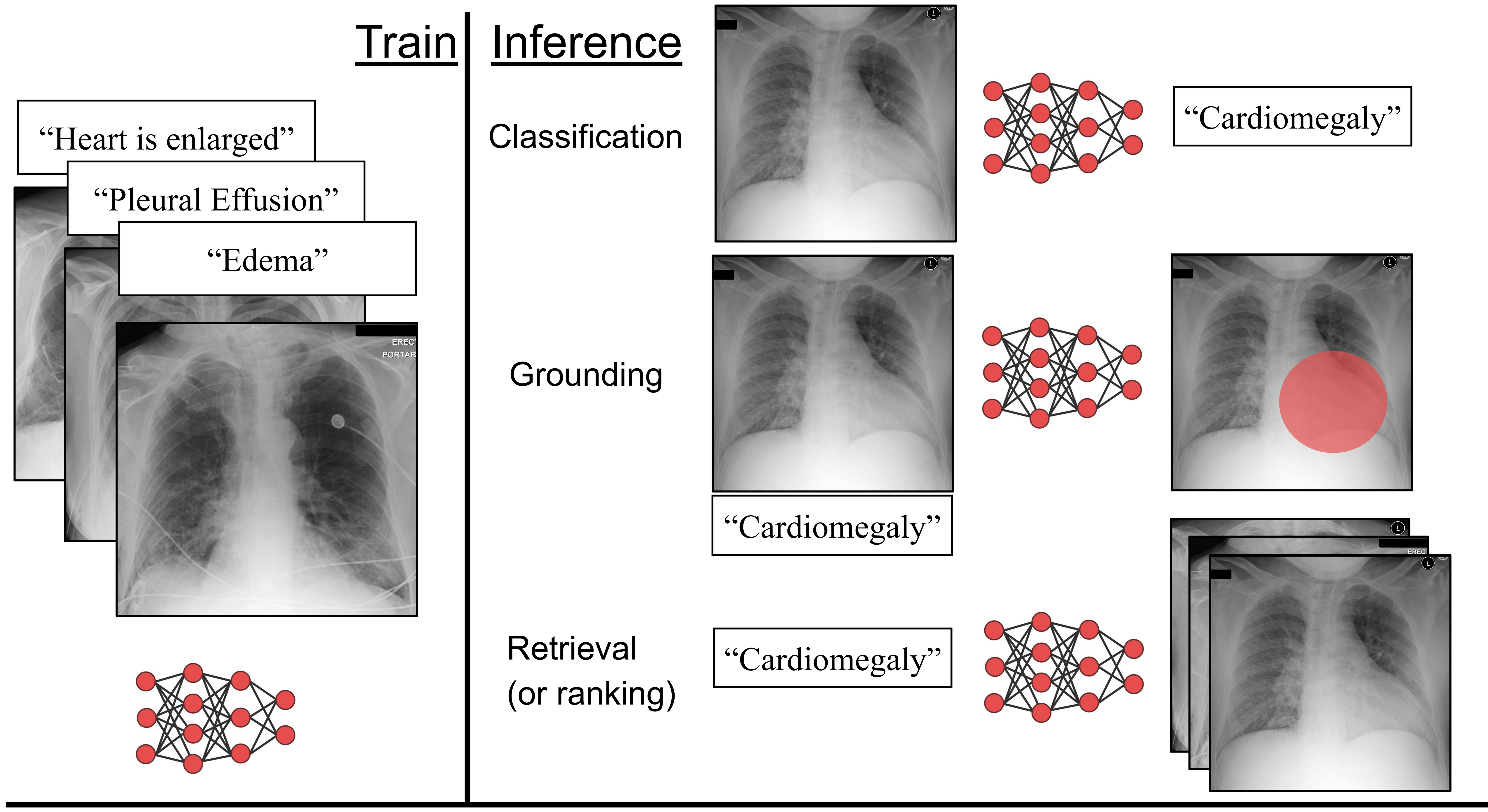
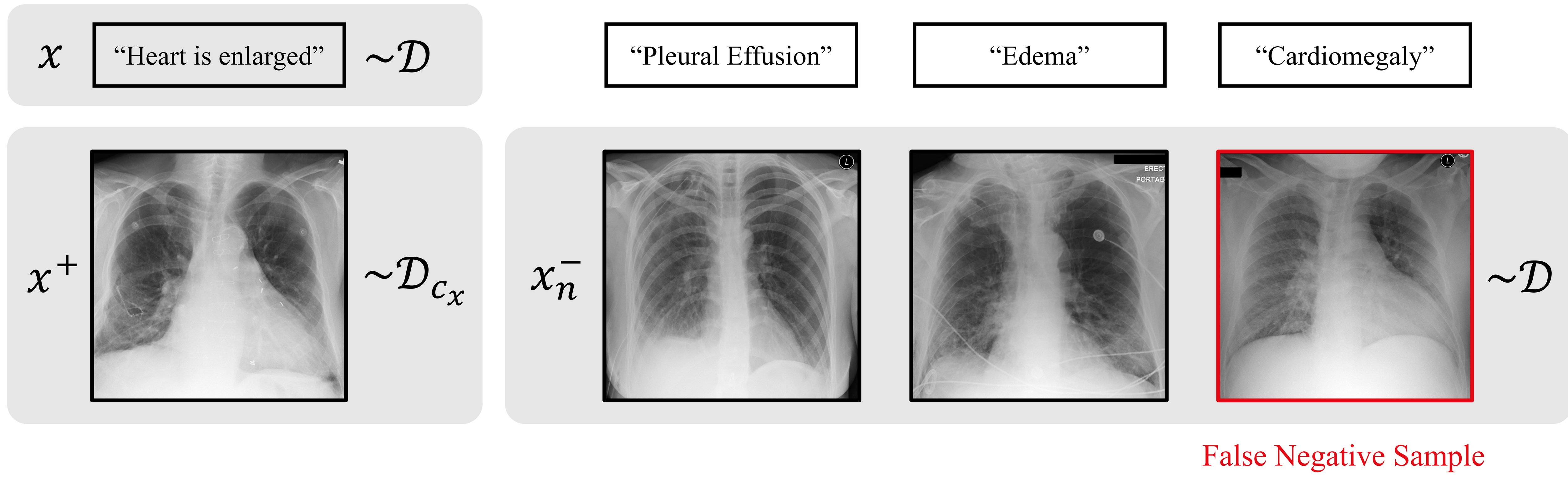
Sample-Specific Debiasing for Better Image-Text Models

Peiqi Wang, Yingcheng Liu, Ching-Yun Ko, William M. Wells, Seth Berkowitz, Steven Horng, and Polina Golland

Setting: Use Contrastive Learning to Train Image-Text Models





$$\mathbb{E}_{(x,x^+)\sim\mathcal{D}_{\text{sim}},\{x_n^-\}\sim\mathcal{D}^N}\left[-\log\frac{e^{s(x,x^+)}}{e^{s(x,x^+)}+\sum_{n=1}^N e^{s(x,x_n^-)}}\right]$$

Problem: False negative samples occur at an **uneven** rate *(e.g., common classes -> more potential for false negatives)* and degrade the performance of image-text models

Method: *Sample-specific* debiasing (removes effect of false negatives) that takes into account relative frequency of latent classes.

	Debiased Contrastive Learning (DCL) [Chuang et al. 2020]	Sample-specific Debiased Contrastive Learning (Ours)
Ideal Loss $\bar{\mathcal{L}}$	$\mathbb{E}_{(x,x^+)\sim\mathcal{D}_{\text{sim}}}\left[-\log\frac{e^{s(x,x^+)}}{e^{s(x,x^+)}+N\,\mathbb{E}_{x^-\sim\mathcal{E}_{c_x}}\left[e^{s(x,x^-)}\right]}\right]$	
Estimator $g(x,\{u_n\},\{v_m\};\eta)$ for $\mathbb{E}_{x^-\sim\mathcal{E}_{c_x}}\left[e^{s(x,x^-)}\right]$	$\frac{1}{1-\eta}\left[\frac{1}{N}\sum_{n=1}^{N}e^{s(x,u_n)}\right]-\frac{\eta}{1-\eta}\left[\frac{1}{M}\sum_{m=1}^{M}e^{(x,v_m)}\right]$	$\frac{1}{1-\eta(x)}\left[\frac{1}{N}\sum_{n=1}^{N}e^{s(x,u_n)}\right]-\frac{\eta(x)}{1-\eta(x)}\left[\frac{1}{M}\sum_{m=1}^{M}e^{(x,v_m)}\right]$
Computable Loss $\mathcal{L}$	$\mathbb{E}_{(x,x^+)\sim\mathcal{D}_{\text{sim}},\{u_n\}\sim\mathcal{D}^N,\{v_m\}\sim\mathcal{D}^M_{c_x}}\left[-\log\frac{e^{s(x,x^+)}}{e^{s(x,x^+)}+Ng(x,\{u_n\},\{v_m\};\eta)}\right]$	

Legend

c_x : Class label of x

ρ : Distribution over latent classes

$\mathcal{E}_{c_x}(x')\triangleq p(x'\mid c\neq c_x)=\sum_{c\neq c_x}\frac{\rho(c)}{1-\rho(c_x)}\mathcal{D}_c(x')$: Truly negative data distribution (w.r.t. x)

$\mathcal{E}_{c_x}(x')=\frac{1}{1-\rho(c_x)}\mathcal{D}(x')-\frac{\rho(c_x)}{1-\rho(c_x)}\mathcal{D}_{c_x}(x')$ implies $\mathcal{E}_{c_x}$ can be readily sampled

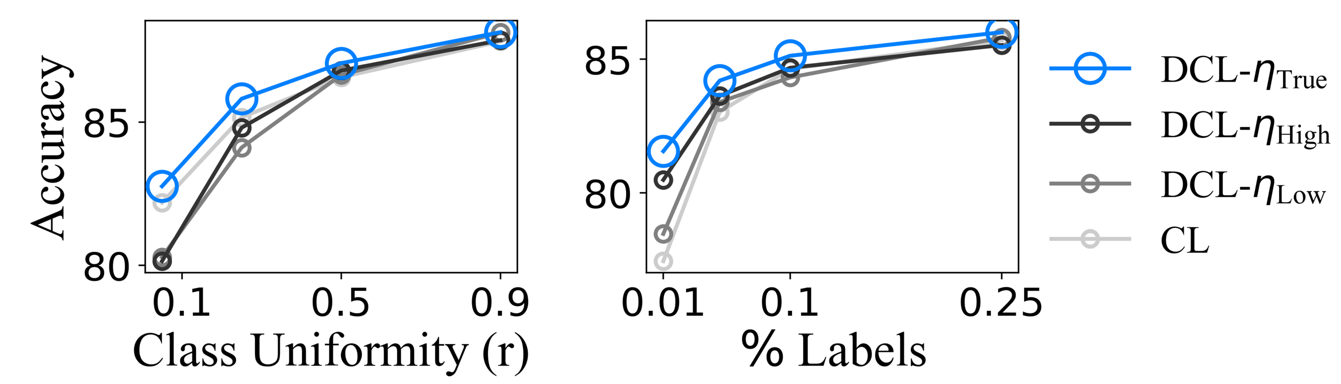
$$|\mathcal{L}-\bar{\mathcal{L}}|\le\frac{3e^2\sqrt{\pi/2}}{\sqrt{N}}\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{1-\rho(c_x)}\right]+\frac{3e^2\sqrt{\pi/2}}{\sqrt{M}}\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{\rho(c_x)}{1-\rho(c_x)}\right]+3e^2\mathbb{E}_{x\sim\mathcal{D}}\left[\left|\frac{1}{1-\eta(x)}-\frac{1}{1-\rho(c_x)}\right|\right],$$

Approximation Error due to
(1) Finite sample approximation using estimator g
(2) Misspecification of the sample-specific class probability function η

Language Model Estimate of Class Probability: We employ the likelihood of the text $p_{\text{LM}}(x)$ provided by a language model (LM) for text x to construct the estimate $\eta(x)$ of the class probability $\rho(c_x)$

$$\eta_{\text{LM}}(x)=a\cdot p_{\text{LM}}(x)^k$$

Results (CIFAR10)

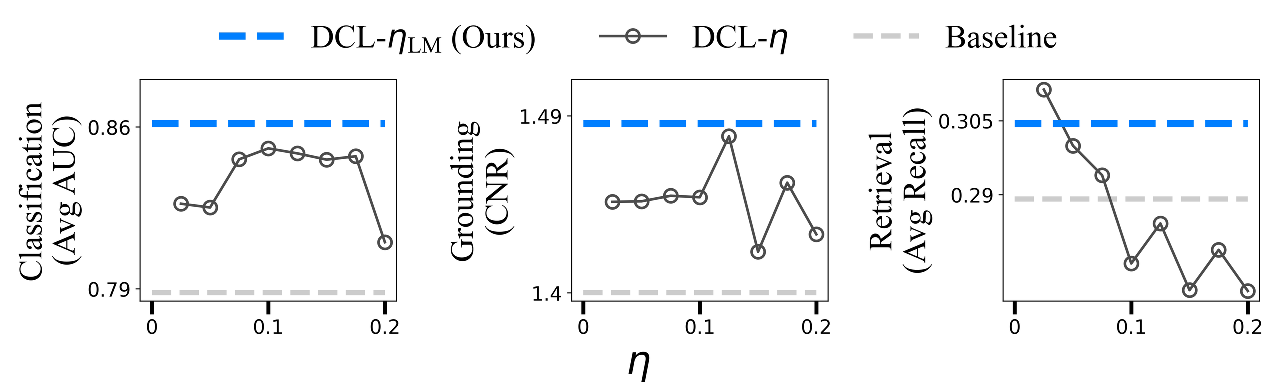


Debiased contrastive learning with true class probabilities ($\text{DCL-}\eta_{\text{True}}$) consistently outperforms baseline (CL) and DCL with misspecified class probabilities ($\text{DCL-}\eta_{\text{Low}}$, $\text{DCL-}\eta_{\text{High}}$).

Synthetic Data

- For each r, generate **CIFAR10-r** as subset of CIFAR10 data set as follows: For 5 selected classes, include random r fraction of images, while keep all images for each of the remaining 5 classes.
- The CIFAR10-r data set has a class probability of $\eta_{\text{Low}}=0.2r/(1+r)$ for each of the 5 selected classes and $\eta_{\text{High}}=0.2/(1+r)$ the remaining classes.

Results (MIMIC-CXR)



Optimal correction factor η varies across downstream tasks, making it challenging to pick one that works for all downstream tasks.

Our method ($\text{DCL-}\eta_{\text{LM}}$) outperforms debiased contrastive learning that applies a fixed amount of correction η to all samples ($\text{DCL-}\eta$) and LSE+NL that does not correct for the false negatives (Baseline).

Method		Classification AUC↑	Classification ACC↑	Grounding CNR↑	Grounding mIoU↑	Retrieval Recall↑	Retrieval MedR↓
No Correction	BioViL	0.78	0.62	1.14	0.17	0.25	148
No Correction	LSE+NL	0.79	0.65	1.40	0.19	0.29	115
Remove	by CheX5 Labels	**0.86**	0.71	1.37	0.19	0.29	113
Resample	by Text Sim.	0.82	0.71	1.37	0.19	**0.30**	111
Remove	by Text Sim.	0.84	**0.72**	1.39	0.19	**0.30**	112
Reweight	by Text Sim.	0.84	0.69	1.40	0.19	0.29	113
DCL-η w/ $\eta=0.05$		0.80	**0.72**	1.46	0.19	**0.30**	**104**
DCL-η w/ $\eta=0.1$		0.85	**0.72**	1.45	0.19	0.29	111
DCL-η_{LM} (Ours)		**0.86**	**0.72**	**1.49**	**0.20**	**0.30**	**104**

Debiased contrastive learning with sample-specific class probability η_{LM} ($\text{DCL-}\eta_{\text{LM}}$) outperforms state- of-the-art baseline methods BioViL and LSE+NL (no false negative correction) and alternative approaches to false negative correction.

Methods that attempt to remove, resample, or reweight likely false negative examples can improve classification performance but offer minimal improvement for and retrieval tasks. We hypothesize that this is because the fine-grain classes that must be also handled for the latter two tasks make identifying false negative samples more error-prone.

Evaluation Tasks

- **Image Classification** on RSNA Pneumonia: zero-shot or fine-tuned with 1% or 100% of labels; ACC is accuracy; AUC is area under curve.
- **Visual Grounding** on MS-CXR: CNR or contrast-noise ratio is a measure of discrepancy between region-sentence scores (heatmap) inside vs. outside the bbox; mIoU measures how well the thresholded heatmap overlap with the bbox.
- **Cross-Modal Retrieval** on MS-CXR: R@K is the fraction of times the correct item was found in the top K results; MedR is the median rank of the correct item in the ranked list.

This work was supported by NIH NIBIB NAC P41EB015902, Philips, Wistron, and MIT Lincoln Laboratory.