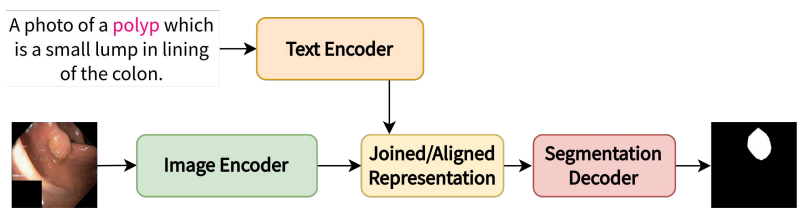
TuneVLSeg: Prompt Tuning Benchmark for Vision-Language Segmentation Models

Rabin Adhikari, Safal Thapaliya, Manish Dhakal, Bishesh Khanal
Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal

ACCV HANOI VIETNAM 2024 DEC 8-12

Motivation

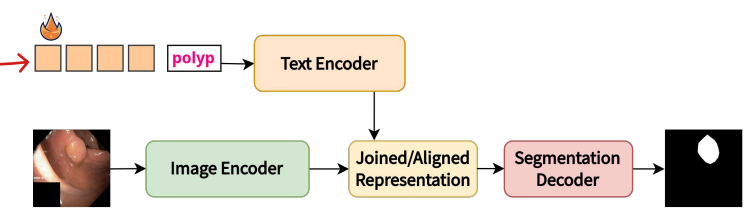


Vision-Language Segmentation Models (VLSMs)

- Leverage image and text data
- Generalize across new datasets and vision tasks

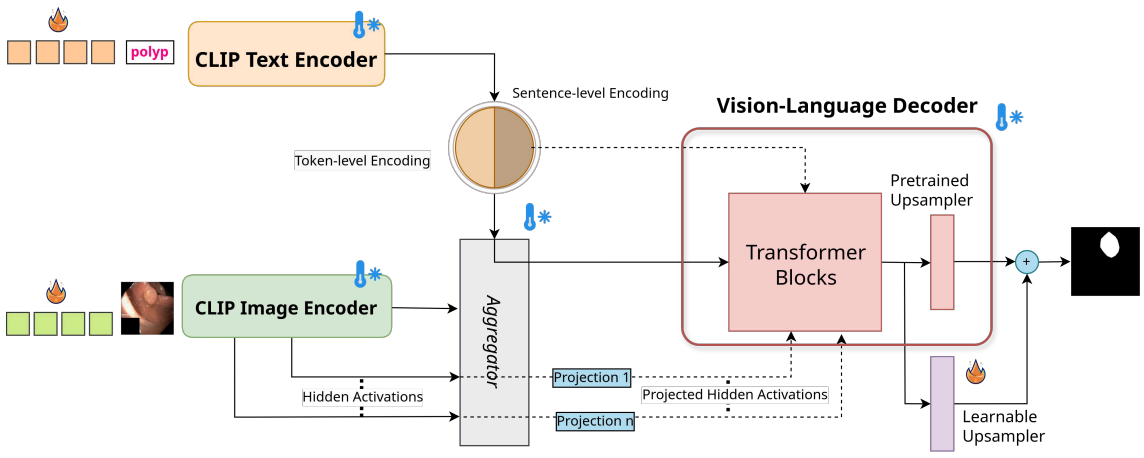
Adapting VLSMs on new medical datasets

Finetuning	Scarcity of labeled data; infeasible
Prompt Engineering	Tedious to determine the right set of prompts
Prompt Tuning	Learnable prompt vectors; effective choice



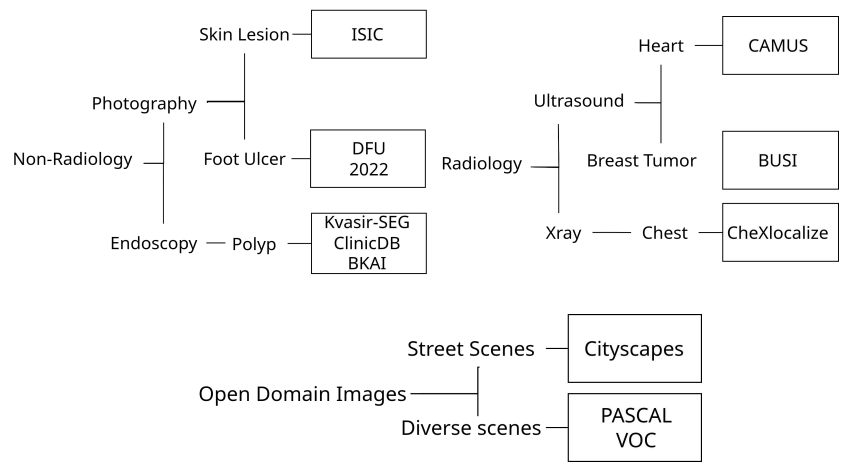
Prompt Tuning benchmark for VLSM

We present the first extensive study of prompt tuning for VLSMs: CRIS and CLIPSeg

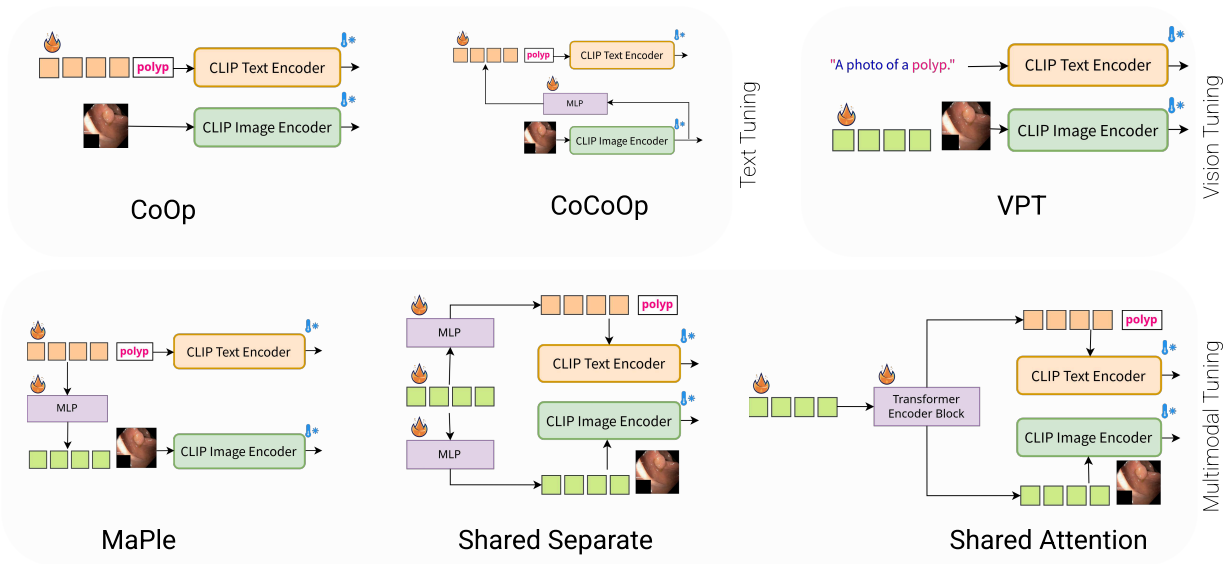


Context vectors can be added at multiple depths of both encoder layers

Datasets



Prompt Tuning Strategies



Experimental Setup

We study the effects of different hyperparameters on the models to gain insights into their performance gain

Hyperparameter	Search Space
Learning rate	$[10^{-5}, 5 \times 10^{-3}]$
Weight decay	$[10^{-5}, 0.01]$
Prompt depth	$[1, 11]$
Intermediate dimension	$\{32, 64, 96, 128\}$
Use LORA	$\{true, false\}$
Transformer: Number of Heads	$\{16, 20, 32\}$
Transformer: Dropout Probability	$[0.1, 0.55]$
Transformer: Feed-Forward Dim	$\{1280, 1420\}$
Transformer: LayerNorm First	$\{true, false\}$
Shared Space Dimension	$\{32, 64\}$

Key Questions

Natural images vs medical images
Is multimodal prompt tuning better than unimodal?
Effects of different hyperparameters on model and tuning performance

Key Results & Takeaways

- Prompt Tuning performance is comparable with end-to-end finetuning
- Vision tuning - on par to multimodal tuning but fewer hyperparameters to tune
- Initialization of context vectors matters
- Using the *Learnable Upsampler* improves performance for segmentation
- Increasing prompt depth: better for multimodal and visual, not for text tuning

Prompt tuning is a promising direction to adapt segmentation models for distribution shifts.

Limitations

- Study is limited for two VLSMs (CRIS, CLIPSeg)
  - Binary segmentation
  - Uses CLIP encoders

Paper link

Code link