XBENCH: A COMPREHENSIVE BENCHMARK FOR VISUAL-LANGUAGE EXPLANATIONS IN CHEST RADIOGRAPHY

Haozhe Luo[1,3] *, Shelley Zixin Shu*[1]*, Ziyu Zhou*[2]*, Sebastian Otálora*[3]*, Mauricio Reyes*[1,4]

[1]ARTORG Center for Biomedical Engineering Research, University of Bern, Switzerland
[2]Shanghai Jiao Tong University, China
[3]Kaiko.AI, Switzerland
[4]Dept. of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Switzerland

1. ABSTRACT

Vision–language models (VLMs) have recently shown remarkable zero-shot performance in medical image understanding, yet their grounding ability, the extent to which textual concepts align with visual evidence, remains underexplored. In the medical domain, however, reliable grounding is essential for interpretability and clinical adoption. In this work, we present the first systematic benchmark for evaluating cross-modal interpretability in chest X-rays across seven CLIP-style VLM variants. We generate visual explanations using cross-attention and similarity-based localization maps, and quantitatively assess their alignment with radiologist-annotated regions across multiple pathologies. Our analysis reveals that: (1) while all VLM variants demonstrate reasonable localization for large and well-defined pathologies, their performance substantially degrades for small or diffuse lesions; (2) models that are pretrained on chest X-ray–specific datasets exhibit improved alignment compared to those trained on general-domain data. (3) The overall recognition ability and grounding ability of the model are strongly correlated. These findings underscore that current VLMs, despite their strong recognition ability, still fall short in clinically reliable grounding, highlighting the need for targeted interpretability benchmarks before deployment in medical practice. XBENCH code is available at https://github.com/Roypic/Benchmarkingattention.

***Index Terms*—** Explainability, Benchmark, VLM, Grounding

2. INTRODUCTION

Deep learning has achieved remarkable progress in medical image analysis, enabling automated interpretation of chest radiographs at expert-level accuracy in certain diagnostic tasks. Recent advances in vision-language models (VLMs) [8, 9, 10] further extend this capability by jointly learning from paired image-text data, demonstrating strong zero-shot recognition and transferability across medical domains. However, in clinical settings, the value of such models extends beyond classification accuracy. Whether models' predictions are grounded in meaningful visual evidence are equally important [11, 12, 5], as reliable grounding, or cross-modal interpretability, is essential for clinical trust, model validation, and regulatory acceptance.

While numerous VLMs[13, 14, 15] have shown promising performance on image-level tasks, their spatial reasoning and localization abilities remain poorly understood. Prior work has revealed that post-hoc saliency methods, though widely adopted for medical explainability, often fail to localize fine-grained or small-scale pathologies compared with radiologist annotations [5]. In particular, benchmarks like CheXlocalize[5] have highlighted large gaps between model-generated heatmaps and expert-drawn regions, underscoring the need for standardized and quantitative evaluation of grounding performance.

To address this gap, we introduce **XBench**, the first comprehensive benchmark for evaluating cross-modal interpretability in chest X-rays. XBENCH integrates the **Dataset**, **Model**, and **Metrics** modules into a unified evaluation framework (Fig. 1), supporting seven representative CLIP-style VLMs spanning pretraining from natural images to chest X-ray–specific data. Grounding performance is assessed with Pointing Game, Dice, and IoU, while AUC, Accuracy, and F1 are jointly reported. Across 36 findings and 12,601 cases, XBENCH reveals systematic explainability patterns: domain-specific pretraining improves grounding for large, well-defined pathologies, but models still underperform on small lesions, obfuscated/overlapping regions, and diffuse or scale-variant findings.

Together, this benchmark establishes a rigorous foundation for evaluating and improving the interpretability of medical vision–language models, paving the way toward clinically reliable multimodal AI.

3. TASK FORMULATION & IMPLEMENTATION DETAILS

We study zero-shot recognition and grounding of C diagnostic concepts $\mathcal{C}$ over pooled datasets $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$. Each image $x \in \mathbb{R}^{H \times W}$ has labels $\mathbf{y} \in \{0,1\}^C$ and, when available, regions $\mathcal{R} = \{R_c \subset \Omega\}_{c \in \mathcal{C}}$. A CLIP-style VLM $f_\theta =$

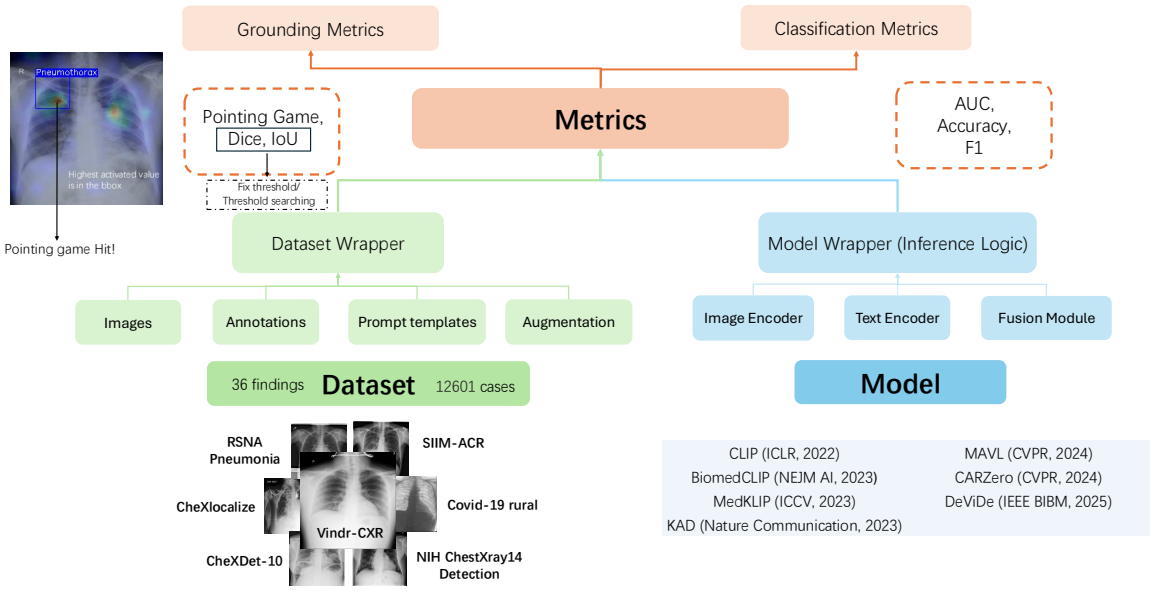

Fig. 1. Overview of the unified evaluation framework for medical vision-language models. The framework integrates three main components: **Dataset**, **Model**, and **Metrics**. The **Dataset Wrapper** organizes multi-source datasets (36 diseases, 12,601 cases) including images, annotations, prompt templates, and augmentations from RSNA Pneumonia[1], Covid19-rural[2], ChestDet-10[3], SIIM[4], CheXlocalize[5], ChestXray14 Detection[6], and Vindr-CXR[7]. The **Model Wrapper** standardizes inference logic across vision-language models, encapsulating the image encoder, text encoder, and fusion module; it supports representative models such as CLIP, BiomedCLIP, MedKILP, KAD, MAVL, CARZero, and DeViDe. The **Metrics** module unifies both **Grounding Metrics** (e.g., Pointing Game, Dice, IoU with fixed or searched thresholds) and **Classification Metrics** (e.g., AUC, Accuracy, F1), enabling comprehensive and comparable evaluation across datasets and models.

(h_θ, g_θ) queries class c via $t_c = \tau(c)$ and computes $s_c(x) = \langle h_\theta(x), g_\theta(t_c)\rangle$, $p_c(x) = \sigma(s_c(x))$, and $\hat{y}_c = \mathbb{I}\{p_c(x) \geq 0.5\}$. Class saliency $M_c(x)$ derives from similarity ϕ_{sim} or cross-attention ϕ_{att}. Grounding uses normalized maps $\widetilde{M}_c(x)$ thresholded as $B_c(x;\tau) = \{u : \widetilde{M}_c(x)_u \geq \tau\}$. We compute Pointing Game $\mathbb{I}\{\arg\max_u M_c(x)_u \in R_c^{(0.5)}\}$, Dice $= \frac{2|B_c \cap R_c|}{|B_c|+|R_c|}$, and IoU $= \frac{|B_c \cap R_c|}{|B_c \cup R_c|}$. Results are reported at fixed $(\gamma, \tau) = (0.5, 0.5)$ and best τ, with recognition metrics (macro AUC, F1, AUPRC, Hamming acc) averaged per class and dataset.

We benchmark on seven CXR datasets: RSNA Pneumonia [1] (1 class), SIIM-ACR Pneumothorax [4] (1), COVID-19 Rural [2] (1), CheXDet-10 [3] (10), CheXlocalize [5] (13), ChestX-ray14 Detection [6] (8), and VinDr-CXR [7] (21). Unless noted, models use batch size 8, input resolution 224×224, and each model's official prompt style for text encoding; grounding uses a best threshold searching strategy from 0 to 1 with step 0.01 (or a fixed threshold $\tau = 0.5$). All experiments run on NVIDIA H200 GPUs (141 GB). XBENCH supports custom component insertion, and only needs to modify the config file to achieve flexible reasoning.

4. RESULTS AND ANALYSIS

4.0.1. Zero-Shot Classification Performance

We evaluate seven CLIP-style VLMs in a zero-shot setting across multi- and single-disease datasets, emphasizing grounding without task-specific fine-tuning. As shown in Table.1,3,4,5 and 2, CARZero is consistently strongest. Gains are pronounced for large, well-defined findings (e.g., cardiomegaly, consolidation), all domain-sepcific models outperform natural-image baselines like CLIP, reflecting the value of CXR-specific pretraining. For emergent conditions such as COVID-19 (Table.1), CARZero also leads, substantially surpassing MedKLIP (0.19) and BioMedCLIP (0.30). Notably, for COVID-19 grounding and recognition, MAVL and MedKLIP outperform KAD and DeViDe, though their in-domain performance is lower—underscoring the importance of detailed query prompts at inference time. Corresponding Attention map visualization are shown in the Fig.2.

4.0.2. Correlation Analysis and Transferability Insights

As shown in Fig. 3, *classification* AUC and *pointing-game* ACC are strongly coupled, as indicated by a high coefficient of determination ($R^2 = 0.92$): recognition gains typically strengthen

Method	COVID-19			Pneumonia			Pneumothorax			CheXDet-10			CheXlocalize			VinDR-CXR			ChestX-ray14		
	Point	Dice	IoU	Point	Dice	IoU	Point	Dice	IoU	Point	Dice	IoU	Point	Dice	IoU	Point	Dice	IoU	Point	Dice	IoU
CLIP[13]	15.62	18.31	10.92	7.39	20.20	11.97	0.46	3.16	1.64	7.10	12.99	8.10	3.22	7.96	4.33	2.47	7.31	4.25	7.42	12.36	7.11
BiomedCLIP[14]	3.12	16.56	9.69	12.64	20.20	11.97	1.01	3.17	1.64	5.41	12.68	7.92	3.24	8.16	4.46	3.95	7.55	4.39	9.05	12.67	7.37
MedKLIP[15]	28.12	19.25	11.24	42.78	33.07	21.25	1.55	3.79	2.00	31.79	23.32	15.38	18.12	17.91	10.98	23.74	19.04	12.02	40.06	27.74	18.21
KAD[8]	6.25	27.45	18.18	70.11	42.06	28.05	2.47	4.18	2.20	32.18	23.26	15.47	24.21	18.73	11.61	18.35	18.55	11.80	43.61	29.65	19.41
MAVL[16]	15.62	16.45	9.61	29.31	20.14	11.94	2.78	4.09	2.25	26.12	19.19	12.58	17.49	13.31	7.88	15.89	16.53	10.32	31.31	20.83	13.12
CARZero[17]	**53.12**	**36.64**	**24.26**	**83.66**	**50.47**	**36.45**	**5.56**	**4.94**	**2.79**	**48.38**	**31.35**	**22.40**	**33.35**	**23.20**	**15.54**	**39.07**	**31.01**	**22.28**	**61.57**	**39.44**	**28.01**
DeViDe[9]	3.12	28.36	18.56	70.77	40.16	26.48	3.40	4.36	2.33	35.26	26.56	18.12	27.02	18.22	11.22	21.47	20.43	13.03	49.16	30.57	20.04

Table 1. **Grounding metrics across datasets.** Each dataset group shows the mean Pointing Game and the best-threshold Dice/IoU. Single-disease datasets focus on one pathology; multi-disease show averages over classes. All values are percentages. Best and second-best in each column are in **bold** and underline, respectively.

Table 2. **Per-class Pointing Game performance on VinDR-CXR.** Each model spans two rows to display all classes. The best result per class is shown in **bold**, and the second best is underlined. For most findings, the grounding performance of all models remains below 50%.

Model	Mean	Classes										
		Aortic enl.	Atelectasis	Calcif.	Cardiomegaly	Clav. fract.	Consol.	Emphysema	Enl. PA	ILD	Infiltration	Lung Opac.
		Lung cavity	Lung cyst	Mediast. shift	Nodule/Mass	Other lesion	Pleural eff.	Pleural thick.	Pneumothorax	Pulm. fibrosis	Rib fract.	
CLIP	2.47	5.73	0.00	0.60	9.26	0.00	6.90	0.00	0.00	15.10	1.89	0.00
		0.00	0.00	0.00	1.34	6.17	4.26	0.66	0.00	0.00	0.00	
BioMedCLIP	3.95	7.73	1.20	0.00	7.90	0.00	0.00	0.00	0.00	2.80	3.51	0.00
		0.00	**50.00**	0.00	1.25	1.12	4.85	0.62	0.00	1.91	0.00	
MedKLIP	23.74	30.21	24.68	7.14	58.52	0.00	50.57	**33.33**	**42.86**	**40.62**	32.08	32.00
		25.00	0.00	25.00	24.16	18.52	14.89	0.00	20.00	18.88	0.00	
KAD	18.35	12.08	21.69	14.44	86.60	0.00	62.37	0.00	0.00	1.40	19.30	27.50
		0.00	0.00	0.00	34.38	3.37	38.83	2.47	22.22	2.39	**36.36**	
MAVL	15.89	30.73	12.99	3.57	12.59	0.00	34.48	**33.33**	14.29	33.85	18.87	26.67
		12.50	0.00	6.25	9.40	18.52	5.32	0.00	**33.33**	17.86	9.09	
CARZero	**39.07**	**77.60**	**41.56**	26.19	76.30	**100.00**	**77.01**	**33.33**	28.57	2.08	**52.83**	**38.67**
		25.00	0.00	**37.50**	**37.58**	**51.06**	15.23	**33.33**	**33.33**	**20.92**	27.27	
DeViDe	21.47	10.63	31.33	**34.44**	**88.32**	0.00	73.12	0.00	14.29	2.80	17.54	31.25
		0.00	0.00	**37.50**	37.50	7.87	**41.75**	1.23	16.67	14.83	27.27	

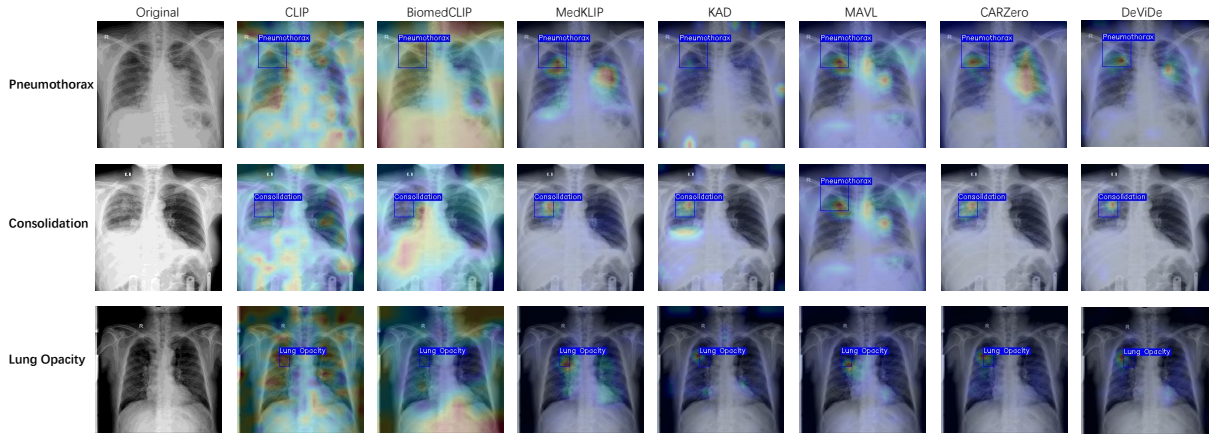

Fig. 2. Visual comparison of disease localization across six vision-language models on chest X-rays. Blue boxes mark ground-truth regions. CARZero and DeViDe show more accurate and focused attention for Pneumothorax, Consolidation, and Lung Opacity.

grounding. Three pretraining regimes are: (i) direct contrastive alignment (CLIP, BioMedCLIP) with modest AUC and weak localization; (ii) structured alignment (MedKLIP, MAVL) in the mid-range; and (iii) cross-attention alignment (DeViDe, KAD, CARZero) in the upper-right with the best joint performance. Notably, CARZero lies above the trend, translating recognition performance into spatial evidence more efficiently. Overall, the monotonicity implies that strong classification performance often carries over to grounding.

Table 3. Per-class Pointing game results on ChestX-ray14 dataset.

Model	Mean	ATE	CARD	EFF	INF	MASS	NOD	PNEU	PTX
CLIP	7.42	3.33	20.55	11.11	11.38	4.71	2.53	1.67	4.08
BioMedCLIP	9.05	2.78	42.47	0.65	4.88	7.06	0.0	12.5	2.04
MedKLIP	40.06	32.78	81.51	26.14	51.22	35.29	11.39	56.67	25.51
KAD	43.61	38.33	90.41	47.71	23.58	52.94	17.72	60.83	17.35
MAVL	31.31	33.33	52.74	9.8	40.65	37.65	7.59	35.0	33.67
CARZero	**61.57**	**50.56**	**99.32**	**60.13**	**74.8**	**65.88**	**24.05**	**75.0**	**42.86**
DeViDe	49.16	43.33	91.78	44.44	47.15	54.12	**24.05**	70.0	18.37

Table 4. Per-class Pointing game results on CheXDet-10 dataset.

Model	Mean	ATE	CALC	CONS	EFF	EMPH	FIB	FX	MASS	NOD	PTX
CLIP	7.11	27.08	0.00	6.14	7.00	10.81	18.67	0.00	0.00	1.35	0.00
BioMedCLIP	5.54	12.50	2.70	7.22	5.76	18.92	4.00	0.00	0.00	0.00	3.03
MedKLIP	33.09	37.50	5.41	67.15	40.74	**59.46**	34.67	4.48	31.03	16.22	21.21
KAD	32.00	58.33	0.00	74.73	59.26	32.43	8.00	8.96	51.72	16.22	12.12
MAVL	26.09	41.67	2.70	42.96	24.69	45.95	30.67	0.00	51.72	2.70	18.18
CARZero	**48.96**	**66.67**	**14.29**	**81.01**	**70.74**	61.76	**45.71**	**20.31**	**64.29**	**23.29**	**35.71**
DeViDe	34.27	**66.67**	0.00	**78.34**	63.79	29.73	22.67	7.46	58.62	16.22	9.09

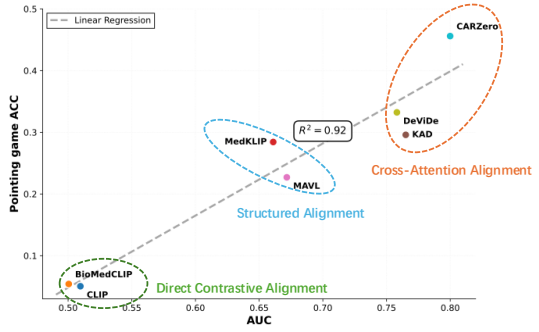

Fig. 3. Correlation between disease classification and grounding accuracy across vision-language models.

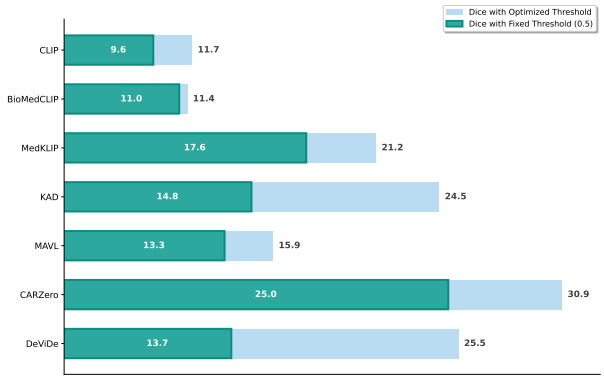

Fig. 4. Comparison of Dice coefficients under fixed (0.5) and optimized thresholds across seven vision-language models. CARZero achieves the highest Dice scores in both settings, indicating superior lesion localization consistency.

4.0.3. Threshold Sensitivity and Calibration Insights

Fig. 4 juxtaposes Dice scores at a fixed threshold ($\tau = 0.5$) against threshold searched optimal value, highlighting calibration gaps $\Delta\text{Dice} = \text{Dice}_{\text{opt}} - \text{Dice}_{0.5}$. DeViDe and KAD show the largest discrepancies (11.8%, 9.7%), reflecting strong separability but skewed distributions near $\tau = 0.5$; CARZero is moderate (5.9%); MedKLIP, MAVL, and CLIP narrower (3.6%, 2.6%, 2.1%); BioMedCLIP minimal (0.4%). Smaller gaps enable deployment with little tuning, while larger ones demand post-hoc calibration or class-specific thresholds. Notably, high Dice_{opt} models with big ΔDice (e.g., DeViDe, KAD) pinpoint score calibration, not discriminability, as the key issue. Thus, report both fixed and optimized metrics, and prioritize calibration in tuning for better interpretability.

Table 5. Per-class Pointing game results on CheXlocalize dataset.

Model	Mean	Air. Opac.	ATE	CARD	CONS	EDEMA	Enl. CARD	FX	Lung Les.	EFF	Ple. Oth.	PNEU	PTX	Sup. Dev.
CLIP	6.88	5.05	1.15	13.37	0.00	9.09	8.87	0.00	0.00	1.71	0.00	0.00	0.00	2.59
BioMedCLIP	4.49	1.44	1.72	13.95	0.00	3.90	8.53	0.00	0.00	0.00	0.00	10.00	0.00	2.59
MedKLIP	17.28	33.57	4.60	34.88	3.45	22.08	39.93	**16.67**	21.43	5.98	0.00	40.00	0.00	12.94
KAD	21.37	10.11	13.22	73.84	24.14	23.38	51.54	**16.67**	**42.86**	15.38	0.00	40.00	0.00	3.56
MAVL	17.94	34.30	2.30	38.37	6.90	22.08	62.12	**16.67**	7.14	2.56	0.00	20.00	0.00	14.89
CARZero	**33.38**	**40.43**	**14.37**	**86.63**	**37.93**	**35.06**	**66.55**	0.00	**42.86**	**23.08**	0.00	**50.00**	**18.18**	**18.45**
DeViDe	25.67	18.05	16.67	77.91	27.59	27.27	48.46	**16.67**	35.71	15.38	0.00	**50.00**	9.09	8.41

4.0.4. Inconsistency between Grounding, Classification, and Recognition Difficulty

A per-class analysis on VinDR-CXR uncovers that the intuitive correlation "improved recognition yields enhanced grounding" does not hold uniformly across pathologies. Notably, small or scale-variant lesions such as *Pneumothorax*, *Calcification*, and *Nodule/Mass* reveal a stark recognition-grounding discrepancy: VLMs attain robust classification performance but falter in providing faithful spatial cues (e.g., CARZero's Pointing Game performance are 0.33/0.26/0.38). Conversely, larger, shape-salient abnormalities like *Cardiomegaly* elicit reliable localization (CARZero 0.76; DeViDe 0.88). Such patterns imply that prevailing VLMs excessively leverage global contextual priors while remaining vulnerable to lesion-scale ambiguities.

5. CONCLUSION

We present XBENCH, a unified benchmark for recognition and grounding in chest radiography. Across seven VLMs, we observe a strong model-level coupling between AUC and pointing accuracy, yet persistent per-class mismatches for small or scale-variant lesions, and notable calibration gaps between fixed and optimized thresholds. These results indicate that current medical VLMs still rely on global context and lack robust, size-aware spatial evidence. While our analysis centers on CLIP-style VLMs, recent domain-adapted MLLMs (e.g., a 7B LLaVA-Rad trained on 697k radiograph–report pairs) have outperformed much larger general models (GPT-4V) on factual report generation. We'll further incorporaete such MLLM baselines in XBENCH to reveal whether their free-form explanations align better with radiologist annotations and how far foundation models have progressed.

6. REFERENCES

[1] "Rsna pneumonia detection challenge (2018)," https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.

[2] Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Piroon Jenjaroenpun, Thomas Powell, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri Blake, et al., "Chest imaging representing a covid-19 positive rural us population," *Scientific data*, vol. 7, no. 1, pp. 414, 2020.

[3] Jingyu Liu, Jie Lian, and Yizhou Yu, "Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities," *arXiv preprint arXiv:2006.10550*, 2020.

[4] Carol Wu Anna Zawacki, Julia Elliott George Shih, ParasLakhani Mikhail Fomitchev, Mohannad Hussain, and Shunxing Bao Phil Culliton, "Siim-acr pneumothorax segmentation," https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019.

[5] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al., "Benchmarking saliency methods for chest x-ray interpretation," *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022.

[6] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[7] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al., "Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, pp. 429, 2022.

[8] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Communications*, vol. 14, no. 1, pp. 4542, 2023.

[9] Haozhe Luo, Ziyu Zhou, Corentin Royer, Anjany Sekuboyina, and Bjoern Menze, "Devide: Faceted medical knowledge for improved medical vision-language pre-training," *arXiv preprint arXiv:2404.03618*, 2024.

[10] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun, "Medclip: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[11] Haozhe Luo, Aurélie Pahud de Mortanges, Oana Inel, and Mauricio Reyes, "Dwarf: Disease-weighted network for attention map refinement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 59–68.

[12] Aurélie Pahud de Mortanges, Haozhe Luo, Shelley Zixin Shu, Amith Kamath, Yannick Suter, Mohamed Shelan, Alexander Pöllinger, and Mauricio Reyes, "Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging," *NPJ digital medicine*, vol. 7, no. 1, pp. 195, 2024.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al., "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[15] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21372–21383.

[16] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans, "Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11492–11501.

[17] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou, "Carzero: Cross-attention alignment for radiology zero-shot classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11137–11146.