VLSM-Adapter: Finetuning Vision-Language Segmentation Efficiently with Lightweight Blocks

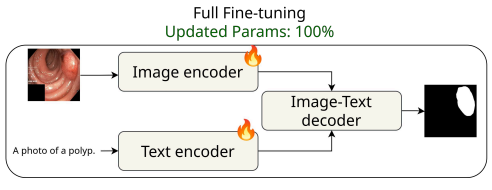
Manish Dhakal, Rabin Adhikari, Safal Thapaliya, Bishesh Khanal

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

Finetuning foundational models is resource intensive

Vision Language Segmentation Models (VLSMs)

- Allows text prompts at inference to guide segmentation
- Mostly trained in open-domain datasets
- Requires end-to-end finetuning for medical datasets



Resource consuming and expensive

Adapters can efficiently finetune foundational models

Adapters

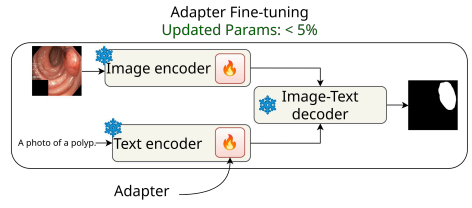
- Linear layers with dimensions smaller than pretrained models
- Can be plugged into existing pretrained architectures
- Can be used to efficiently finetune VLSMs for medical datasets
- Not studied for VLSMs

Adapter formulation

$$f' = Adapter(f) = f + \sigma(\psi(f.W_1).W_2)$$

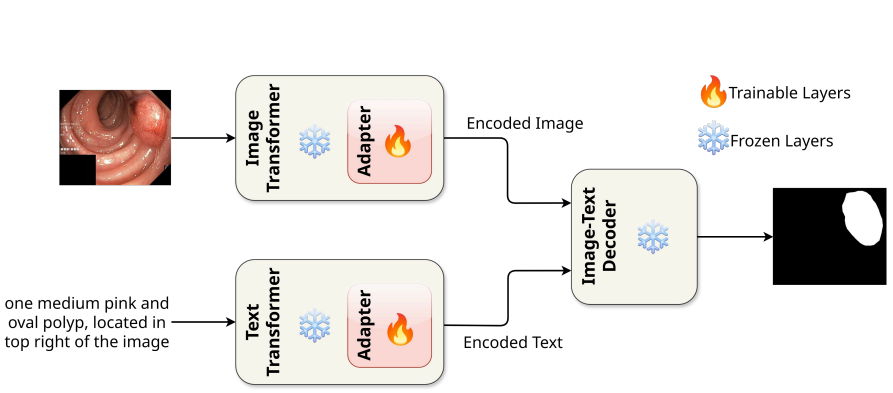
$$W_1 \in \mathbb{R}^{d \times d'} \quad W_2 \in \mathbb{R}^{d' \times d} \quad d' \leq d$$

σ , ψ are non-linear activation functions.

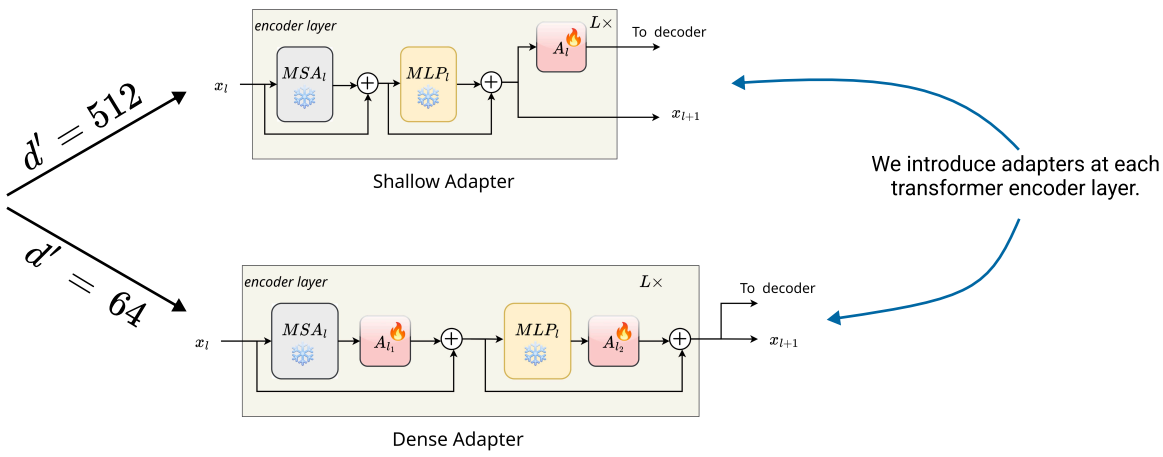


Reduced computing resources

Learnable adapter modules to finetune pretrained VLSMs



Vision Language Segmentation Model with Adapter

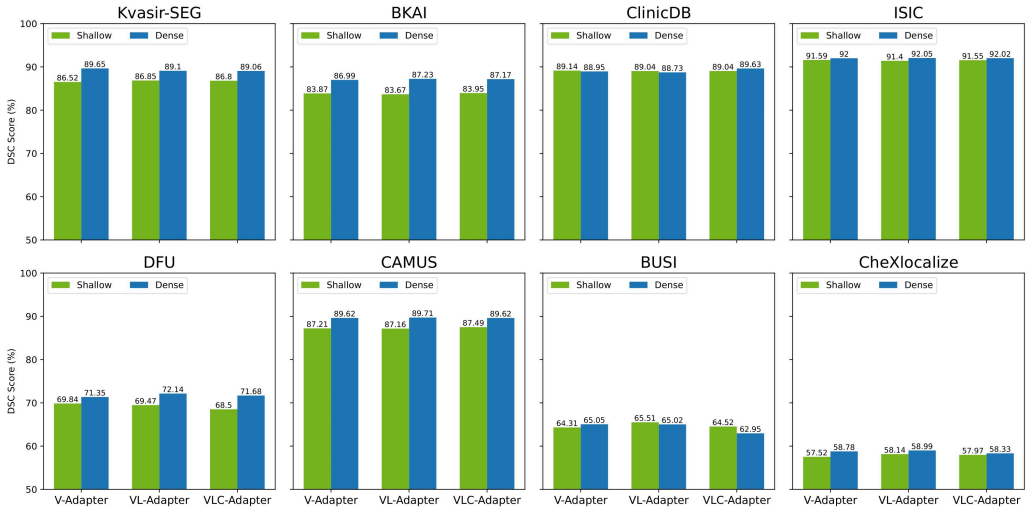


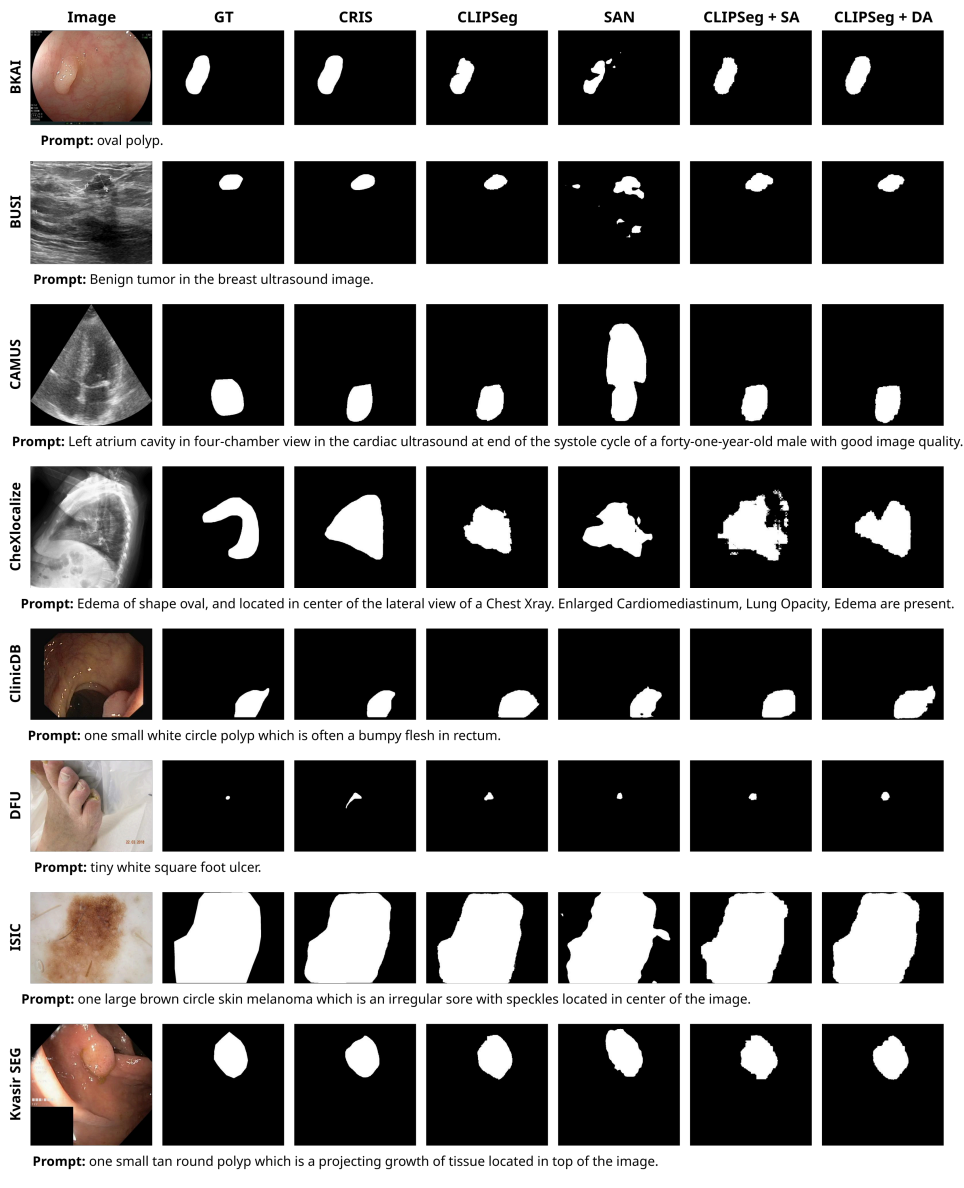
How different adapter configurations perform when finetuning VLSMs on 2D medical datasets?

Adapters finetuning matches end-to-end finetuning

Datasets	Metrics	Upper Bound CLIPSeg 150M	Upper Bound CRIS 147M	Adapter Fine-tune SAN 8.4M	Adapter Fine-tune CLIPSeg SA (**Ours**) 4.2M	Adapter Fine-tune CLIPSeg DA (**Ours**) **3M**
Kvasir-SEG	DSC (%) ↑	87.69	89.43	69.58	86.85	**89.10**
	IoU (%) ↑	81.72	83.37	58.05	79.26	**82.39**
	HD95 ↓	54.02	55.23	130.75	52.18	**47.79**
BKAI	DSC (%) ↑	85.59	92.62	66.26	83.67	**87.23**
	IoU (%) ↑	77.52	88.30	54.58	75.02	**79.81**
	HD95 ↓	87.91	49.80	224.37	87.79	**70.02**
ClinicDB	DSC (%) ↑	88.58	93.63	81.36	**89.04**	88.73
	IoU (%) ↑	81.51	88.74	72.61	**81.93**	81.84
	HD95 ↓	19.30	12.36	38.42	**18.03**	18.76
ISIC-16	DSC (%) ↑	91.88	91.49	90.39	91.40	**92.05**
	IoU (%) ↑	85.76	85.41	83.61	85.05	**85.98**
	HD95 ↓	60.93	64.39	87.25	60.29	**54.38**
DFU	DSC (%) ↑	72.12	74.01	63.38	69.47	**72.14**
	IoU (%) ↑	61.61	64.31	51.63	58.27	**61.42**
	HD95 ↓	38.24	41.92	60.10	**38.75**	38.79
CAMUS	DSC (%) ↑	88.93	91.29	46.42	87.16	**89.71**
	IoU (%) ↑	80.69	84.42	31.81	78.01	**81.85**
	HD95 ↓	16.69	12.33	175.81	19.14	**14.16**
BUSI	DSC (%) ↑	62.91	67.50	45.61	**65.51**	65.02
	IoU (%) ↑	55.52	60.90	35.27	**58.19**	57.20
	HD95 ↓	72.98	50.63	152.10	**63.36**	64.37
CheXlocalize	DSC (%) ↑	58.51	60.76	44.37	58.14	**58.99**
	IoU (%) ↑	45.45	47.99	31.97	44.84	**46.01**
	HD95 ↓	537.57	519.21	724.55	**533.04**	535.97

Using adapters at both text and image encoders works the best





Limitations and Future direction

- Performance does not match with vision-only models
- Can be extended to 3D medical image datasets
- Opens pathways for continual and multi-task learning

Read paper