Boosting Medical Visual Understanding from Multi-Granular Language Learning (ICLR 2026)

Zihan Li[*1] Yiqing Wang[*2] Sina Farsiu[†2] Paul Kinahan[†1]
[1]University of Washington [2]Duke University

[*]Equal Contribution [†]Co-corresponding Author

The Problem: CLIP's Single-Granularity Limitation

Standard CLIP

Single-label, single-granularity

Image → "Chest X-ray"

Problems:

- Ignores multi-label nature
- Loses hierarchical structure
- Poor feature transferability
- Clinically meaningless attention

MGLL (Ours)

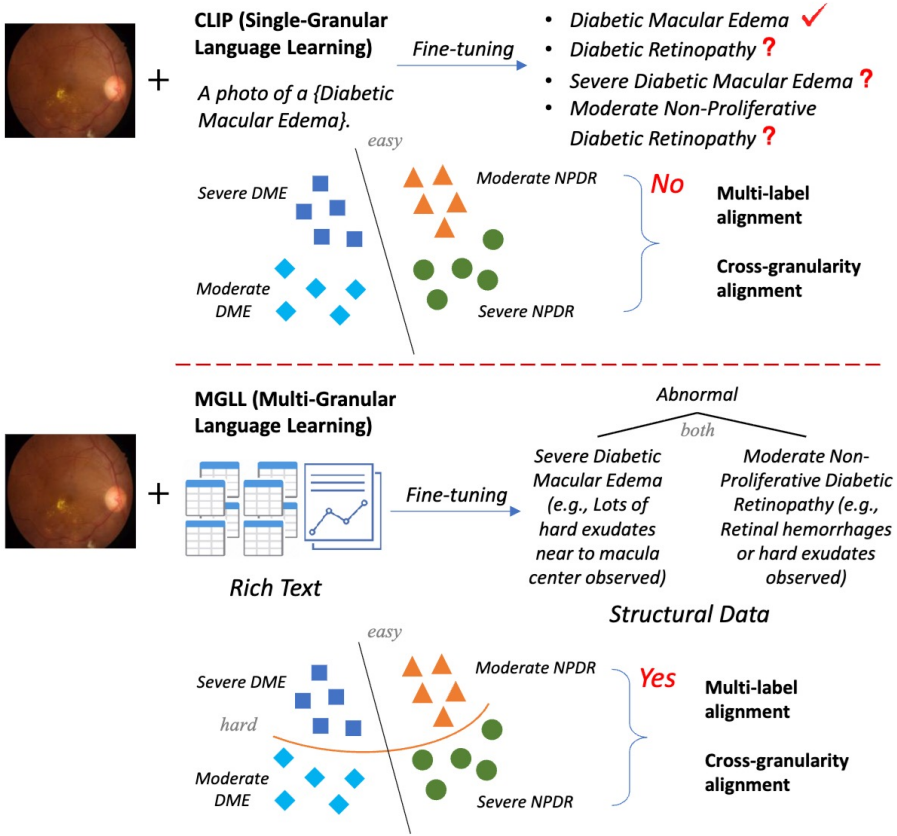
Multi-label, multi-granularity

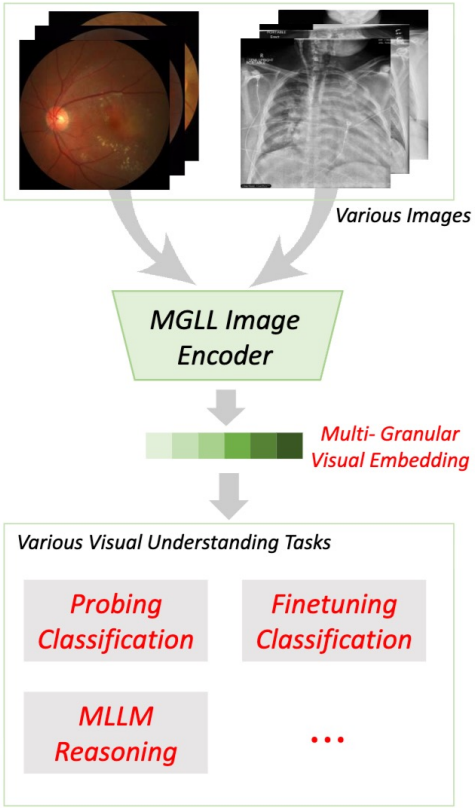
Granularity 1: "DX" (modality)
Granularity 2: "Chest X-ray" (study)
Granularity 3: "PA View" (series)

Advantages:

- Hierarchical alignment
- Plug-and-play, ~ 3% overhead

The Problem: CLIP's Single-Granularity Limitation





Multi-Granularity in Medical Imaging

Retinal Fundus (3 Granularities)

G1: Binary Status

"Abnormal"

G2: Disease Category

" Diabetic Retinopathy "

G3: Clinical Explanation

"Retinal hemorrhages and hard exudates observed in macula"

Chest X-ray (3 Granularities)

G1: Modality

"DX" (Digital Radiography)

G2: Study Description

"Chest X-ray"

G3: Series Description

"PA View" (posteroanterior)

All textual metadata extracted from DICOM headers — no manual annotation required

MGLL Framework: Three Key Components

Soft CLIP Loss

Extends contrastive loss with soft labels from label co-occurrence. At optimality, image feature → weighted centroid of text features.

+

Point-wise Constraint

Binary CE ensuring correct pairwise ranking between positive and negative image-text pairs at each granularity.

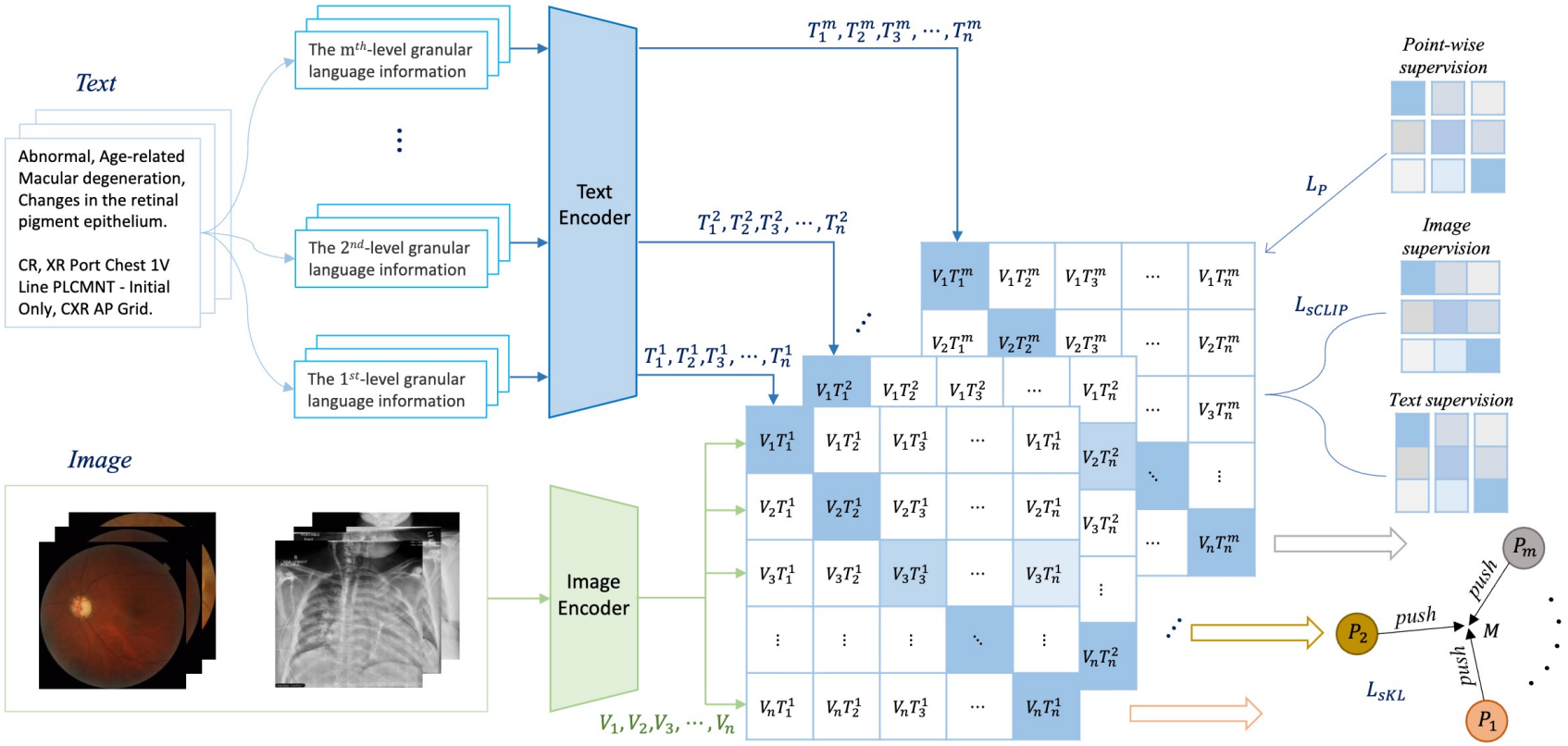
+

Smooth KL Divergence

Cross-granularity consistency: fine predictions must be compatible with coarse ones. Smooth KL avoids numerical instability.

Paper: https://arxiv.org/pdf/2511.15943
Code: https://github.com/HUANGLIZI/MGLL

MGLL Framework: Three Key Components



Key Theoretical Insight

At optimality under Soft CLIP Loss, the image feature converges to the weighted centroid of its associated text features:

$$V_i = \Sigma_k w_{ik} \cdot T_{ik} \, |\Sigma_k w_{ik} \cdot T_{ik}|$$

Why this matters:

- CLIP forces Visual representation to match a single text T — biased representation
- MGLL allows Visual embedding to sit at the semantic center of ALL its labels — balanced representation
- Weights w_{ik} ensure each label contributes proportionally to its clinical relevance
- This principled formulation explains why soft alignment is fundamentally more appropriate for multi-label medical domains

Large-Scale Multi-Granular Datasets

MGLL-Fundus

246,389
image-text pairs

- 49 public datasets aggregated
- 50+ disease categories
- 2 granularity levels
- Clinical explanations from EyeWiki

MGLL-Xray

190,882
X-ray images from MIDRC

- DICOM → PNG with metadata
- 3 granularity levels from headers
- CR + DX modality types
- Zero manual annotation needed

Results: State-of-the-Art Across 10+ Benchmarks

+11%
AUC on ADAM
(LP: 90% vs 79%)

+16%
AUC on RFMiD
(LP: 80% vs 63%)

99.1%
AUC on MIDRC-XR
(FT, vs 94% UniMed)

+34%
LLaVA-Med gain
with MGLL encoder

Dataset	CLIP	Best Baseline	MGLL (Ours)	Improvement
ADAM (LP AUC)	61.43%	79.33%	**90.02%**	**+10.69%**
REFUGE (LP AUC)	56.82%	84.59%	**92.42%**	**+7.83%**
MIDRC-XR (FT AUC)	88.52%	94.15%	**99.08%**	**+4.93%**
MIDRC-Portable (FT AUC)	91.83%	97.26%	**99.75%**	**+2.49%**
RFMiD Multi-label (LP AUC)	44.66%	63.38%	**79.62%**	**+16.24%**

Results: State-of-the-Art Across 10+ Benchmarks

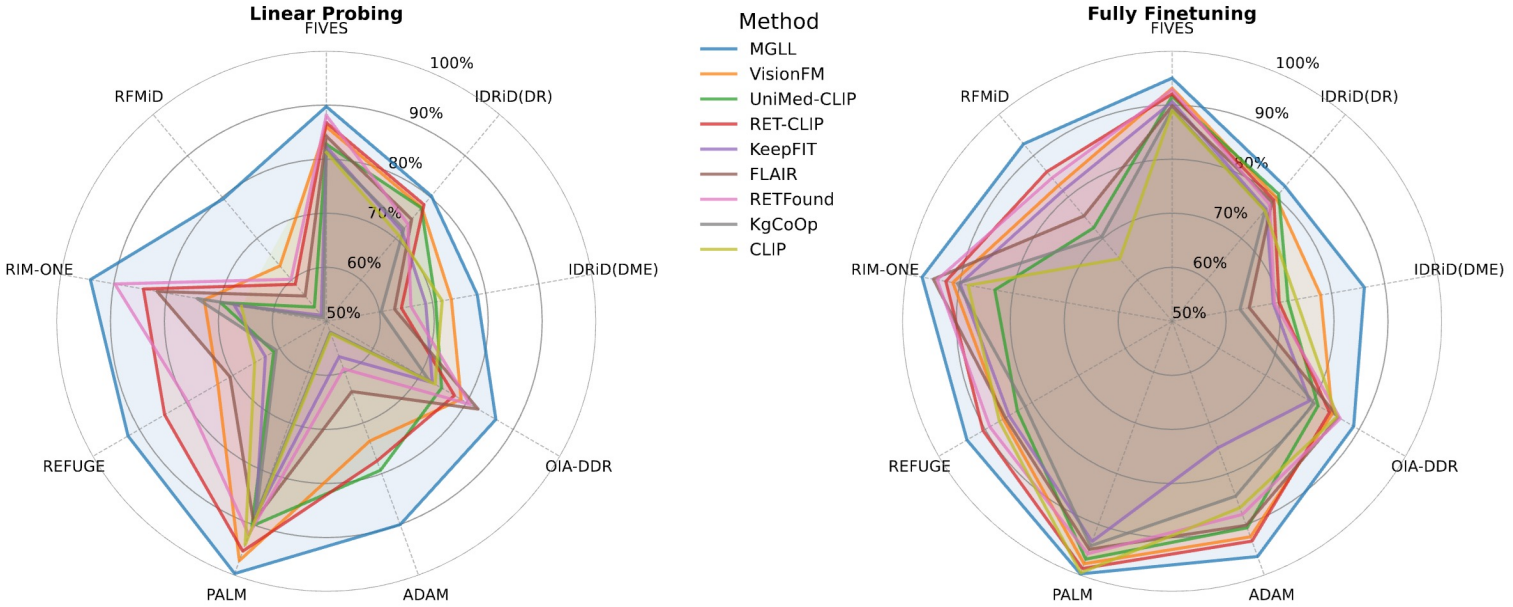


Figure 3: The quantitative comparison (AUC) between baseline methods and proposed MGLL on nine fundus downstream datasets.

Integration with Multimodal LLMs

Replacing CLIP encoders with MGLL in 7 different MLLMs consistently improved performance — even for medical-specialized models.

Up to **34.1% improvement** for LLaVA-Med with MGLL encoder swap

Key takeaway:

- MGLL's plug-and-play design means better visual features transfer directly to downstream models
- Works across architectures: LLaVA, Med-Flamingo, LLaVA-Med, and 4 others
- Even medical-specialized LLMs (trained on medical data) benefit from MGLL's multi-granular features
- No architectural changes needed — just swap the vision encoder

Integration with Multimodal LLMs

Table 2: Comparison of multiple-choice accuracy with MGLL in multimodal large language models on selected ten representative diseases.

Method	AMD	Cataract	CSR	DR	Glaucoma	Media Haze	Myopia	Retinitis	DME	Tessellation	**Average ↑**
InstructBLIP Dai et al. (2023)	80.17%	80.00%	0.00%	76.51%	59.30%	16.13%	44.25%	11.11%	63.79%	41.67%	47.29%
+ MGLL	**83.63%**	**85.00%**	**28.57%**	**82.55%**	**65.43%**	**45.16%**	**52.65%**	**44.44%**	**74.14%**	**58.33%**	**61.99%** (14.7% ↑)
Mini-Gemini Li et al. (2024)	76.61%	85.00%	14.29%	79.87%	67.90%	38.71%	58.41%	33.33%	60.34%	33.33%	54.78%
+ MGLL	**82.46%**	**85.00%**	**42.86%**	**84.56%**	**72.22%**	**58.06%**	**64.16%**	**55.56%**	**65.52%**	**41.67%**	**65.21%** (10.4% ↑)
Qwen-VL Bai et al. (2023)	81.87%	75.00%	28.57%	80.54%	78.40%	54.84%	76.55%	22.22%	84.48%	25.00%	60.75%
+ MGLL	**85.96%**	**80.00%**	**42.86%**	**89.93%**	**87.04%**	**70.97%**	**80.97%**	**33.33%**	**89.66%**	**41.67%**	**70.24%** (9.5% ↑)
InternVL Chen et al. (2024a)	81.29%	85.00%	71.43%	94.63%	89.51%	64.52%	88.05%	44.44%	87.93%	66.67%	77.35%
+ MGLL	**86.55%**	**90.00%**	**71.43%**	**96.64%**	**90.74%**	**67.74%**	**91.15%**	**55.56%**	**94.83%**	**75.00%**	**81.96%** (4.6% ↑)
LLaVA Liu et al. (2023)	83.04%	90.00%	42.86%	87.25%	91.36%	48.39%	88.50%	44.44%	93.10%	58.33%	72.73%
+ MGLL	**84.80%**	**90.00%**	**57.14%**	**93.96%**	**91.98%**	**61.29%**	**90.71%**	**66.67%**	**96.55%**	**66.67%**	**79.98%** (7.3% ↑)
LLaVA-Med Li et al. (2023)	16.37%	15.00%	42.86%	26.85%	25.31%	25.81%	23.89%	33.33%	16.67%	16.67%	24.28%
+ MGLL	**58.48%**	**65.00%**	**57.14%**	**77.18%**	**59.26%**	**51.61%**	**57.08%**	**44.44%**	**55.17%**	**58.33%**	**58.37%** (34.1% ↑)
Med-Flamingo Moor et al. (2023)	25.73%	30.00%	57.14%	36.91%	24.07%	22.58%	18.58%	22.22%	24.14%	8.33%	26.97%
+ MGLL	**69.01%**	**75.00%**	**71.43%**	**80.54%**	**61.11%**	**54.84%**	**45.58%**	**44.44%**	**51.72%**	**33.33%**	**58.70%** (31.7% ↑)
Janus-Pro Chen et al. (2025)	88.30%	75.00%	42.86%	93.29%	90.74%	58.06%	87.17%	33.33%	62.07%	58.33%	68.92%
+ MGLL	**90.64%**	**85.00%**	**71.43%**	**96.64%**	**95.06%**	**67.74%**	**90.27%**	**55.56%**	**70.69%**	**75.00%**	**79.80%** (10.88% ↑)

Connecting MGLL to Clinical Research

Clinical Foundation Models

Multi-granular alignment for clinical notes

focus: Foundation models for clinical text using private unique data

Multimodal Disease Prediction

MGLL handles multi-label conditions at different stages of progression naturally

focus: Predicting disease onset from multimodal data with uncertainty

Data Harmonization at Scale

DICOM/LOINC harmonization as multi-granular alignment — directly transferable to EHR unification

focus: Unifying data across millions of patients and providers

MGLL is plug-and-play, scalable, and designed for real clinical workflows

Summary

Problem:
CLIP's single-granularity alignment fails for complex medical images

Solution:
MGLL — multi-label + cross-granularity alignment via 3 complementary losses

Impact:
SOTA on 10+ benchmarks, 34% MLLM improvement, clinically grounded features

Key Properties:
Architecture-agnostic, plug-and-play, minimal computational overhead

Zihan Li | University of Washington | Imaging Research Laboratory