ACCV HANOI VIETNAM 2024 DEC 8-12

NAAMII

TuneVLSeg: Prompt Tuning Benchmark for Vision-Language Segmentation Models

17th Asian Conference on Computer Vision

Hanoi, Vietnam

10th December 2024

Rabin Adhikari, **Safal Thapaliya**, Manish Dhakal, Bishesh Khanal




NAAMII, Nepal

Outline

- Vision Language Models (VLMs) and Segmentation models (VLSMs)
- Adapting foundational VLMs and VLSMs
- Prompt Tuning
- TuneVLSeg Benchmark Framework
- Key Results

Outline

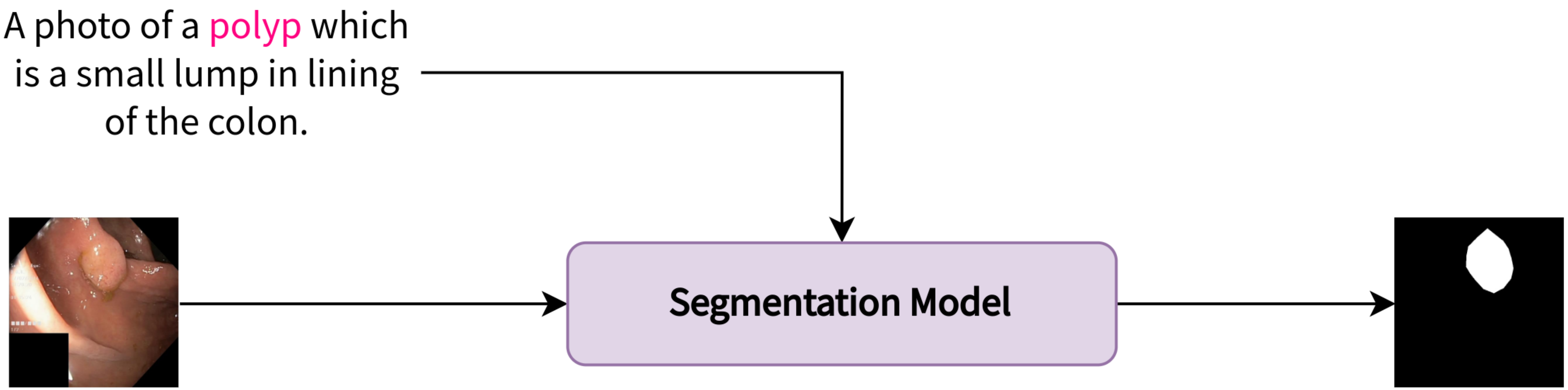
- **Vision Language Models (VLMs) and Segmentation models (VLSMs)**
- Adapting foundational VLMs and VLSMs
- Prompt Tuning
- TuneVLSeg Benchmark Framework
- Key Results

Segmentation in Medical Images

- Crucial for diagnosis, prognosis and surgery planning
- Recent segmentation models:
  - Excellent performance on curated datasets
  - Lack generalization across image modalities and datasets
  - Requires retraining when new classes are introduced

Segmentation with prompts

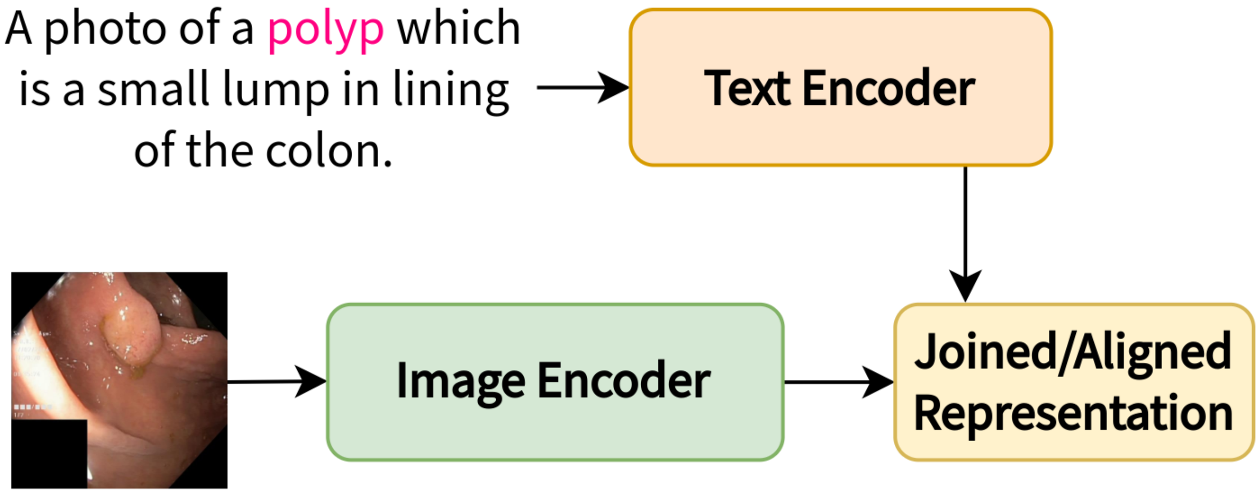
- Enables human interaction by describing the target structure
- Open vocabulary segmentation on new classes
- Easier to adapt models to new image modalities and datasets



Segmentation with prompts

- Enables human interaction by describing the target structure
- Open vocabulary segmentation on new classes
- Easier to adapt models to new image modalities and datasets

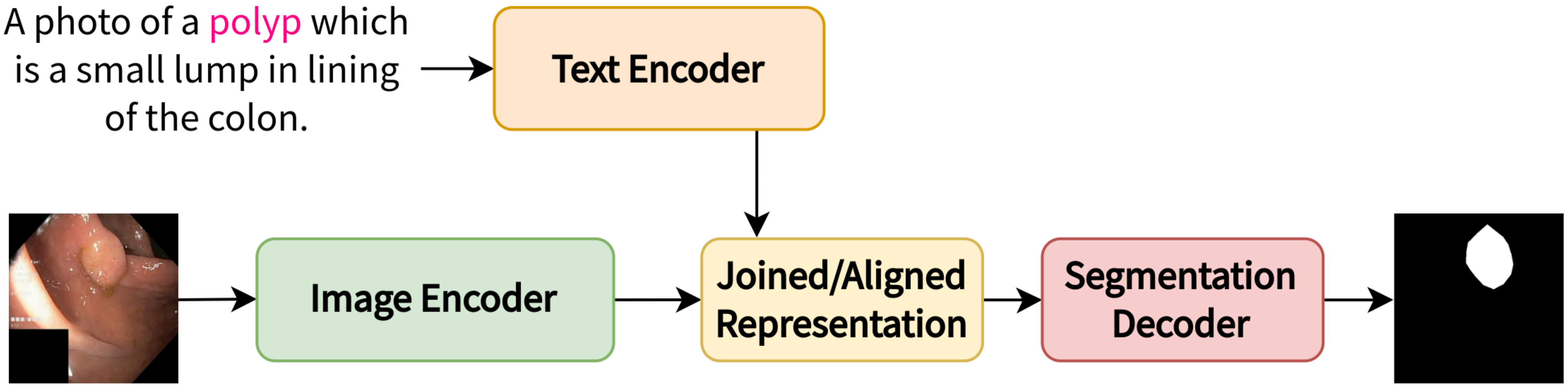
Vision-Language Model (VLM)



Segmentation with prompts

- Enables human interaction by describing the target structure
- Open vocabulary segmentation on new classes
- Easier to adapt models to new image modalities and datasets

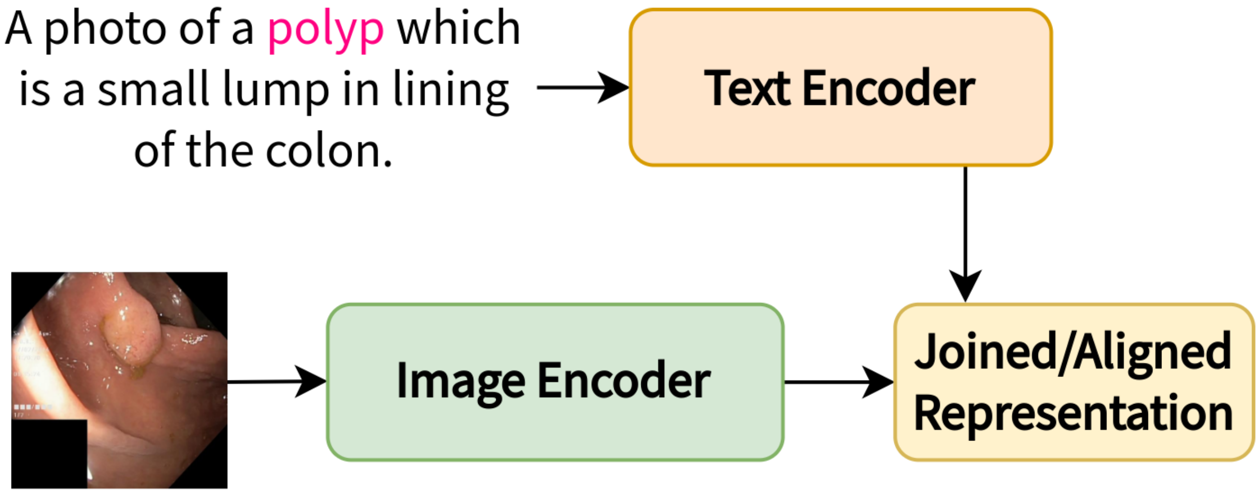
Vision-Language Segmentation Model (VLSM)



Foundational VLMs

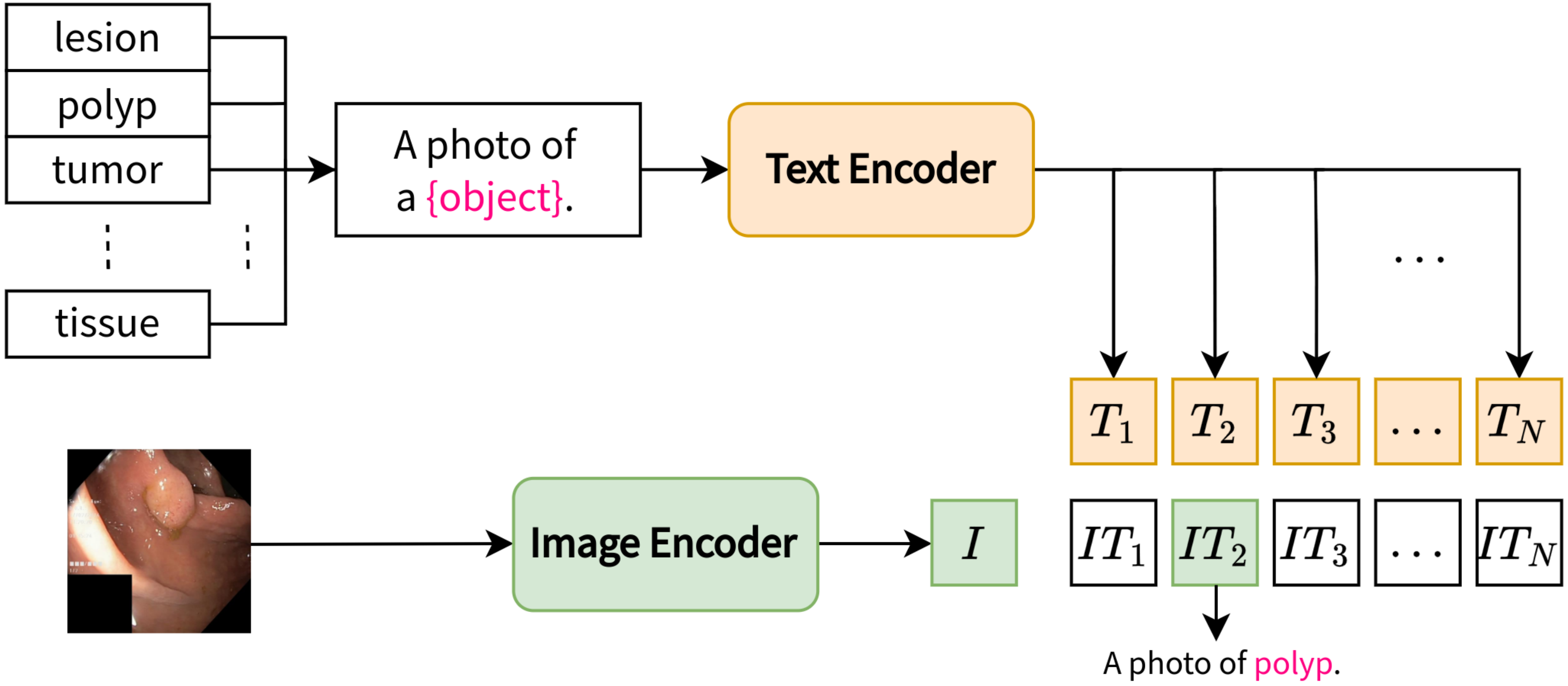
- Large scale pretraining to align text and image representations
- Millions of image-text pairs

Vision-Language Model (VLM)



Foundational VLMs: CLIP

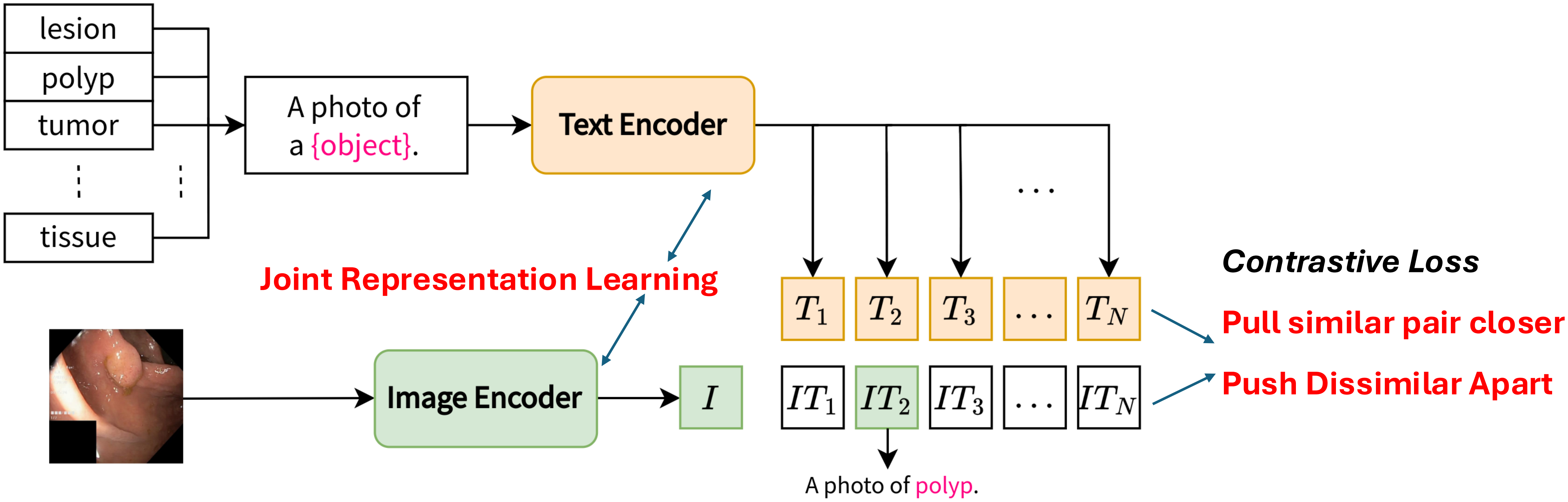
The most popular vision language model trained on 400 million image-text pairs



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.

Foundational VLMs: CLIP

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

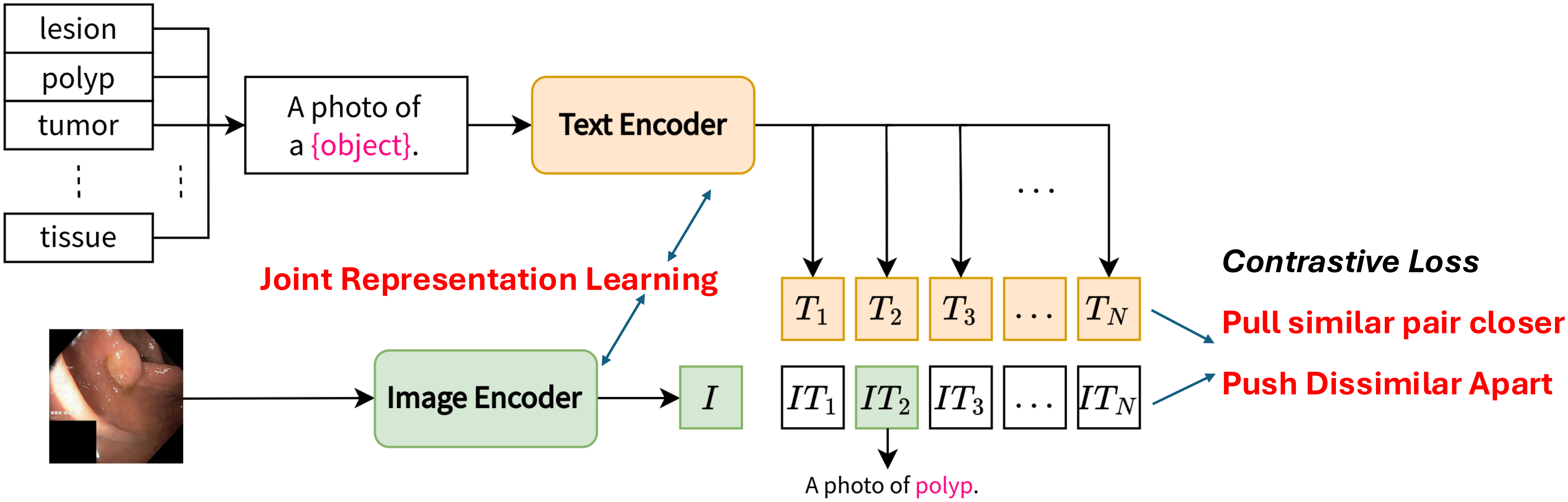


Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Foundational VLMs: CLIP

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

Reusing the encoders that have learnt powerful representations for building VLSMs

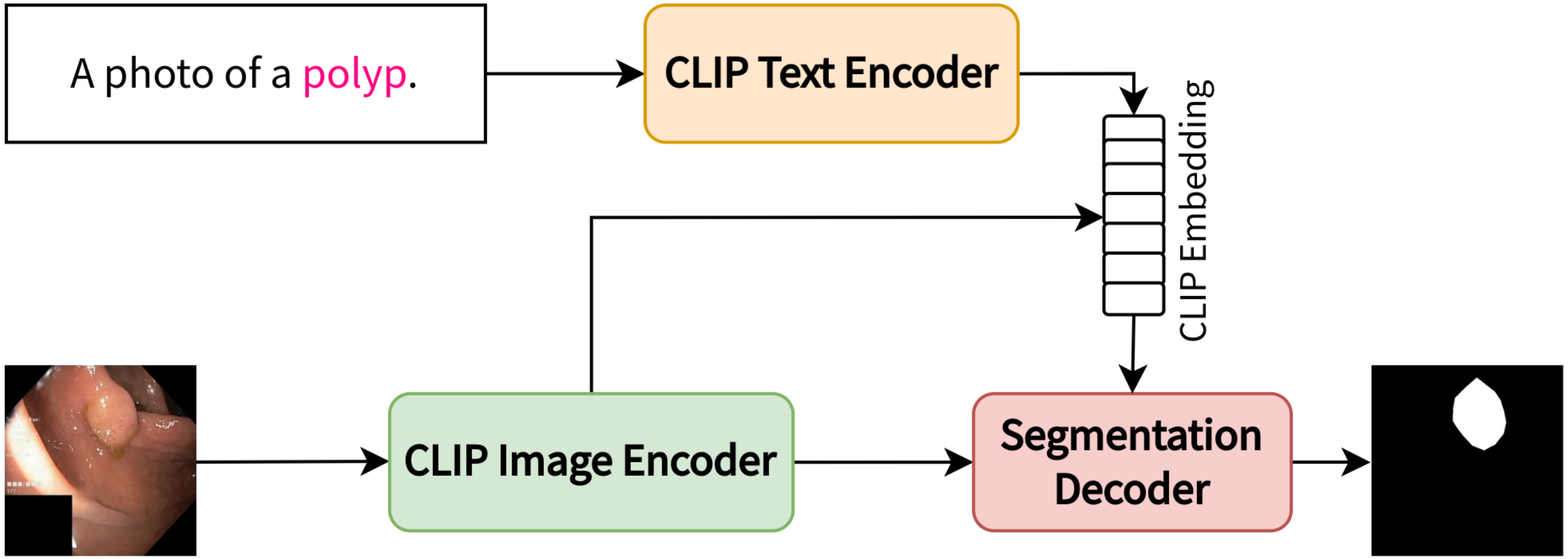


Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

Foundational VLSM: CLIPSeg

- Trained on PhraseCut Dataset with 340,000 image-text pairs
- Excellent zero-shot and few-shot performance on natural image segmentation
  - Due to the *prompts*

Both encoders are Transformer models



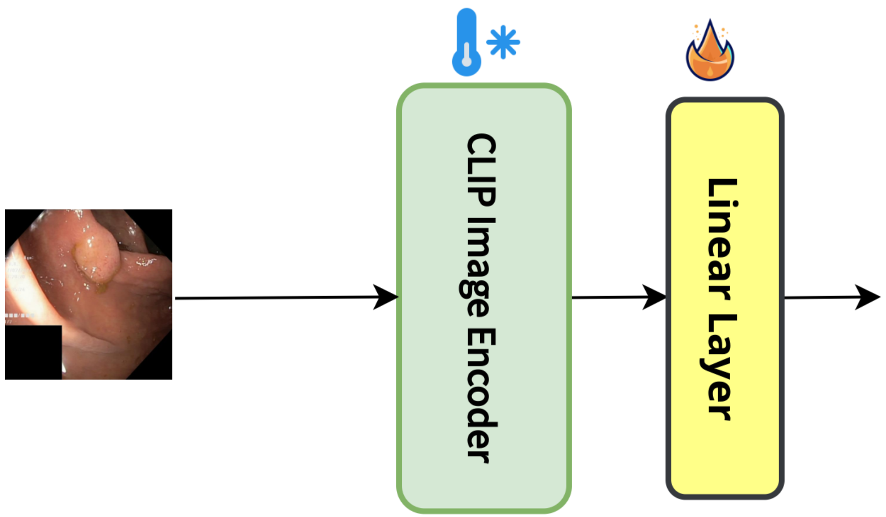
Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7086-7096).

Outline

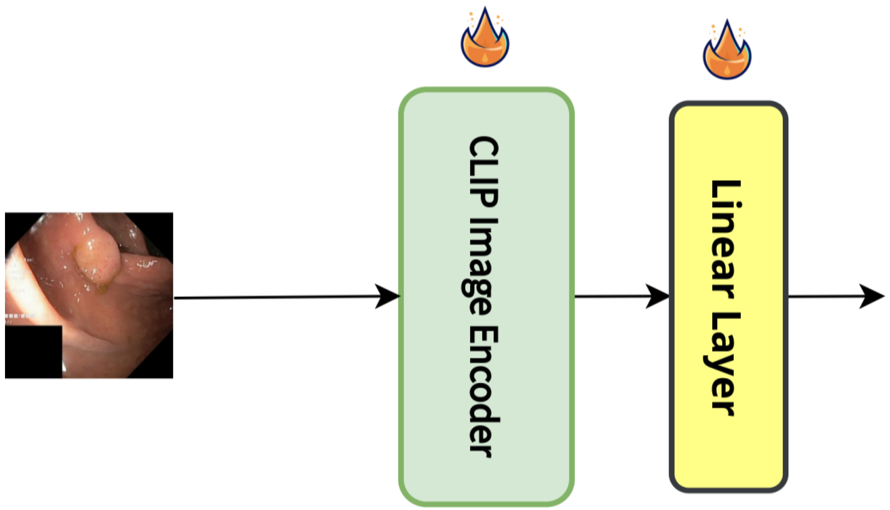
- Vision Language Models (VLMs) and Segmentation models (VLSMs)
- **Adapting foundational VLMs and VLSMs**
- Prompt Tuning
- TuneVLSeg Benchmark Framework
- Key Results

Adapting foundational VLMs for medical images

- Scarce labeled medical datasets
- Massive scale of models
- Finetuning these models is infeasible for medical images



(i) Linear Probing

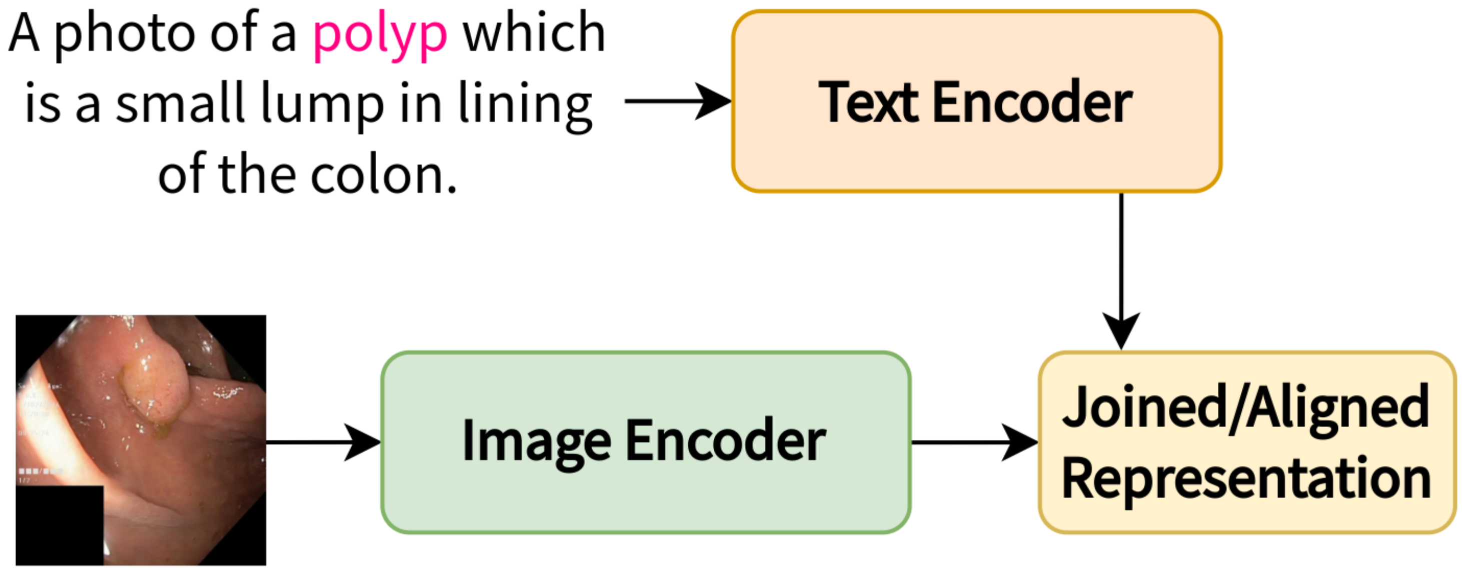


(ii) Full Fine-tuning

Adapting foundational VLMs for medical images

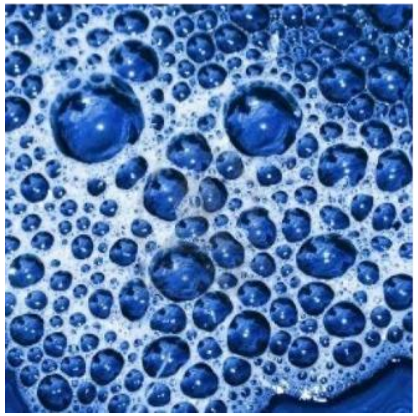
Prompt Engineering

Try out multiple text prompts



Prompt Engineering in VLMs improves performance

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32

Different prompts perform differently due to inherent bias in dataset

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56

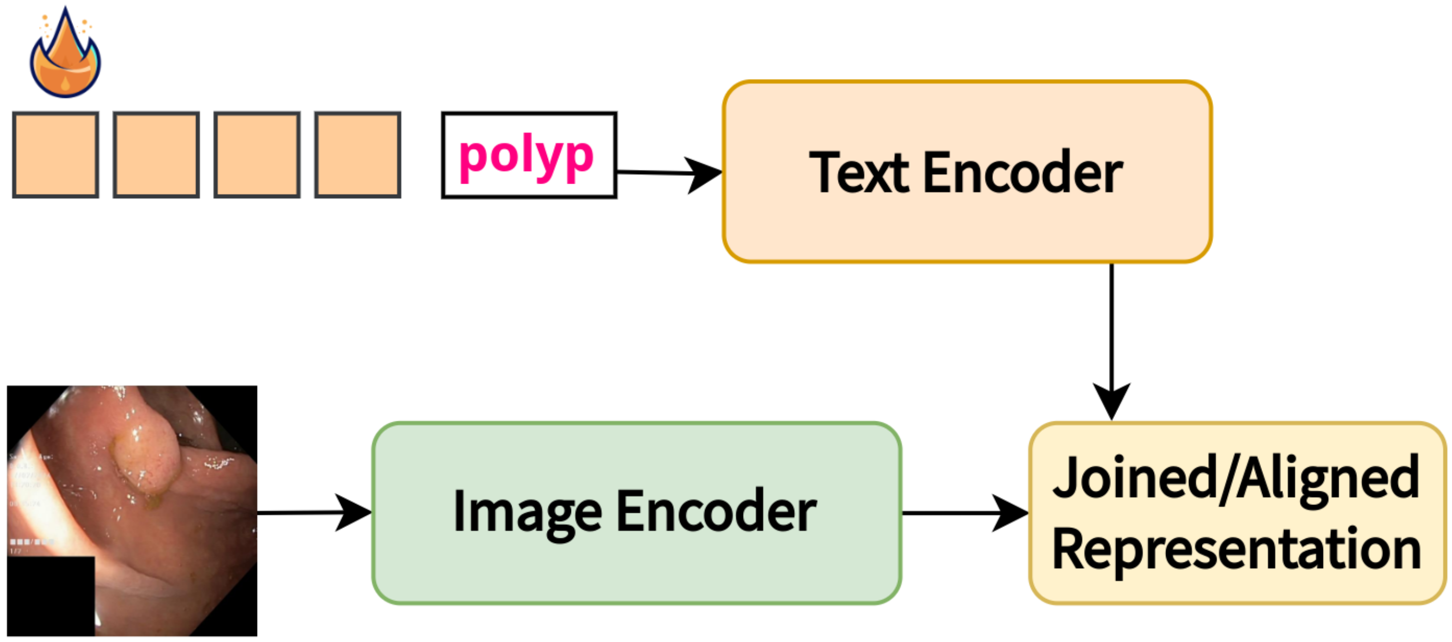
It is hard to find the right set of prompts

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337-2348.

Adapting foundational VLMs for medical images

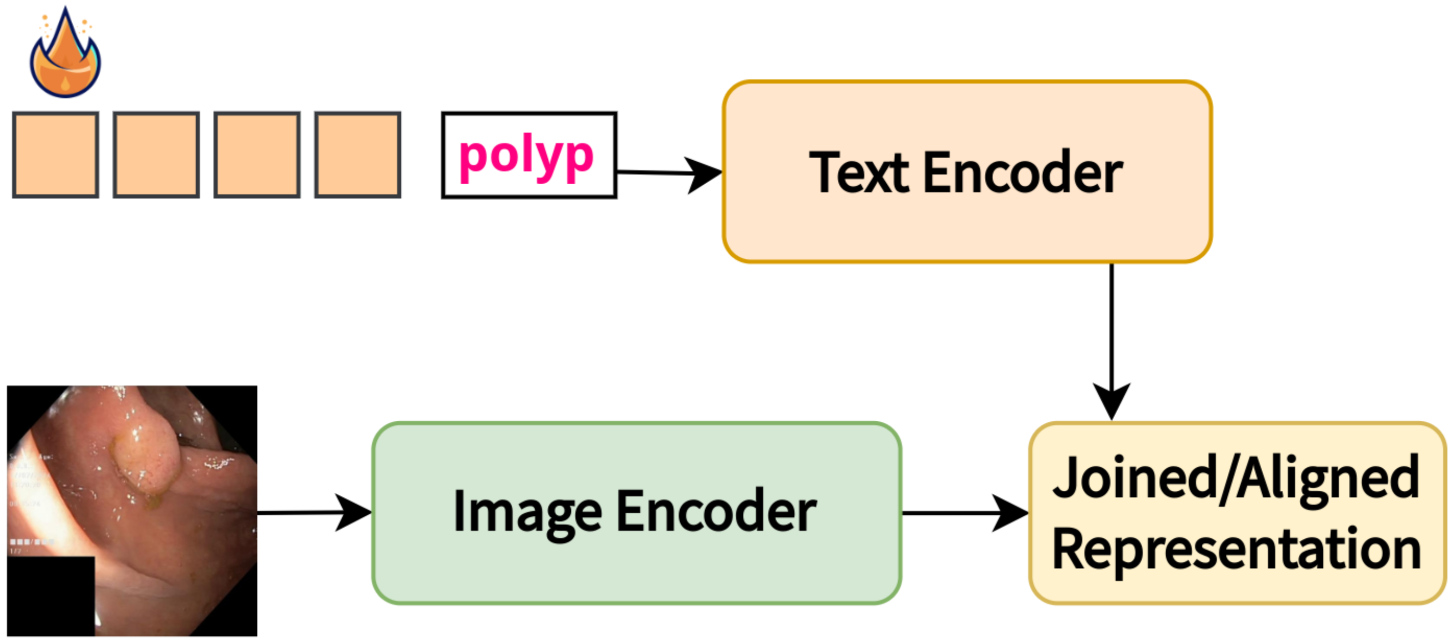
Prompt Tuning

Introduce learnable context vectors instead of text prompts



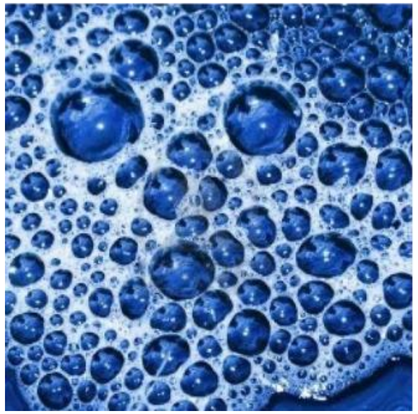
Prompt Tuning

- Adapts VLMs to new datasets by updating only the context vectors
- Automatically *learns* prompts for downstream tasks



Prompt Tuning in VLMs gives excellent performance

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**63.58**

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**83.53**

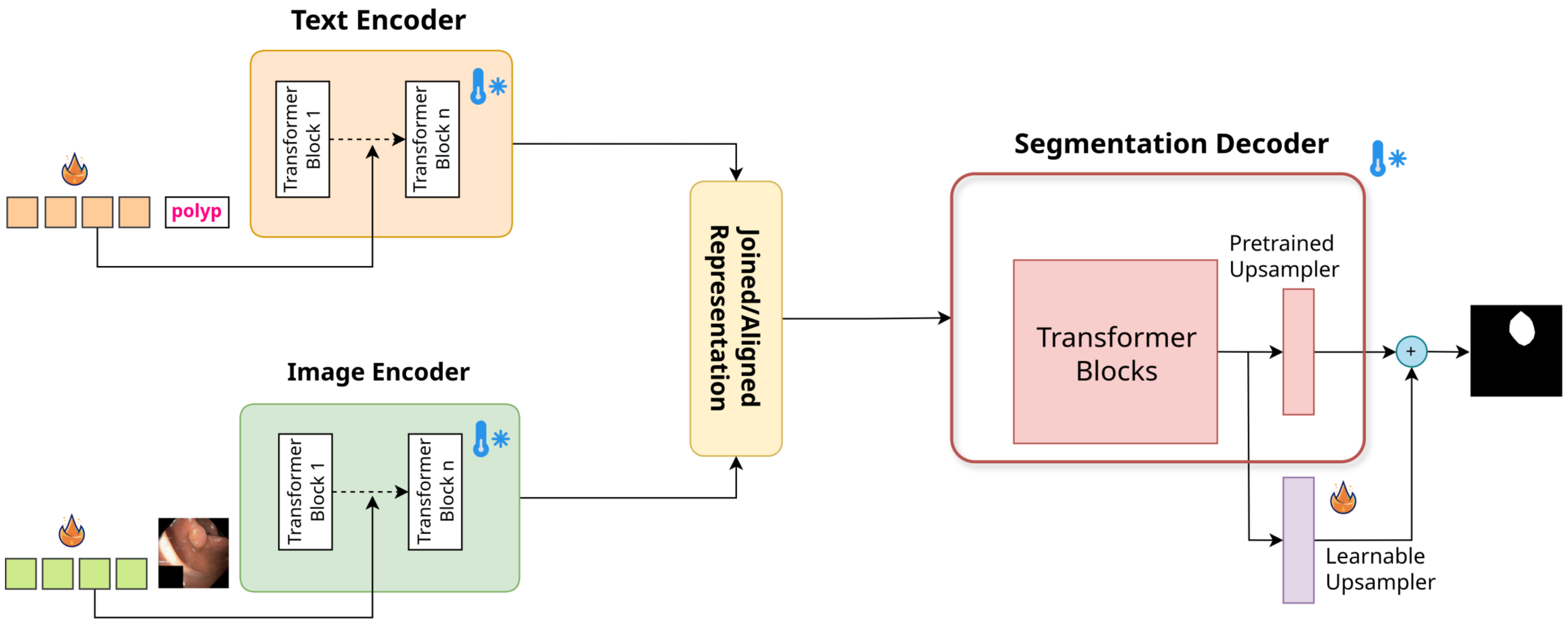
Significant performance improvement

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337-2348.

Outline

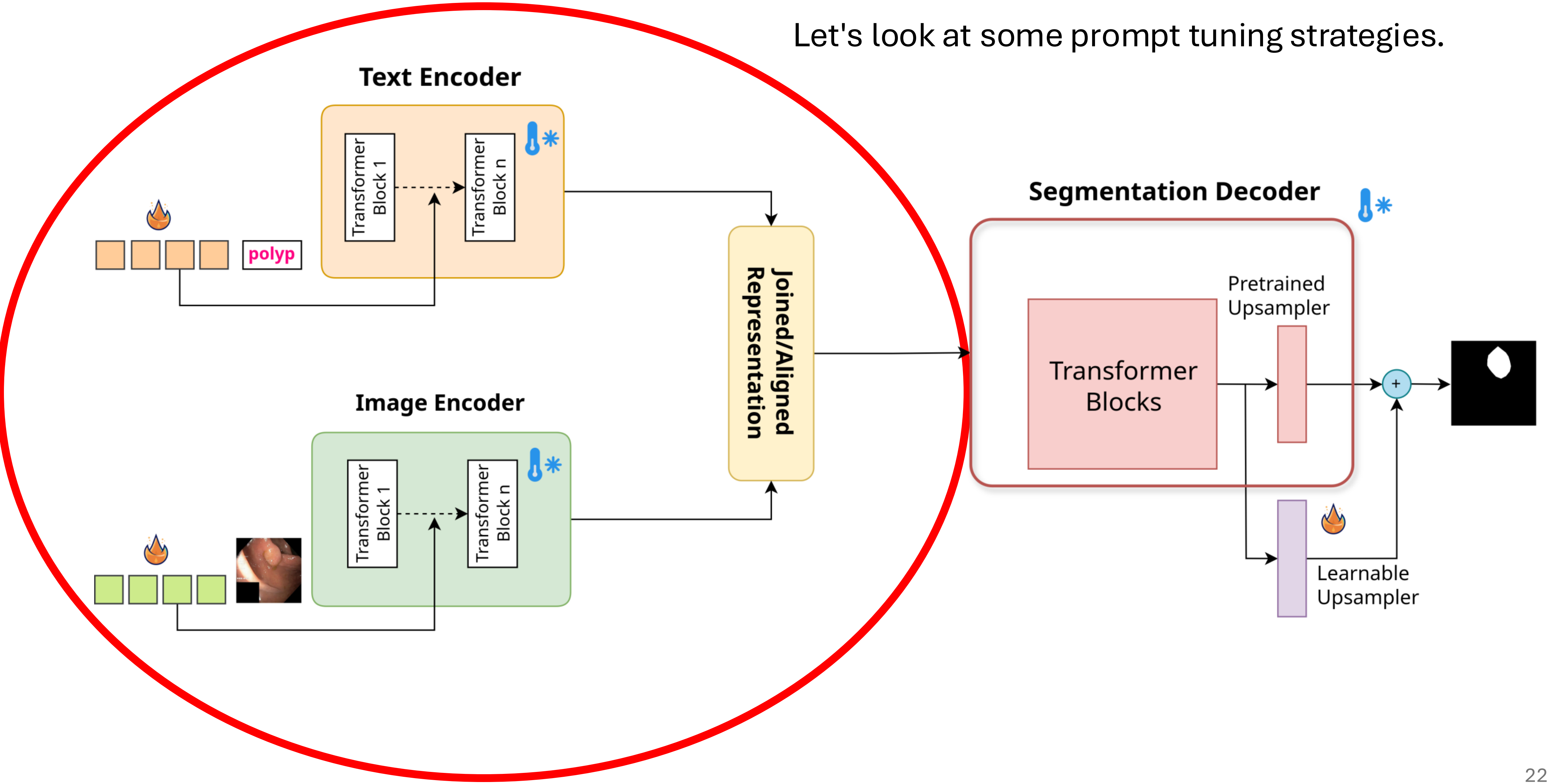
- Vision Language Models (VLMs) and Segmentation models (VLSMs)
- Adapting foundational VLMs and VLSMs
- **Prompt Tuning**
- TuneVLSeg Benchmark Framework
- Key Results

A closer look at Prompt Tuning in VLSMs



A closer look at Prompt Tuning in VLSMs

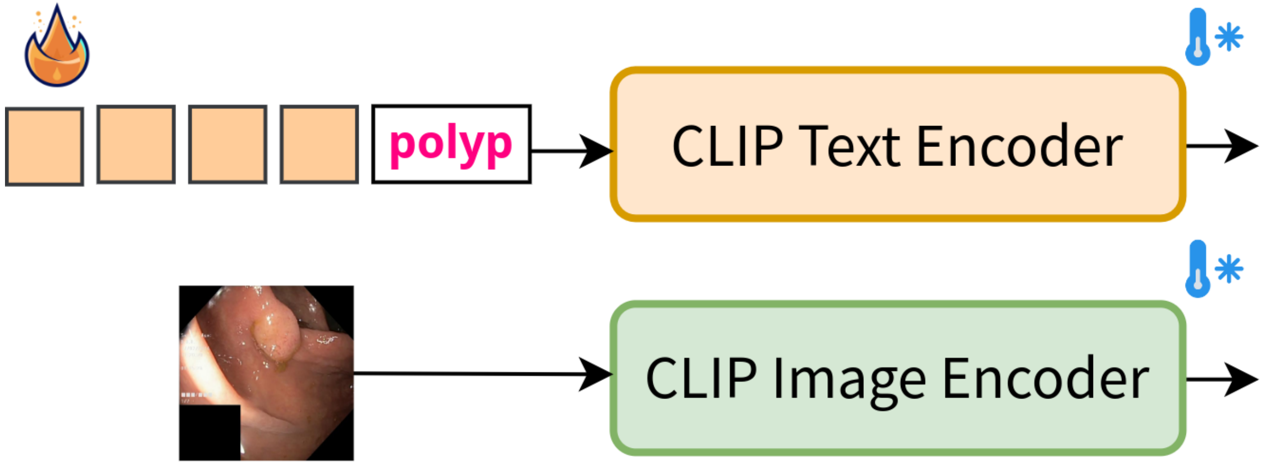
Let's look at some prompt tuning strategies.



Prompt Tuning Strategies: Unimodal

Introducing the context vectors at text branch

One set of vectors for the whole dataset or class



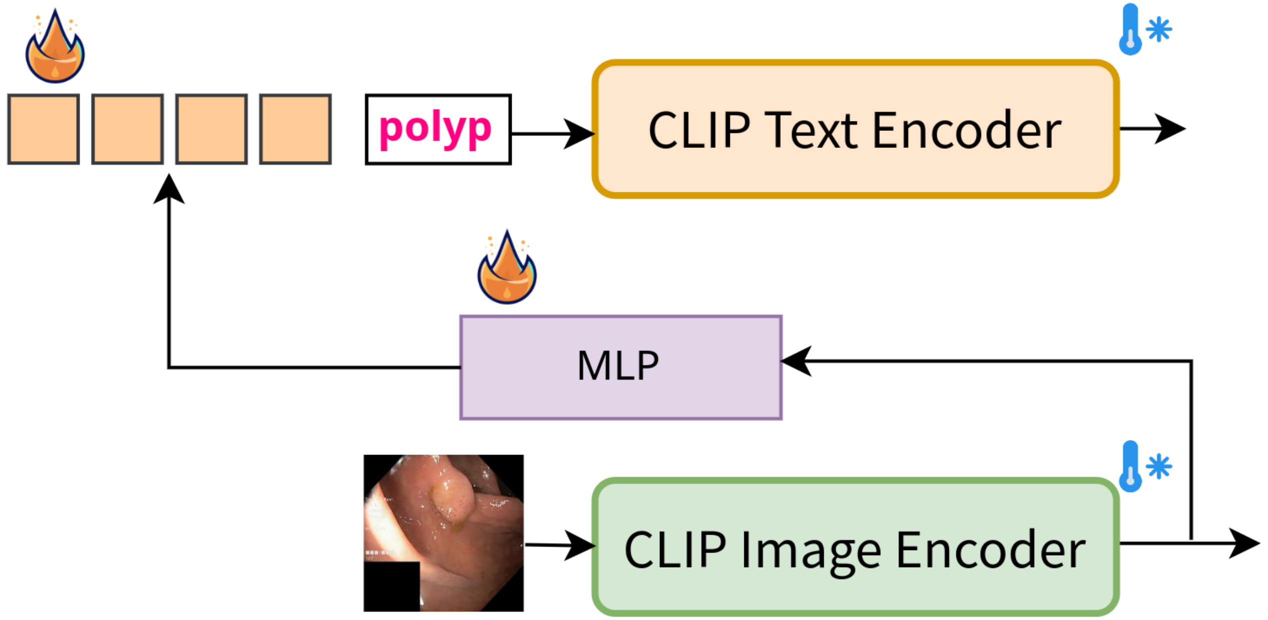
Context Optimization (CoOp)

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337-2348.

Prompt Tuning Strategies: Unimodal

Image instance conditions the text context vectors

Different prompt vectors for each instance



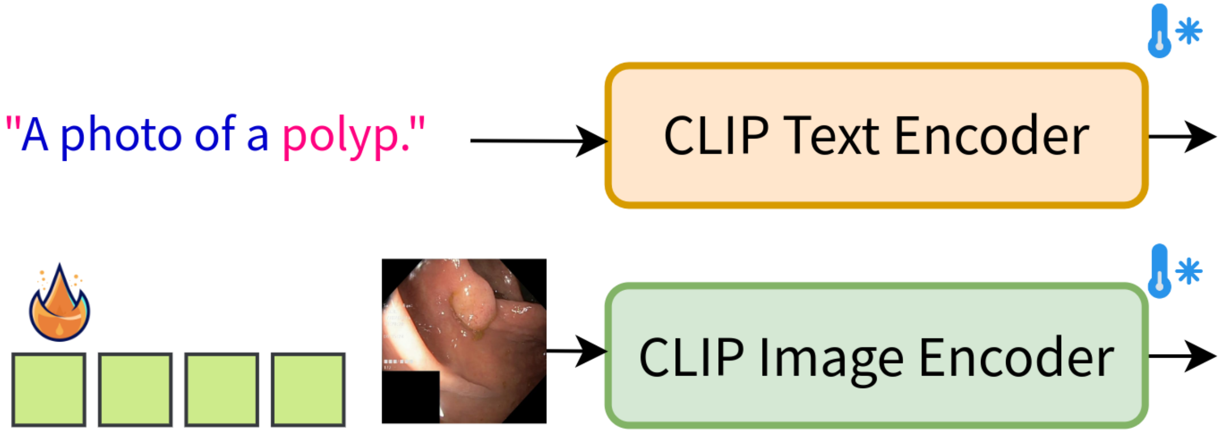
Conditional Context Optimization (CoCoOp)

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816-16825).

Prompt Tuning Strategies: Unimodal

Introducing the context vectors at vision branch

Works for transformer models.

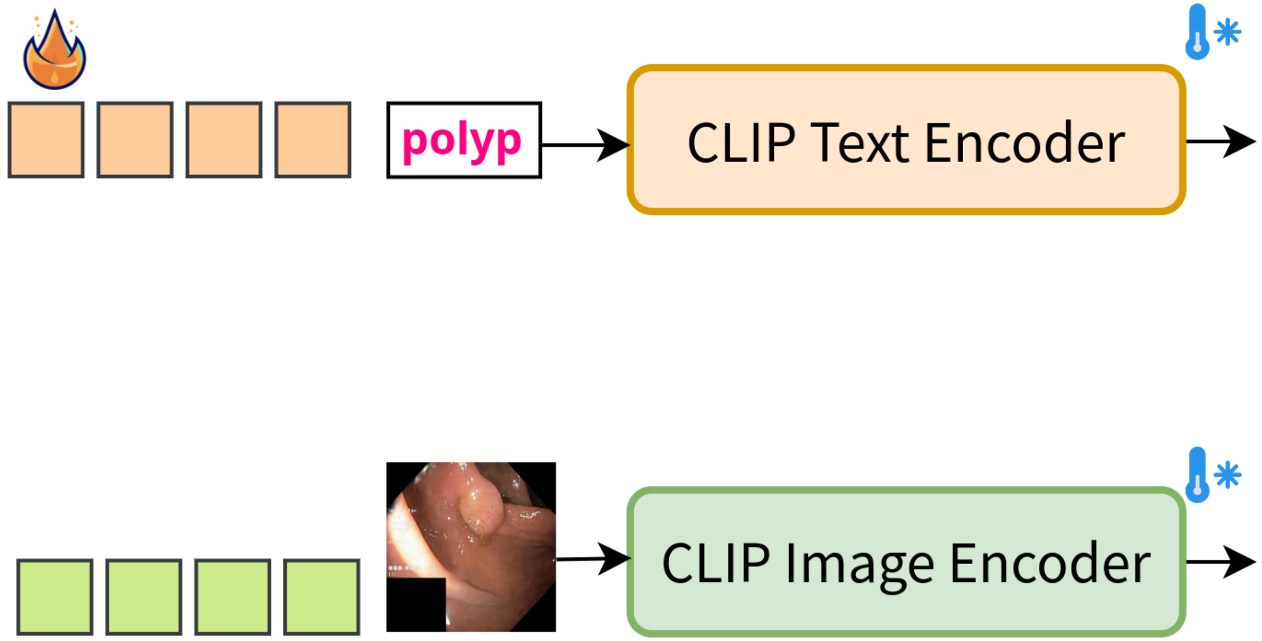


Visual Prompt Tuning (VPT)

Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022, October). Visual prompt tuning. In *European Conference on Computer Vision* (pp. 709-727). Cham: Springer Nature Switzerland.

Prompt Tuning Strategies: Multimodal

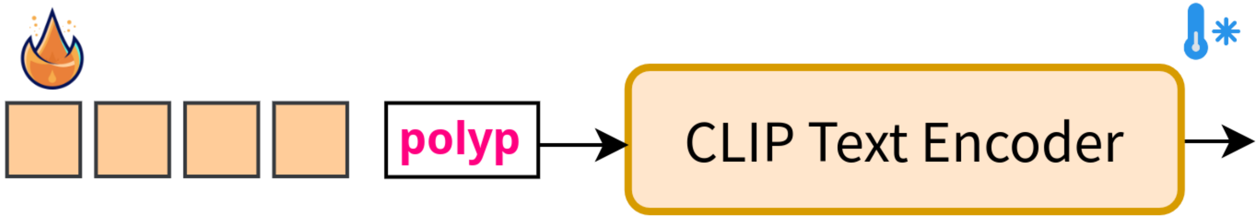
Introducing the context vectors at both at text and vision branch



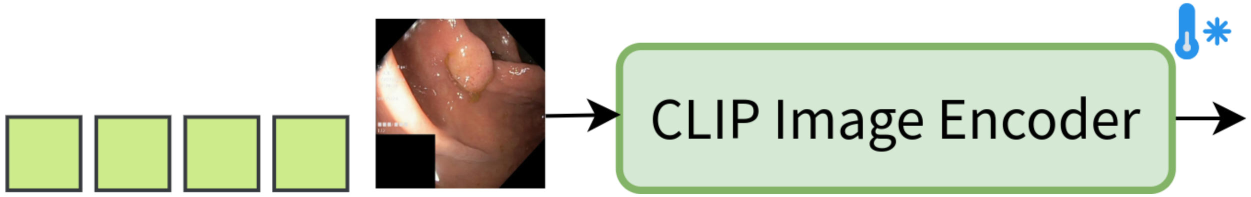
Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19113-19122).

Prompt Tuning Strategies: Multimodal

Introducing the context vectors at both at text and vision branch



No interaction between text and image ----> Suboptimal performance

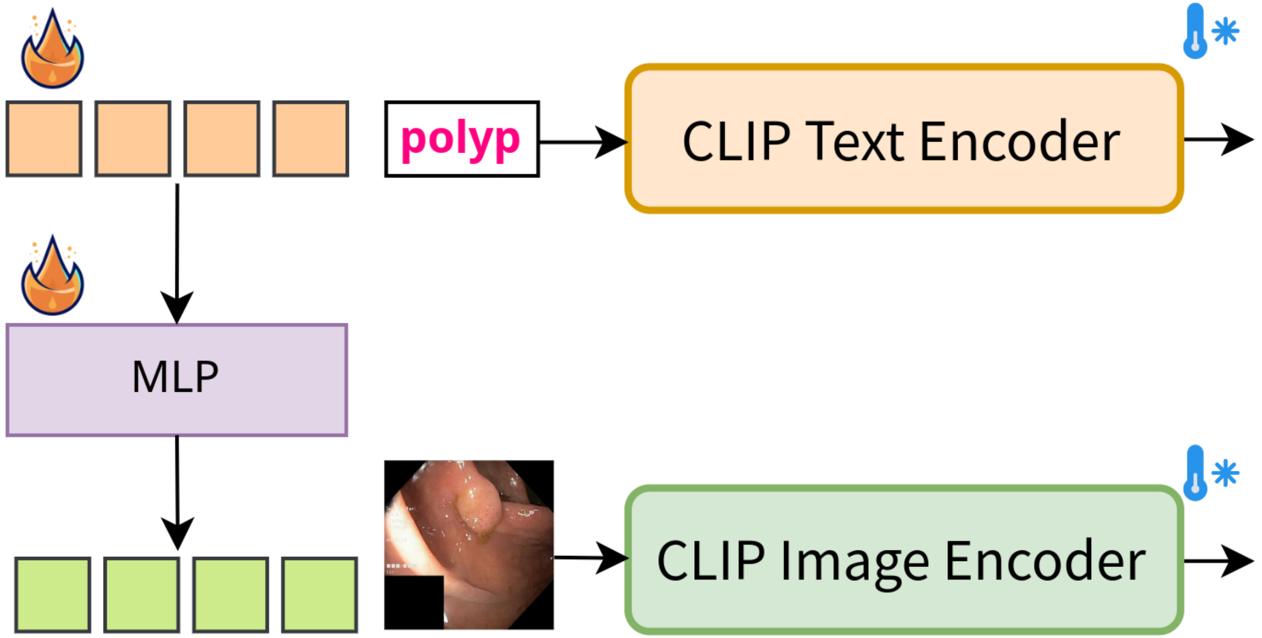


Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19113-19122).

Prompt Tuning Strategies: Multimodal

Introducing the context vectors at both at text and vision branch

Prompts are initialized in text embedding space



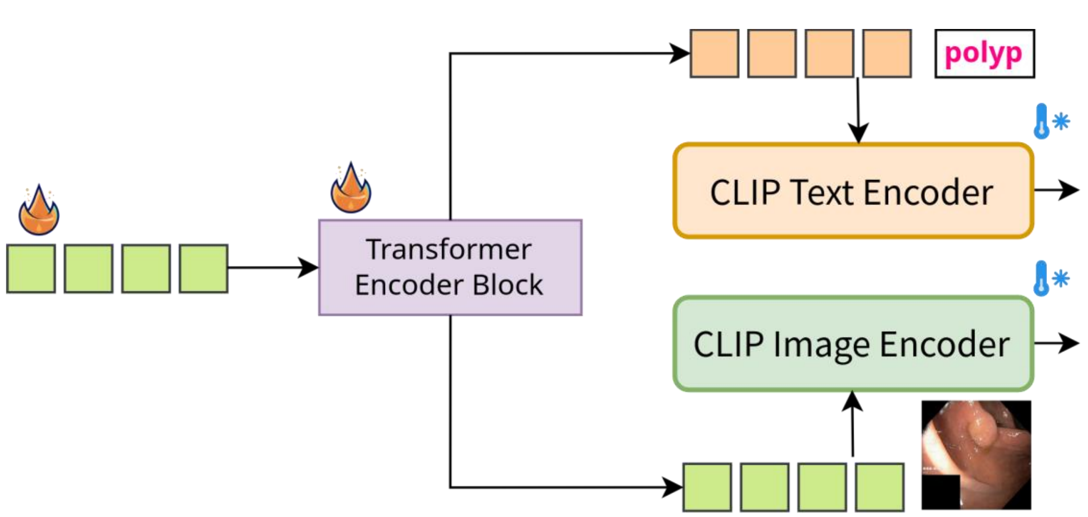
MaPLe

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19113-19122).

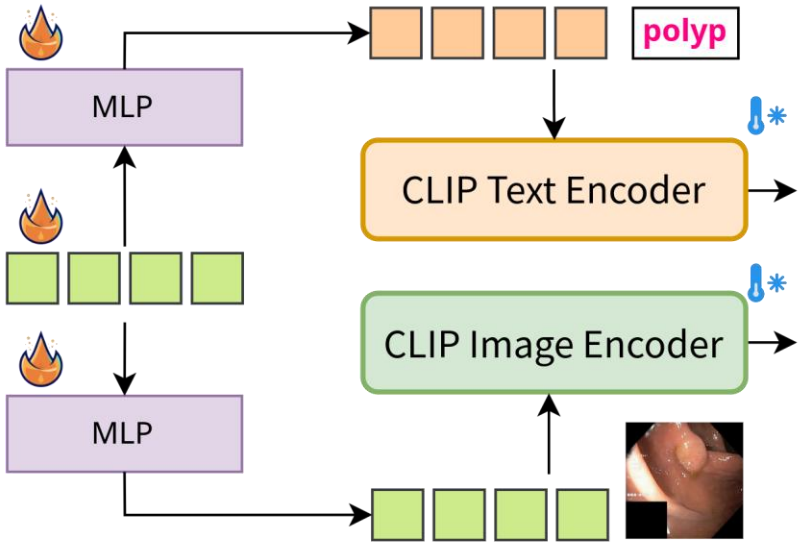
Prompt Tuning Strategies: Multimodal

Introducing the context vectors at both at text and vision branch

Prompts are initialized in shared embedding space



Shared Attention

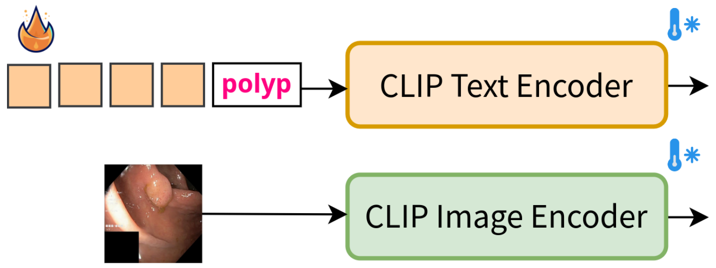


Shared Separate

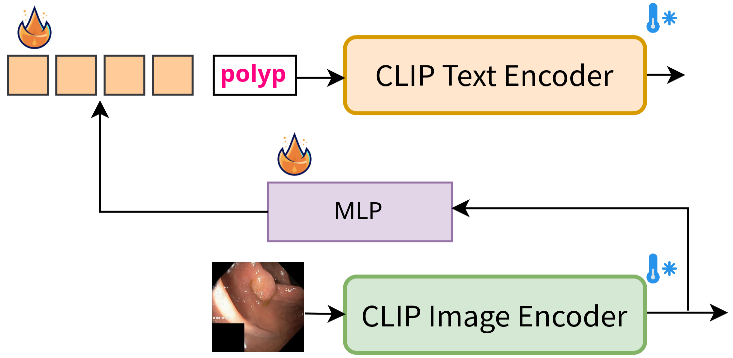
Unified Prompts

Zang, Y., Li, W., Zhou, K., Huang, C., & Loy, C. C. (2022). Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225.

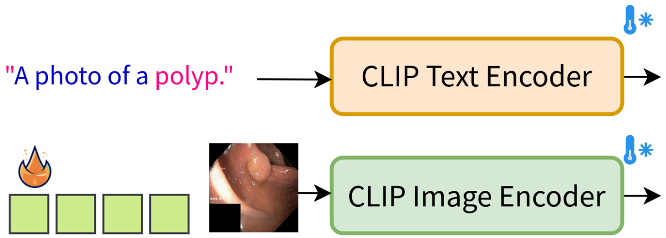
Prompt Tuning Strategies: Overview



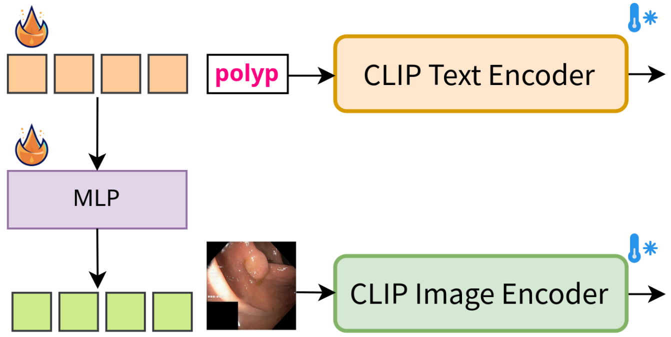
CoOp



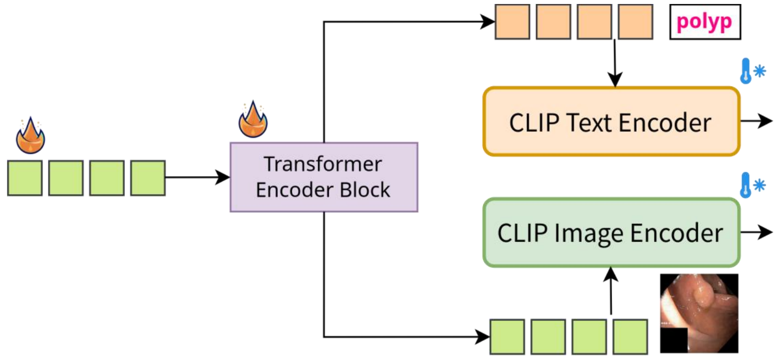
CoCoOp



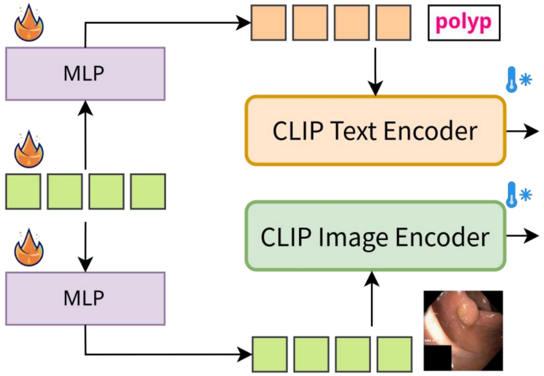
VPT



MaPLe

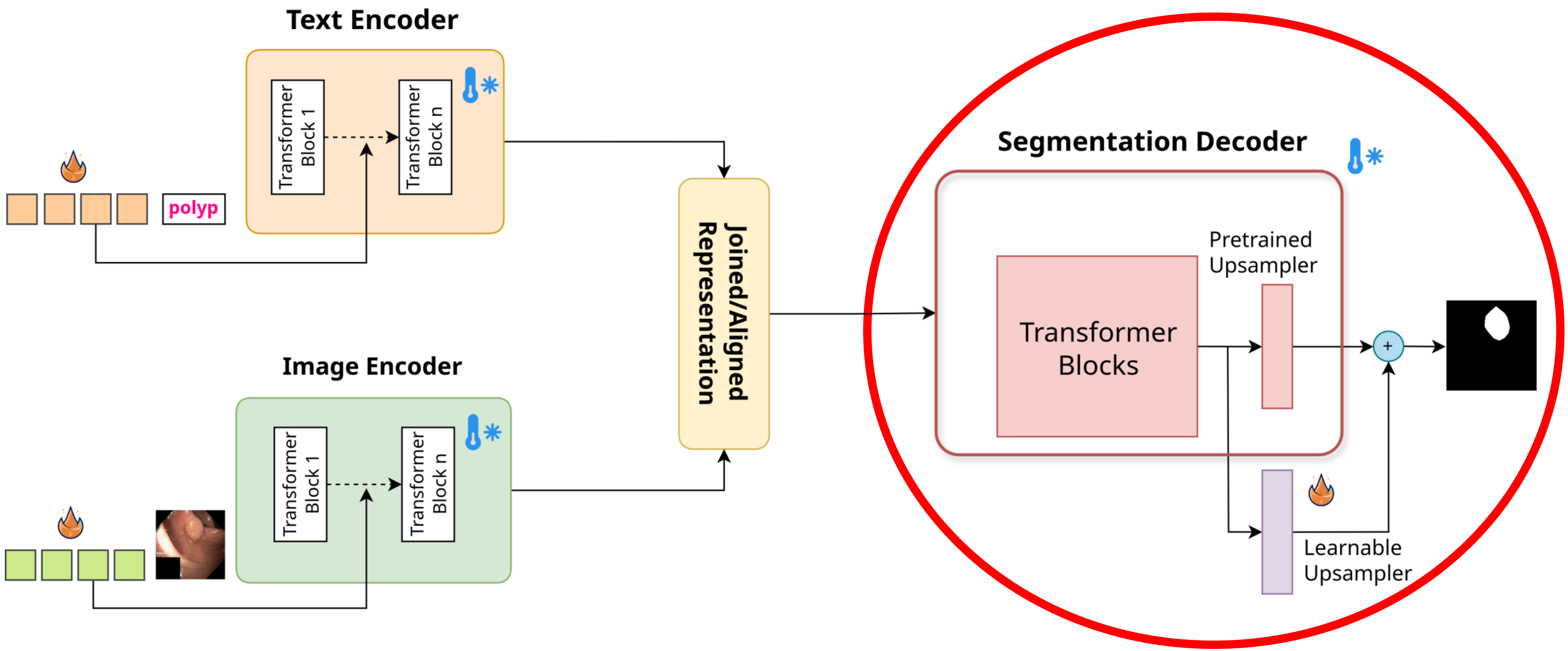


Shared Attention

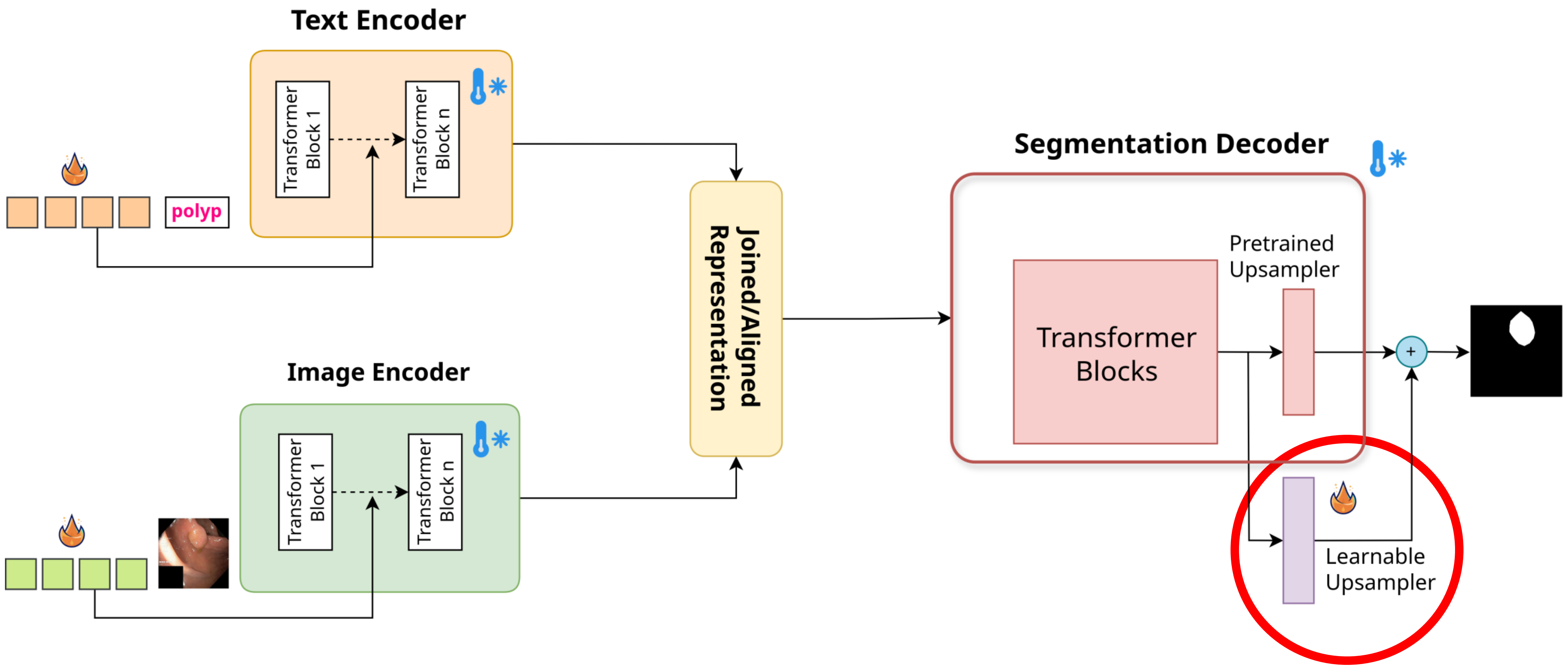


Shared Separate

A closer look at Prompt Tuning in VLSMs



A closer look at Prompt Tuning in VLSMs



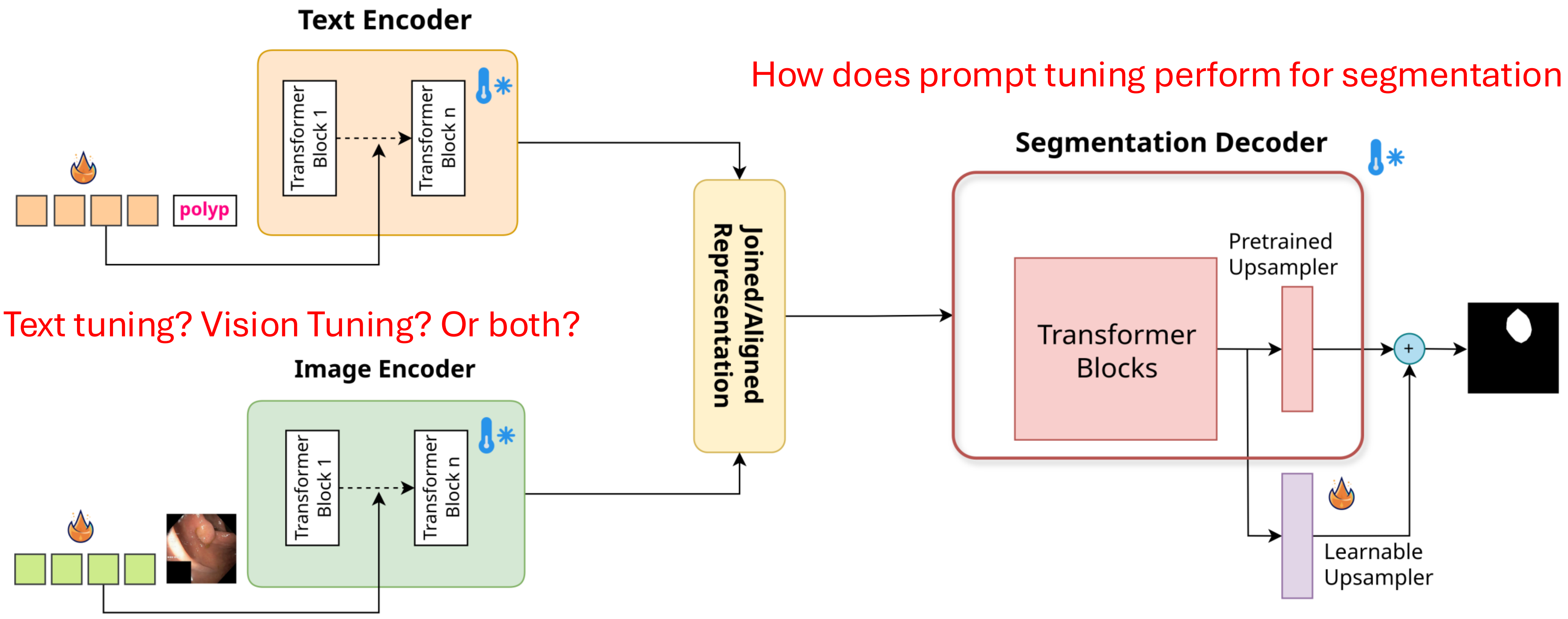
We added this to see if it makes a difference in segmentation performance.

This is inspired by VPT, which shows good performance when final layer is trained.

A closer look at Prompt Tuning in VLSMs

How does prompt tuning perform for segmentation tasks?

Text tuning? Vision Tuning? Or both?



What should be the prompt depth?

What if the dataset is completely different from pretraining dataset?

Outline

- Vision Language Models (VLMs) and Segmentation models (VLSMs)
- Adapting foundational VLMs and VLSMs
- Prompt Tuning
- **TuneVLSeg Benchmark Framework**
- Key Results

TuneVLSeg Benchmarking Framework

Prompt Tuning Strategies

Text Tuning: **CoOp, CoCoOp**
Visual Tuning: **VPT**
Multimodal Prompt Tuning: **MaPle, Shared Attention, Shared Separate**

TuneVLSeg Benchmarking Framework

Prompt Tuning Strategies

Text Tuning: **CoOp, CoCoOp**
Visual Tuning: **VPT**
Multimodal Prompt Tuning: **MaPle, Shared Attention, Shared Separate**

Key Questions

- Performance of different prompt tuning strategies in segmentation
- Effects of adding context vectors at multiple depths for text and image encoders?
- Is multimodal prompt tuning better than unimodal?
- Natural images vs medical images

TuneVLSeg Benchmarking Framework

Prompt Tuning Strategies

Text Tuning: **CoOp, CoCoOp**
Visual Tuning: **VPT**
Multimodal Prompt Tuning: **MaPle, Shared Attention, Shared Separate**

Key Questions

- Performance of different prompt tuning strategies in segmentation
- Effects of adding context vectors at multiple depths for text and image encoders?
- Is multimodal prompt tuning better than unimodal?
- Natural images vs medical images

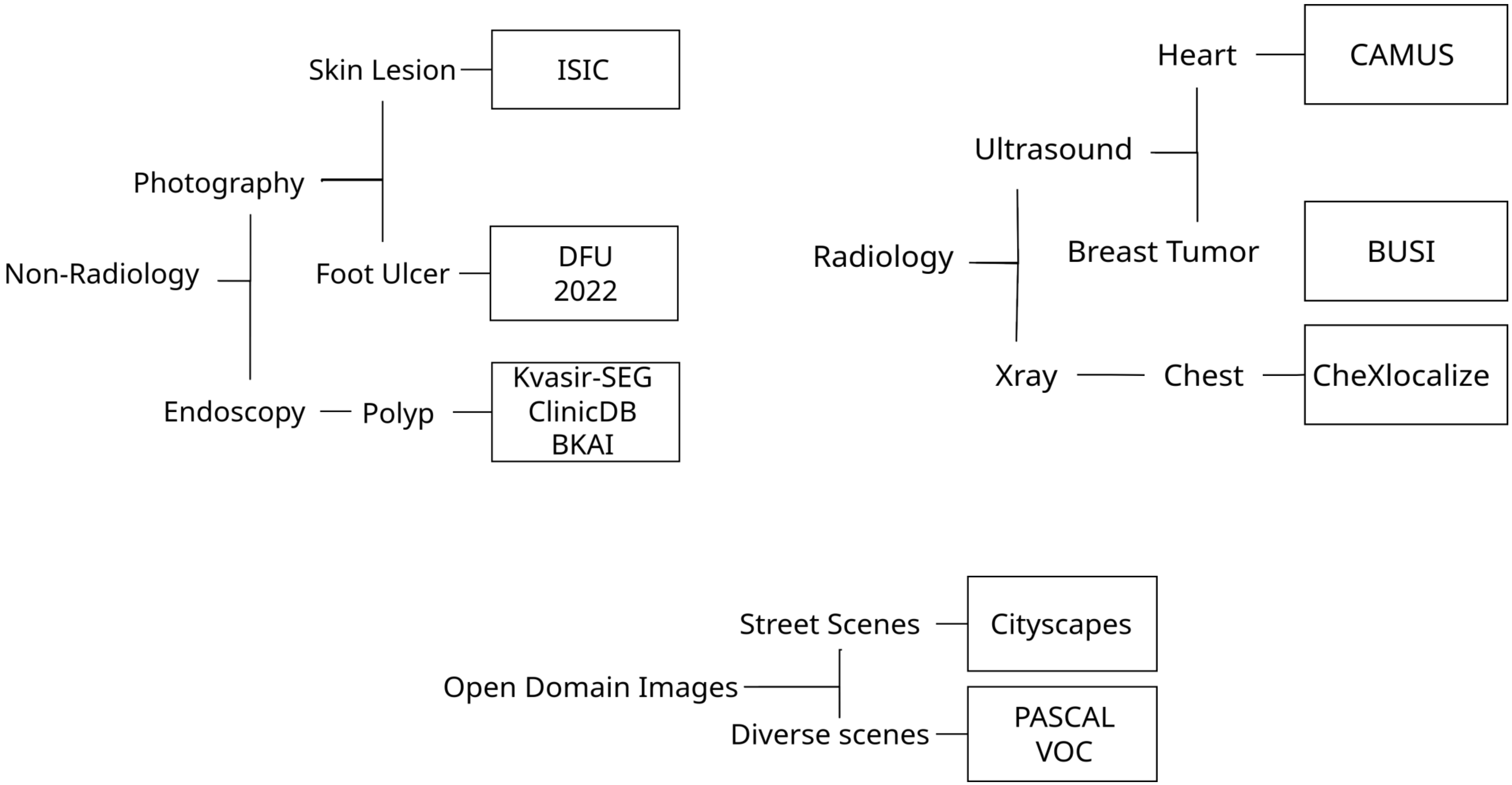
Models

- 2 class-agnostic VLSMs: **CLIPSeg, CRIS**

Datasets

- 8 medical datasets: 3 radiology, 5 non-radiology
- 2 open domain datasets

Datasets



Skin Lesion
ISIC
Photography
Non-Radiology
Foot Ulcer
DFU 2022
Endoscopy
Polyp
Kvasir-SEG
ClinicDB
BKAI
Heart
CAMUS
Ultrasound
Radiology
Breast Tumor
BUSI
Xray
Chest
CheXlocalize
Street Scenes
Cityscapes
Open Domain Images
Diverse scenes
PASCAL VOC

Experimental setup

Hyperparameter	Search Space	Applicable for	Space Type
Learning rate	$[10^{-5}, 5x10^{-3}]$	ALL	Log
Weight decay	$[10^{-5}, 0.01]$	ALL	Log
Prompt depth	[1, 11]	ALL	Integer
Intermediate dimension	32, 64, 96, 128	CoCoOp, Maple	Choice
Transformer: Number of Heads	16, 20, 32	Shared Attention	Choice
Transformer: Dropout Probability	[0.1, 0.55]	Shared Attention	Linear
Transformer: Feed-Forward Dim	1280, 1420	Shared Attention	Choice
Transformer: LayerNorm First	true, false	Shared Attention	Choice
Shared Space Dimension	32, 64	Shared Separate	Choice

Experimental setup

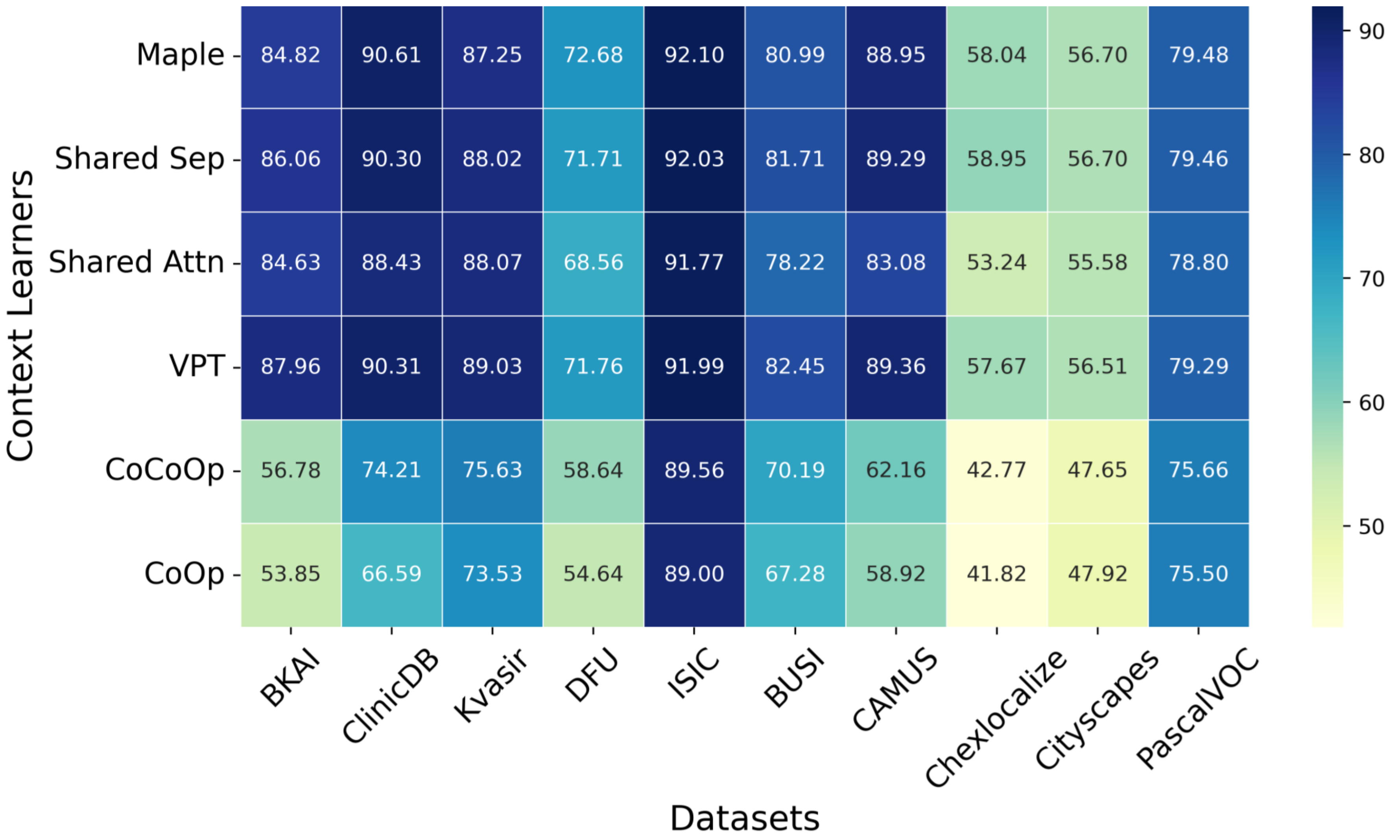
Hyperparameter	Search Space	Applicable for	Space Type
Learning rate	$[10^{-5}, 5x10^{-3}]$	ALL	Log
Weight decay	$[10^{-5}, 0.01]$	ALL	Log
Prompt depth	[1, 11]	ALL	Integer
Intermediate dimension	32, 64, 96, 128	CoCoOp, Maple	Choice
Transformer: Number of Heads	16, 20, 32	Shared Attention	Choice
Transformer: Dropout Probability	[0.1, 0.55]	Shared Attention	Linear
Transformer: Feed-Forward Dim	1280, 1420	Shared Attention	Choice
Transformer: LayerNorm First	true, false	Shared Attention	Choice
Shared Space Dimension	32, 64	Shared Separate	Choice

We ran each experiment 20 times with the search space for each parameter

Outline

- Vision Language Models (VLMs) and Segmentation models (VLSMs)
- Adapting foundational VLMs and VLSMs
- Prompt Tuning
- TuneVLSeg Benchmark Framework
- **Key Results**

Choice of Prompt Tuning Techniques

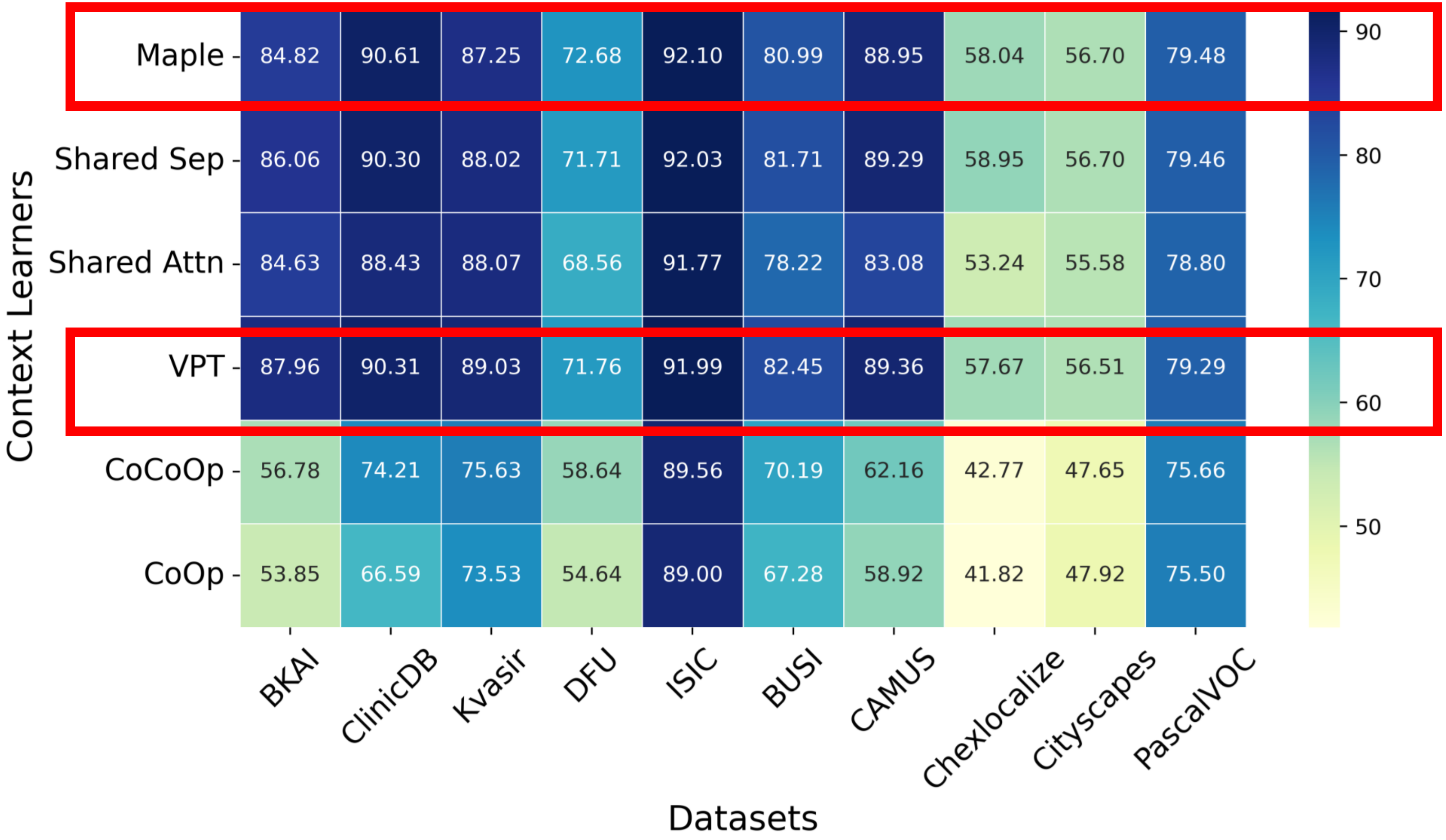


Choice of Prompt Tuning Techniques

Context Learners \ Datasets	BKAI	ClinicDB	Kvasir	DFU	ISIC	BUSI	CAMUS	Chexlocalize	Cityscapes	PascalVOC
Maple	84.82	90.61	87.25	72.68	92.10	80.99	88.95	58.04	56.70	79.48
Shared Sep	86.06	90.30	88.02	71.71	92.03	81.71	89.29	58.95	56.70	79.46
Shared Attn	84.63	88.43	88.07	68.56	91.77	78.22	83.08	53.24	55.58	78.80
VPT	87.96	90.31	89.03	71.76	91.99	82.45	89.36	57.67	56.51	79.29
CoCoOp	56.78	74.21	75.63	58.64	89.56	70.19	62.16	42.77	47.65	75.66
CoOp	53.85	66.59	73.53	54.64	89.00	67.28	58.92	41.82	47.92	75.50

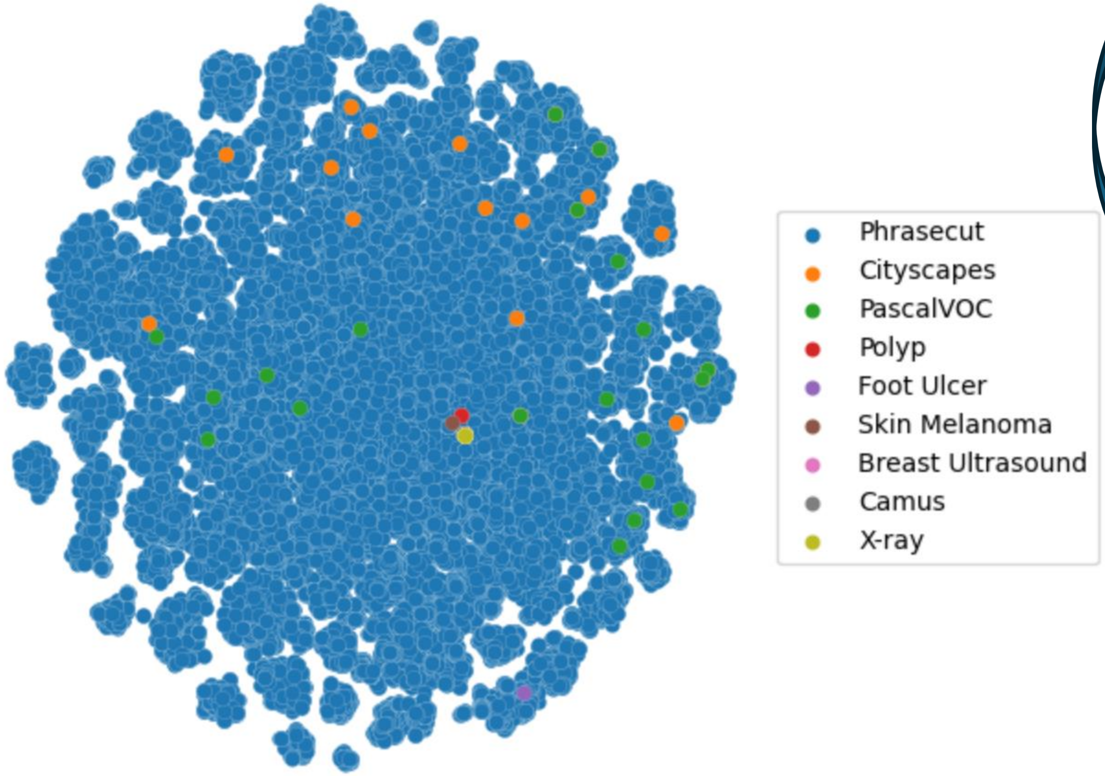
Text tuning does not perform well.

Choice of Prompt Tuning Techniques

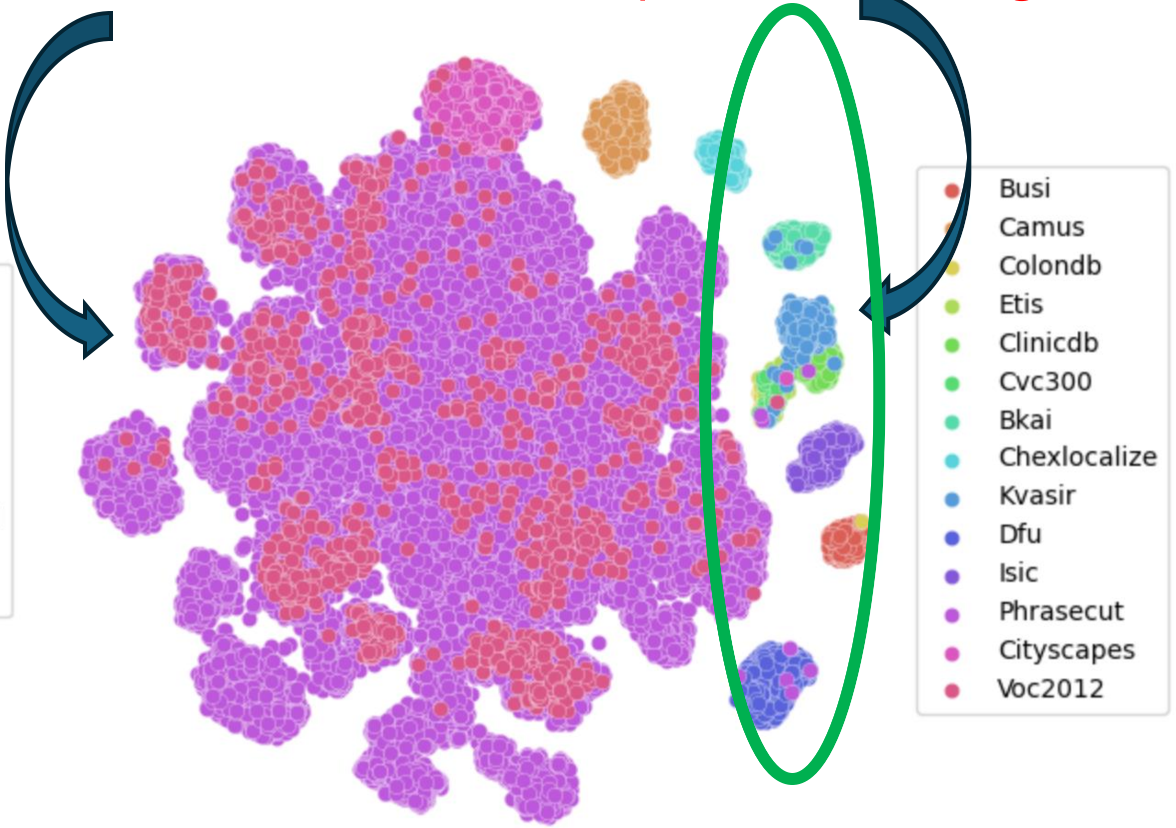


Is VPT's performance due to diversity of images and prompts in datasets ?

Separate clusters for medical and open domain images



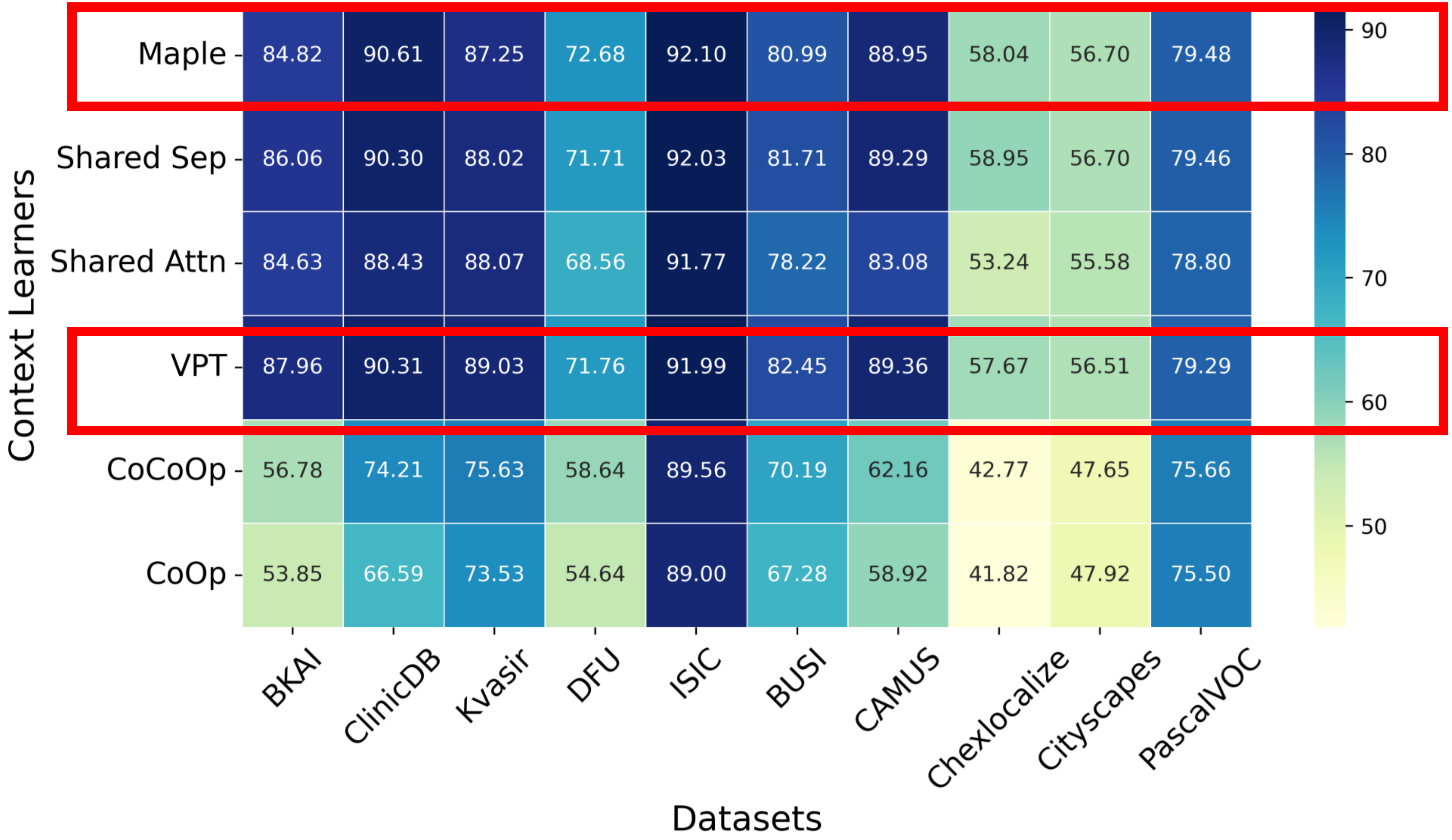
Phrases (Text Prompts)



Images

Significant distribution shift in images than prompts might be the reason for VPT's better performance.

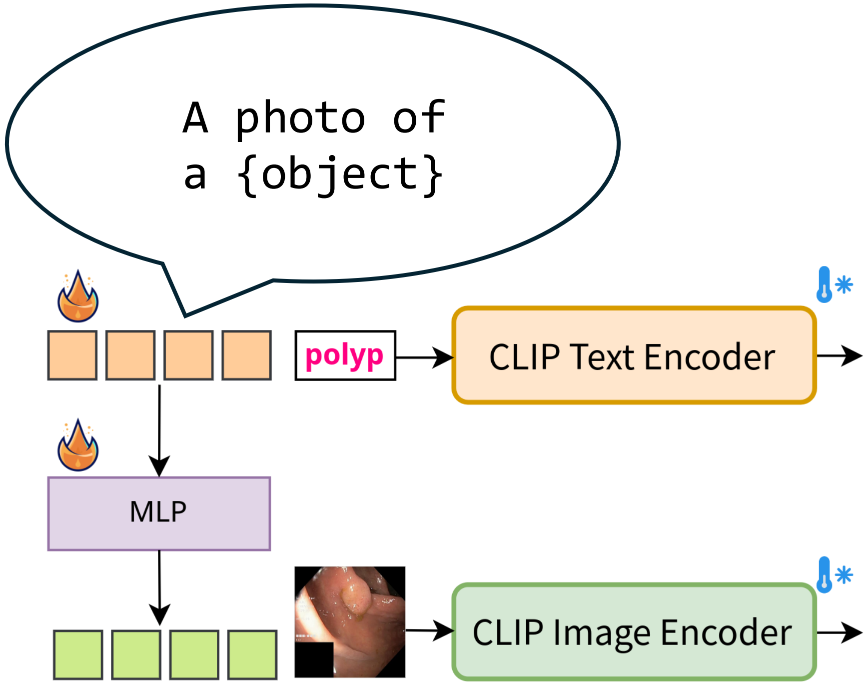
Choice of Prompt Tuning Techniques



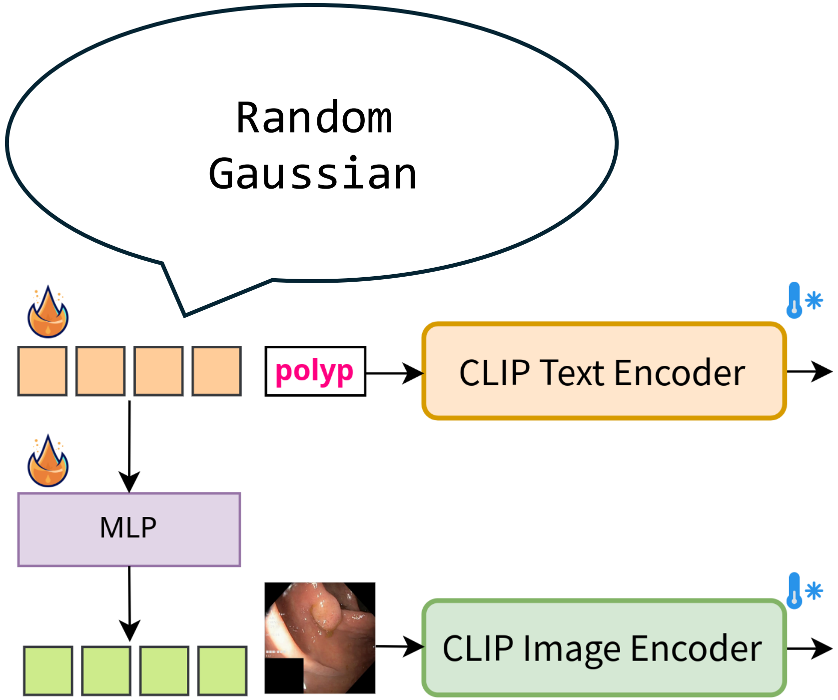
VPT has fewer hyperparameters to tune; smaller search space; can be a good starting choice for good results

Context Vector Initialization

The context vectors of Maple can either be heuristically initialized or randomly.

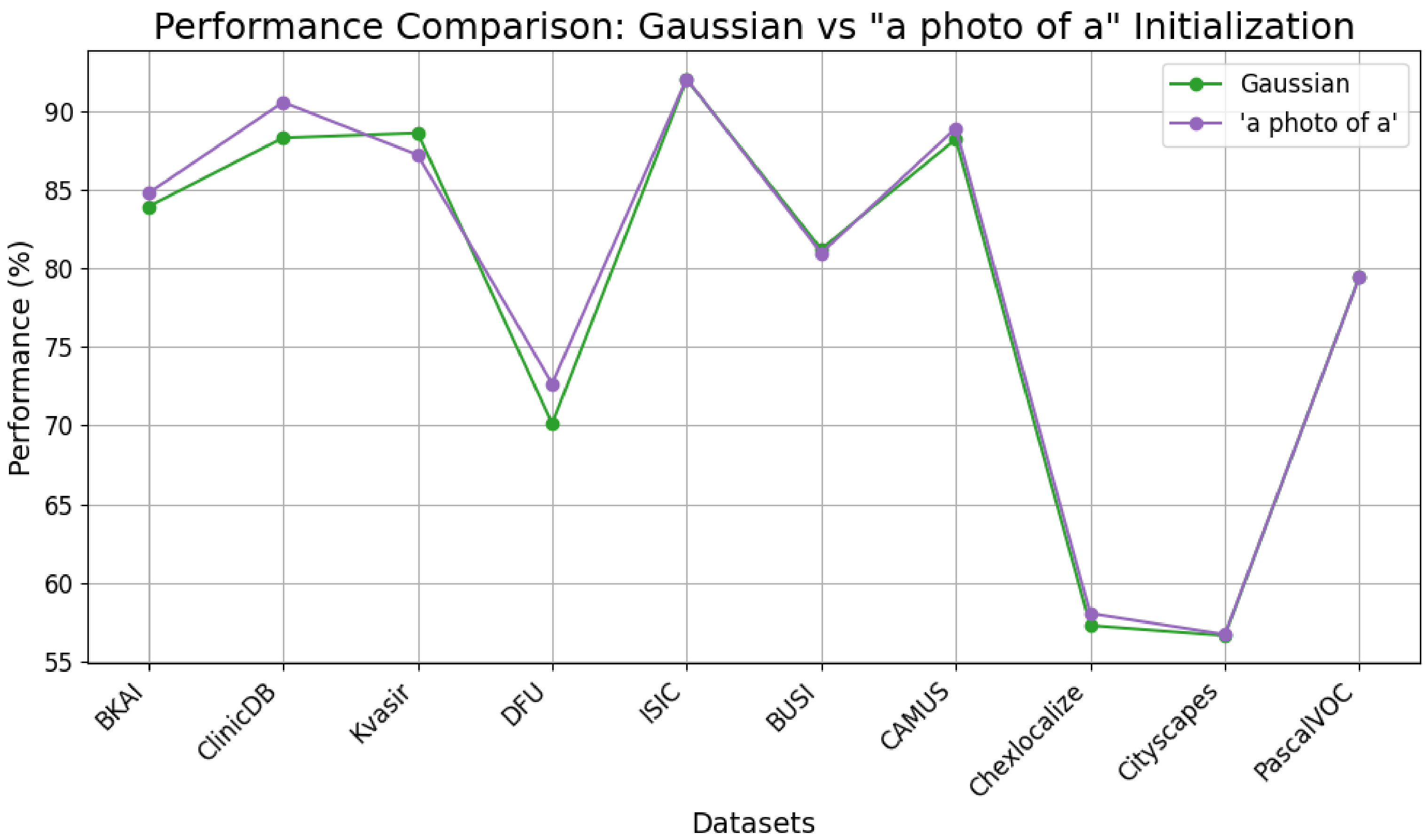


Maple with Heuristic Initialization



Maple with Random Initialization

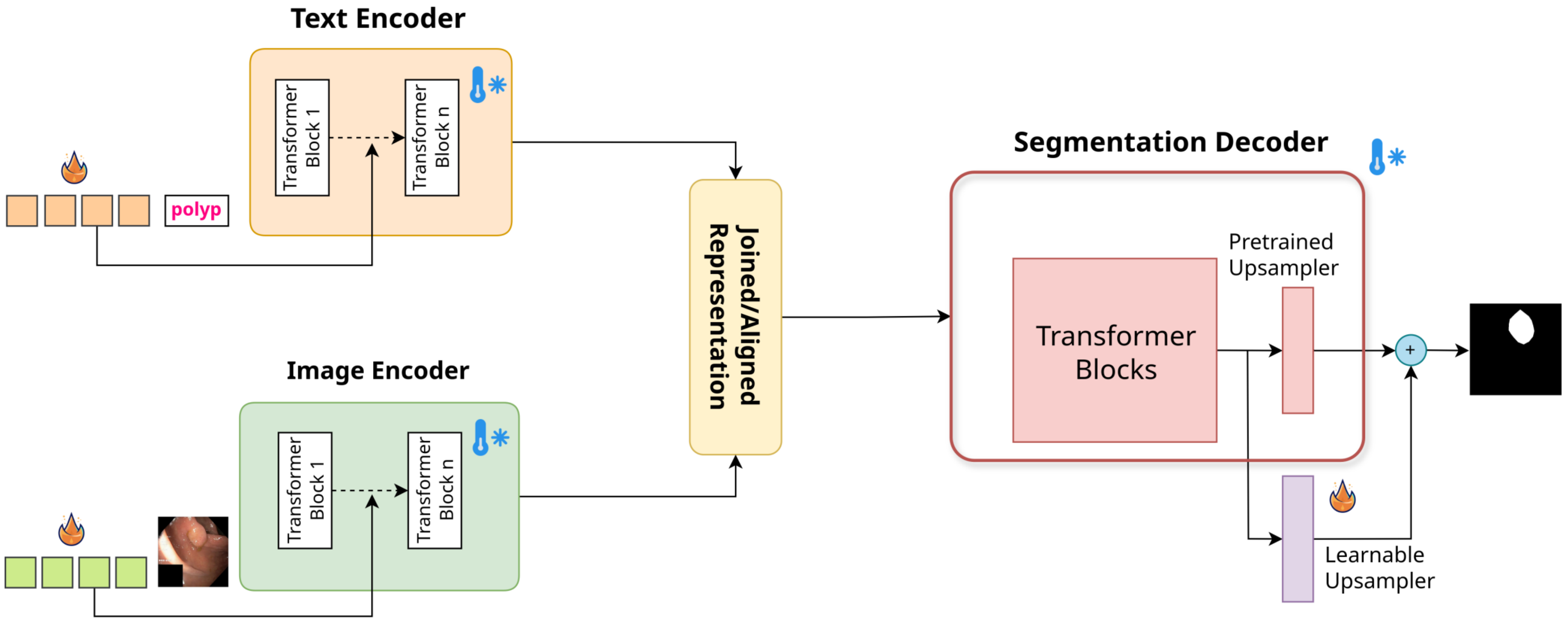
Context Vector Initialization



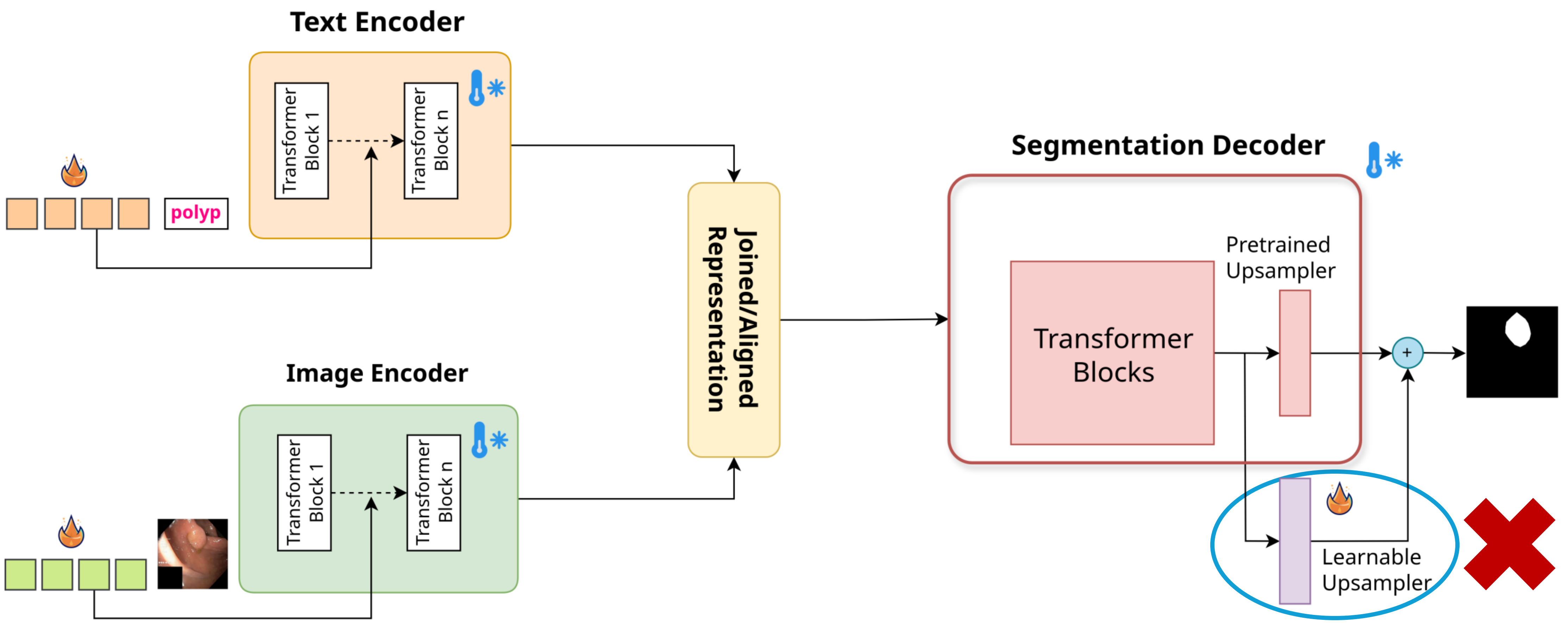
It *might* be a good idea to initialize the context vectors with embeddings of "a photo of a".

Might be because CLIP is trained on the prompt template "a photo of a <CLS>".

Is the performance of context learners due to learnable upsampler?

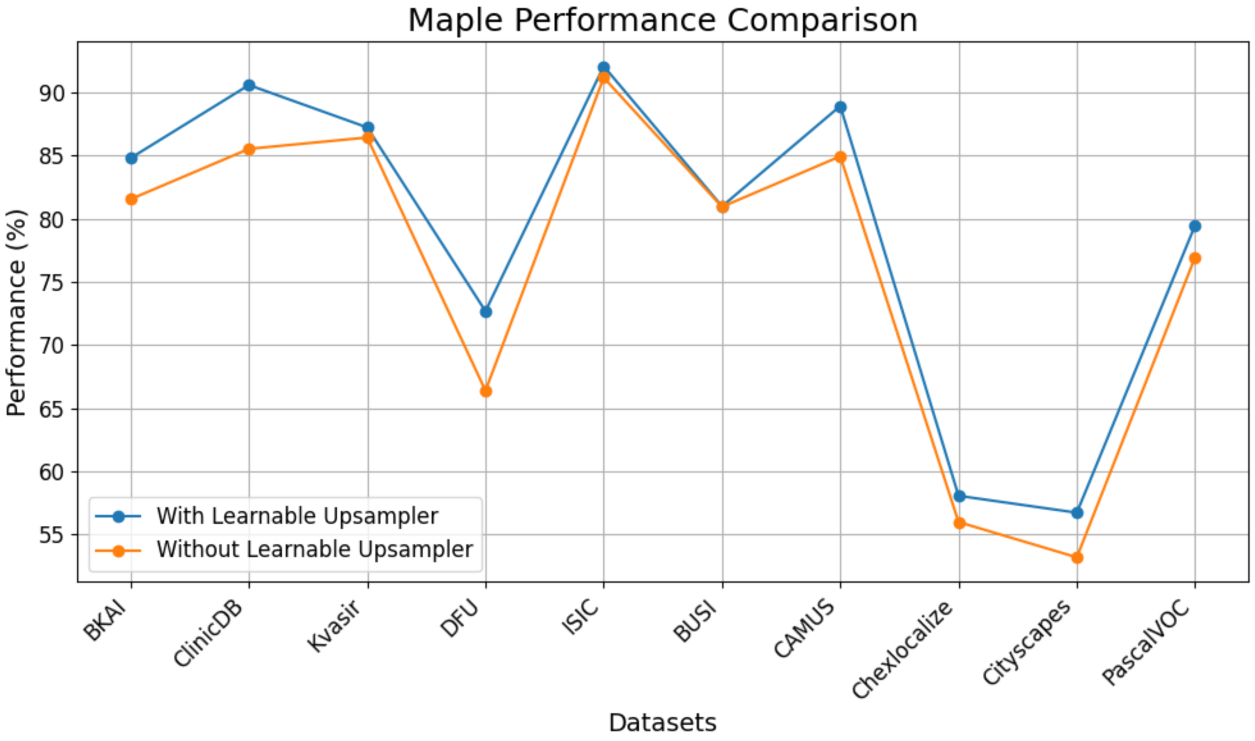


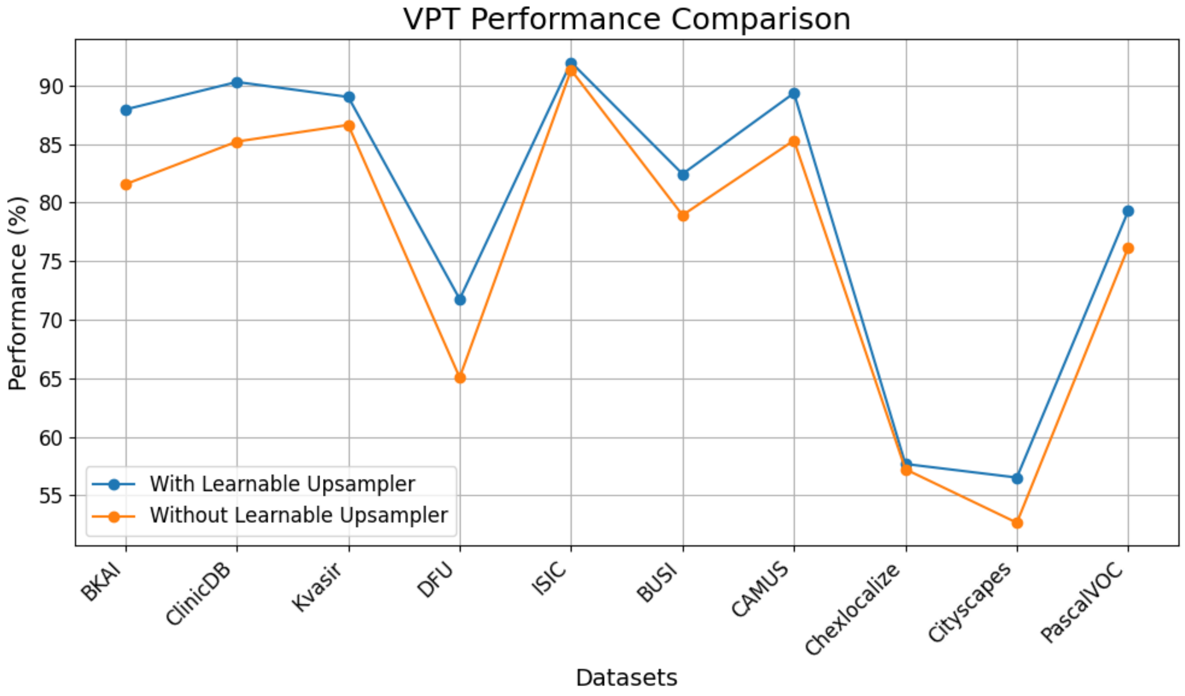
Is the performance of context learners due to learnable upsampler?



We trained the models by removing this block.

Is the performance of context learners due to learnable upsampler?

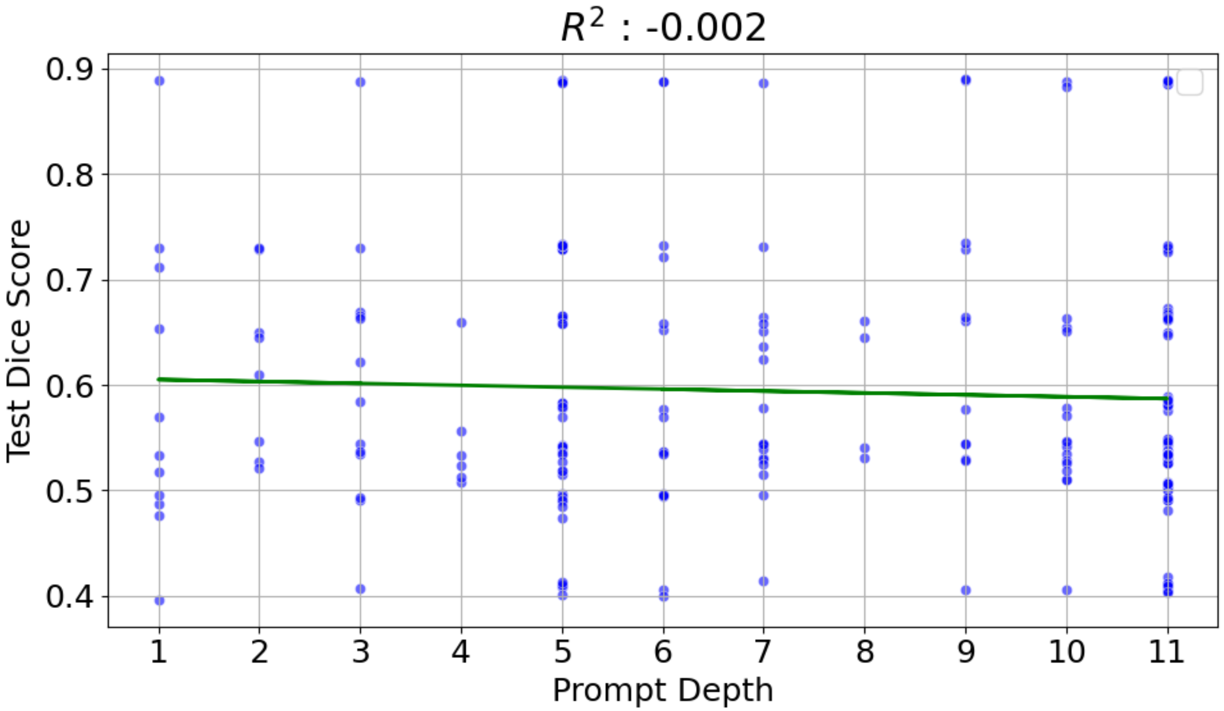




Using the learnable upsampler clearly has benefits.

What should the prompt depth be?

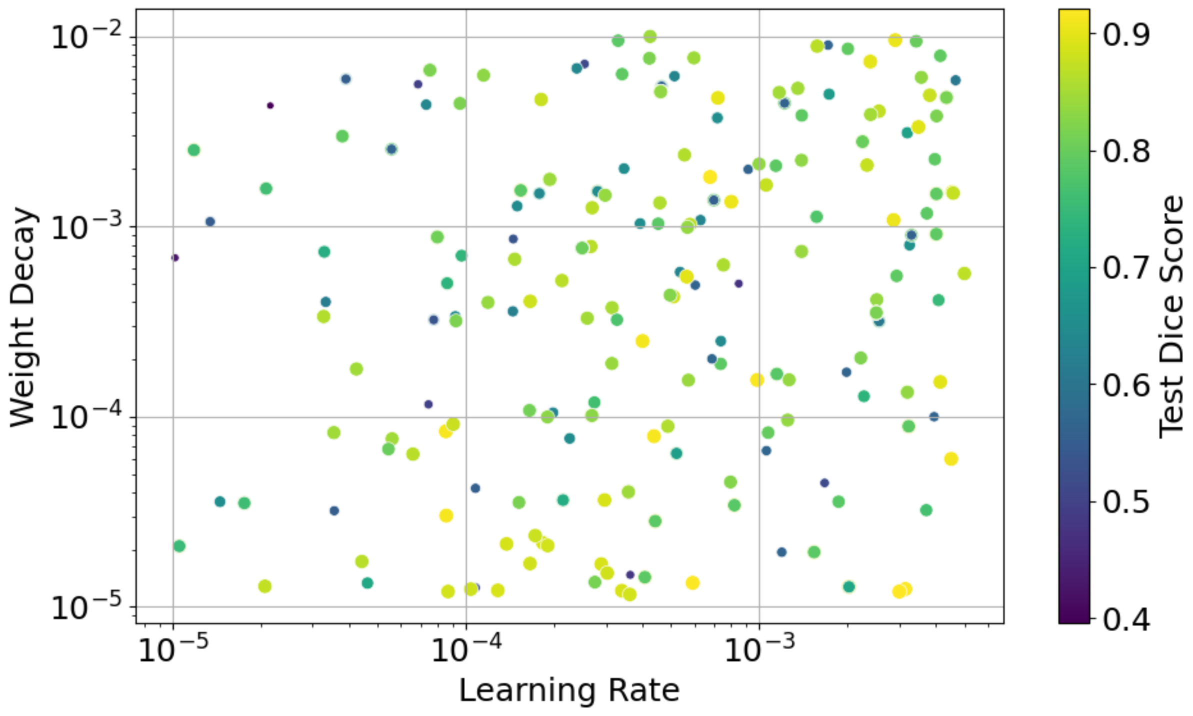
- There is no strong correlation between the prompt depth and dice score.
- Increasing prompt depth may not always increase the dice score.



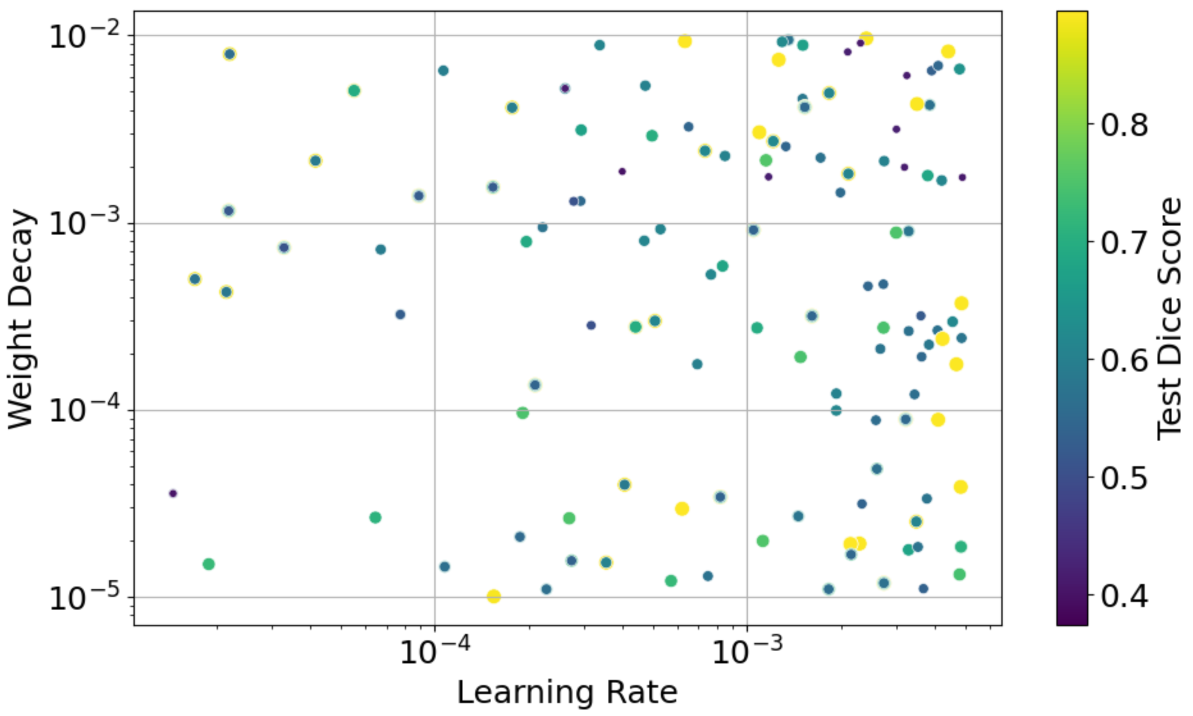
This is for text tuning methods.

Any specific choice for learning rate and weight decay?

- There is no strong correlation between the Learning Rate/Weight Decay and dice score.



Maple



CoCoOp

Wrapping up…

- We performed benchmark evaluation on:
  - 2 CLIP-based VLSMs
  - 8 medical segmentation datasets
  - 2 open domain datasets
  - 6 prompt tuning strategies
- Our framework can be extended to other VLSMs and prompt tuning methods.

Scan to read paper

Prompt tuning is an effective strategy to adapt VLSMs for domain-specific segmentation tasks.

But we need to consider the caveats that comes with tuning different parameters of these methods.