DSL501

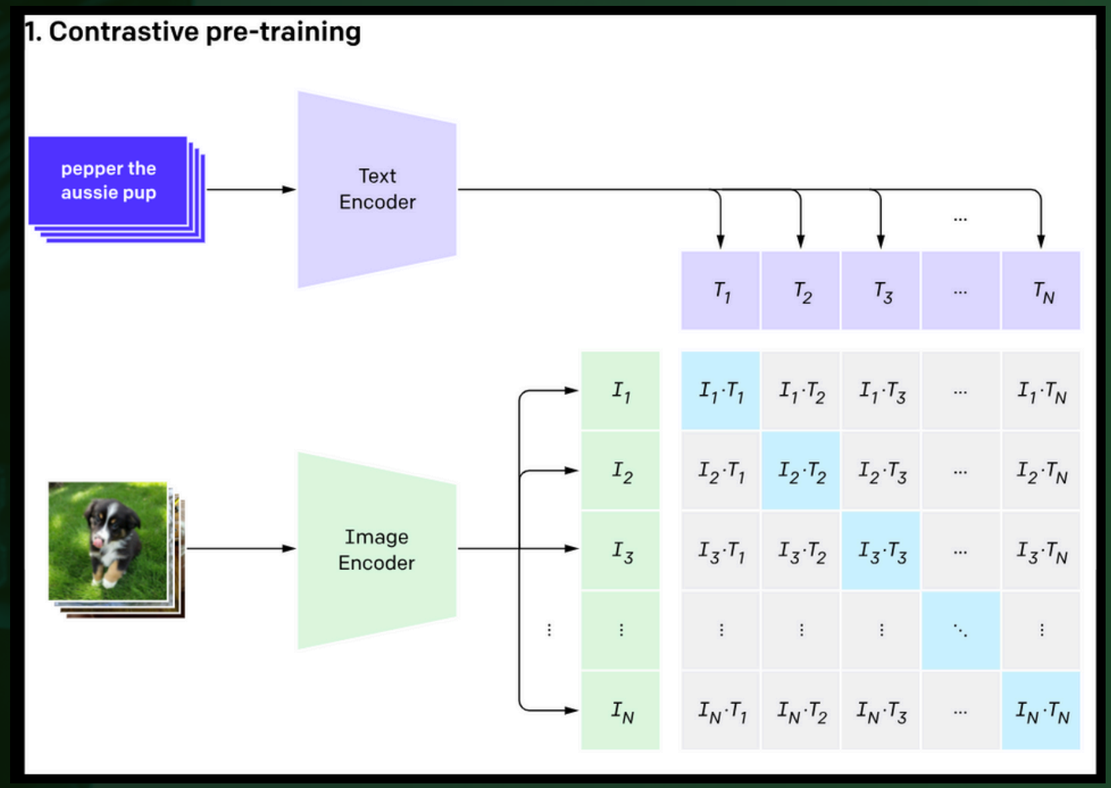
Region-Aware MedCLIP for Medical Image–Text Understanding

BY SHASHANK YADAV
3RD-YEAR B.TECH
DSAI

Introduction

OpenAi CLIP

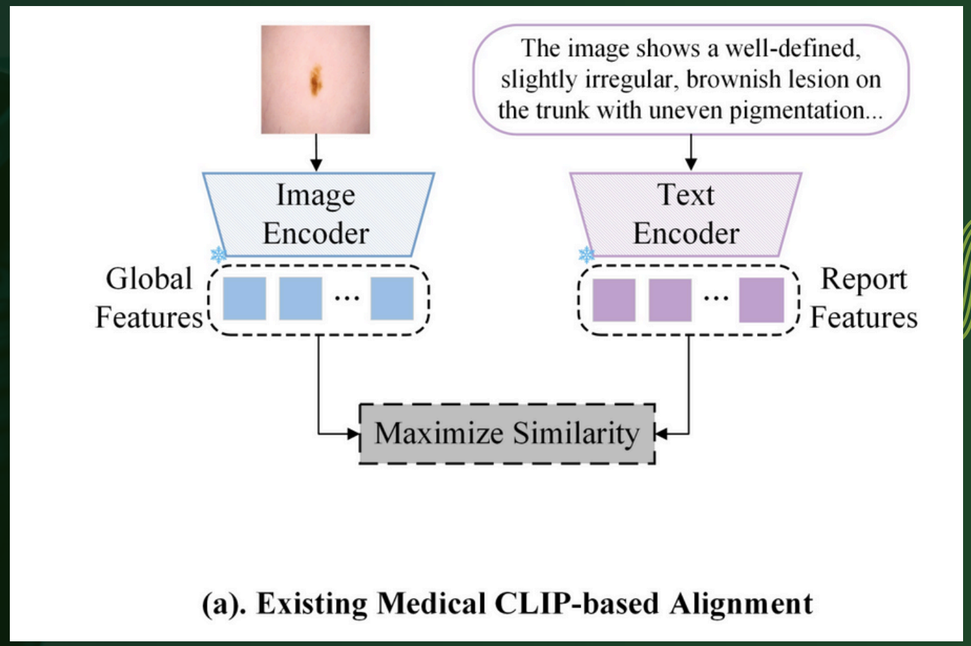
- What is Contrastive Language-Image Pre-Training?
- Why CLIP is good (large-scale, general knowledge)?



Problem

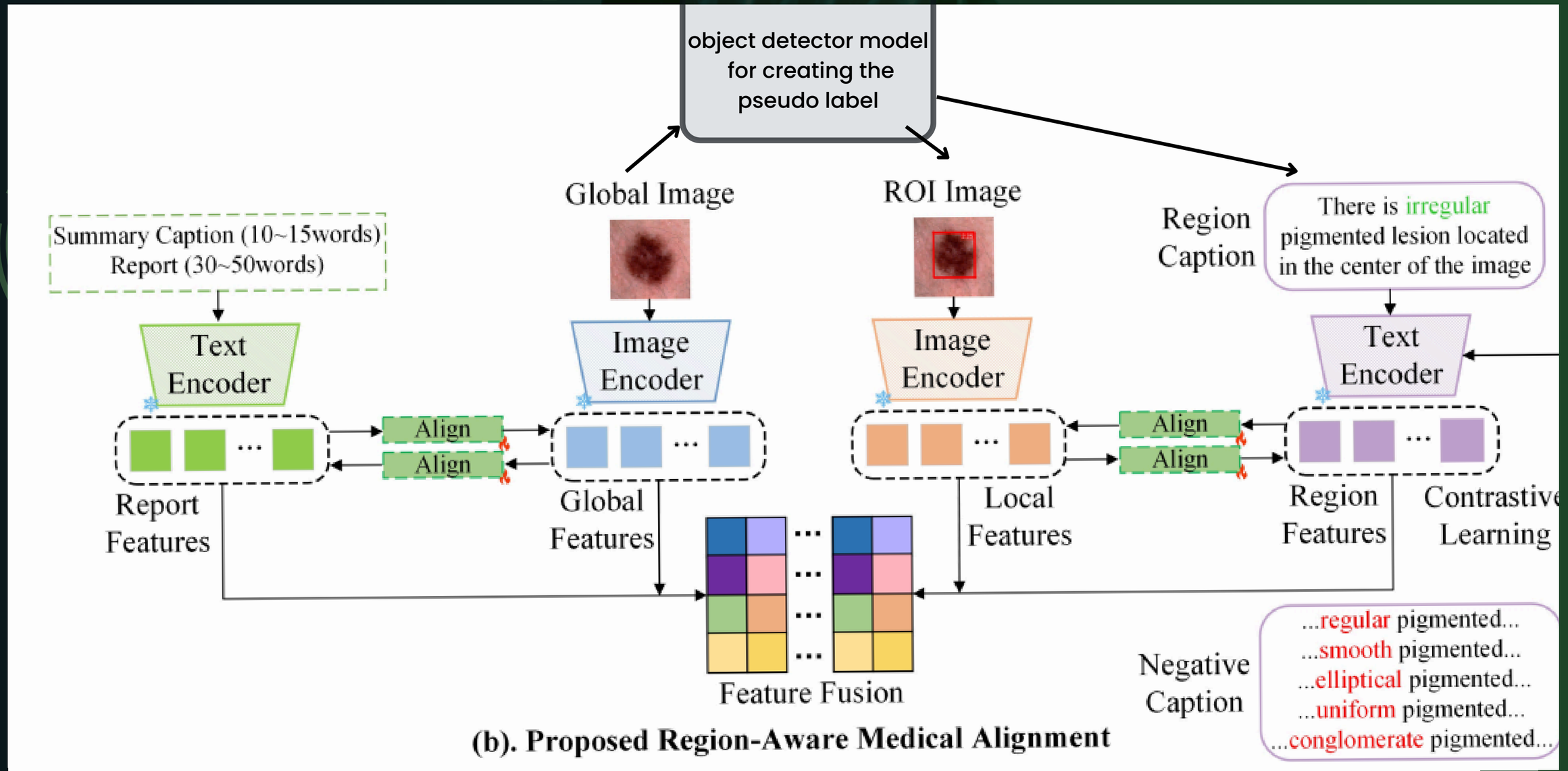
- Why CLIP is limited in medical imaging
- Medical images require fine-grained reasoning.
- Regions matter (lesions, nodules, abnormalities).

?



(a). Existing Medical CLIP-based Alignment

Proposed workflow



(b). Proposed Region-Aware Medical Alignment

TRAINING

DATA PREPARATION
ROI and REGION-based pseudo label through template

cvpr region clip paper
NIH Chest X-rays
or MIMIC

Microsoft's BIOMEDCLIP on our data for inference

Teacher model

Student Model (RegionMed-CLIP – Trainable)

Dual streams:
Global image → global features
ROI crops → local features
Feature Fusion: global + local
Outputs: similarity scores (ROI ↔ text, global ↔ report)

FORWARD PASS

- vg = global image features
- Vr =ROI image features

$zv = Wvf \quad , zt = Wtt$

$$s(zv, zt) = \frac{Zv \cdot Zt}{||zv||\ ||zt||}$$

BACK PROPOGATION

custom LOSS founction

LOSS Fn

Global Contrastive Loss (InfoNCE style):

$$L_{global} = -\log \frac{\exp(s(z_g, z_t^g)/\tau)}{\sum_j \exp(s(z_g, z_t^j)/\tau)}$$

ROI Contrastive Loss:

For each ROI–caption pair:

$$L_{ROI} = -\sum_i \log \frac{\exp(s(z_r^i, z_t^i)/\tau)}{\sum_j \exp(s(z_r^i, z_t^j)/\tau)}$$

Distillation (KL divergence):

$$L_{distill} = \sum_i KL(p_{teacher}(T|I_r^i) \parallel p_{student}(T|I_r^i))$$

Negative Caption Loss:

Push ROI embeddings away from mismatched captions:

$$L_{neg} = -\sum_i \log \frac{\exp(s(z_r^i, z_t^i)/\tau)}{\exp(s(z_r^i, z_t^i)/\tau) + \sum_{t_{neg}} \exp(s(z_r^i, z_{t_{neg}})/\tau)}$$

$$L_{total} = \alpha L_{global} + \beta L_{ROI} + \gamma L_{neg} + \lambda L_{distill}$$

where $\alpha, \beta, \gamma, \lambda$ control contributions (often set =1 except distill ~0.3–0.5).

Expected outcome

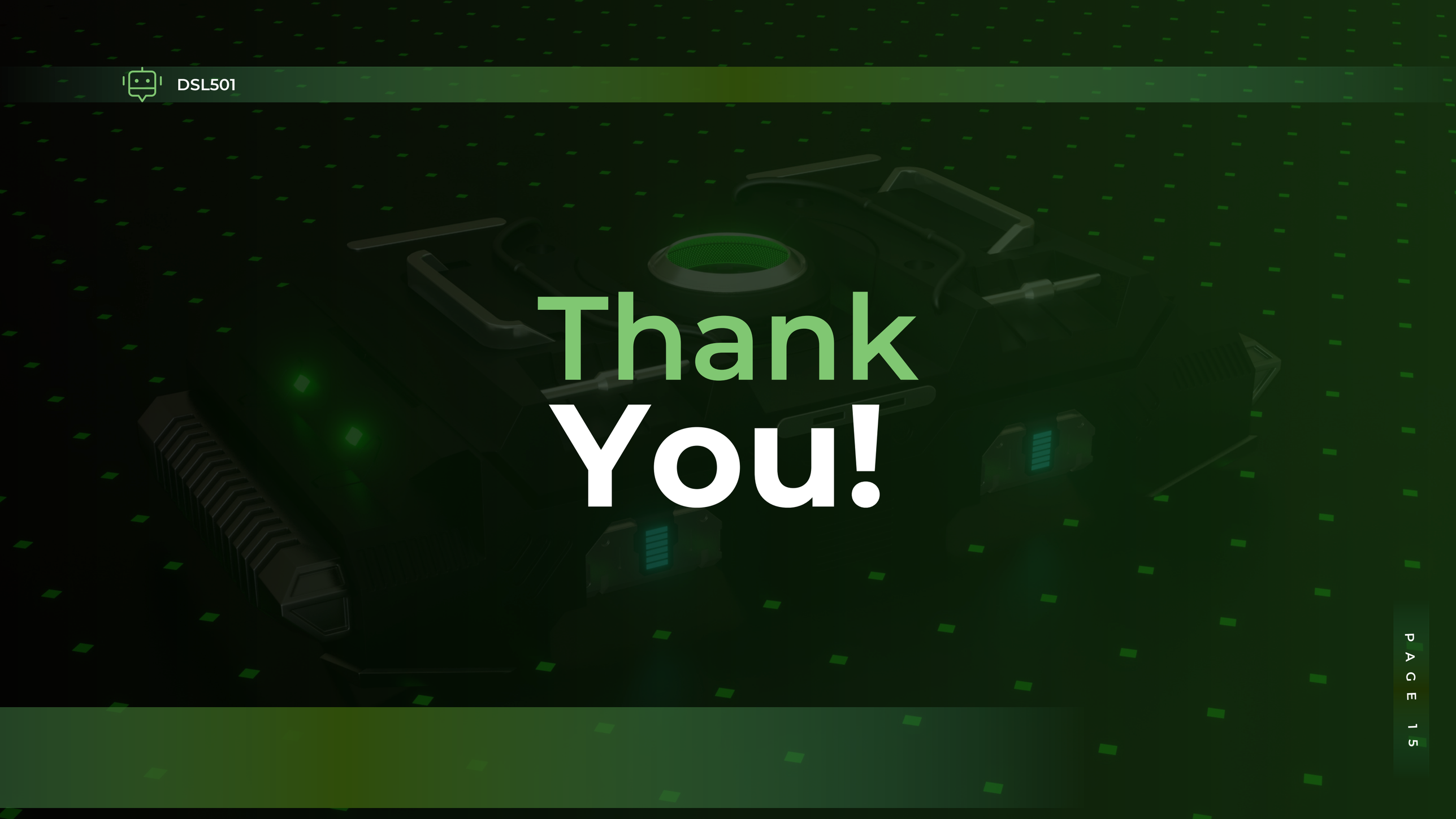
Zero-shot disease classification

- Higher AUROC & accuracy vs. CLIP/MedCLIP.

Image ↔ Report retrieval

Region ↔ Caption matching

Fine-grained lesion description retrieval.



Thank You!