Efficient Few-Shot Clinical Task Adaptation with Large Language Model

Kaipeng Zheng, Weiran Huang, Lichao Sun

Content

- Foundation Model Selection
- Efficient Fine-tuning with Partial Freezing
- LLM-Contextualized Semantic Guidance
- Early Stopping in Few-Shot Multi-Label Classification

Foundation Model Selection

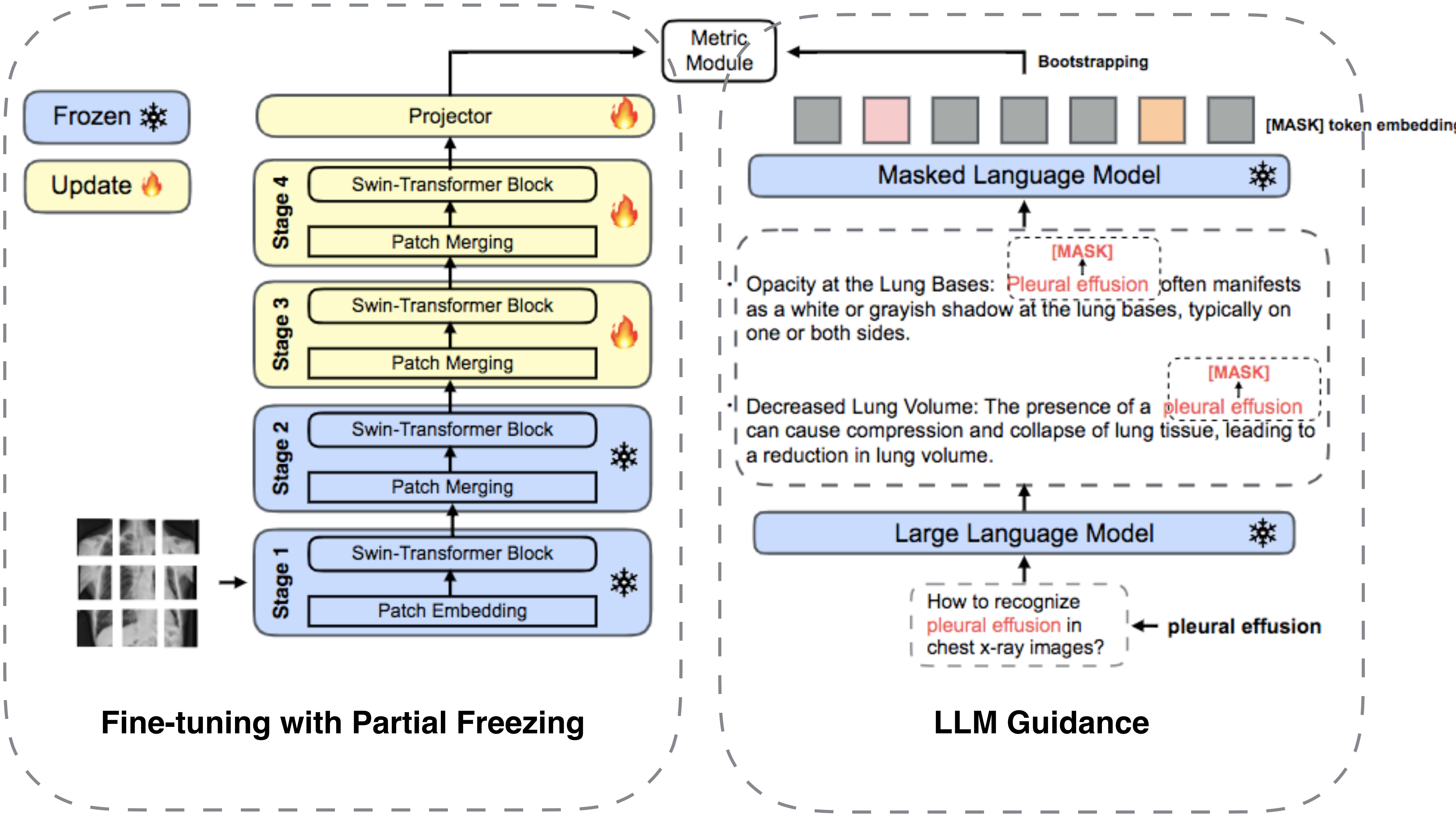
Method	Chest			Colon			Endo		
	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
CLIP (ViT-B)	51.52 ± 1.35	52.03 ± 0.51	52.34 ± 1.09	78.22 ± 6.41	83.85 ± 5.86	85.53 ± 1.98	48.93 ± 4.18	54.69 ± 5.86	59.53 ± 1.55
MAE (ViT-B)	54.64 ± 1.59	63.47 ± 0.60	66.08 ± 1.41	82.50 ± 4.60	90.62 ± 2.74	94.27 ± 2.22	58.19 ± 5.20	65.66 ± 2.05	69.43 ± 1.01
Sup ViT-B	55.13 ± 1.72	63.28 ± 0.63	65.16 ± 0.56	84.87 ± 3.32	93.26 ± 2.03	95.51 ± 1.25	57.02 ± 8.75	65.24 ± 3.24	68.51 ± 1.38
GLIP (Swin-L)	54.87 ± 1.49	61.87 ± 0.79	64.85 ± 1.54	86.41 ± 3.32	92.18 ± 1.98	95.57 ± 1.86	60.17 ± 4.61	65.99 ± 3.52	70.69 ± 0.85
Sup Swin-L	56.21 ± 1.67	64.07 ± 1.16	65.94 ± 0.75	85.35 ± 5.27	94.54 ± 2.51	96.66 ± 1.39	61.49 ± 4.65	66.87 ± 2.02	72.39 ± 0.92
DAViT	53.73 ± 1.35	62.63 ± 0.85	64.80 ± 0.55	85.36 ± 4.39	91.29 ± 4.08	95.21 ± 1.72	56.18 ± 6.00	64.34 ± 5.11	68.26 ± 2.73

Table 1. Comparison with baselines in 1-shot, 5-shot, and 10-shot settings across all tasks in MedFMC. The average mAUC on the validation set is reported.

observations:

- **Swin-Transformer** pre-trained on ImageNet-21K consistently outperforms others.
- Advanced foundation models trained on natural images **do not** exhibit superiority in adaptation to few-shot clinical tasks.

Core Techniques

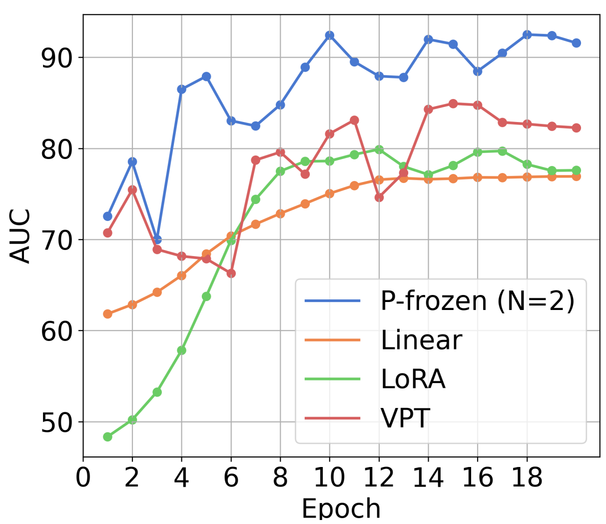


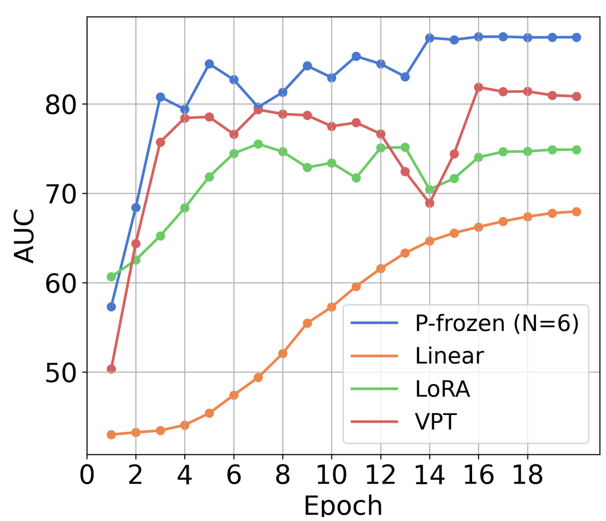
Implementation

- **Vision backbone:** The officially implemented Swin-Transformer pre-trained on ImageNet-21K (named as swin-large_in21k-pre-3rdparty_in1k-384* in MMPretrain).
- **Language supervision:** GPT-4 for contextualization, BERT pretrained on PubMed as the masked language model.
- **Optimizer:** AdamW, batch size = 4, learning rate = 1e−4. Validation set for model selection.
- **Data augmentation:** center cropping, random cropping, and random horizontal flipping.
- **Ensemble:** Averaging the output scores of 2-4 models trained with different random seeds, class-wise ensemble, no TTA.

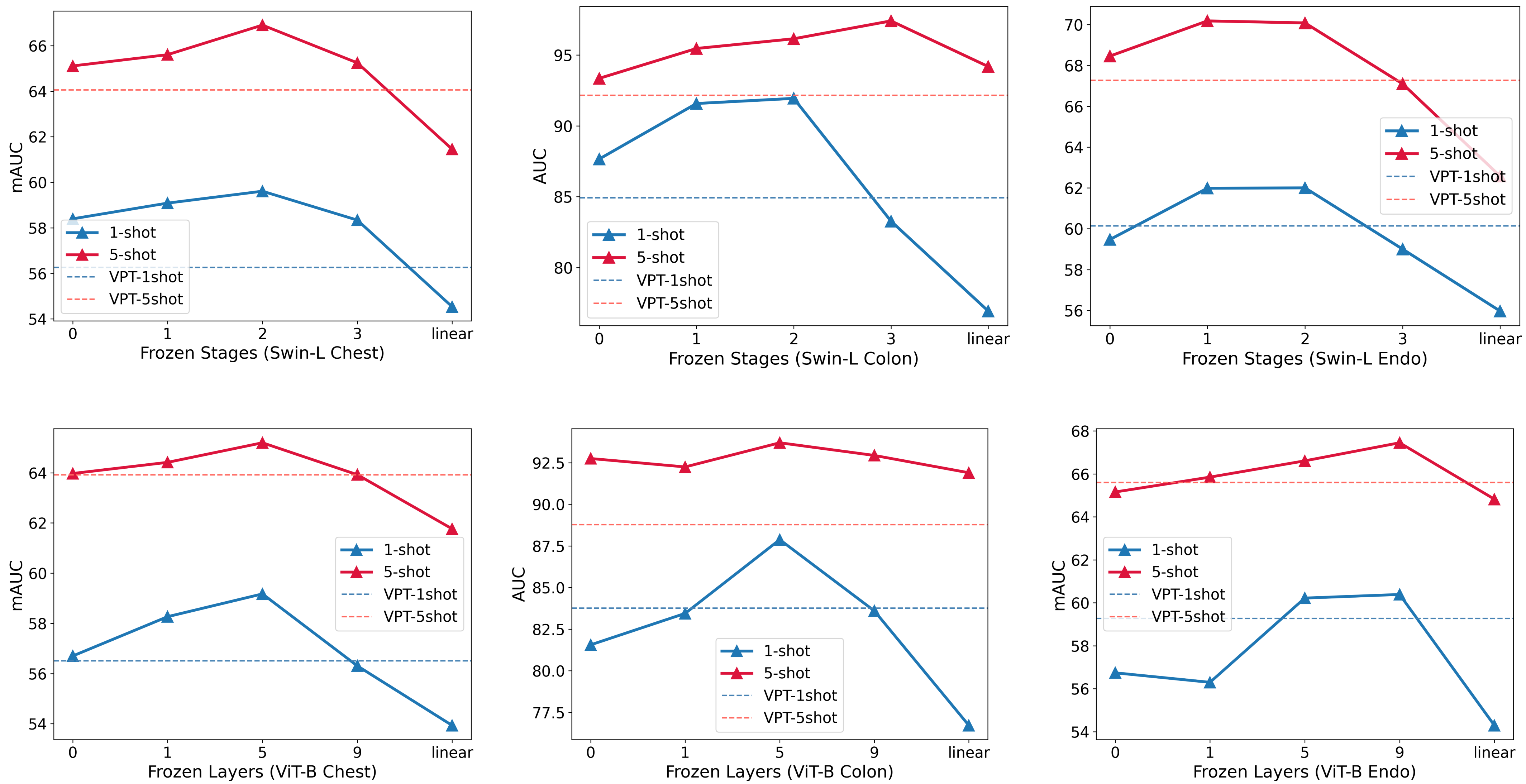
Efficient Fine-tuning with Partial Freezing

Backbone	Adaption Method	Chest 1-shot	Chest 5-shot	Chest 10-shot	Colon 1-shot	Colon 5-shot	Colon 10-shot	Endo 1-shot	Endo 5-shot	Endo 10-shot
Swin-L	Full-Model Fine-Tuning	58.89 ± 0.74	65.10 ± 0.82	67.25 ± 0.65	87.65 ± 2.82	94.21 ± 3.81	96.35 ± 0.95	59.35 ± 5.02	68.21 ± 2.76	69.72 ± 1.52
	Linear Probe	54.18 ± 1.19	61.16 ± 0.89	64.30 ± 0.66	84.36 ± 5.21	94.50 ± 1.79	95.67 ± 0.73	55.55 ± 6.53	64.53 ± 3.72	72.15 ± 0.67
	Adapter	56.21 ± 1.67	64.07 ± 1.16	65.94 ± 0.75	85.35 ± 5.27	94.54 ± 2.51	96.66 ± 1.39	61.49 ± 4.65	66.87 ± 2.02	72.39 ± 0.92
	LoRA	53.43 ± 1.64	65.54 ± 0.79	69.49 ± 0.13	87.84 ± 5.63	97.21 ± 1.23	97.49 ± 1.80	58.98 ± 4.95	67.69 ± 3.01	74.11 ± 0.53
	Visual Prompt Tuning	56.37 ± 1.15	64.51 ± 1.34	64.29 ± 1.63	81.67 ± 3.80	92.13 ± 3.55	95.42 ± 0.56	60.13 ± 6.13	67.26 ± 2.23	70.99 ± 1.31
	Ours	**59.92** ± 1.29	**66.90** ± 0.62	**69.73** ± 0.32	**91.50** ± 2.47	**97.51** ± 1.16	**97.85** ± 1.25	**62.51** ± 4.65	**70.13** ± 2.37	**74.85** ± 0.61
ViT-B	Full-Model Fine-Tuning	57.06 ± 1.83	63.96 ± 0.20	66.12 ± 0.41	81.55 ± 3.62	92.74 ± 3.41	95.23 ± 0.72	58.59 ± 7.44	64.02 ± 5.91	70.04 ± 1.39
	Linear Probe	53.69 ± 0.97	61.12 ± 1.60	65.14 ± 0.28	78.88 ± 7.89	92.72 ± 1.84	96.03 ± 1.36	55.34 ± 6.83	63.46 ± 5.59	70.36 ± 0.67
	Adapter	55.13 ± 1.72	63.28 ± 0.63	65.16 ± 0.56	84.87 ± 3.32	93.26 ± 2.03	95.51 ± 1.25	57.02 ± 8.75	65.24 ± 3.24	68.51 ± 1.38
	LoRA	54.84 ± 1.67	65.20 ± 0.55	67.92 ± 0.70	78.19 ± 6.50	93.70 ± 2.38	94.95 ± 1.80	55.91 ± 9.36	66.97 ± 1.72	72.12 ± 1.27
	Visual Prompt Tuning	56.87 ± 2.00	63.89 ± 0.47	65.60 ± 0.34	83.95 ± 5.58	88.90 ± 1.51	91.29 ± 3.43	59.54 ± 8.28	64.55 ± 5.04	67.52 ± 1.88
	Ours	**59.14** ± 1.31	**65.47** ± 0.61	**68.35** ± 0.65	**87.69** ± 3.39	**94.54** ± 1.79	**96.22** ± 1.54	**61.01** ± 6.80	**67.62** ± 2.13	**72.65** ± 1.61





Efficient Fine-tuning with Partial Freezing



From One-hot Labels to LLM-Contextualized Semantic Guidance

Figure 5. The inter-class correlation matrices obtained by different language supervision methods on the ChestDR task. Left: Encoding category names. Middle: a template method using masked language models without contextualization. Right: our method leveraging large language models for label contextualization. It can be observed that the context generated by large language models plays a crucial role in fine-grained category distinguishing.

From One-hot Labels to LLM-Contextualized Semantic Guidance

Method	Adaption	Semantic Supervision	1-shot	5-shot	10-shot
Swin-L	fine-tuning (frozen-stage = 1)	None	59.92 $\pm$ 1.29	66.90 $\pm$ 0.62	69.73 $\pm$ 0.32
Swin-L	fine-tuning (frozen-stage = 1)	class-name	59.08 $\pm$ 1.22	68.02 $\pm$ 0.66	70.37 $\pm$ 0.78
Swin-L	fine-tuning (frozen-stage = 1)	template	57.32 $\pm$ 4.59	68.38 $\pm$ 0.50	70.72 $\pm$ 0.63
Swin-L	fine-tuning (frozen-stage = 1)	context (ours)	**62.95** $\pm$ 0.19	**69.24** $\pm$ 0.60	**71.41** $\pm$ 0.37
ViT-B/16	visual prompt tuning	None	56.87 $\pm$ 2.00	63.89 $\pm$ 0.47	65.60 $\pm$ 0.34
ViT-B/16	visual prompt tuning	class-name	58.64 $\pm$ 1.35	65.72 $\pm$ 0.93	67.35 $\pm$ 0.98
ViT-B/16	visual prompt tuning	template	54.51 $\pm$ 3.30	65.94 $\pm$ 0.70	67.96 $\pm$ 0.51
ViT-B/16	visual prompt tuning	context (ours)	**59.66** $\pm$ 1.69	**66.31** $\pm$ 0.94	**68.58** $\pm$ 0.40

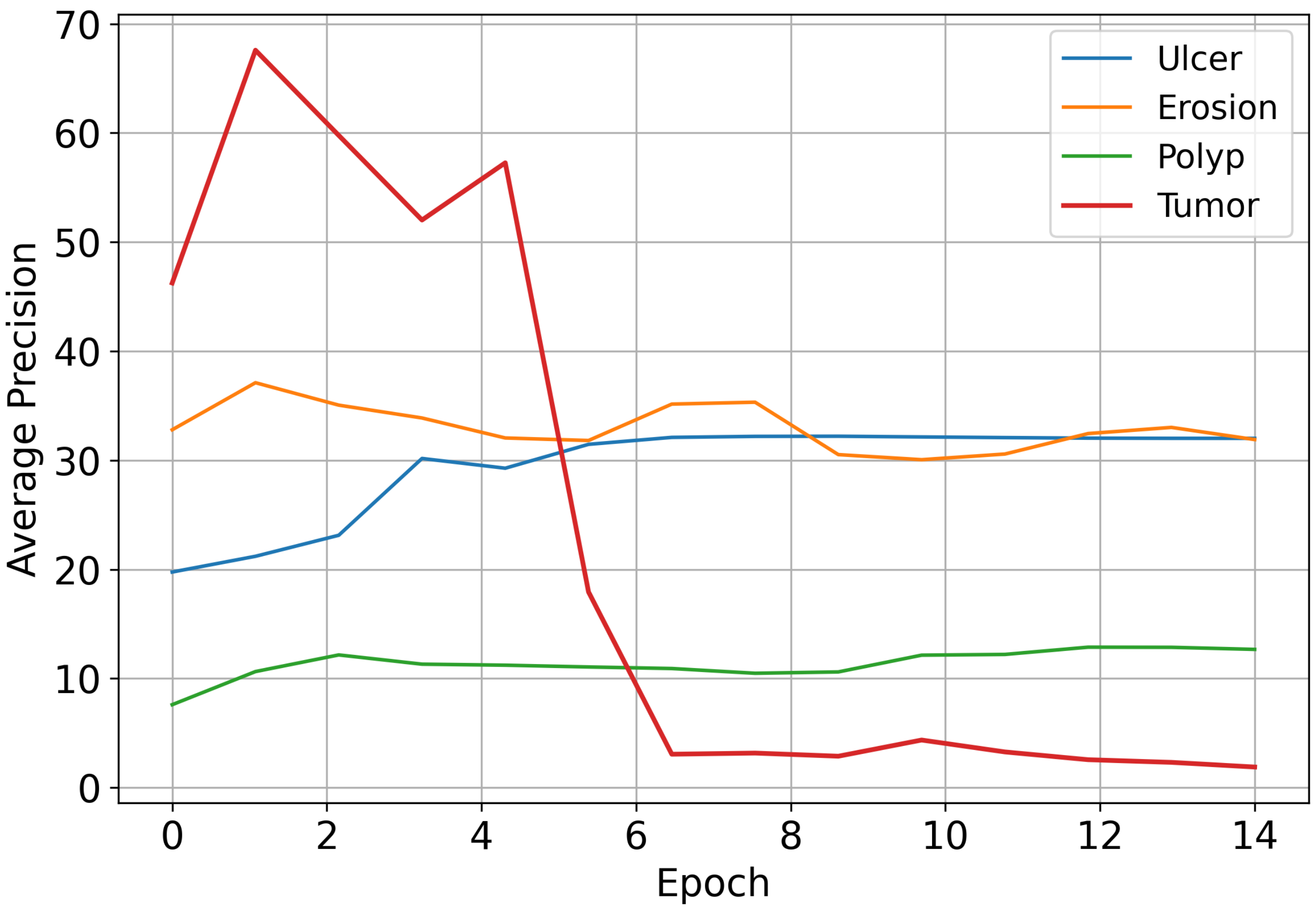
Table 3. Comparison of different language supervision methods on the ChestDR task in 1-shot, 5-shot, and 10-shot settings, where "None" indicates the use of only one-hot labels. The average mAUC on the validation set is reported.

		1-shot	
Method	exp	mAp	mAUC
vpt	1	19.23	57.70
vpt+semantic label	**1**	**21.63**	**61.19**
swin	1	20.33	59.22
swin+semantic label	**1**	**22.35**	**61.90**
vpt	2	15.45	55.60
vpt+semantic label	**2**	**17.62**	**59.88**
swin	2	17.79	59.06
swin+semantic label	**2**	**18.04**	**62.01**
swin	3	17.26	56.68
swin+semantic label	**3**	**18.18**	**61.07**
swin	4	18.31	58.46
swin+semantic label	**4**	**18.72**	**60.04**
swin	5	18.15	60.24
swin+semantic label	**5**	**19.04**	**61.28**

Severe Inconsistency in Optimal Early Stopping Time for Different Categories

- **Issue:** In few-shot multi-label classification, some categories achieve optimal performance with very few iterations, and further iterations result in overfitting on those categories. Conversely, other categories may require many iterations to reach optimal performance.

- **Method:** Class-wise Ensemble.



Thanks

http://arxiv.org/abs/2312.07125