Narrative-Aware Domain Knowledge Integration for Annotation-Free Medical Image Localisation

Tahsir Ahmed Munna[1,2,*], Nuno Guimarães[1,2] and Alípio Jorge[1,2]

[1]INESC TEC, Portugal
[2]Faculdade de Ciências da Universidade do Porto (FCUP), Portugal

Abstract

Pixel-level human annotation for chest X-ray image localisation requires substantial expert effort, is time-consuming, and is often impractical at scale, particularly for large clinical imaging repositories. As a result, annotation-free localisation approaches that leverage image–text pairs, such as radiology reports, have emerged as a promising alternative. However, most existing methods treat free-text clinical narratives (reports) as sufficient supervision, despite their inherent subjectivity, stylistic variability, and weak alignment with structured medical knowledge.

In this position paper, we argue that annotation-free localisation can be strengthened by further linking image parts to semantic narrative information, including explicit domain knowledge. Rather than relying solely on raw report text, we introduce disease-centric narrative descriptions that encode clinically meaningful attributes such as disease definition, anatomical location, visual appearance, shape, and extent (e.g., structured narratives for concepts like *lung opacity*). This formulation preserves the narrative nature of clinical language while providing vision-language models with more sophisticated and medically grounded concept information. We present an initial empirical analysis showing that incorporating domain-aware narratives leads to progressive improvements in localisation performance on chest X-ray benchmarks compared to free-text supervision alone. Although the observed gains are modest and not yet statistically significant, they are consistent across metrics and supported by qualitative evidence, indicating that richer narrative descriptions help guide spatial grounding. These findings suggest that moving beyond unstructured report text toward narrative-aware domain knowledge integration is a promising direction for the development of vision-language models for annotation-free chest X-ray localisation. We position this work as an ongoing research effort and outline future directions aimed at refining visual and textual knowledge representation, as well as semantically aligning them. Future developments will achieve more clinically robust and well grounded localisation systems based on scalable narrative-based supervision.

Keywords

Annotation-free chest X-ray localisation, Narrative-based domain knowledge, Vision-language learning, Multi-modal representation learning, Concept-level grounding

1. Introduction

Accurate localisation of pathological regions in chest X-ray images is essential for the development of clinical applications such as computer-aided diagnosis, disease monitoring, and model interpretability [1]. Traditionally, localisation relies on pixel-level annotations provided by a medical expert. Although effective, such annotations are expensive, time-consuming, difficult to scale, and subject to inter-observer variability, particularly for large clinical chest X-ray datasets and rare pathologies [2, 3]. These limitations have motivated growing interest in *annotation-free* and *weakly supervised* chest X-ray image localisation, where models learn to localise disease regions without explicit spatial annotations [4, 5]. A widely adopted strategy exploits paired image-text data, using free-text radiology reports as weak supervision [6, 7, 8]. Because reports are routinely generated in clinical workflows and describe observed findings, they offer a scalable alternative to manual annotation.

However, radiology reports are not designed for precise spatial supervision of chest X-ray findings. They are free text shaped by reporting style, institutional conventions, and contextual priorities [9].

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'26 Workshop, Delft (The Netherlands), 29-March-2026

*Corresponding author.
tahsir.a.munna@inesctec.pt (T. A. Munna); nuno.r.guimaraes@inesctec.pt (N. Guimarães); amjorge@fc.up.pt (A. Jorge)
0000-0001-9269-502X (T. A. Munna); 0000-0003-2854-2891 (N. Guimarães); 0000-0002-5475-1382 (A. Jorge)

Reports may omit visual details, vary in specificity, or use ambiguous language, limiting their reliability as surrogates for structured medical knowledge [4, 10, 11]. While most annotation-free localisation methods implicitly assume that clinical free-text supervision alone is sufficient, clinical reasoning relies heavily on *domain knowledge*, including formal disease definitions, typical anatomical locations, characteristic visual appearances, and expected spatial extent. Current models, however, must largely infer such semantics from unstructured free text [12]. Additionally, although currently ,as our future work, a fully fledged narrative approach to text semantics enables the unfolding of role and temporal relations between concepts, players and events in clinical reports [13].

In this position paper, we argue that annotation-free localisation of findings in chest X-ray images can be strengthened by *integrating explicit domain knowledge in a narrative-aware approach*, where a narrative refers to free-form human-authored text such as radiology reports. Rather than replacing free-text supervision or imposing rigid ontological constraints, we propose enriching clinical free-text with disease-centric descriptions that encode clinically meaningful attributes such as definition, location, appearance, shape, and extent. For example, a finding such as *lung opacity*—a pathology commonly identified in chest X-rays —can be described through a structured approach, capturing its visual density and anatomical distribution. This approach preserves the narrative nature of clinical language while exposing models to more explicit, medically grounded semantic information. Initial empirical observations indicate *progressive improvements in localisation performance* compared to clinical free-text supervision alone. Although the gains are modest and not yet statistically significant, their consistency across metrics and qualitative results suggests that richer narrative-aware knowledge exploration can better guide spatial grounding.

Position and Key Insights

In the spirit of a position paper, this work does not claim a final solution to annotation-free medical image localisation. We are in the process of developing a richer and more effective approach from the outlined narrative-aware principles. The paper advances the following key insights:

- Free-text clinical reports of chest X-rays can be enhanced as reliable supervision for annotation-free localisation.
- Explicit domain knowledge, particularly if expressed in a narrative-aware form, provides complementary semantic grounding for chest X-rays.
- Enriching weak supervision with structured narrative semantic information leads to progressive improvements in localisation quality on chest X-ray benchmarks, as supported by initial empirical evidence.
- Annotation-free chest X-ray localisation systems should explicitly separate narrative variability from domain knowledge to achieve clinically robust performance.

Reframing annotation-free localisation as an opportunity for *jointly modelling clinical narratives and domain knowledge*, we aim to stimulate discussion and future research on knowledge-aware vision-language learning in medical imaging.

2. Related Work

Annotation-free and weakly supervised chest X-ray localisation has emerged as a scalable alternative to pixel-level supervision, driven by the high cost, limited availability, and inter-observer variability of expert annotations [14, 15]. Early work relied on image-level labels and class activation mapping to obtain coarse localisation cues, but such methods often lack semantic grounding and struggle with complex pathologies [16, 17].

The availability of large-scale paired chest X-ray–text datasets has recently enabled vision–language approaches that leverage free-text radiology reports as weak supervision [6, 8]. By aligning medical images with clinical text, these models learn joint representations that support retrieval, report generation,

and annotation-free localisation [6, 4]. In particular, multimodal contrastive learning frameworks that exploit global, sentence-level, and word-level image–text alignment have demonstrated strong localisation performance without explicit bounding box annotations [18, 19]. Among these, the annotation-free localisation framework of Yang et al. [4] represents a strong state-of-the-art baseline.

However, most annotation-free vision–language localisation methods implicitly assume that free-text radiology reports are sufficient carriers of medical knowledge. In practice, reports are stylistically variable, incomplete, and context-dependent, often omitting visual attributes or using ambiguous phrasing [20]. Recent studies further indicate that models trained on chest X-ray reports may overfit to linguistic patterns rather than learning robust concept-level grounding [10, 21].

To mitigate these limitations, knowledge-aware vision–language models have explored incorporating external medical knowledge, such as ontologies, concept embeddings, or textual definitions [12, 22]. Zhang et al. [12] enrich radiology pretraining with ontology-derived concept definitions, and Chen et al. [23] incorporate medical knowledge graphs via an *align, reason and learn* framework that explicitly separates factual knowledge from narrative supervision. At the task level, Huang et al. [24] inject disease-level textual definitions into a U-Transformer for radiology report generation, showing that concept-level grounding leads to more clinically accurate outputs. While effective, many of these approaches rely on rigid ontological structures or direct concatenation of definitions with narrative text, which may limit flexibility and inadequately capture the narrative nature of clinical language—particularly in annotation-free localisation settings.

A complementary direction concerns how knowledge is *represented* rather than merely injected. Li et al. [10] show that augmenting vision–language models with structured knowledge descriptions of abnormalities–covering typical appearance, location, and severity–substantially improves grounding in chest X-ray images, even when such descriptions are absent at inference time. This finding directly motivates our approach: rather than relying on ontology look-ups or rigid templates, we encode disease-centric attributes in free-form narrative language, preserving clinical expressiveness while supplying the structured semantics that raw reports lack.

The use of structured background knowledge to guide visual recognition has a broader history in computer vision. Foundational work by Lampert et al. [25] introduced *semantic attribute learning*, in which per-class visual attribute descriptions—such as shape or anatomical location—are used to condition classifiers at test time without any target-class training images. Knowledge graph-based approaches have since extended this principle to bridge seen and unseen categories in zero-shot and few-shot settings [26, 27]. In the medical domain, ontology-conditioned classifiers have used disease hierarchies from resources such as RadLex and UMLS to improve generalisation to unseen pathologies [28, 29], and trait-guided semantic embeddings of visual attributes have been applied specifically to chest X-ray zero-shot diagnosis [30]. Qin et al. [31] demonstrated that injecting expert-level medical knowledge into prompts of pretrained vision–language models enables fine-grained grounding without additional labelled data, underscoring that knowledge *expressiveness*—not only its availability—is key to generalisation.

Building on these lines of work, we focus on how domain knowledge specific to chest X-ray pathologies is represented and integrated in an annotation-free localisation setting. Rather than relying on discrete attribute vectors or rigid ontological constraints, we propose expressing disease-centric knowledge in a narrative-aware form and treating it as a contrastive training signal. This formulation preserves the continuous, context-sensitive nature of clinical language while providing vision–language models with the structured concept-level grounding that free-text reports alone cannot reliably supply.

3. From Free-Text Reports to Narrative-Aware Domain Knowledge Injection

Baseline Background

The baseline model, namely AFLoc, proposed by Yang et al. [4], addresses annotation-free pathology localisation in chest X-ray images by leveraging paired chest X-ray–report data from MIMIC-CXR [32] as weak supervision, without requiring any explicit spatial annotations. The framework follows a multimodal vision–language learning strategy (Figure 1, baseline method). A ResNet-50 [33] image encoder extracts multi-scale visual features that preserve spatial structure, producing a 19×19 spatial feature map, where each cell corresponds to a local chest X-ray patch. A ClinicalBERT [34] text encoder processes free-text radiology reports to obtain hierarchical representations at the report, sentence, and word levels. Training is driven by three hierarchical contrastive alignment objectives. First, a global-level loss aligns the whole-image embedding with the full report embedding, ensuring that the overall visual representation is semantically consistent with the clinical narrative. Second, a sentence-level loss aligns regional image features with individual report sentences, encouraging the model to associate specific chest X-ray regions with the corresponding clinical findings described in each sentence. Third, a word-level loss aligns local patch embeddings with individual disease-related tokens, providing fine-grained grounding signals that link sub-regional image patches to specific pathological concepts in the text. At inference, localisation maps are produced by computing cosine similarity between each spatial patch embedding and the word-level embeddings of a query phrase, yielding a dense 19×19 similarity map over the chest X-ray. Regions exceeding a similarity threshold are treated as predicted localisation areas and evaluated against ground-truth bounding boxes. Together, these three objectives produce similarity maps that associate pathological findings with corresponding chest X-ray regions in a fully annotation-free manner, without requiring any additional labelled data or architectural modifications at inference time.

In our experiments, our proposed model is initialised from the publicly available AFLoc pre-trained checkpoint. During fine-tuning, the ClinicalBERT encoder weights are kept frozen for the first epoch to stabilise early training, after which both encoders are jointly updated for the remaining steps. We adopt the same training hyperparameters as in the original AFLoc framework, but to enrich the training signal, each free-text radiology report is probabilistically augmented with a disease-centric narrative description that encodes clinically meaningful attributes. This narrative-aware augmentation is applied only during training and is never available at inference time, ensuring that localisation remains fully annotation-free.

Why Free-Text Supervision Is Limited?

Clinical reasoning about chest X-ray findings relies on more than narrative descriptions alone. When interpreting chest X-ray images, clinicians implicitly draw on structured domain knowledge, including formal disease definitions, typical anatomical locations, characteristic visual appearances, and expected spatial extent. This knowledge shapes how visual evidence is interpreted, even when such details are not explicitly stated in a report. In contrast, annotation-free localisation models trained solely on free-text radiology reports must implicitly infer this domain knowledge from narrative patterns. Free-text reports are written for clinical communication rather than visual supervision and often omit precise spatial details, vary in phrasing across radiologists, and emphasise clinical context over visual characteristics. As a result, models relying exclusively on free-text supervision may develop weak or unstable text-image alignment, become sensitive to reporting style, or rely on textual shortcuts rather than visual evidence. This mismatch between narrative language and structured medical knowledge motivates our central position.

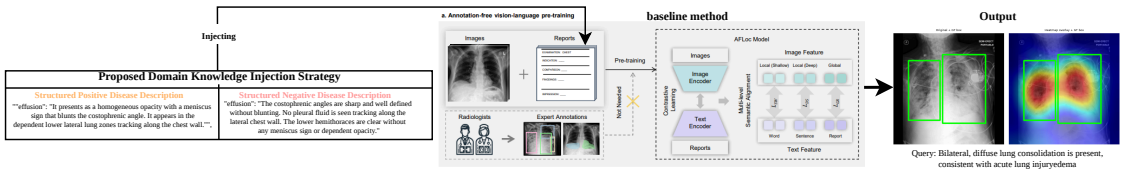

Figure 1: Domain Knowledge Injection Strategy Applied to the Baseline [4]

Proposed Domain Knowledge Injection Strategy

In our proposed strategy, domain knowledge specific to chest X-ray pathologies is injected in a narrative-aware manner during training, as shown in Figure 1. For each of the eight cardiopulmonary pathologies in MIMIC-CXR, we manually construct a compact disease-centric narrative description following the approach of MedKLIP [35]. Rather than extracting descriptions from ontologies such as MeSH [36] or UMLS [37], or generating them automatically with a language model, each description is hand-crafted to encode only *visually verifiable* attributes — what the finding looks like on a chest X-ray and where it is anatomically located. Pathophysiological mechanisms and clinical context are deliberately excluded, as they cannot be verified visually and would introduce distribution shifts relative to the report language on which the model was pre-trained. Each description follows a two-sentence Appearance–Location template: the first sentence describes visual appearance (opacity type, density, and key radiographic signs), and the second specifies anatomical location and spatial distribution (see Table 1). Restricting each description to approximately 40–45 ClinicalBERT tokens ensures that the 97-token context window is not dominated by injected text, leaving sufficient budget for the original report sentence.

Table 1
Disease-centric narrative descriptions for the eight MIMIC-CXR cardiopulmonary pathologies, each encoding visual appearance and anatomical location in a two-sentence Appearance–Location template.

Pathology	Description	Pathology	Description
Atelectasis	Linear or plate-like opacities with volume loss and ipsilateral diaphragm elevation; lower lobes, particularly the left base.	Cardiomegaly	Enlarged cardiac silhouette with cardiothoracic ratio >0.5; central mediastinum projecting into both lower lung zones.
Consolidation	Dense homogeneous airspace opacity with air bronchograms; lower lobes, particularly the right lower lobe.	Edema	Bilateral perihilar haziness with butterfly distribution and Kerley B lines; central hilar regions and both lower zones.
Effusion	Homogeneous opacity with meniscus sign blunting the costophrenic angle; dependent lower lateral lung zones.	Emphysema	Bilateral hyperlucency with flattened diaphragms and decreased vascular markings; upper and middle lung zones.
Pneumonia	Segmental or lobar airspace opacity with air bronchograms and indistinct borders; lower lobes and right middle lobe.	Pneumothorax	Visible visceral pleural line with absent peripheral lung markings; apex and upper lateral pleural space on PA radiograph.

At training time, each report sentence is scanned for disease mentions using a negation-aware longest-match alias lookup over a fixed vocabulary of the eight pathologies and their common synonyms (e.g., *hydrothorax* → Effusion, *hyperinflation* → Emphysema). A negation check inspects a ±5-word window around each match for cues such as *"no"*, *"without"*, and *"ruled out"*; only positively-affirmed mentions proceed. When a disease is detected, the corresponding narrative description is injected with probability $p_{\text{inj}} = 0.5$. The narrative and the report sentence are then encoded as a BERT sentence pair: the narrative is assigned to sequence slot A (protected from truncation) and the report to slot B (truncated from the right if the combined length exceeds 97 tokens). No additional loss term is introduced; the injected narrative acts purely as an enriched input to the existing AFLoc alignment objectives, steering the text encoder toward spatially grounded representations without modifying the training procedure.

At inference, we follow the AFLoc query strategy of using precise disease descriptions [4] as the

text query, rather than raw report sentences, without any modification to the inference pipeline. No narrative description is injected at inference time, ensuring that localisation remains fully annotation-free and consistent with the original AFLoc inference design. The choice of precise disease descriptions as the query is deliberate: since our model has been trained with domain-aware narrative descriptions encoding explicit visual and anatomical attributes, it is particularly well-suited to this query design, enabling the similarity map to more accurately highlight the corresponding pathological region in the chest X-ray. In this way, the benefit of narrative-aware training is carried implicitly into inference through enriched visual-semantic representations, rather than through any modification to the query itself.

4. Position: Implications & Evaluation for Localisation

Having described the narrative-aware injection pipeline above, we now report its effect on chest X-ray localisation performance. We fine-tune the AFLoc baseline on 171,008 frontal-view chest X-ray–report pairs from MIMIC-CXR. Following the same strategy as AFLoc, we restrict training to frontal-view images, as the paired radiology reports in MIMIC-CXR correspond exclusively to frontal projections, ensuring consistency between the visual and textual modalities and enabling a fair comparison with the baseline. We evaluate on the full MS-CXR benchmark [38], which consists exclusively of frontal-view chest X-ray images and provides manually annotated bounding boxes for pathological findings, ensuring end-to-end consistency across training and evaluation.

Localisation performance is measured using IoU [39] and Dice [40] for spatial overlap, and CNR [41] for contrast separation between predicted pathological regions and surrounding tissue. IoU measures the ratio of the intersection to the union of the predicted and ground-truth bounding boxes, while Dice computes the harmonic mean of precision and recall over the predicted region, providing a complementary perspective on spatial overlap. CNR quantifies how distinctly the predicted activation region stands out from the surrounding background tissue, reflecting the sharpness and specificity of the localisation map rather than its spatial extent alone. Together, these metrics capture both spatial accuracy and the quality of region activation. As shown in Table 2, the narrative-augmented model (Ours) consistently improves over the AFLoc baseline across all three metrics. While the gains are modest and not yet statistically significant given the scale of evaluation, their consistency across metrics supports our central position that enriching free-text supervision with structured narrative domain knowledge is a promising direction for strengthening annotation-free localisation in chest X-ray images. We view this formulation as an intermediate step toward more clinically grounded vision-language learning, rather than a final solution.

Evidence from Preliminary Experiments

Table 2 reports results on the full MS-CXR evaluation set using mean values across thresholds {0.1, 0.2, 0.3, 0.4, 0.5}, where each threshold binarises the cosine similarity map into a predicted region for IoU and Dice computation; averaging across thresholds ensures the evaluation is robust to operating-point selection rather than optimised for a single cutoff. The results have been compared against two additional vision-language baselines: BioViL [42] and GLoRIA [6], along with AFLOC. All methods are evaluated under the *precise description* setting, which uses detailed phrase queries that more closely match the kind of structured knowledge our approach provides. Narrative-aware knowledge injection improves mean IoU from 0.324 to 0.330, mean Dice from 0.462 to 0.470, and CNR from 1.636 to 1.671 over the AFLoc baseline, while consistently outperforming both BioViL and GLoRIA across all three metrics. Although the gains over AFLoc are modest in absolute terms, they are consistent across all metrics and represent a directional improvement achieved purely through narrative enrichment of the text input without any architectural change or additional annotation.

These results should be interpreted as evidence of directional progress rather than definitive performance gains. The improvements are consistent across metrics and evaluation scales, and we acknowledge that the current gains over the AFLoc baseline are not yet statistically significant. We are actively

extending this work with longer training schedules, refined knowledge representations, and larger-scale statistical evaluation to further strengthen localisation performance and reduce variance.

Table 2
Localisation performance on the full MS-CXR benchmark (1,162 chest X-ray–phrase pairs, 8 cardiopulmonary pathologies), *precise description* setting. Mean values are averaged across thresholds {0.1, 0.2, 0.3, 0.4, 0.5}. **Bold** indicates best result per metric.

Method	IoU	Dice	CNR
BioViL [42]	0.228	0.342	1.083
GLoRIA [6]	0.268	0.392	1.287
AFLoc [4]	0.324	0.462	1.636
Ours	**0.330**	**0.470**	**1.671**

5. Conclusion and Limitations

This position paper argues that annotation-free chest X-ray localisation should move beyond treating free-text radiology reports as implicit carriers of structured medical knowledge. While reports offer scalable weak supervision, their narrative and subjective nature limit robust spatial grounding when used in isolation. We show that explicitly introducing domain knowledge—specific to chest X-ray pathologies—in a narrative-aware manner provides a principled way to complement free-text supervision without replacing it. Our initial experiments on the full MS-CXR benchmark indicate consistent, though modest, improvements across all localisation metrics, suggesting that separating narrative variability from domain knowledge is a promising direction, even if the current gains are not yet statistically conclusive. The primary limitation of this work lies in the early-stage nature of the knowledge-injection design and the modest scale of empirical validation. Performance remains sensitive to design choices such as knowledge placement and feature pooling, and we emphasise that this study represents an intermediate step towards a more robust solution. As a key future direction, we plan to enrich narrative-aware semantic alignment by incorporating the patient diagnostic journey [43], where temporally structured clinical narratives provide semantically precise grounding cues derived from a patient's longitudinal diagnostic history—offering a richer and more contextually grounded supervision signal than isolated report sentences. We are actively experimenting with improved knowledge representations, stronger narrative–knowledge separation, and refined alignment strategies, and we hope this work encourages future research on clinically grounded, narrative-aware, annotation-free chest X-ray localisation beyond report-centric learning.

Acknowledgments

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project HfPT, with reference 41.

Declaration on Generative AI

The authors utilised Claude AI for camera-ready manuscript refinement and Writefull (via Overleaf) to enhance writing quality and readability. All AI-assisted content was subsequently reviewed and revised by the authors, who bear full responsibility for the accuracy and integrity of the published work.

References

[1] A. M. Al-Zoghby, A. Ismail Ebada, A. S. Saleh, M. Abdelhay, W. A. Awad, A comprehensive review of multimodal deep learning for enhanced medical diagnostics., Computers, Materials & Continua 84 (2025).

[2] Z. Huang, Z. Wang, T. Zhao, X. Ding, X. Yang, Toward high-quality pseudo masks from noisy or weak annotations for robust medical image segmentation, Neural Networks 181 (2025) 106850.

[3] A. M. Freire, E. M. Rodrigues, J. V. Sousa, M. Gouveia, D. Ferreira-Santos, T. Pereira, H. P. Oliveira, P. Sousa, A. C. Silva, M. S. Fernandes, et al., Clinical annotation and medical image anonymization for ai model training in lung cancer detection, in: International Conference on Human-Computer Interaction, Springer, 2025, pp. 309–325.

[4] H. Yang, H.-Y. Zhou, J. Liu, W. Huang, C. Li, Z. Li, Y. Gao, Q. Liu, Y. Liang, Q. Yang, et al., A multimodal vision–language model for generalizable annotation-free pathology localization, Nature Biomedical Engineering (2026) 1–15.

[5] M. A. Shawkat, M. Hasan, T. Hasan, Weakly supervised tuberculosis localization in chest x-rays through knowledge distillation, arXiv preprint arXiv:2512.11057 (2025).

[6] X. Wang, et al., Gloria: A multimodal global-local representation learning framework, ICCV (2021).

[7] M. A. Shaaban, A. Khan, M. Yaqub, Medpromptx: Grounded multimodal prompting for chest x-ray diagnosis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 211–222.

[8] K. Zou, Y. Bai, Z. Chen, Y. Zhou, Y. Chen, K. Ren, M. Wang, X. Yuan, X. Shen, H. Fu, Medrg: Medical report grounding with multi-modal large language model, arXiv preprint arXiv:2404.06798 (2024).

[9] U. Kulsoom, F. G. Glavin, M. Bendechache, Natural language processing and machine learning for analysis of radiology reports-a systematic review, IEEE Access (2025).

[10] J. Li, C. Liu, W. Bai, R. Arcucci, C. I. Bercea, J. A. Schnabel, Enhancing abnormality grounding for vision language models with knowledge descriptions, arXiv preprint arXiv:2503.03278 (2025).

[11] J. Sato, K. Sugimoto, Y. Suzuki, T. Wataya, K. Kita, D. Nishigaki, M. Tomiyama, Y. Hiraoka, M. Hori, T. Takeda, et al., Annotation-free multi-organ anomaly detection in abdominal ct using free-text radiology reports: a multi-centre retrospective study, EBioMedicine 110 (2024).

[12] Y. Zhang, et al., Knowledge-enhanced vision–language pretraining for medical imaging, IEEE Transactions on Medical Imaging (2023).

[13] A. L. Fernandes, P. Silvano, N. Guimarães, R. Rb-Silva, T. A. Munna, L. F. Cunha, A. Leal, R. Campos, A. Jorge, Human experts vs. large language models: Evaluating annotation scheme and guidelines development for clinical narratives, Proceedings of the Text2Story (2025) 149–160.

[14] G. Litjens, et al., A survey on deep learning in medical image analysis, Medical Image Analysis (2017).

[15] J. Irvin, et al., Chexpert: A large chest radiograph dataset, AAAI (2019).

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[17] R. R. Selvaraju, M. Cogswell, A. Das, et al., Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[18] M. Imran, Y. Lee, Multimodal vision–language models in medical imaging: A survey of retrieval, interpretability, and trust, IEEE Access (2026).

[19] Y. Nam, D. Y. Kim, S. Kyung, J. Seo, J. M. Song, J. Kwon, J. Kim, W. Jo, H. Park, J. Sung, et al., Multimodal large language models in medical imaging: current state and future directions, Korean Journal of Radiology 26 (2025) 900.

[20] D. Demner-Fushman, et al., Preparing a corpus of clinical notes, Journal of Biomedical Informatics (2016).

[21] S. Jain, et al., A two-stage framework for weakly supervised medical image localization, Medical

Image Analysis 70 (2021) 102001.
[22] A. Singhal, et al., Towards knowledge-grounded vision–language models for medical imaging, Nature Machine Intelligence (2023).
[23] Z. Chen, G. Li, X. Wan, Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5152–5161. doi:10.1145/3503161.3547948.
[24] Z. Huang, X. Zhang, S. Zhang, KiUT: Knowledge-injected U-transformer for radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19809–19818.
[25] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 453–465. doi:10.1109/TPAMI.2013.140.
[26] X. Wang, Y. Ye, A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6857–6866.
[27] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, E. P. Xing, Rethinking knowledge graph propagation for zero-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 11487–11496.
[28] Y. Zhang, et al., Knowledge-enhanced visual-language pre-training on chest radiology images, Nature Communications 14 (2023) 4542. doi:10.1038/s41467-023-40260-7.
[29] D. Mahapatra, et al., Multi-label generalized zero shot chest X-ray classification by combining image-text information with feature disentanglement, IEEE Transactions on Medical Imaging 44 (2025) 31–43. doi:10.1109/TMI.2024.3429471.
[30] U. Hayat, L. Shen, Generalized zero-shot chest X-ray diagnosis through trait-guided multi-view semantic embedding with self-training, arXiv preprint arXiv:2102.09916 (2021).
[31] Z. Qin, H. Yi, Q. Lao, K. Li, Medical image understanding with pretrained vision language models: A comprehensive study, in: International Conference on Learning Representations (ICLR), 2023.
[32] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, S. Horng, MIMIC-CXR Database, PhysioNet (2024). URL: https://doi.org/10.13026/4jqj-jw95. doi:10.13026/4jqj-jw95, version 2.1.0.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[34] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
[35] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. 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.
[36] National Library of Medicine, Medical Subject Headings (MeSH), https://www.nlm.nih.gov/mesh, 2024.
[37] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.
[38] S. Bannur, S. Hyland, Q. Liu, F. Pérez-García, M. Ilse, D. Coelho de Castro, B. Boecking, H. Sharma, K. Bouzid, A. Schwaighofer, M. T. Wetscherek, H. Richardson, T. Naumann, J. Alvarez Valle, O. Oktay, MS-CXR-T: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing, PhysioNet (2023). URL: https://doi.org/10.13026/pg10-j984. doi:10.13026/pg10-j984, version 1.0.0.
[39] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.
[40] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, M. B. Blaschko, Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index, IEEE transactions on medical imaging 39 (2020) 3679–3690.

[41] N. Oberhofer, G. Compagnone, E. Moroder, Use of cnr as a metric for optimisation in digital radiology, in: World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany: Vol. 25/2 Diagnostic Imaging, Springer, 2009, pp. 296–299.

[42] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, et al., Learning to exploit temporal structure for biomedical vision-language processing, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 15016–15027.

[43] T. A. Munna, A. L. Fernandes, P. Silvano, N. Guimarães, A. Jorge, Using llms to generate patient journeys in portuguese: an experiment, in: Proceedings of Text2Story-Eighth Workshop on Narrative Extraction From Texts held in conjunction with the 47th European Conference on Information Retrieval (ECIR 2025) CEUR, 2025.