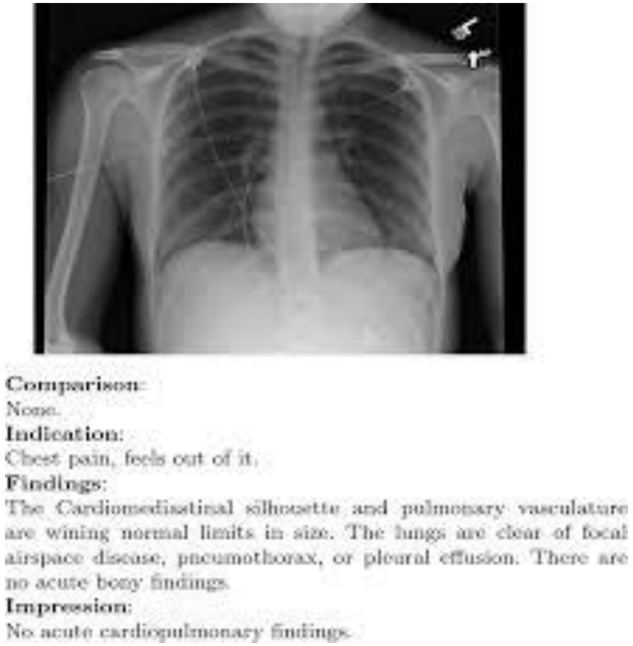
Distribution-Based Masked Medical Vision-Language Model Using Structured Reports

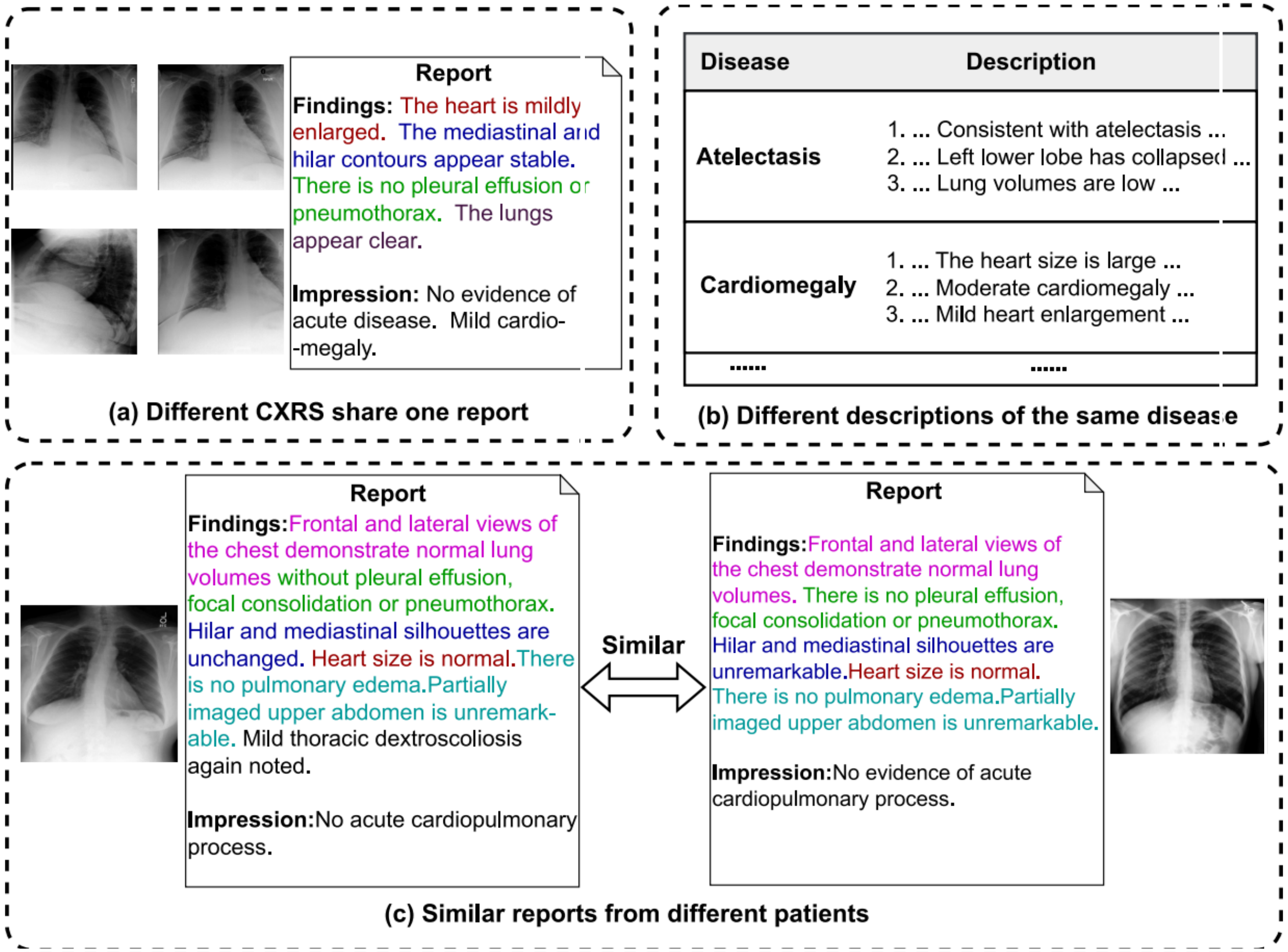
SHREYANK N GOWDA, RUICHI ZHANG, XIAO GU, YING WENG, LU YANG

Motivation

- Vision-language pretraining has been successful in natural images
- Medical images pose different challenges:
  - Labels are costly, disease distribution is imbalanced, medical language is unique
- Growing interest in vision-language pretraining for medical image analysis



Motivation


Report
Findings: The heart is mildly enlarged. The mediastinal and hilar contours appear stable. There is no pleural effusion or pneumothorax. The lungs appear clear.
Impression: No evidence of acute disease. Mild cardio--megaly.
(a) Different CXRS share one report
Disease
Description
Atelectasis
1. ... Consistent with atelectasis ...
2. ... Left lower lobe has collapsed ...
3. ... Lung volumes are low ...
Cardiomegaly
1. ... The heart size is large ...
2. ... Moderate cardiomegaly ...
3. ... Mild heart enlargement ...
......
......
(b) Different descriptions of the same disease
Report
Findings:Frontal and lateral views of the chest demonstrate normal lung volumes without pleural effusion, focal consolidation or pneumothorax. Hilar and mediastinal silhouettes are unchanged. Heart size is normal.There is no pulmonary edema.Partially imaged upper abdomen is unremark-able. Mild thoracic dextroscoliosis again noted.
Impression:No acute cardiopulmonary process.
Similar
Report
Findings:Frontal and lateral views of the chest demonstrate normal lung volumes. There is no pleural effusion, focal consolidation or pneumothorax. Hilar and mediastinal silhouettes are unremarkable.Heart size is normal. There is no pulmonary edema.Partially imaged upper abdomen is unremarkable.
Impression:No evidence of acute cardiopulmonary process.
(c) Similar reports from different patients

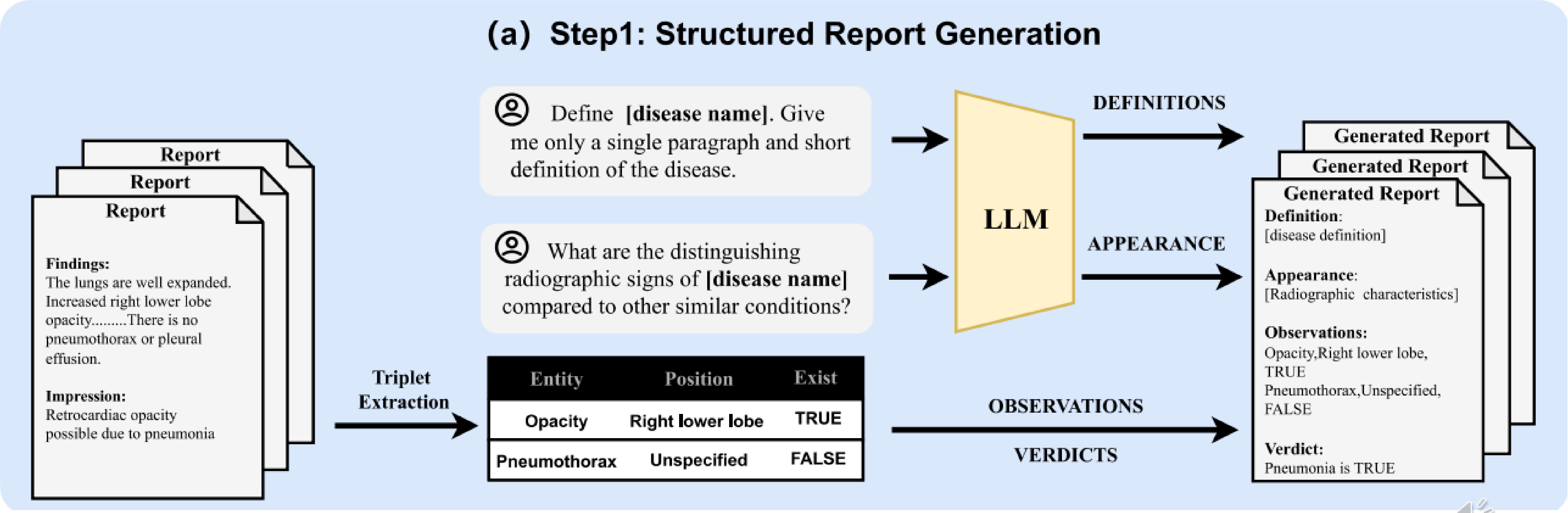
Motivation

- ⚙️ **Goal**: Train models that generalize across: Low-label regimes, Unseen diseases and Multi-task settings
- 🔍 **Key idea**: Explicitly model **uncertainty i**n both the **image** and the **report i.e.** Represent features as probabilistic distributions
- 📃 **Structured, LLM-generated reports**
- 🎯 Enables robust, uncertainty-aware vision-language modeling

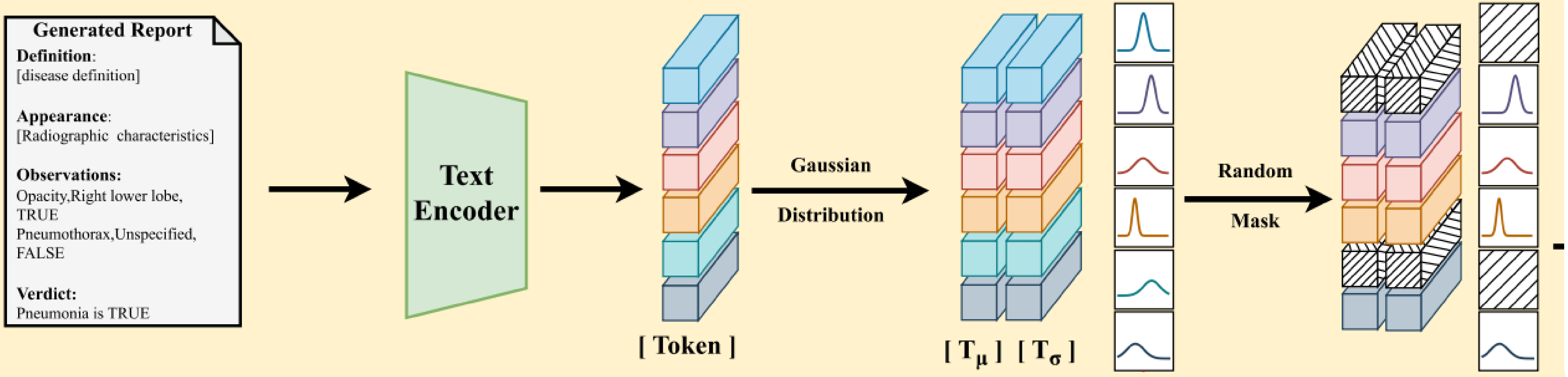
Research Questions

1. How can uncertainty in medical vision-language data be explicitly modeled to improve generalization across downstream clinical tasks?
2. Can structured language model–generated reports reduce semantic inconsistencies in medical datasets and improve image-text alignment?
3. How does distribution-based masking influence the learning of clinically relevant features in vision-language pretraining?

D-MLM Overview Step 1

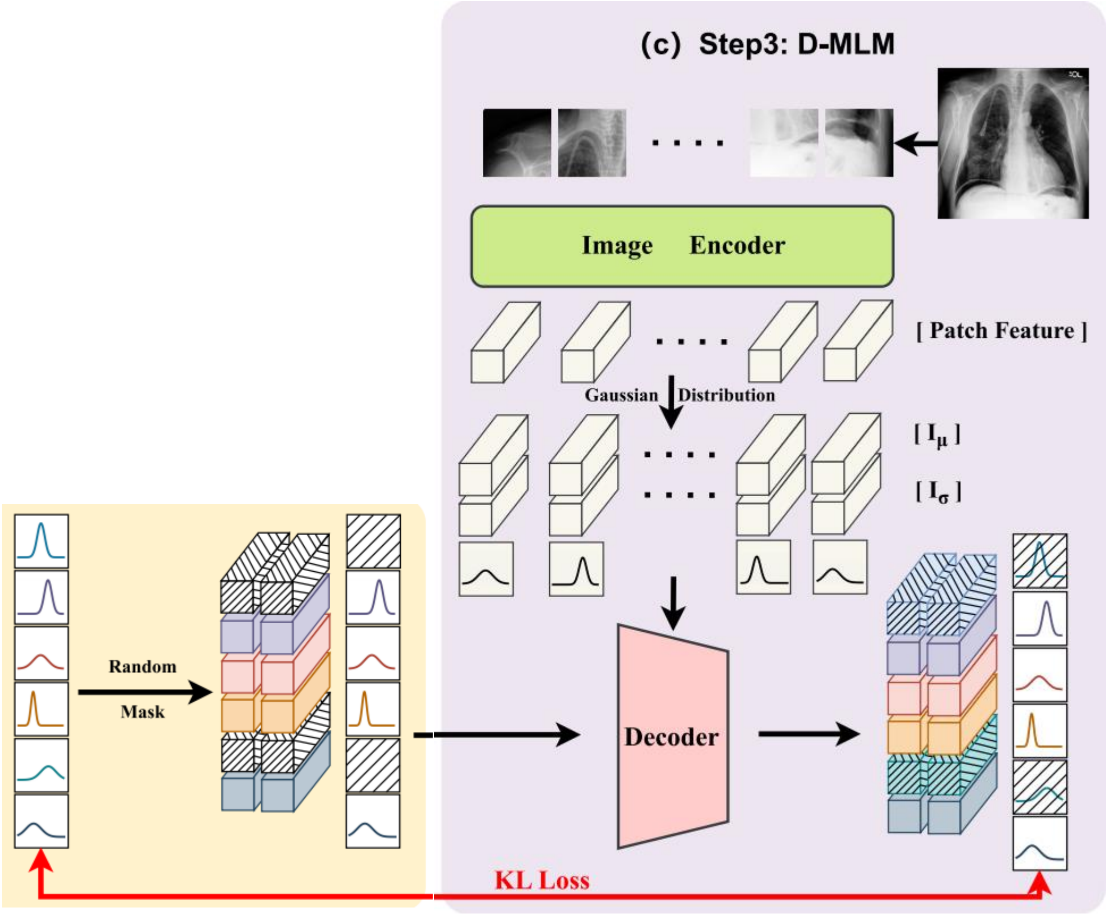


D-MLM Overview Step 2



(b) Step2: Distribution-Based Masking

D-MLM Overview Step 3



$$h_i = N(\mu_i, \sigma_i^2)$$

$$p(h_i | \mathbf{I}, \mathbf{T}_{\backslash i}) = N(\hat{\mu}_i, \hat{\sigma}_i^2)$$

$$\mathcal{L}_{\text{D-MLM}} = \mathbf{E}_{(\mathbf{I},\mathbf{T})\sim D}\left[\sum_{t\in\mathcal{M}_{\text{text}}} \mathrm{KL}\left(N(\hat{\mu}_t, \hat{\sigma}_t^2) \parallel N(\mu_t, \sigma_t^2)\right) + \sum_{p\in\mathcal{M}_{\text{image}}} \mathrm{KL}\left(N(\hat{\mu}_p, \hat{\sigma}_p^2) \parallel N(\mu_p, \sigma_p^2)\right)\right]$$

$$\mathcal{L}_{\text{align}} = \sum_{(h_i^{\{\text{text}\}}, h_j^{\{\text{image}\}})\in\mathcal{A}} W\Big(N(\mu_i^{\{\text{text}\}}, \sigma_i^{\{\text{text}\}^2}), N(\mu_j^{\{\text{image}\}}, \sigma_j^{\{\text{image}\}^2})\Big)$$

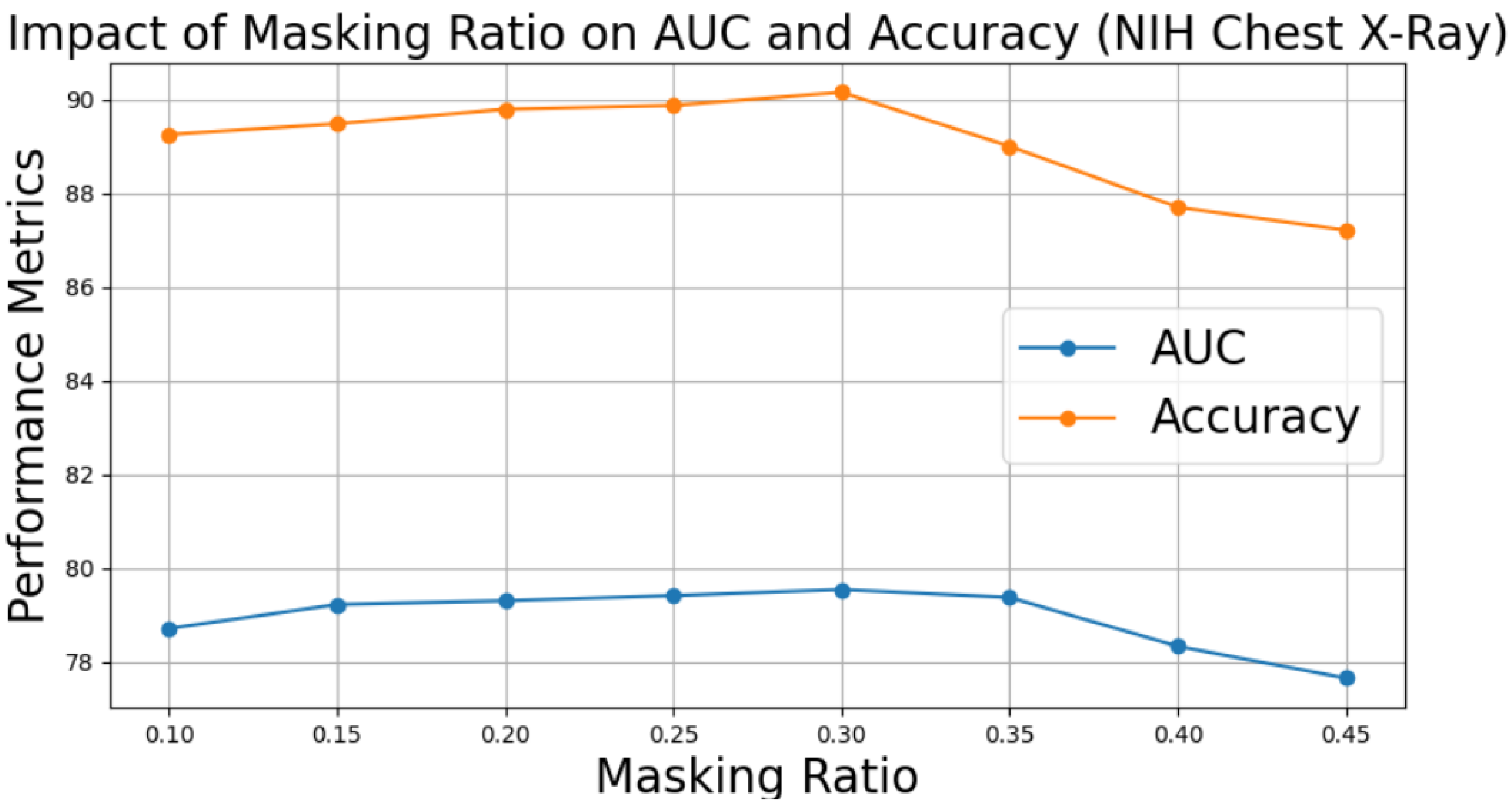
Need for structured reports

Methods	AUC↑	F1↑	ACC↑
Original Report	69.95	20.04	77.71
Triplet	73.48	24.42	82.89
KE-Triplet	76.84	26.11	86.55
M&M [16]	77.92	27.55	88.52
Structured Reports (Ours)	**79.54**	**28.81**	**90.15**

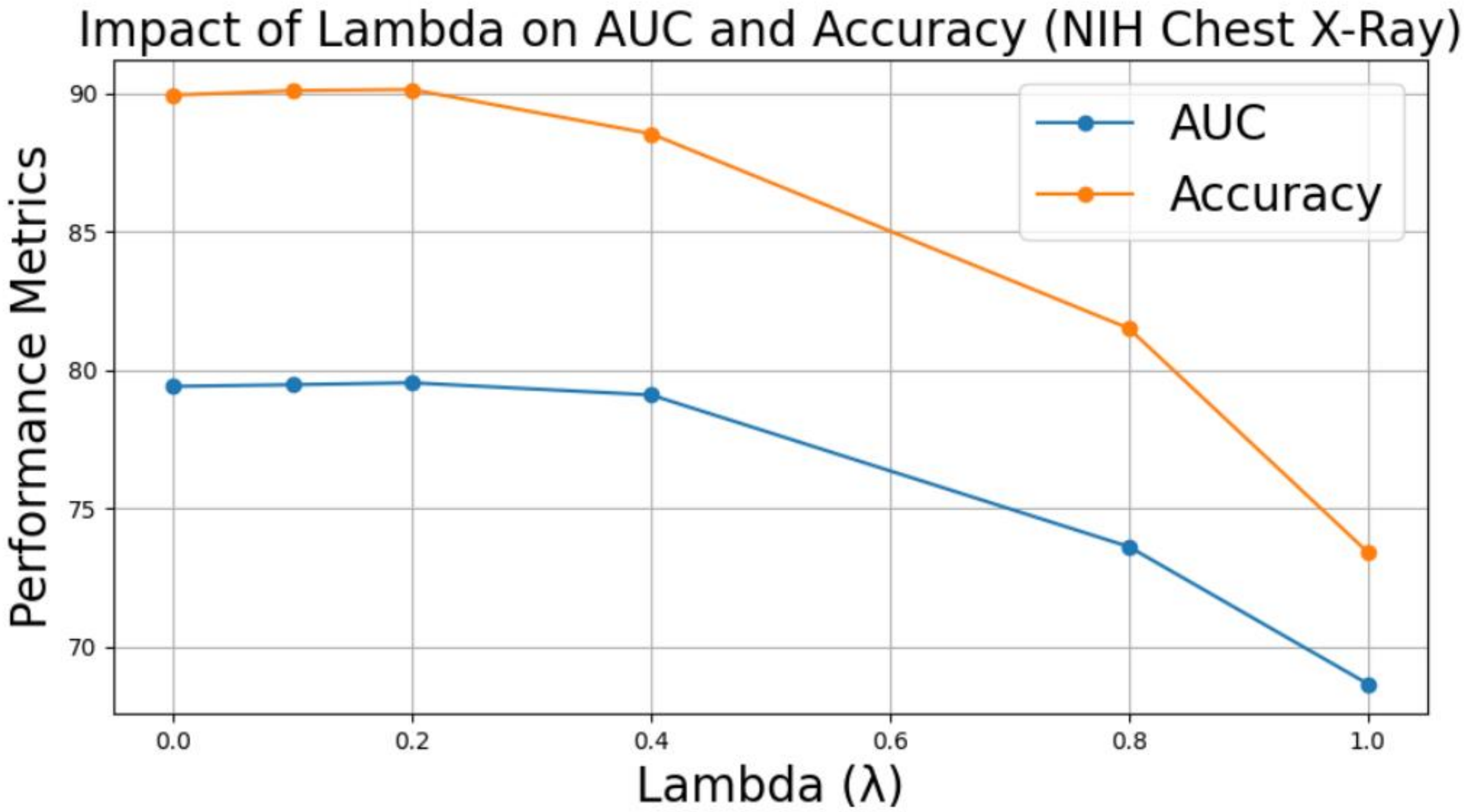
Need for D-MLM

Methods	AUC↑	F1↑	ACC↑
No Masking	61.48	16.33	70.54
MAE [18]	68.84	18.85	75.59
MaskVLM [26]	58.87	14.96	66.69
M&M [16]	77.92	27.55	88.52
D-MLM (Ours)	**79.54**	**28.81**	**90.15**

Impact of masking ratio



Impact of loss functions



Results – SSL and SL

	RSNA Pneumonia			SIIM-ACR			CheXpert		
Methods	1%	10%	100%	1%	10%	100%	1%	10%	100%
MoCo [16]	82.33	85.22	87.90	75.49	81.01	88.43	78.00	86.27	87.24
SimCLR [6]	80.18	84.60	88.07	74.97	83.21	88.72	67.41	86.74	87.97
ConVIRT [35]	83.98	85.62	87.61	84.17	85.66	91.50	85.02	87.58	88.21
GLoRIA [19]	84.12	86.83	89.13	85.05	88.51	92.11	83.61	87.40	88.34
BioViL [4]	81.95	85.37	88.62	79.89	81.62	90.48	80.77	87.56	88.41
LoVT [26]	85.51	86.53	89.27	85.47	88.50	92.16	85.13	88.05	88.27
PRIOR [7]	85.74	87.08	89.22	87.27	89.13	92.39	86.16	88.31	88.61
MedKLIP [33]	87.31	87.99	89.31	85.27	90.71	91.88	86.24	88.14	88.68
M&M [13]	88.11	89.44	91.91	88.81	91.15	93.88	88.45	90.02	90.88
MLIP [23]	89.30	90.04	90.81	-	-	-	89.03	89.44	90.04
UniMedI [18]	90.02	90.41	91.47	-	-	-	89.44	89.72	90.51
IMITATE [24]	91.73	92.85	93.46	-	-	-	89.13	89.49	89.66
D-MLM (Ours)	**91.94**	**92.91**	**93.84**	**91.11**	**92.44**	**95.18**	**89.80**	**90.41**	**91.45**

Results – ZSL

Methods	RSNA Pneumonia AUC↑	F1↑	ACC↑	SIIM-ACR AUC↑	F1↑	ACC↑	NIH Chest X-Ray AUC↑	F1↑	ACC↑
ConVIRT [35]	80.42	58.42	76.11	64.31	43.29	57.00	61.01	16.28	71.02
GLoRIA [19]	71.45	49.01	71.29	53.42	38.23	40.47	66.10	17.32	77.00
BioViL [4]	82.80	58.33	76.69	70.79	48.55	69.09	69.12	19.31	79.16
PRIOR [7]	85.58	62.91	77.85	86.62	70.11	84.44	74.51	23.29	84.41
MedKLIP [33]	86.94	63.42	80.02	89.24	68.33	84.28	76.76	25.25	86.19
M&M [13]	88.91	66.58	83.14	91.15	71.58	86.15	77.92	27.55	88.52
D-MLM (Ours)	**90.15**	**68.42**	**85.11**	**91.45**	**72.18**	**86.88**	**79.54**	**28.81**	**90.15**

Conclusion

1. We introduced a new pretraining framework (D-MLM) that models medical images and text as **probabilistic distributions**, capturing both inter- and intra-modal uncertainty.
2. To improve clinical grounding, we used **LLM-generated structured reports** with consistent sections such as definitions, appearances, observations, and verdicts. This helps align visual features with medical semantics.
3. Our **adaptive masking strategy**, guided by the structured text, focuses learning on the most diagnostically relevant parts of the image and report.
4. The model delivers **strong performance across tasks** — including classification, grading, and segmentation — and is especially effective in **low-label** and **zero-shot** settings.


THANK
YOU