MedTri: Structured Medical Report Normalization for Enhanced Vision–Language Pretraining

Yuetan Chu[1], Xinghua Ma[1], Xinran Jin[1], Gongning Luo[1], and Xin Gao[1]✉

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
xin.gao@kaust.edu.sa

Abstract. Medical vision–language pretraining increasingly relies on medical reports as large-scale supervisory signals; however, raw reports often exhibit substantial stylistic heterogeneity, variable length, and a considerable amount of image-irrelevant content. Although text normalization is frequently adopted as a preprocessing step in prior work, its design principles and empirical impact on vision-language pretraining remain insufficiently and systematically examined. In this study, we present **MedTri**, a deployable normalization framework for medical vision-language pretraining that converts free-text reports into a unified *[Anatomical Entity: Radiologic Description + Diagnosis Category]* triplet. This structured, anatomy-grounded normalization preserves essential morphological and spatial information while removing stylistic noise and image-irrelevant content, providing consistent and image-grounded textual supervision at scale. Across multiple datasets spanning both X-ray and computed tomography (CT) modalities, we demonstrate that *structured, anatomy-grounded text normalization is an important factor in medical vision–language pretraining quality*, yielding consistent improvements over raw reports and existing normalization baselines. In addition, we illustrate how this normalization can easily support targeted text-level augmentation strategies, including knowledge enrichment and anatomy-grounded counterfactual supervision, which provide complementary gains in robustness and generalization without altering the core normalization process. Together, our results position structured text normalization as a critical and generalizable preprocessing component for medical vision–language learning, and establish MedTri as a practical implementation of this approach. The source code is available at https://github.com/Arturia-Pendragon-Iris/MedTri.

Keywords: Medical vision-language pretraining · Large-language model · Contrastive learning

1 Introduction

vision-language pretraining (VLP) that leverages naturally paired medical images and medical reports has emerged as a powerful paradigm in medical image

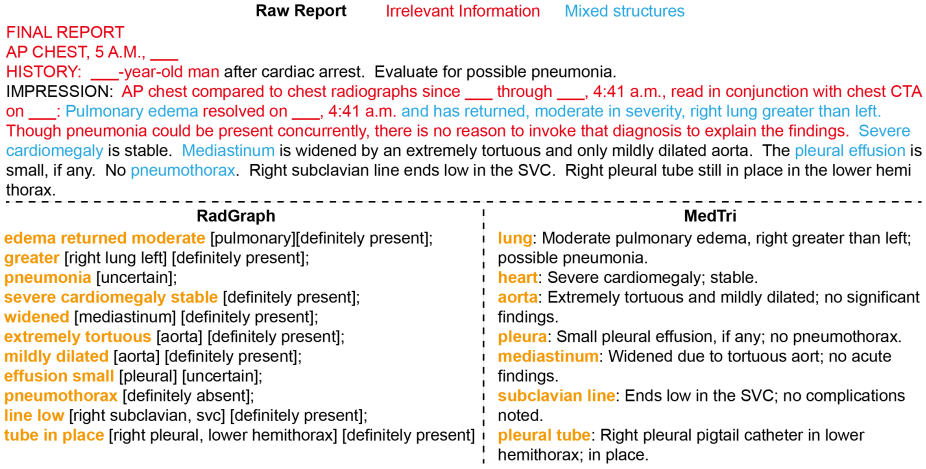

Fig. 1. Illustration of normalization differences across raw reports, RadGraph, and MedTri. The raw clinical report contains irrelevant content and mixed structures. RadGraph extracts only diagnostic entities with limited imaging descriptions. MedTri produces anatomically anchored, image-grounded triplets that preserve morphological and spatial detail while removing stylistic and clinically irrelevant text.

analysis [14][22]. Such pairs provide large-scale semantic supervision without additional annotation effort [28][25], as reports contain expert interpretations grounded in patient-specific visual findings. By jointly modeling visual content and textual descriptions, medical VLP strategies have demonstrated strong capabilities in capturing disease-relevant semantics, improving representation quality, and enhancing generalization across diverse downstream tasks [19]. This synergy positions vision-language alignment as a foundational component for modern medical image pretraining.

Despite the success of VLP, the textual supervision provided by raw clinical reports introduces several obstacles for effective multimodal alignment. Medical reports vary widely in style, verbosity, and structure, often mixing image-grounded observations with unrelated clinical history or management recommendations [17]. This heterogeneity reduces image-relevant signals, inflates sequence length, and weakens the fine-grained correspondence between visual findings and textual descriptions. As a result, recent studies increasingly adopt text normalization to standardize report content and enhance the stability of vision-language learning [19][20]. However, text normalization is often introduced as a preprocessing component and combined with other architectural designs, while its standalone design principles and empirical impact on vision–language pretraining remain insufficiently examined. Moreover, existing normalization methods, such as schema-based or NER-driven systems (e.g., RadGraph [13]), primarily focus on entity extraction, whereas cloud-based LLM methods rely on large-scale generative rewriting at the expense of increased computational overhead and potential

privacy concerns. Consequently, the field still lacks a lightweight, anatomically expressive, and locally deployable normalization solution.

In this study, we present **MedTri**, a structured normalization approach for medical vision–language pretraining. MedTri converts free-text radiology reports into a unified *[Anatomical Entity: Radiologic Description + Diagnosis Category]* triplet that preserves essential morphological and spatial information while removing stylistic noise and image-irrelevant content [8][9]. This design enables consistent, efficient, and privacy-preserving report normalization suitable for large-scale pretraining. Beyond the normalization itself, MedTri further provides a modular interface that allows additional text-level augmentation on top of the normalized triplet. Using this interface, we instantiate two optional examples: knowledge enrichment and anatomy-grounded counterfactual augmentation. Across multiple datasets covering both X-ray and CT modalities, we systematically demonstrate that *structured, anatomy-grounded text normalization is an important factor in medical vision–language pretraining quality*, with MedTri consistently outperforming raw reports and existing normalization approaches. The optional augmentation modules provide further, complementary gains in performance and generalization. Together, these results position MedTri as a practical and anatomically expressive normalization approach for medical vision–language learning.

Dataset	# Reports	Image Modality	Body Region
MIMIC-CXR [14]	65,000	X-ray	Chest
CT-RATE [11]	20,000	CT	Chest
CT-INSPECTION [12]	10,000	CT	Chest
AbdomenAtlas [18]	5,000	CT	Abdomen
Parrot [17]	2,658	CT, MRI, X-ray, etc	Head, abdomen, chest, etc

Table 1. Overview of the report dataset used to develop and evaluate the MedTri normalization platform.

2 Method

2.1 Structured Triplet

To support stable normalization for medical vision-language pretraining, we adopt a structured triplet that decomposes each report into a set of clinically grounded triplets:

$$[\textit{Anatomical Entity: Radiologic Description} + \textit{Diagnosis Category}]$$

The schema captures the minimal semantic unit of radiologic reasoning, consisting of an anatomical anchor, objective imaging attributes, and an associated diagnostic interpretation when present. By explicitly disentangling these components, MedTri converts heterogeneous free-text reports into anatomy-level alignment units, reducing lexical and stylistic variability while preserving semantically discriminative, image-grounded information for vision-language pretraining.

2.2 Local Model Development

Clinical reports from different institutions exhibit substantial variability in style, grammar, and diagnostic phrasing. To ensure that MedTri remains robust to this heterogeneity, we established a dataset with more than 100,000 reports selected from multiple publicly available datasets covering diverse imaging modalities and anatomical regions (Table 1). Using this dataset, we first generated structured reference summaries through a standardized ChatGPT-5.1 prompt specifically designed for our triplet schema. The prompt was iteratively refined on approximately 1,000 cases and then applied to the full dataset, yielding paired samples (x_i, y_i) in which x_i is the original report and y_i is its normalized triplet (Table 2).

To enable scalable and privacy-preserving deployment, we train a lightweight biomedical encoder–decoder model (BioBart[27]) to approximate the cloud LLM-distilled supervision. The model is fine-tuned on the paired dataset using standard cross-entropy loss, following a traditional Seq2Seq formulation [15][26]. This locally deployable text-transfer model forms the core of the MedTri approach, allowing rapid and consistent normalization without relying on large cloud-based language models. A total of 500 reports were randomly sampled from the established dataset for model testing, whereas the remaining reports were used for model training and validation.

Input:	IMPRESSION: 1. No evidence of pulmonary embolism. No acute intrathoracic pathology. 2. Mild bronchial thickening, nonspecific, but may reflect infectious or inflammatory process. 3. Main pulmonary artery is dilated measuring 39 mm, possibly reflecting background pulmonary hypertension.
Prompt:	You will receive one paragraph containing radiology findings (image-based description) and a final diagnosis. Your ONLY task is to convert this information into a structured radiology summary using the specified format. Each line must follow this exact format: [Anatomical Term]: [Radiologic Description + Diagnosis Category]. List only the anatomical regions mentioned in the findings or diagnosis. Maintain the anatomical hierarchy (e.g., lung before airway, heart before aorta) if multiple structures are present. For each anatomical term, include radiologic findings (location, size, attenuation, density, texture, etc.) and diagnosis category if explicitly stated. Use objective imaging language only. Use direct descriptive phrasing. Keep each anatomical line concise and single-sentence, using semicolons to separate multiple findings.
Output:	Lung: No evidence of pulmonary embolism; no acute intrathoracic pathology. Bronchi: Mild bronchial thickening; nonspecific; may reflect infectious or inflammatory process. Main pulmonary artery: Dilated; measures 39 mm; possibly reflecting background pulmonary hypertension.

Table 2. Example of the standardized ChatGPT-5.1 prompt used to generate structured triplets from free-text reports.

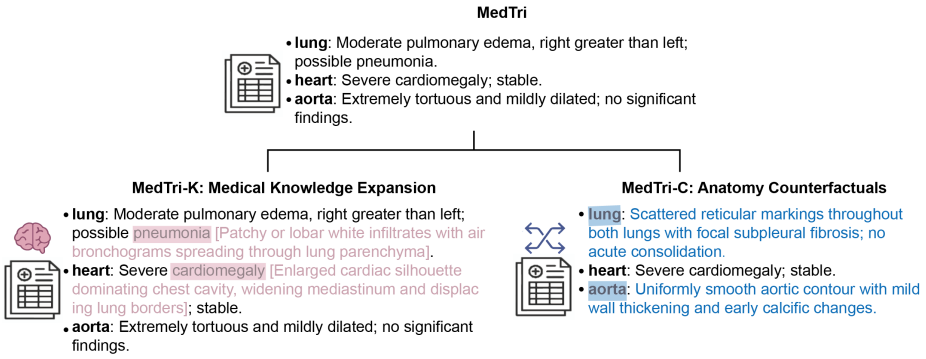

Fig. 2. Illustration of MedTri's optional augmentation modules. MedTri-K enriches each triplet with clinically validated radiologic signatures to improve visual interpretability. MedTri-C generates anatomically inconsistent counterfactuals through controlled perturbations to strengthen fine-grained visual–semantic discrimination.

2.3 Optional Text-Level Augmentation on MedTri

Medical Knowledge Expansion (MedTri-K) Prior work in vision–language learning has shown that explicitly grounding diagnostic terms in visual attributes can improve semantic alignment [19]. However, such strategies can be difficult to apply reliably to raw radiology reports due to stylistic variability and entangled narrative structure. Leveraging the structured triplet produced by MedTri, we can integrate a lightweight knowledge expansion mechanism (Fig. 2, left). For each normalized triplet, the diagnosis is augmented with a concise description of its characteristic radiological appearance, retrieved from our created dictionary of standard medical definitions covering over 100 common radiological findings and diagnoses. For example, pneumonia is associated with parenchymal consolidation or high-attenuation opacity within the affected lobe. These dictionary entries are created by board-certified radiologists and normalized for terminological consistency, allowing them to be seamlessly integrated into MedTri without altering its underlying structure. Importantly, this module operates entirely on normalized text and does not modify the core normalization process.

Anatomy-Grounded Counterfactuals (MedTri-C) Counterfactual supervision and hard negative sampling have been widely explored to encourage fine-grained discrimination in vision–language models [16][4][6]. The anatomically grounded triplet structure produced by MedTri enables a controlled and fine-grained instantiation of counterfactual text augmentation through localized perturbations. Specifically, we generate counterfactual reports by modifying the descriptions of several anatomical entities (n=2 in our study) at a time, replacing them with semantically incompatible counterparts randomly sampled from other normalized triplets (Fig. 2, right). Replacement is constrained within the same anatomical hierarchy level to ensure structural consistency while altering

local semantic alignment. These substitutions preserve the overall syntactic format and global report semantics, while deliberately disrupting the local factual alignment between anatomy and imaging findings.

The resulting counterfactual texts are paired with the original images and treated as hard negatives during contrastive training. By introducing fine-grained, anatomy-level inconsistencies rather than global semantic shifts, this strategy forces the model to attend to localized visual evidence and fine-grained anatomical features, instead of relying on coarse diagnostic signals. As with MedTri-K, this module is optional and applied only during training, without affecting the normalization backbone.

3 Experiments and Results

3.1 Evaluation of MedTri for Normalization

To assess the quality and deployability of the proposed approach, we compare MedTri against two representative baselines: (1) ChatGPT-5.1-based structured rewriting, which serves as a high-quality but non-deployable reference for distilled supervision, and (2) Qwen2.5-14B [1], a compact open-source model commonly used in local summarization workflows. Quantitative results are summarized in Table 3.

To evaluate clinical validity beyond surface-level textual similarity and to mitigate potential reference bias, we conduct a physician expert evaluation as the primary assessment. Twenty board-certified physicians independently rated 50 randomly selected cases in a double-blind manner using a five-point Likert scale, focusing on anatomical correctness and image groundedness. As shown in Table 3, MedTri achieves expert scores comparable to the cloud-based LLM reference and substantially higher than the open-source baseline, indicating that it preserves clinically meaningful and image-grounded information while remaining deployable in typical research and clinical settings. The evaluation also showed high inter-rater reliability, with ICC values of 0.983 and 0.990 for anatomical correctness and image groundedness, respectively, and Fleiss' kappa values of 0.656 and 0.634, supporting both the reliability of averaged scores and agreement among physicians.

We also report automatic text similarity metrics, including BERT score, BLEU, and ROUGE [7][10][21], computed against ChatGPT-5.1-generated references. These metrics primarily measure the degree to which MedTri approximates the distilled reference normalization and are therefore used as proxy indicators of consistency rather than as direct measures of clinical correctness. Quantitative results for the open-source baseline are included for reference.

3.2 MedTri for Improving Downstream Tasks

Experiment Settings and Datasets We adopt Swin Transformer and Vision Transformer (ViT) [3][5] as the image encoders, and BiomedVLP-CXR [2]

Metric	ChatGPT-5.1	Qwen2.5	MedTri
Anatomical correctness (1-5)	4.888	3.363	4.838
Image groundedness (1-5)	4.775	2.938	4.763
VRAM usage (GB)	–	15.04	1.50
Time/report (s)	3.25±2.14	1.21±0.67	0.45±0.18
BERT score	–	0.734±0.073	0.906±0.072
BLEU	–	0.097	0.541
ROUGE-1	–	0.566±0.127	0.824±0.146
ROUGE-2	–	0.307±0.122	0.720±0.190

Table 3. Comparison of computational efficiency, expert evaluation, and normalization accuracy across different systems.

as the text encoder. All reports were truncated to a fixed maximum length of 512 tokens, which sufficiently covers the majority of radiology reports [11] in our datasets and avoids disproportionately truncating raw reports, ensuring a fair comparison. Model training is conducted using the InfoNCE objective for contrastive learning. We apply data augmentation, including random horizontal flipping and Gaussian noise injection (sigma=0.05). All experiments are executed on an Ubuntu workstation equipped with a single NVIDIA A6000 GPU.

Pretraining is conducted separately for two imaging modalities. For 2D radiographs, we use the MIMIC-CXR dataset [14] (n=370k), while for 3D volumetric imaging, we adopt the CT-RATE dataset [11] (n=42k) for CT pretraining.

For X-ray downstream evaluation, we conduct multi-label classification on three datasets: MIMIC-CXR, NIH ChestX-ray14 [24] (n=112k), and RSNA-Pneumonia [23] (n=30k). From each dataset, we randomly sample 1,024 studies for evaluation.

CT downstream tasks are evaluated on the CT-RATE dataset, with the training and testing splits following the official protocol described in [11]. Due to the severe class imbalance commonly observed in medical imaging datasets, we determine the decision threshold that maximizes the F1 score and report the corresponding accuracy. Both F1 score and accuracy are computed on a per-label basis and then macro-averaged across all labels [19].

SwinT	MIMIC						NIH						Pneumonia					
	1%		10%		100%		1%		10%		100%		1%		10%		100%	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Raw report	0.612	0.218	0.701	0.263	0.781	0.302	0.655	0.221	0.722	0.266	0.762	0.293	0.812	0.673	0.833	0.699	0.848	0.722
RadGraph	0.584	0.205	0.673	0.251	0.748	0.284	0.628	0.207	0.703	0.249	0.747	0.265	0.804	0.665	0.822	0.690	0.840	0.715
MedTri	0.628	0.229	0.716	0.279	0.792	0.314	0.672	0.233	0.739	0.281	0.779	0.304	0.817	0.676	0.836	0.701	0.850	0.723
MedTri-K	**0.645**	**0.241**	**0.732**	0.286	0.784	0.305	**0.689**	**0.242**	0.746	0.284	0.771	0.292	0.823	**0.681**	**0.840**	**0.703**	0.851	0.722
MedTri-C	0.629	0.226	0.731	**0.291**	**0.803**	**0.318**	0.674	0.227	**0.751**	**0.290**	**0.801**	**0.321**	**0.824**	0.679	0.838	0.701	**0.855**	**0.728**

ViT	MIMIC						NIH						Pneumonia					
	1%		10%		100%		1%		10%		100%		1%		10%		100%	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Raw report	0.609	0.215	0.703	0.265	0.774	0.282	0.655	0.221	0.720	0.262	0.787	0.284	0.794	0.644	0.827	0.683	0.847	0.718
RadGraph	0.587	0.207	0.674	0.250	0.741	0.262	0.631	0.209	0.702	0.248	0.738	0.266	0.785	0.649	0.813	0.678	0.834	0.707
MedTri	0.629	0.229	0.717	0.280	0.787	0.299	0.673	0.235	0.738	0.281	0.792	0.306	0.801	0.660	0.832	0.693	0.846	0.721
MedTri-K	**0.647**	**0.241**	**0.734**	0.288	0.782	0.289	**0.690**	**0.244**	**0.746**	0.285	0.792	0.306	**0.806**	**0.671**	0.841	**0.706**	0.842	0.723
MedTri-C	0.625	0.226	**0.734**	**0.291**	**0.801**	**0.307**	0.677	0.238	0.744	**0.289**	**0.808**	**0.312**	0.803	0.669	**0.846**	0.704	**0.855**	**0.728**

Table 4. Downstream classification performance of Swin Transformer (SwinT) and ViT on different X-ray datasets under different text preprocessing strategies. Best performance is marked in bold.

Downstream Task Performances The quantitative results are presented in Table 4 and 5, respectively. Across all X-ray and CT benchmarks, MedTri and its variants consistently outperform raw reports and RadGraph across different data scales and visual backbones. The improvements are observed for both accuracy and F1 score, with particularly pronounced gains in low-data regimes (1% and 10%), indicating that structured normalization substantially improves sample efficiency for vision-language pretraining. The consistent performance gains across datasets and architectures demonstrate that the benefits of MedTri are robust and stem primarily from improved textual supervision. Wilcoxon signed-rank tests further showed significant improvements of MedTri over raw reports and RadGraph on X-ray ($p = 0.000231$ and $p = 0.000194$, respectively) and on CT ($p = 0.03125$).

	SwinT						ViT					
	1%		10%		100%		1%		10%		100%	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Raw report	0.719	0.493	0.741	0.518	0.775	0.525	0.696	0.473	0.733	0.511	0.745	0.518
RadGraph	0.714	0.480	0.735	0.513	0.744	0.523	0.694	0.471	0.722	0.508	0.716	0.509
MedTri	0.723	0.499	0.753	0.520	0.785	0.538	0.704	0.488	0.745	0.518	0.775	0.537
MedTri-K	**0.732**	**0.510**	0.757	0.519	0.789	0.536	**0.718**	**0.498**	0.743	**0.519**	0.775	0.530
MedTri-C	0.726	0.501	**0.767**	**0.523**	**0.794**	**0.557**	0.704	0.485	**0.751**	**0.519**	**0.782**	**0.543**

Table 5. Downstream classification performance of Swin Transformer (SwinT) and ViT on CT-RATE datasets under different text preprocessing strategies. Best performance is marked in bold.

The two optional augmentation modules exhibit complementary behaviors. MedTri-K (knowledge expansion) tends to achieve the best or near-best performance under limited data settings (1% and 10%), suggesting that explicitly linking diagnostic terms to characteristic imaging appearances provides additional semantic grounding when training data is scarce. However, its gains diminish at full-data scale (100%), where the model can already learn such associations implicitly from abundant image–text pairs. In contrast, MedTri-C (counterfactual construction) shows limited improvement in the 1% setting, likely because extremely limited data prevents the model from effectively exploiting fine-grained counterfactual distinctions. As data availability increases (10% and 100%), counterfactual supervision becomes more effective, leading to stronger gains in medium- and full-data regimes by encouraging finer-grained visual-semantic discrimination. The improvements of MedTri-K and MedTri-C over MedTri were also statistically significant ($p = 0.00733$ and $p = 0.000245$).

4 Discussion and Conclusion

In this work, we present MedTri, a lightweight and deployable approach that normalizes free-text medical reports into structured, anatomically grounded triplets for vision-language pretraining. Experiments on X-ray and CT datasets show that structured normalization consistently improves downstream performance

across data scales, datasets, and visual backbones, offering a practical alternative to cloud LLM-dependent pipelines for scalable and privacy-preserving clinical and research use.

This study has several limitations. Our evaluation focuses on CLIP-style contrastive vision-language pretraining and does not cover generative, instruction-tuned, or task-specific multimodal frameworks. MedTri is also evaluated only on radiology reports and imaging modalities, leaving its applicability to other medical domains or non-radiological clinical narratives unexplored. Finally, the proposed triplet schema is one principled design choice, and alternative schemas or decomposition strategies warrant further investigation.

Acknowledgments. This work was supported by KAUST Office of ORA under Awards RGC/3/6655-01-01, RGC/3/6208-01-01, RGC/3/5679-01-01, REI/1/5289-01-01, REI/1/5992-01-01, URF/1/6713-01-01, FCC/1/5932-12-11, URF/1/6599-01-02, RGC/3/6295-01-01, URF/1/6993-01-01, KCSH Award 5932, and Center of Excellence on Generative AI Award 5940.

Disclosure of Interests. The authors declare no competing interests.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision-language processing (2022)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Chu, Y., Wang, J., Luo, P., Chen, H., Zhang, Z., Zhang, J., Zhang, Y., Ju, Y., Xiong, Y., Luo, X., et al.: Ct-based ai system for quantitative and integrated management of acute respiratory distress syndrome in critical care. npj Digital Medicine (2026)
5. Chu, Y., Zhang, Y., Han, Z., Yang, C., Zhou, L., Luo, G., Huang, C., Gao, X.: Improving representation of high-frequency components for medical visual foundation models. IEEE Transactions on Medical Imaging (2025)
6. Feng, C., Patras, I.: Maskcon: Masked contrastive learning for coarse-labelled dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (6 2023)
7. Feng, C., Shen, M., Balashankar, A., Gerner-Beuerle, C., Rodrigues, M.R.D.: Noisy but valid: Robust statistical evaluation of llms with imperfect judges. In: International Conference on Learning Representations (ICLR) (2026)
8. Feng, C., Tzimiropoulos, G., Patras, I.: Clipcleaner: Cleaning noisy labels with clip. In: Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM) (10 2024)

9. Feng, C., Tzimiropoulos, G., Patras, I.: Noisebox: Towards more efficient and effective learning with noisy labels. IEEE Transactions on Circuits and Systems for Video Technology (7 2024)
10. Gao, Y., Lu, Y., Zhang, Z., Nie, J., Yu, S., Xuan, Q.: Dspc: Dual-stage progressive compression framework for efficient long-context reasoning. arXiv preprint arXiv:2509.13723 (2025)
11. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., et al.: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
12. Huang, S.C., Huo, Z., Steinberg, E., Chiang, C.C., Lungren, M.P., Langlotz, C., Yeung, S., Shah, N., Fries, J.A.: Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track
13. Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)
14. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
15. Ke, Z., Cao, Y., Chen, Z., Yin, Y., He, S., Cheng, Y.: Early warning of cryptocurrency reversal risks via multi-source data. Finance Research Letters p. 107890 (2025)
16. Lai, C., Song, S., Yan, S., Hu, G.: Improving vision and language concepts understanding with multimodal counterfactual samples. In: European Conference on Computer Vision. pp. 174–191. Springer (2024)
17. Le Guellec, B., Adambounou, K., Adams, L.C., Agripnidis, T., Ahn, S.S., Ait Chalal, R., D'Antonoli, T.A., Amouyel, P., Andersson, H., Bentegeac, R., et al.: Parrot, an open multilingual radiology reports dataset. European Journal of Radiology Artificial Intelligence p. 100066 (2025)
18. Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis **97**, 103285 (2024)
19. Liang, X., Li, X., Li, F., Jiang, J., Dong, Q., Wang, W., Wang, K., Dong, S., Luo, G., Li, S.: Medfilip: Medical fine-grained language-image pre-training. IEEE Journal of Biomedical and Health Informatics **29**(5), 3587–3597 (2025)
20. Lin, J., Xia, Y., Zhang, J., Yan, K., Cao, K., Lu, L., Luo, J., Zhang, L.: Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. arXiv preprint arXiv:2404.15272 (2024)
21. Lu, Y., Cheng, H., Fang, Y., Wang, Z., Wei, J., Xu, D., Xuan, Q., Yang, X., Zhu, Z.: Reassessing layer pruning in llms: New insights and methods. arXiv preprint arXiv:2411.15558 (2024)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

23. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
24. Summers, R.: Nih chest x-ray dataset of 14 common thorax disease categories. NIH Clinical Center: Bethesda, MD, USA (2019)
25. Xiao, C., Xu, T., Ma, S., Jiang, Y., Gao, H., Wu, Y.: Reversible primitive–composition alignment for continual vision–language learning. In: The Fourteenth International Conference on Learning Representations (2026)
26. Yang, Y., Zeng, S., Lin, T., Chang, X., Qi, D., Xiao, J., Liu, H., Chen, R., Chen, Y., Huo, D., et al.: Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. arXiv preprint arXiv:2602.11236 (2026)
27. Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., Yu, S.: Biobart: Pretraining and evaluation of a biomedical generative language model. In: Proceedings of the 21st Workshop on Biomedical Language Processing. pp. 97–109 (2022)
28. Zhou, H., Tang, J., Zhang, J., Li, Y., Xiao, C., Hou, L., Ke, Z., Yao, J.: Comem: Compositional concept-graph memory for vision–language adaptation. In: The Fourteenth International Conference on Learning Representations (2026)