Dynamic Decision Learning

Test-Time Evolution for Abnormality Grounding in Rare Diseases

Jun Li[1,2], Mingxuan Liu[3], Jiazhen Pan[1,2], Che Liu[4], Wenjia Bai[4], Cosmin I. Bercea[1,2], Julia A. Schnabel[1,2,5,6]

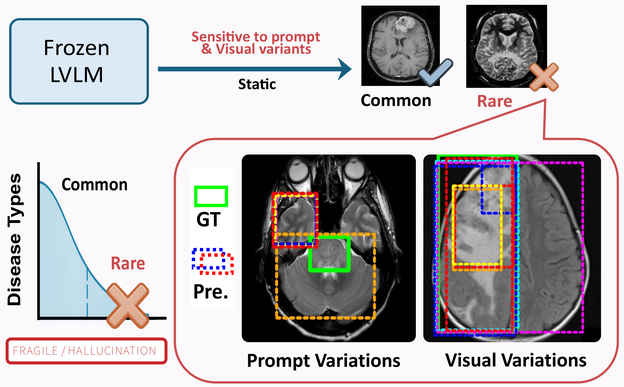
ICML International Conference On Machine Learning

Without finetuning, how can frozen LVLMs improve abnormality grounding in long-tailed clinical scenarios?

Munich Center for Machine Learning

1 Static inference breaks on rare diseases

- Rare diseases suffer from severe data scarcity.
- Frozen LVLM grounding is sensitive to prompt wording and mild visual perturbations.
- Hallucinated detections drift across views.



2 Dynamic Decision Learning (DDL)

Language Path: ***Distribution-Aware Prompt Evolution***

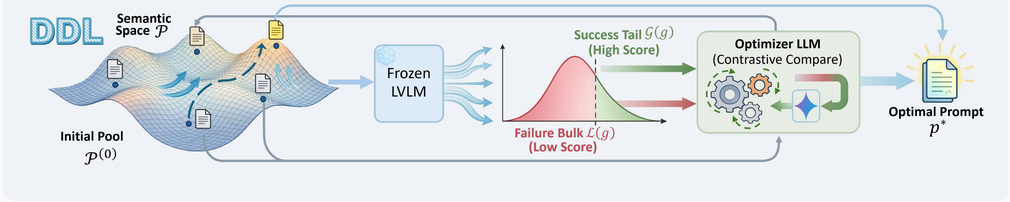
Instruction Seeding
Build initial $P^{(0)}$ from vanilla prompt.

Performance History
Maintain $H^{(g)} = \{(p_i, y_i)\}$

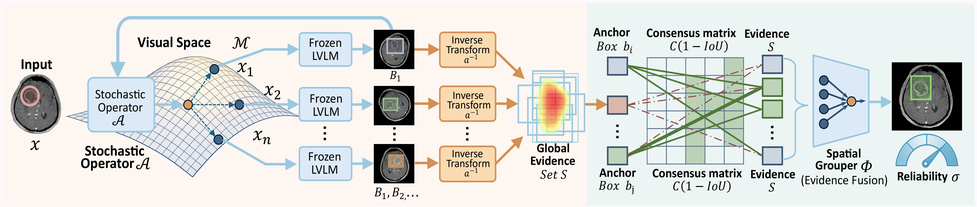
Success–Failure Partition
Split candidates by threshold y^* : success $G^{(g)}$ and failure $L^{(g)}$.

Contrastive Refinement
Optimizer LLM Ψ compares $G^{(g)}$ vs. $L^{(g)}$ to generate $p^{(g+1)}$.

Prompt Convergence
Iterate until the success tail converges to the optimized prompt p^*.



Visual Path: ***Visual-Perception Uncertainty Probing*** ***Referenced Hungarian Consolidation***



Visual Perturbation
Generate different spatial views $x_m = a_m(x)$.

Multi-View Grounding
Predict $B_m = f_\theta(x_m, p^*)$ across views.

Evidence Back-Projection
Map $\hat{B}_m$ to the original coordinate space.

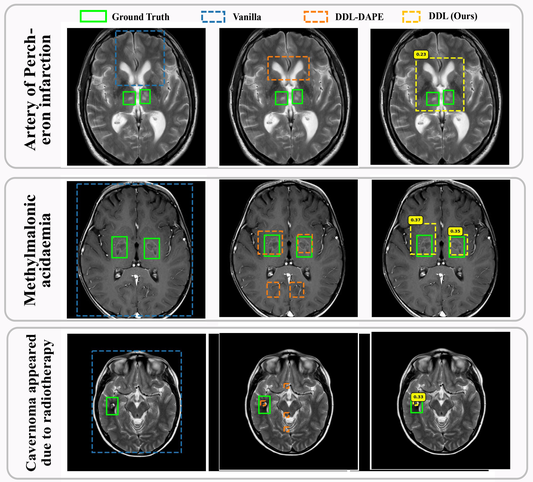
Hungarian Consolidation
Match projected boxes to B_{ref} via $-\text{IoU}$.

Reliability Scoring
Estimate σ_j by consensus and consistency.

Consolidated Output
Return reliable detections $B^* = \{(b_j, \sigma_j)\}$.

3 What DDL delivers

- Both language and visual improve grounding.
- Up to **105%** improvement on mAP@75.



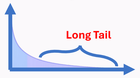
A Benchmarks and Setup

NOVA (rare / long-tail)

- 281 rare pathology types
- 100 development samples
- 806 evaluation samples
- heterogeneous MRI protocols

BTD (common pathologies)

- 3 tumor types
- 100 development samples
- 293 evaluation samples
- comparative in-distribution



Metric: mAP@25, mAP@50, mAP@75
Setup: Backbones Qwen2.5-VL 3B / 7B / 32B / 72B

B Quantitative Gains (mAP@75)

DDL enhances mAP@75 **on NOVA by 65% to 105%** across all model sizes, demonstrating its robustness.

Model Size	Vanilla	DDL (Ours)	Gain
3B	0.040	0.066	+66%
7B	0.036	0.075	+105%
32B	0.058	0.096	+66%
72B	0.065	0.107	+65%

DDL also demonstrates notable progress on **the BTD dataset**, where the baseline performance is already relatively high.

Model Size	Vanilla	DDL (Ours)	Gain
32B	0.160	0.206	+28.7%
72B	0.306	0.346	+13.1%

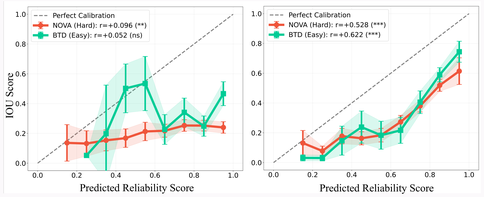
DDL outperforms SFT on NOVA-3B, particularly in long-tail cases.

Method	mAP@25	mAP@50	mAP@75
SFT	0.283	0.128	0.046
DDL (Ours)	**0.298**	**0.150**	**0.066**

DDL delivers robust, scalable long-tail improvements across models and datasets, without any parameter training.

C Calibration Analysis

Calibration strengthens with model scale: small LVLMs align on hard cases, while large LVLMs align across tasks.



[1] Technical University of Munich [2] Munich Center for Machine Learning [3] University of Trento [4] Imperial College London [5] Helmholtz Munich [6] King's College London