DWARF: **D**isease-**w**eighted network for **a**ttention map **ref**inement

Haozhe Luo [1,2]

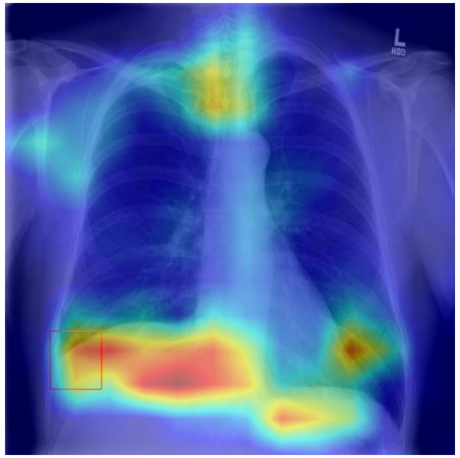
Supervisor: Oana Inel [1], Abraham Bernstein [1], Mauricio Reyes [2]

[1] University of Zurich [2] ARTORG Center for Biomedical Engineering Research, University of Bern

Background

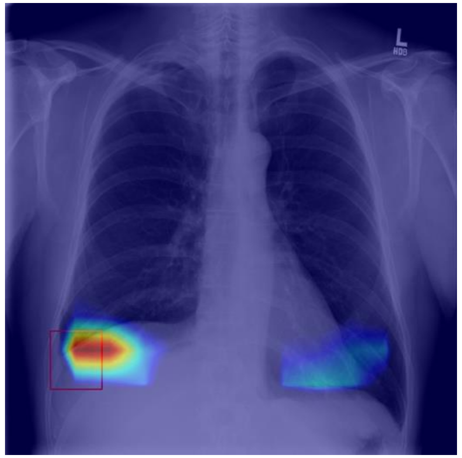
- **The accuracy of AI assisted diagnostics is improving**
- The interpretability of AI diagnostics is still dissatisfying ❎

Unaligned attention



Attention is **not well-focused** on relevant diagnostic areas.

Aligned attention



Attention **aligns well** with relevant diagnostic areas.

Previous work:

(1) Unsupervised attention alignment

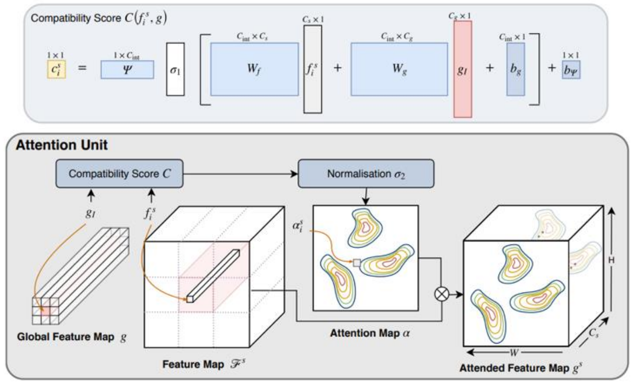
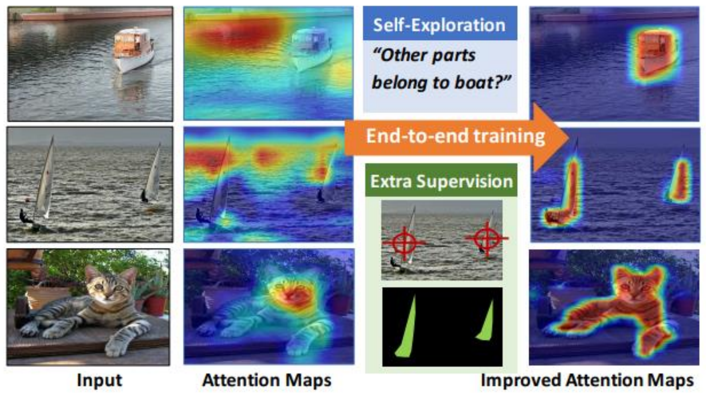


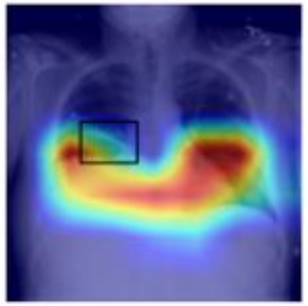
Figure 2: The proposed grid attention block. The global gate signal g_I is shared for the region indicated in red. Tensor dimensions for the compatibility score computation are shown.

(2) Supervised attention alignment for Categorical Classification

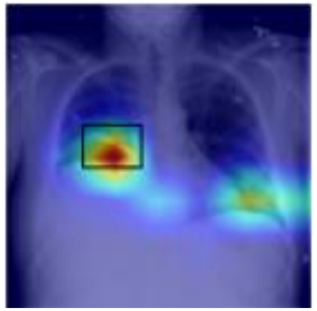


Hypothesis

Improving attention alignment also enhances classification performance.



Attention guidance



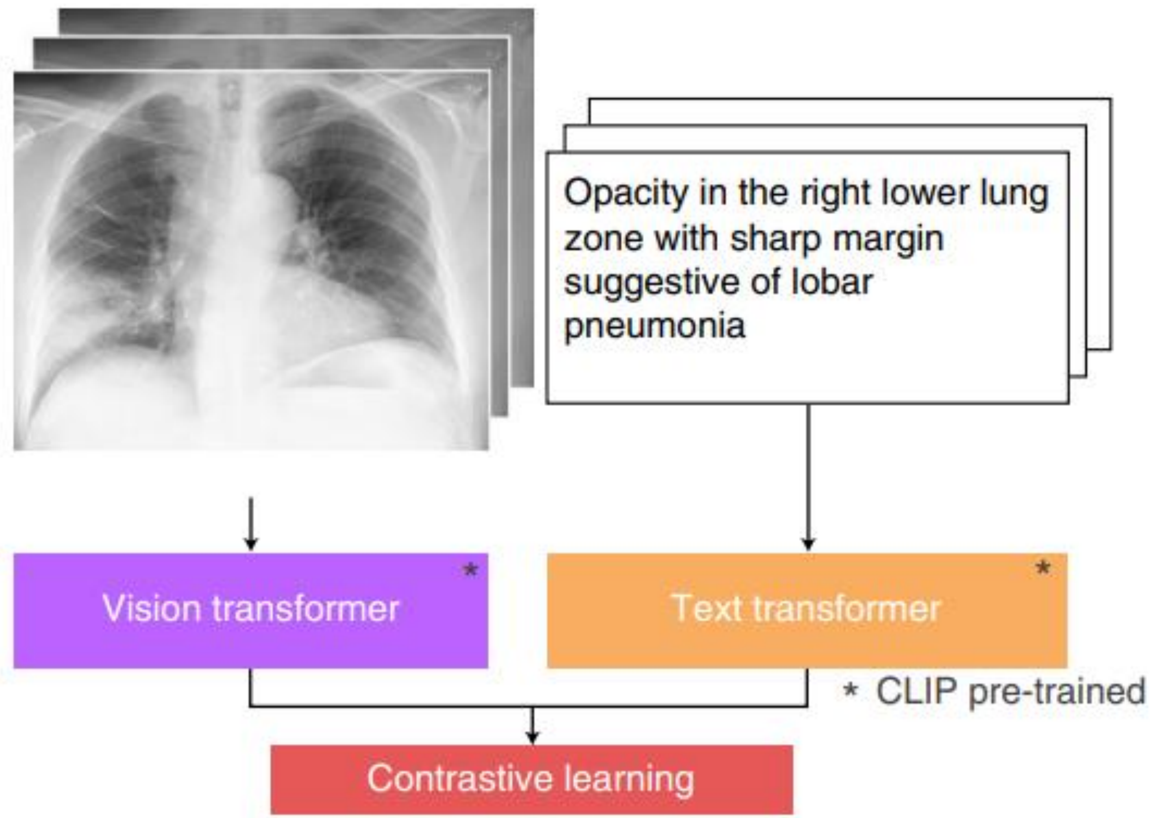
Unaligned attention map

Aligned attention map!

CLS performance improves!

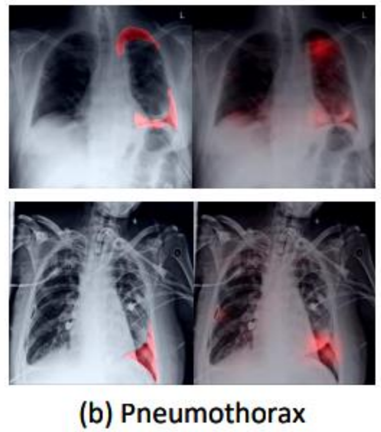
Related work: Vision Language Model

[1]



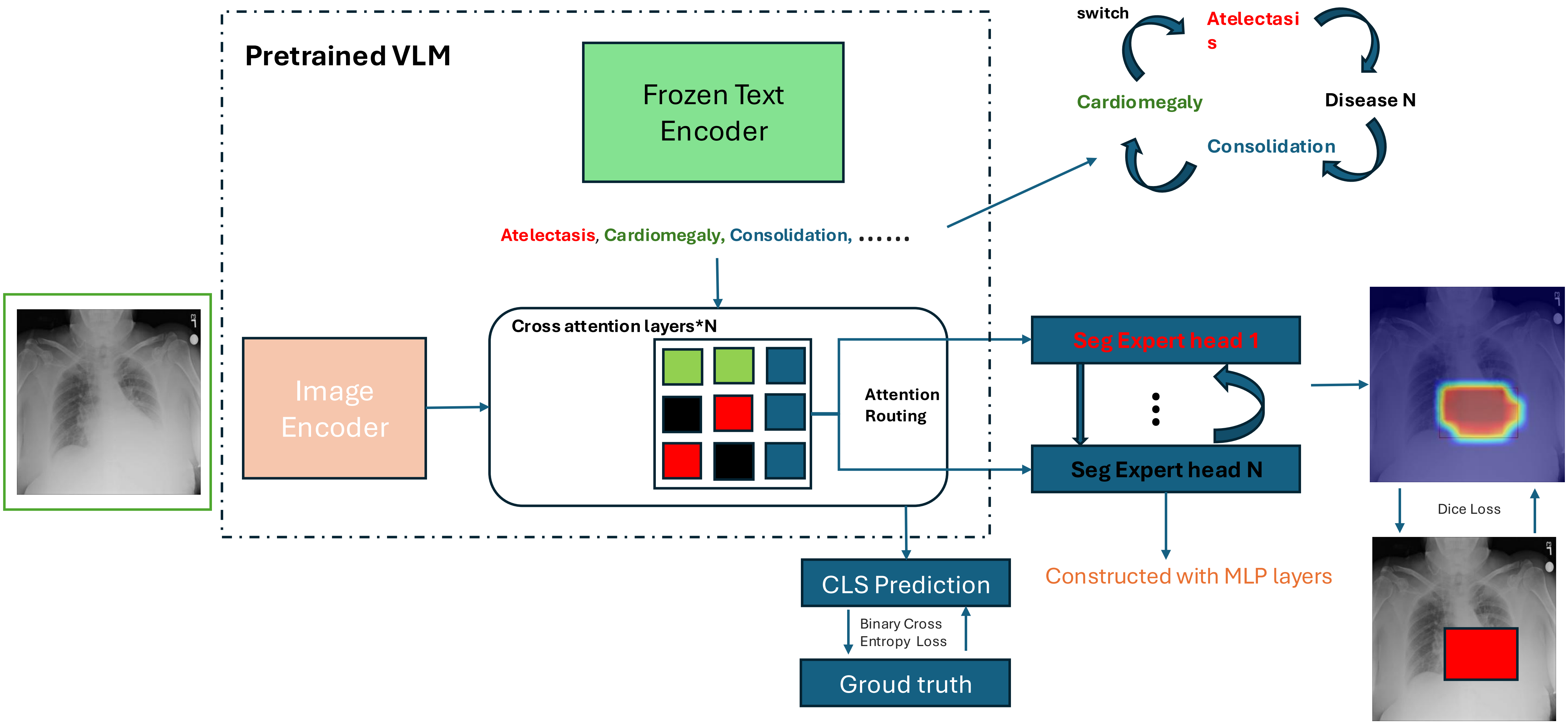
VLM (Vision-Language Model) aligns visual and language information for **cross-modal understanding** and offers accurate finding **attention performance**.

[2]



[1] Tiu E, Talius E, Patel P, et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning[J]. Nature Biomedical Engineering, 2022, 6(12): 1399-1406.
[2] Wu C, Zhang X, Zhang Y, et al. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 21372-21383.

Dwarf: Disease-weighted attention map refinement network



Problem Definition

The goal is to optimize a **multi-label classification** model in medical imaging, incorporating the use of cross-attention feature maps to enhance interpretability. The optimization problem can be formulated as follows:

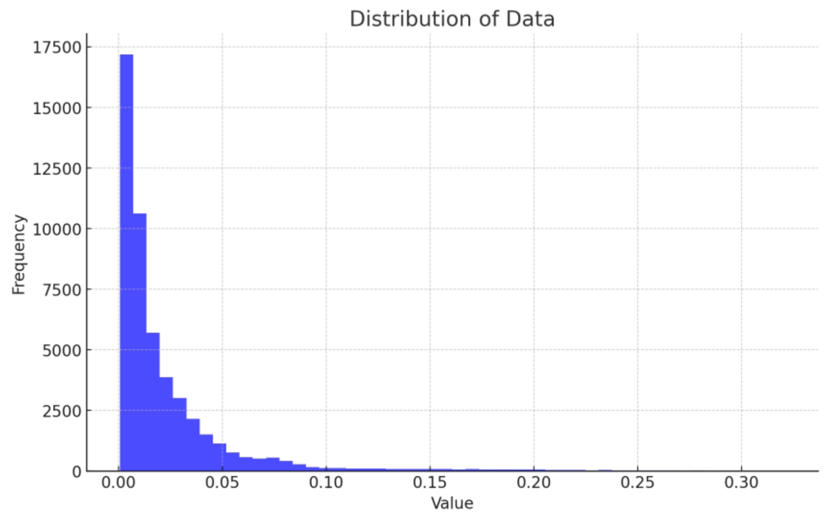
$$\min_{\theta} L\_\text{total}(\theta) = \lambda L\_\text{cls}(\theta) + (1 - \lambda) L\_\text{atten}(\theta)$$

where:

- θ represents the parameters of the model.
- L_total(θ) is the total loss function to be minimized.
- L_cls(θ) is the loss function associated with the multi-label classification accuracy.
- L_atten(θ) corresponds to the loss function for reducing distance between generated attention map and clinicians' attention.
- λ is a loss weight hyperparameter

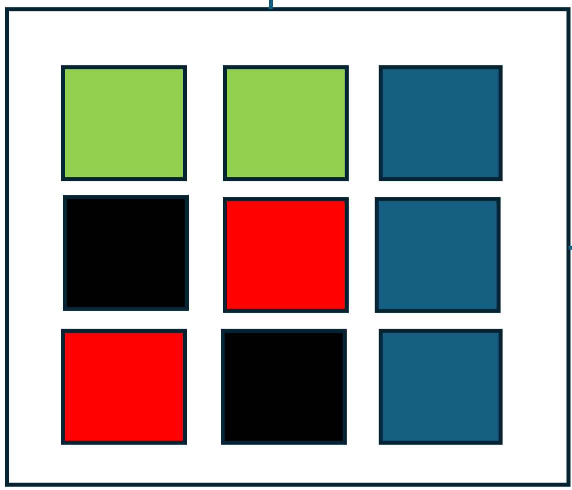
The objective is to simultaneously enhance the **classification** performance and the **grounding** performance provided by the cross-attention maps.

Introducing additional heads



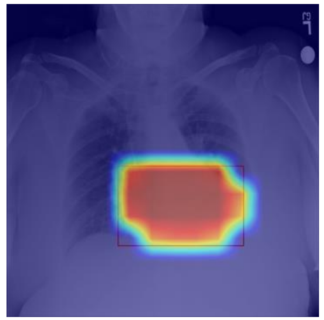
Directly optimize? Attention score scale varies

Cross attention map



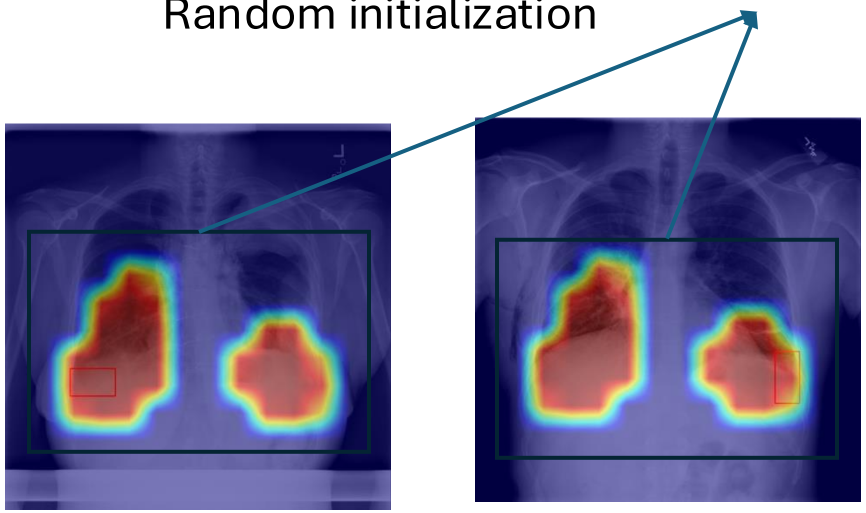
Seg Expert head 1

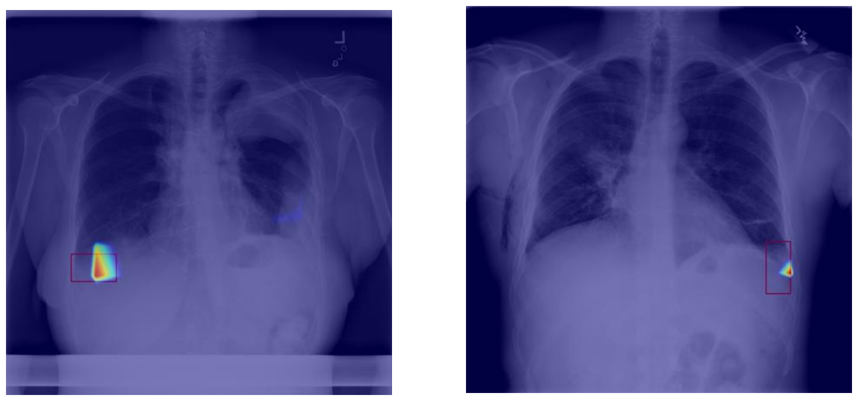
Additional heads to inject



Identity Enhanced Initialization, IEI

- With **randomly initialized heads**, we observed that the heads are encouraged to learn fixed pattern. (trivial solution)
- We propose Identity Enhancement Initialization (IEI) for different disease's heads' parameter initialization.
- With IEI the expert head's parameters are initialized with **Identity Matrix** (Identity mapping)





False Positive Suppression

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \cdot X \cap Y + \alpha + \epsilon}{X + Y + \alpha + \epsilon}$$

$$\text{adjustedY} = \text{Y} + (w_{\text{FP}} - 1) \cdot \text{FP}$$

Final Loss:

$$\mathcal{L}_{\text{seg}} = 1 - \frac{2 \cdot (X \cap Y) + \alpha + \epsilon}{X + \text{adjusted}Y + \alpha + \epsilon}$$

Positive samples were used during optimization for attention guidance.

Network tend to **overestimate**, propose FPS to suppress that issue.

Atelectasis

Datasets:

ChestX-Det

13 pathologies:
1. Atelectasis 2. Calcification 3. Cardiomegaly 4. Consolidation 5. Diffuse Nodule 6. Effusion 7. Emphysema 8. Fibrosis 9. Fracture 10. Mass 11. Nodule 12. Pleural Thickening 13. Pneumothorax

Annotation types: bbox, polygons

Dataset size: 3578 images

Vindr-CXR

13 pathologies:
1. Atelectasis 2. Calcification 3. Cardiomegaly 4. Consolidation 5. Diffuse Nodule 6. Effusion 7. Emphysema 8. Fibrosis 9. Fracture 10. Mass 11. Nodule 12. Pleural Thickening 13. Pneumothorax

Annotation types: bbox

Dataset size: 15000 Images

CheXlocalize

10 pathologies:
1. Airspace opacity 2. Atelectasis 3. Cardiomegaly 4. Consolidation 5. Edema 6. Enlarged cardiomediastinum 7. Lung lesion 8. Pleural effusion 9. Pneumothorax 10. Support devices

Annotation types: bbox, polygons

Dataset size: 234 images

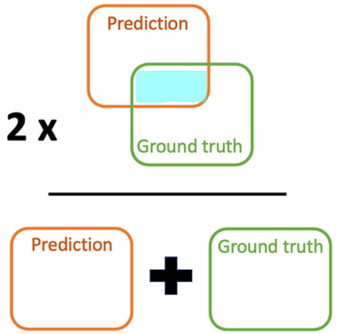
Baselines:

Pretrained Vision-language Model: KAD, DeViDe
Finetuned Vision-language Model: GAIN

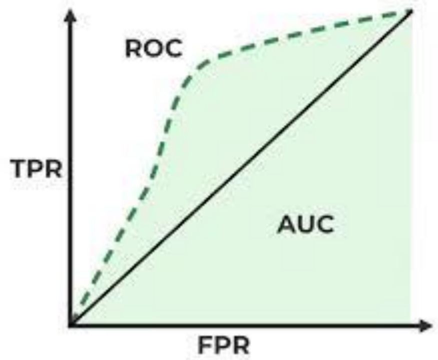
Metrics

(1) Dice:

$$\text{Dice} = \frac{2\ X\ Area\ of\ overlap}{Total\ area} =$$



(2) AUC:



(3) F1 score:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

(4) MCC:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Results of different datasets

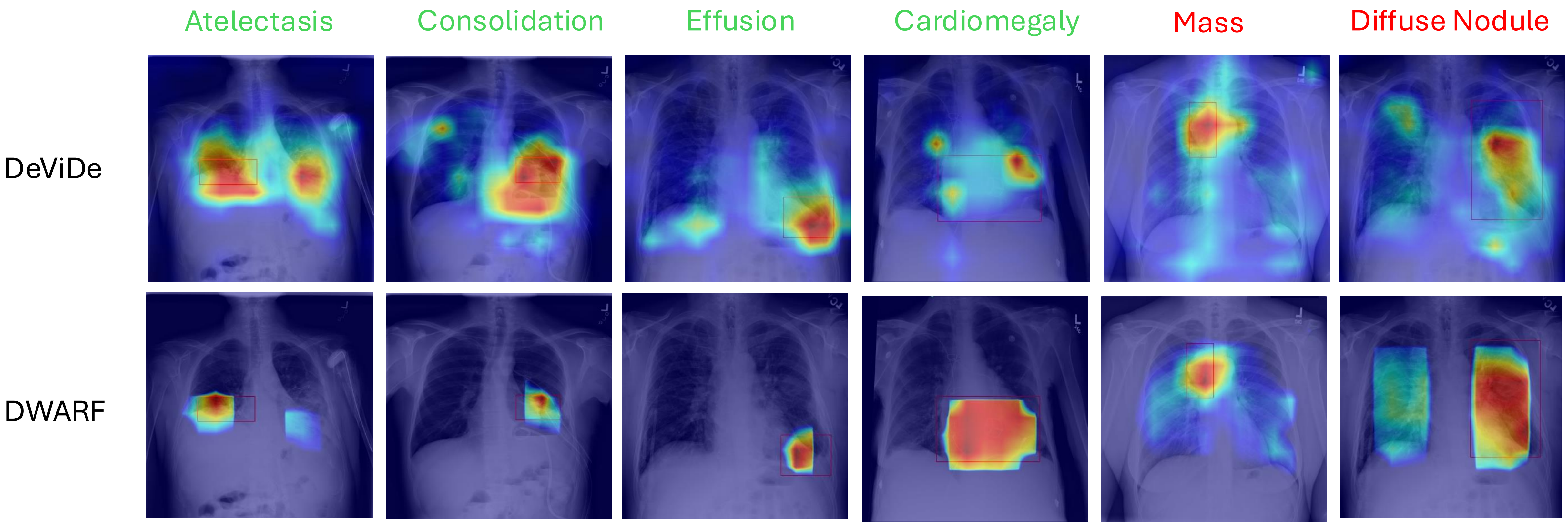
[1] Luo H, Zhou Z, Royer C, et al. DeViDe: Faceted medical knowledge for improved medical vision-language pre-training[J].
[2] Zhang X, Wu C, Zhang Y, et al. Knowledge-enhanced visual-language pre-training on chest radiology images[J]. Nature Communications, 2023, 14(1): 4542.
[3] Li K, Wu Z, Peng K C, et al. Tell me where to look: Guided attention inference network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9215-9223.

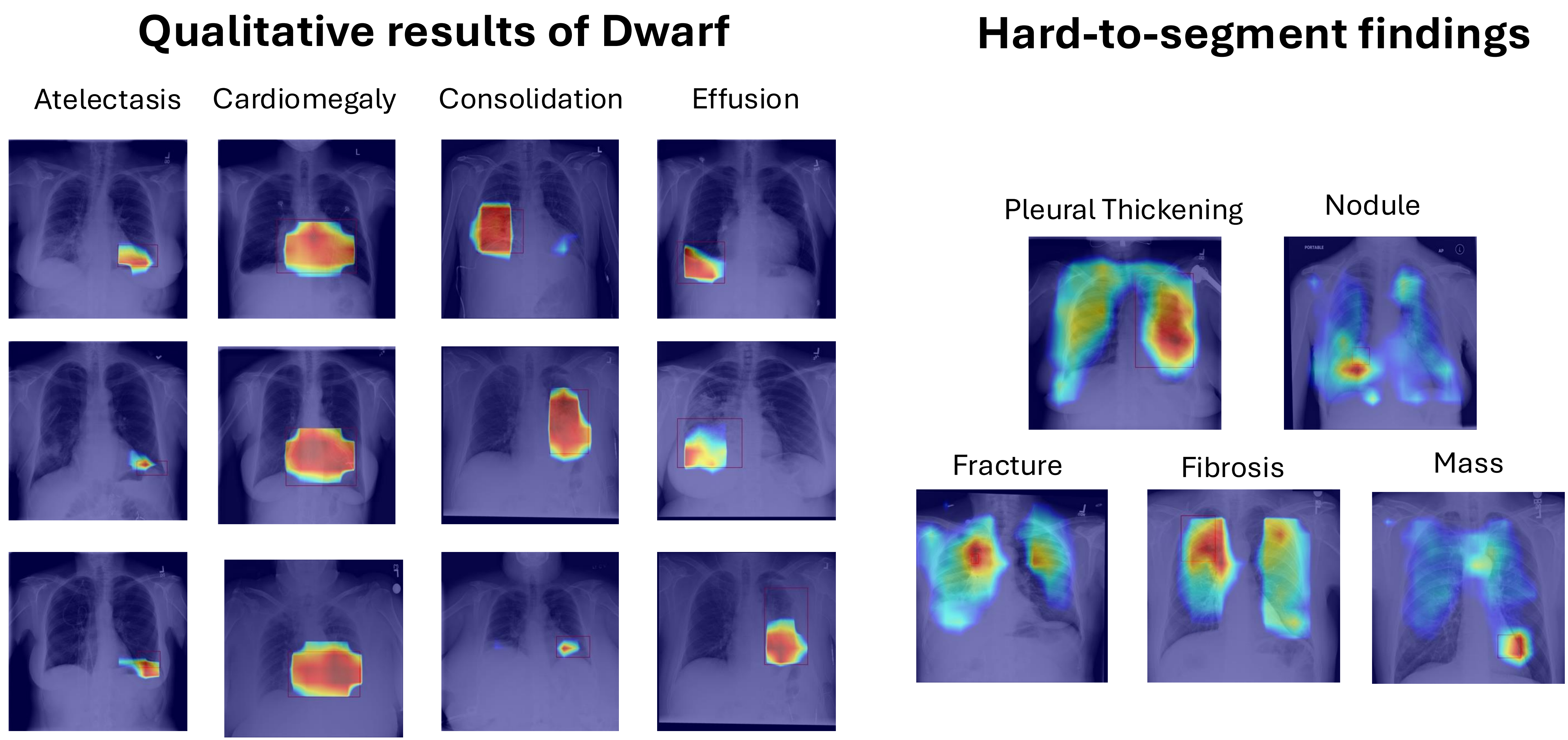
Method	Dataset	AUC (%)	F1 Score (%)	MCC (%)	Dice (%)	Model Type
DeViDe[1]	ChestX Det	74.24	42.66	34.29	13.66	Pretrained VLM
KAD[2]	ChestX Det	73.81	40.07	31.48	13.89	Pretrained VLM
GAIN[3]	ChestX Det	80.9	58.97	**62.65**	13.09	Finetuned VLM
DWARF	ChestX Det	**81.94**	**59.13**	49.87	**18.24**	Finetuned VLM
DeViDe	cheXlocalize	72.26	41.66	30.79	59.83	Pretrained VLM
KAD	cheXlocalize	72.22	41.52	31.53	11.58	Pretrained VLM
GAIN	cheXlocalize	84.44	62.68	58.82	11.91	Finetuned VLM
DWARF	cheXlocalize	**84.83**	**63.44**	**63.14**	**13.4**	Finetuned VLM
DeViDe	Vindr CXR	72.92	41.01	27.98	7.06	Pretrained VLM
KAD	Vindr CXR	73.19	40.22	30.73	7.19	Pretrained VLM
GAIN	Vindr CXR	78.51	45.77	35.91	7.23	Finetuned VLM
DWARF	Vindr CXR	**80.01**	**47.05**	**39.55**	**10.21**	Finetuned VLM

DWARF achieves enhanced Stability across disease numbers

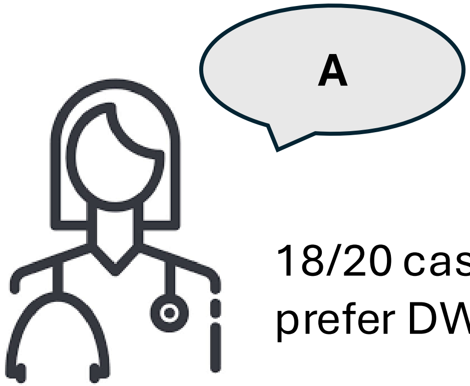
Method	Disease numbers (with same total epochs)	AUC	Max DICE	F1/MCC
GAIN	4	0.8680	0.1438	-
GAIN	7	0.8519	0.1903	-
GAIN	13	0.8090	0.1390	-
DWARF	4	**0.8871**	**0.4147**	-
DWARF	7	**0.8717**	**0.3559**	0.6017/0.5201
DWARF	13	**0.8157**	**0.1805**	0.5344/0.4992

DWARF achieves stable improvement across different disease numbers

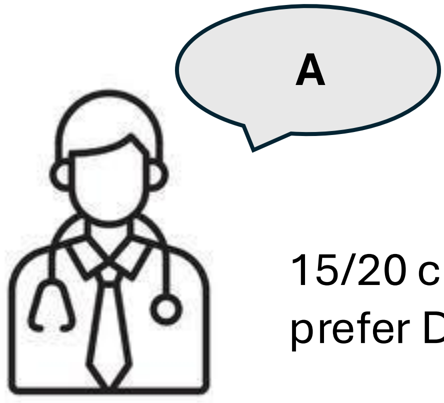

Atelectasis
Consolidation
Effusion
Cardiomegaly
Mass
Diffuse Nodule
DeViDe
DWARF


Qualitative results of Dwarf
Atelectasis
Cardiomegaly
Consolidation
Effusion
Hard-to-segment findings
Pleural Thickening
Nodule
Fracture
Fibrosis
Mass

DWARF Enhances Clinicians' Confidence in Classification Models



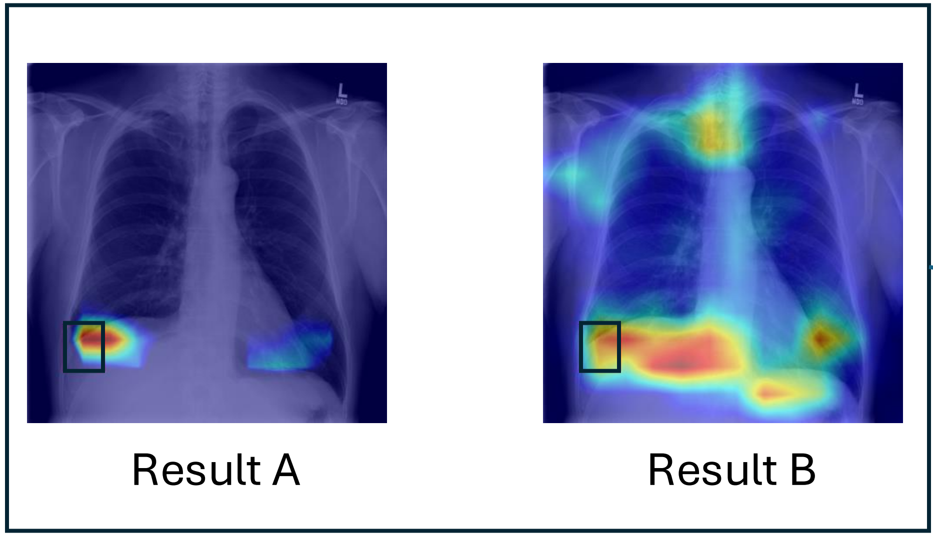
18/20 cases
prefer DWARF



15/20 cases
prefer DWARF

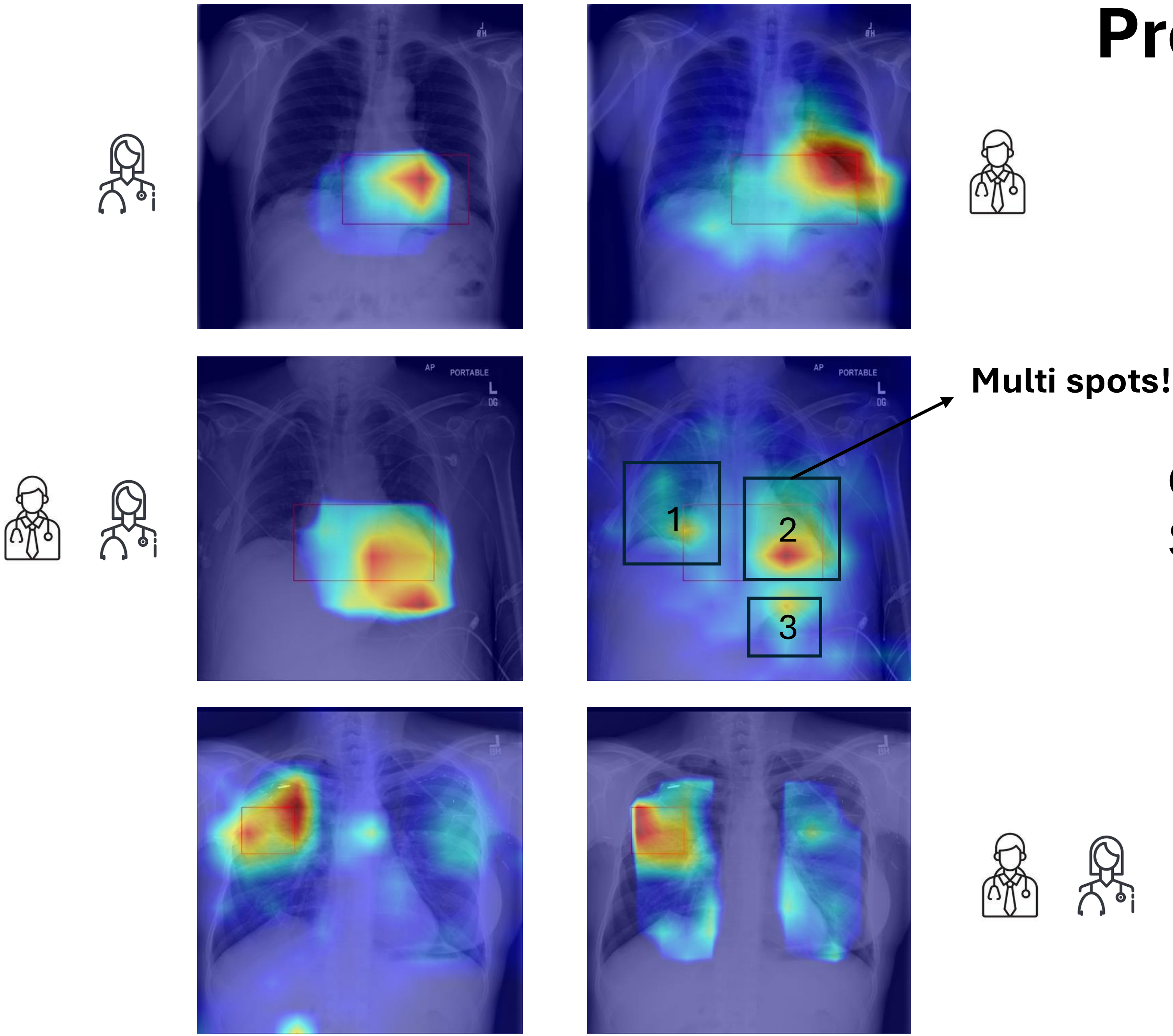
Which one is better?

Principle : Accuracy and Specificity



20 comparisons from ChestXDet dataset covering **Atelectasis**, **Cardiomegaly, Consolidation, and Effusion.** Clinicians preferred DWARF in 82.5% of cases.

Preference of clinicians?

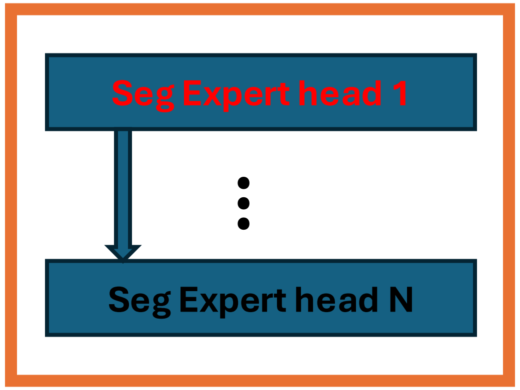


Compared to the Intersection,
Specificity is important!

Ablations

Segmentation expert heads

Method	Max DICE	Max AUC
Directly optimize	0.2288	0.8663
Introducing segmentation expert heads	**0.3559**	**0.8732**



Prompts

Method	Dataset	Max DICE	Max AUC
Finding name	ChestX-Det	**0.1805**	**0.8157**
Finding Description	ChestX-Det	0.1769	0.8125

Example:

- **Finding name:** Effusion
- **Description:** Excess fluid around the lungs

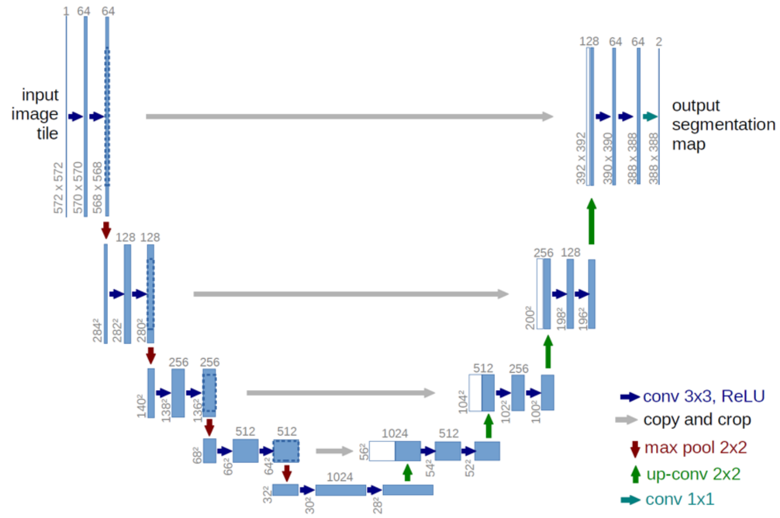
Segmentation teachers' performance

4 diseases

7 diseases

Split the 13 findings to 3 versions according to its performance and morphological differences

Finding name	Dice Score
Atelectasis	0.3031
Calcification	0.0073
Cardiomegaly	0.8342
Consolidation	0.4288
Diffuse Nodule	0.4441
Effusion	0.3525
Emphysema	0.4355
Fibrosis	0.1621
Fracture	0.0755
Mass	0.4583
Nodule	0.0355
Pleural Thickening	0.1355
Pneumothorax	0.0707

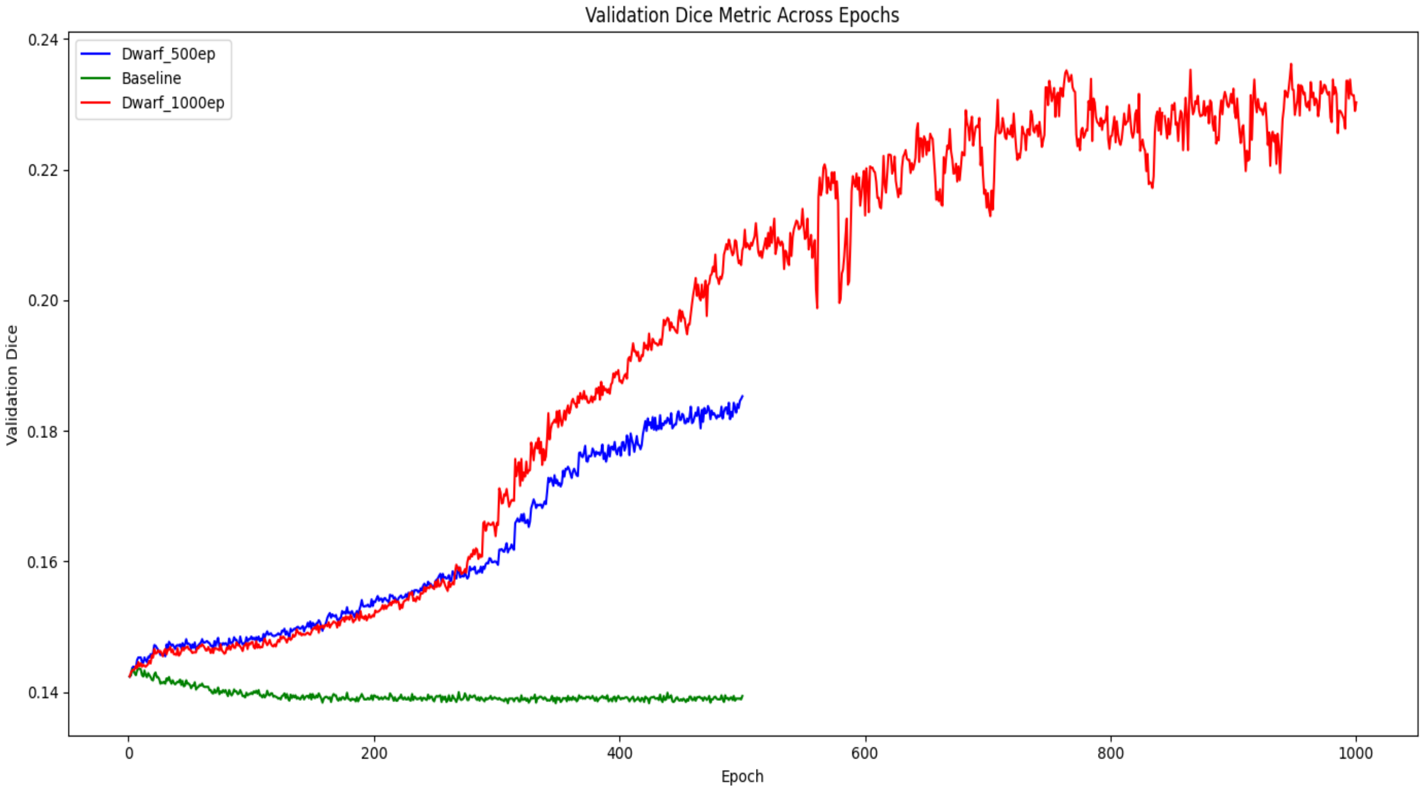


The segmentation teachers are trained with UNet architecture

Ablations about Training with segmentation model teachers

Method	Dataset	Disease numbers	Attention map DICE (Mean DICE across all diseases)	AUC	Max DICE	Max AUC
GAIN	ChestX-Det	7	0.1438	0.8680	0.1438	0.8680
DWARF (expert teachers)	ChestX-Det	7	0.3171	0.8473	0.3694	0.8757
DWARF	ChestX-Det	7	**0.3856**	**0.8578**	**0.3911**	**0.8766**

Ablations about the scalability



Method	Dataset	Epoch number	Finding numbers	DICE	AUC
DWARF	ChestX-Det	500	13	0.1805	0.8157
DWARF	ChestX-Det	1000	13	**0.2302**	**0.8231**

Contributions:

- A novel training framework that optimize **both classification** and **attention maps**;
- A **stable** and **scalable** framework which could also be optimized with **pseudo** labels;
- A performance that surpasses existing **SoTA pretrained/finetuned** baselines.
- **Throughout designed validation** to narrow the gap between DWARF and clinical application;

Limitations:

- More findings need to be validated on, currently we only evaluate around 20 findings in total;
- More baselines need to be finetuned with our method and validate;
- Samples for clinicians' preference test are limited.

Thanks for the listening!