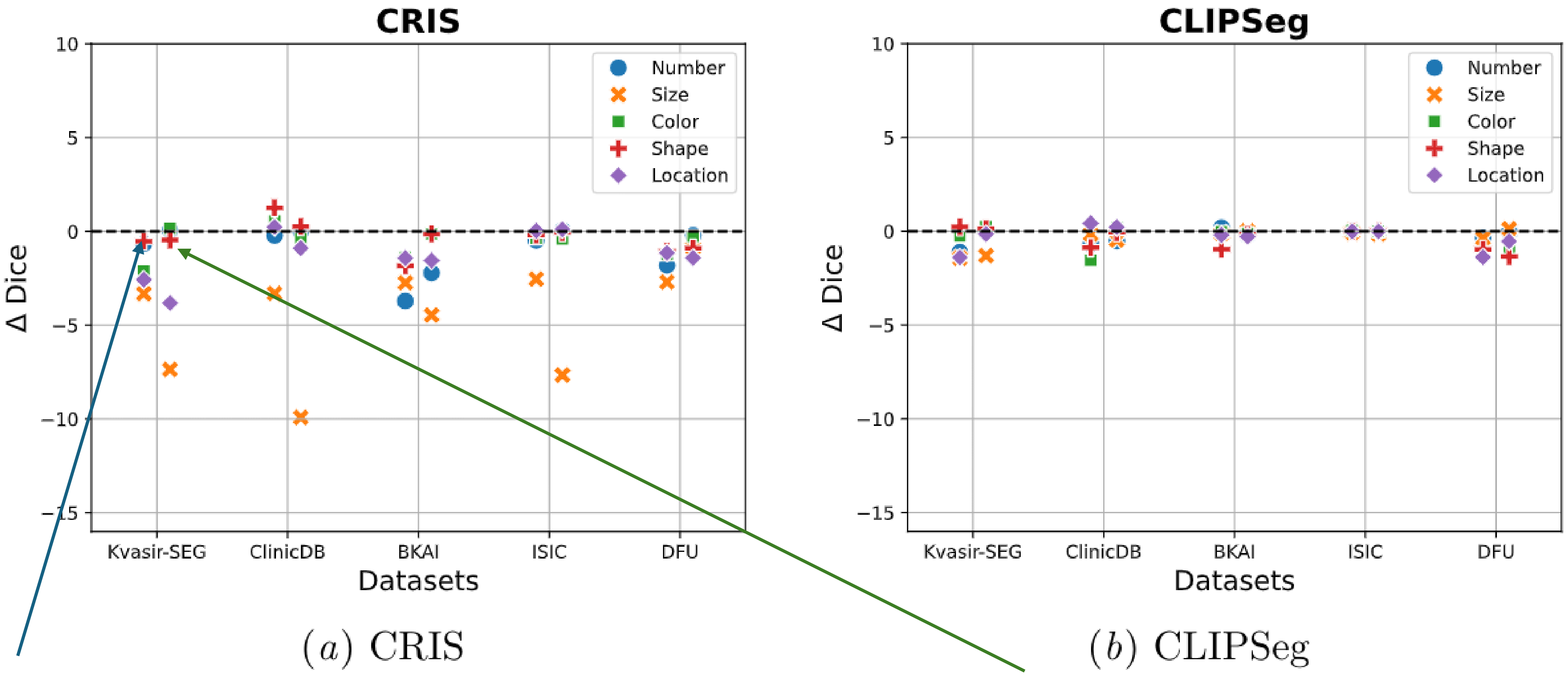
(b) CLIPSeg

Replace attribute values by uncommon English words(replace "large" with "xenogeny")

Replace attribute values by semantically opposite words (replace "large" with "small")

Do models capture language semantics well?

- Altering attributes notably deteriorates CRIS's performance, particularly for: size and location attributes
- More significant decline for semantically opposite words
- Hinting that CRIS is more influenced (or is better able to capture) text semantics

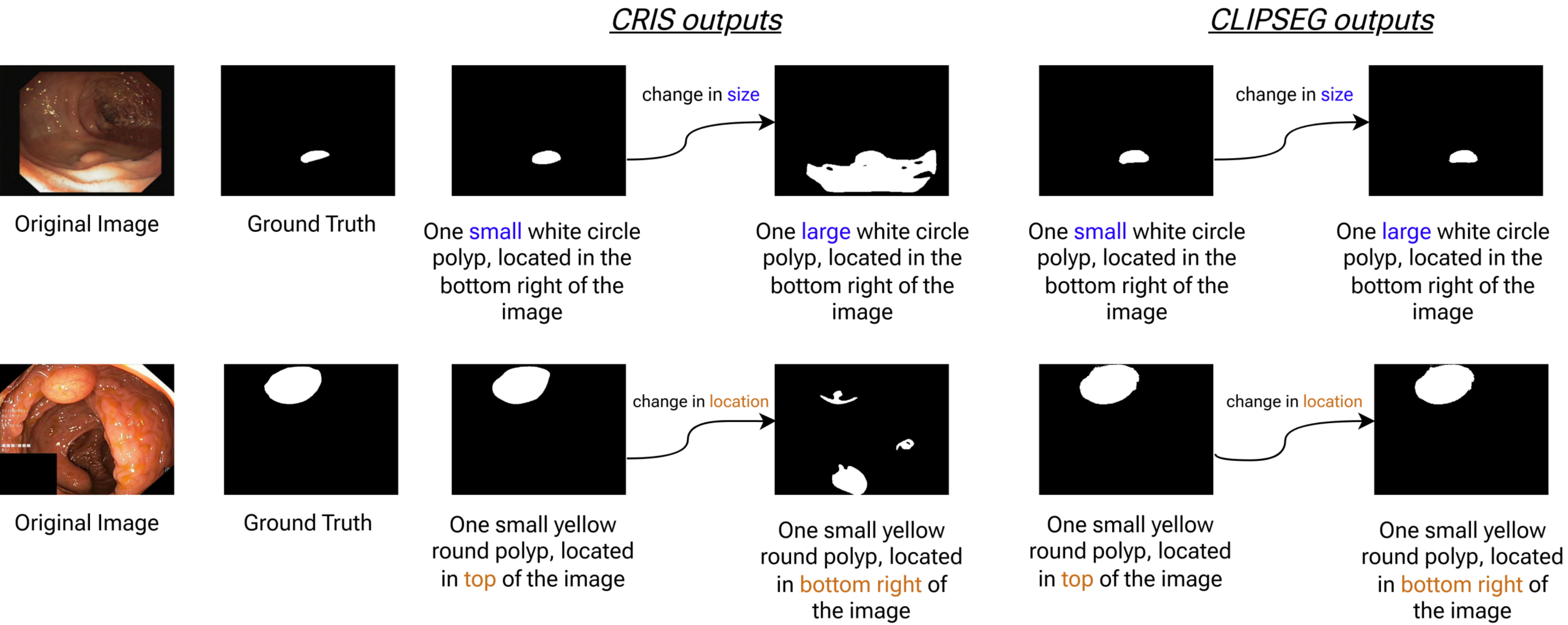


(a) CRIS (b) CLIPSeg

Replace attribute values by uncommon English words(replace "large" with "xenogeny")

Replace attribute values by semantically opposite words (replace "large" with "small")

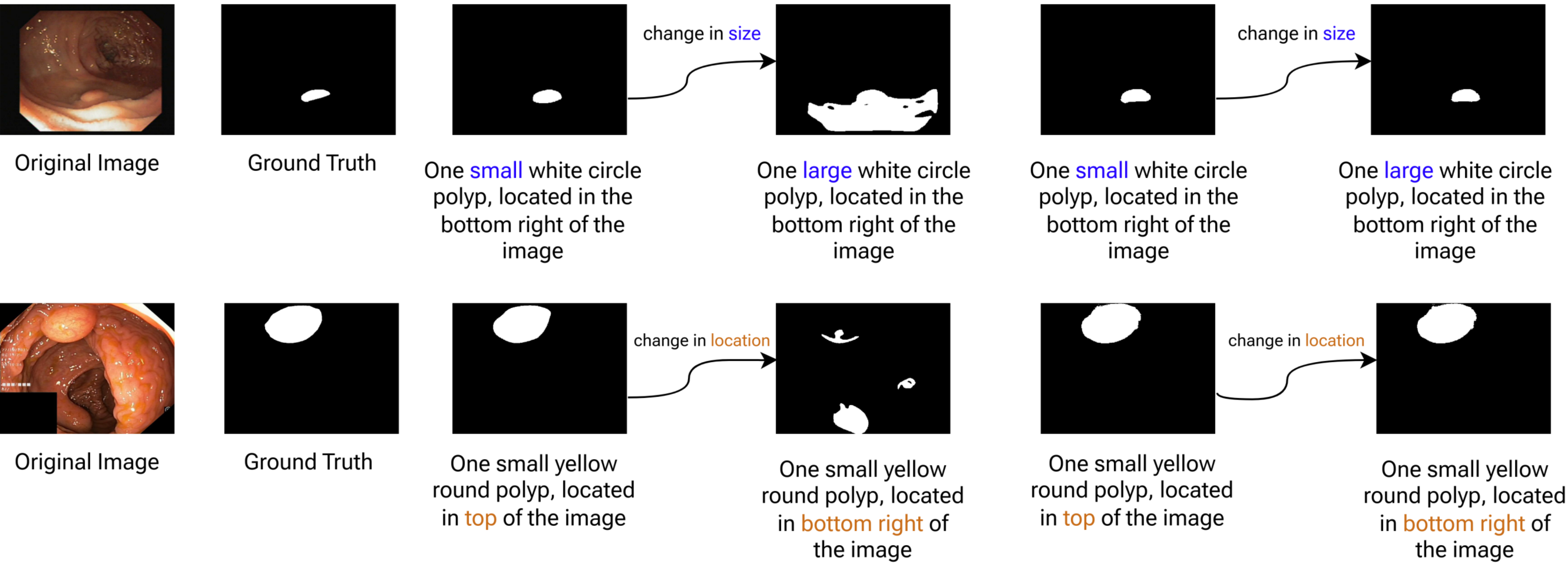
Do models capture language semantics well?



Do models capture language semantics well?

CRIS demonstrates stronger effects of text semantics for location and size attributes

CRIS outputs | *CLIPSEG outputs*



Are VLSMs robust to out-of-distribution data?

Are VLSMs robust to out-of-distribution data?

Tested on → Finetuned on ↓	Model ↓	Kvasir-SEG
Kvasir-SEG	CRIS	**<u>91.39</u>**
	CLIPSeg	89.51
	UNet	84.77
	UNet++	84.70
	DeepLabv3+	84.11
ClinicDB	CRIS	82.66
	CLIPSeg	**84.02**
	UNet	65.80
	UNet++	61.93
	DeepLabv3+	66.63
BKAI	CRIS	**83.74**
	CLIPSeg	83.70
	UNet	68.42
	UNet++	70.64
	DeepLabv3+	69.02

Are VLSMs robust to out-of-distribution data?

VLSMs are robust to out-of-distribution data compared to conventional models.

Tested on → / Finetuned on ↓	Model ↓	Kvasir-SEG
Kvasir-SEG	CRIS	**<u>91.39</u>**
	CLIPSeg	89.51
	UNet	84.77
	UNet++	84.70
	DeepLabv3+	84.11
ClinicDB	CRIS	82.66
	CLIPSeg	**84.02**
	UNet	65.80
	UNet++	61.93
	DeepLabv3+	66.63
BKAI	CRIS	**83.74**
	CLIPSeg	83.70
	UNet	68.42
	UNet++	70.64
	DeepLabv3+	69.02

Key takeaways

- Just adding new attributes to prompts does not help segmentation performance during finetuning
- BiomedCLIP based segmentation models performed worse than CLIP based segmentation models
- CRIS captures better language semantics compared to CLIPSeg
- VLSMs adapt better to distribution shift than conventional models

Key takeaways

- Just adding new attributes to prompts does not help segmentation performance during finetuning
- BiomedCLIP based segmentation models performed worse than CLIP based segmentation models
- CRIS captures better language semantics compared to CLIPSeg
- VLSMs adapt better to distribution shift than conventional models

Text prompts are powerful, but more work is needed in building models that can leverage its power

Thank you!

MIDL Paris 2024





NAAMII

Scan to read paper