Benchmarking Foundation Models for Medical Imaging with Contextual Integration of DICOM Metadata

Nirosh Sivanesan

Supervisor(s): Prof. Dr. Mauricio Reyes, Luo Haozhe
Institution(s): University of Bern, ARTORG Center for Biomedical Engineering Research

Introduction

Medical vision–language models (VLMs) promise open-vocabulary radiology by aligning chest X-rays (CXRs) with free-text reports, enabling flexible zero-shot querying beyond fixed label sets. Yet, despite strong in-distribution results, current CXR VLMs often generalize poorly across institutions. A key obstacle is that radiographs encode not only pathology but also the acquisition process like view geometry (AP/PA), exposure regime, detector response, and vendor-specific post-processing, which varies systematically across sites. When acquisition signatures correlate with clinical labels, contrastive pretraining can exploit them as stable shortcuts, yielding representations where clinical semantics and acquisition "style" are entangled and brittle under domain shift.

Materials and Methods

We pretrain a CXR VLM on MIMIC-CXR and evaluate **zero-shot multi-label** transfer with a fixed prompt set using macro-averaged AUROC across multiple large-scale public cohorts. We introduce **KAT-InfoNCE (Kernel-Adjusted Topology)**, a metadata-conditioned contrastive objective that uses DICOM acquisition parameters only at training time to reshape batch interactions while preserving metadata-free deployment. KAT-InfoNCE couples (i) soft-target attraction, which allocates limited probability mass to clinically overlapping neighbors to reduce multi-label false negatives, and (ii) semantic-first kernel-adjusted repulsion, which discounts technically confounded negatives in the softmax denominator via a hierarchical kernel that prioritizes pathology overlap and bounds acquisition influence. We additionally evaluate an optional query-gated cross-attention module for prompt-specific feature selection.

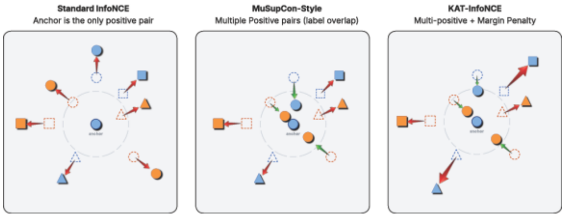
Standard InfoNCE
Anchor is the only positive pair
anchor
MuSupCon-Style
Multiple Positive pairs (label overlap)
anchor
KAT-InfoNCE
Multi-positive + Margin Penalty
anchor

Fig. 1 KAT-InfoNCE overview (kat_schema_3panel): training-time metadata conditions contrastive interactions via soft-target attraction and kernel-adjusted repulsion, while inference remains metadata-free.

Results

Across the benchmark, KAT-InfoNCE improves zero-shot transfer and yields the most consistent external performance among evaluated variants. In particular, it improves macro AUROC over the SOTA baseline CARZero across all evaluated **cohorts**, demonstrating robustness gains that transfer beyond the training distribution. Stratified robustness analysis further shows reduced sensitivity to a clinically meaningful acquisition regime, decreasing the AP-PA AUROC gap on NIH ChestX-ray14 by 28.9%.

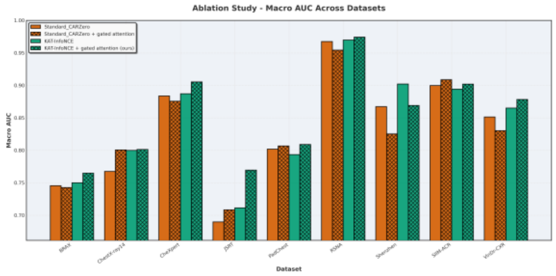

Fig. 2 External-cohort ablation (ablation_bar_plot.png): KAT-InfoNCE (and the full variant with gated attention) improves macro AUROC consistently over the CARZero InfoNCE baseline across cohorts.

Discussion

These results support functional robustness: training-time physics supervision can mitigate acquisition-driven shortcut learning without requiring metadata at inference. By reshaping contrastive competition by reducing pressure to separate samples primarily by acquisition "style", KAT-InfoNCE stabilizes performance under protocol shift while preserving the deployment simplicity of standard VLMs. Future work should extend robustness evaluation to broader acquisition factors (vendor, exposure quantiles) and address calibration under prevalence shift.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Dr. Mauricio Reyes and Luo Haozhe for their guidance, shared discussions, and supervision throughout this thesis.

MSc Artificial Intelligence in Medicine

UNIVERSITÄT BERN