MRI2Rep: Autoregressive Structured Report Generation for 3D Liver MRI

Xinran Li[1], Junlin Yang[2], Annabella Shewarega[2], Zongwei Zhou[4,5], Julius Chapiro[2], James S. Duncan[1,2,3], and Lawrence H. Staib[1,2,3⋆]

[1] Department of Biomedical Engineering
[2] Department of Radiology & Biomedical Imaging
[3] Department of Electrical Engineering
Yale University, New Haven, CT 06520, USA
[4] Johns Hopkins University, Baltimore, MD 21218, USA
[5] Johns Hopkins Medicine, Baltimore, MD 21287, USA
lawrence.staib@yale.edu

Abstract. Manual reporting of 3D MRI studies is time-consuming, yet end-to-end, structured report generation for 3D liver MRI remains underexplored due to volumetric complexity and scarce paired data. We propose MRI2Rep, an autoregressive framework for liver MRI report generation. From 3,929 real-world MRI–report pairs acquired over a 10-year single-institution cohort, a Report-to-Label Canonicalization (RLC) module converts the free-text reports into structured, closed-vocabulary diagnostic sequences without lesion-level annotations. On a held-out test set, MRI2Rep achieves 76.0% case-level sensitivity, 29.4% lesion-level F1 (vs. $\leq$8.3% for adapted medical vision–language baselines), and 82.4% liver-level accuracy. In a blinded reader study, two radiologists rated 75%/70% of AI-generated reports clinically acceptable (vs. 95%/100% for originals), while our automated LLM-based judge (LLM-Eval, 61.8%) applies a stricter standard, validating it as a conservative proxy. To our knowledge, this is the first end-to-end LI-RADS-structured reporting system for 3D liver MRI. The code is available at https://github.com/Alena-Xinran/MRI2Rep.git

Keywords: 3D Medical Image · Radiology Reports · Report Generation

1 Introduction

Radiology reports serve as the primary communication between radiologists and clinicians, translating complex imaging findings into actionable clinical insights. For 3D MRI studies, manual interpretation and report writing is time-intensive and can take radiologists 10–15 minutes per case [4]. Such per-case reporting time directly constrains clinical throughput and is associated with substantial burnout among radiologists (reported prevalence up to 49%) [8]. While recent 3D-MRI vision–language models target representation learning, classification,

⋆ Corresponding author.

or VQA [26,28,29,32], *end-to-end* generation of clinically structured, guideline-conformant reports for 3D liver MRI remains largely unaddressed.

The key challenges are twofold. First, report-level supervision is noisy: terminology varies and negation is common (e.g., "no evidence of HCC" implies a negative label) [14,24]—tolerable in large 2D datasets [18] but a major overfitting risk for much smaller 3D MRI cohorts. Second, 3D MRI adds substantial computational and representational complexity: models must encode high-dimensional volumes and capture subtle cross-sequence cues while modelling long-range 3D spatial relationships under tight memory and compute budgets.

To address these challenges, we propose MRI2Rep, an autoregressive framework that directly translates 3D MRI volumes into structured radiology reports. MRI2Rep first converts noisy free-text reports into clean, closed-vocabulary supervision via a Report-to-Label Canonicalization (RLC) module (§3.1), then trains an encoder–decoder model (§3.2) to predict structured diagnostic sequences from volumetric inputs, and finally renders the predicted sequences into professional reports via deterministic templates (§3.3). Crucially, MRI2Rep requires no lesion-level spatial annotations (bounding boxes or segmentation masks), which are expensive and rarely available in routine clinical data; lesion location is instead inferred purely from volumetric features and report-derived supervision, making the framework directly applicable to retrospective clinical cohorts. We validate MRI2Rep on unseen test data using both automated metrics and a blinded clinical reader study, with full results reported in §4. Our main contributions are:

- A dataset of 3,929 liver MRI report–image pairs, of which 3,830 include complete multi-phase sequences with canonicalized RLC labels.
- RLC, a LI-RADS-guided canonicalization module that converts free-text reports into auditable, closed-vocabulary sequences by resolving negation, uncertainty, and phrasing variation.
- An encoder–decoder model that autoregressively predicts structured diagnostic sequences from multi-phase 3D MRI and renders them into standardized reports.

2 Related Work

Medical Report Generation. Early work focused on 2D chest X-rays using MIMIC-CXR [18] and CheXpert [14] with encoder–decoder architectures [7,17]. More recent 2D methods inject structured findings as an auxiliary task or report basis (ORGAN [13], structured multi-task learning [20]) and provide grounded datasets (PadChest-GR [6]), but remain 2D without LI-RADS reasoning. CT and 3D vision–language models (Merlin [5], CT2Rep [12], M3D [2], RadFM [31], Triad [29], Decipher-MR [32]) target generic reporting, representation learning, or classification; liver-MRI models address lesion classification or VQA (Liver-VLM [26], HepaPathGPT [28]). RadGPT [3] generates reports from 3D tumour masks but needs lesion-level annotations unavailable in routine MRI. None en-

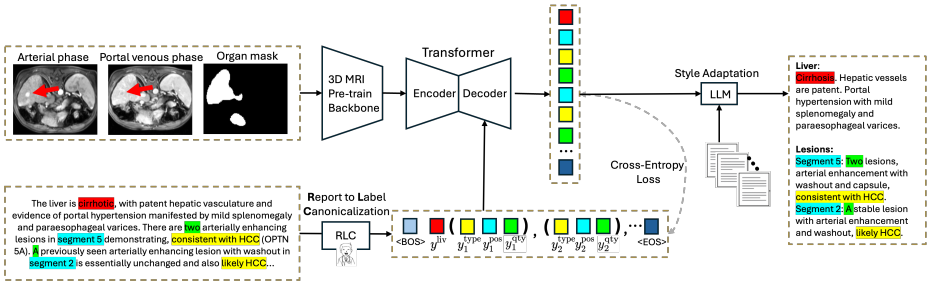

Fig. 1. Overview of MRI2Rep: given a 3D liver MRI volume, the model autoregressively generates a structured radiology report by predicting a sequence of diagnostic findings directly from the image. An auxiliary classification head (detailed in §3.2) provides an additional lesion-level supervision signal to encourage the visual encoder to retain discriminative features.

force LI-RADS decision rules or produce closed-vocabulary structured output without spatial annotations.

LI-RADS and Structured Liver Imaging. LI-RADS [25] provides a standardized algorithm for categorizing liver observations. Prior computational work focuses on classification of pre-segmented lesions [21, 22] rather than end-to-end report generation. MRI2Rep is the first to embed LI-RADS reasoning into both supervision and output vocabulary, using 9 Couinaud sectors [10] without lesion-level annotations.

Structured Label Extraction. Negation, uncertainty, and phrasing variation complicate label extraction [14, 24]. Rule-based systems address this for chest radiology; LLM-based chain-of-thought approaches [19, 30] improve accuracy. RLC extends this to liver MRI with LI-RADS decision logic.

3 MRI2Rep

MRI2Rep operates in three stages—RLC canonicalization (§3.1), autoregressive encoder–decoder prediction (§3.2), and deterministic template rendering (§3.3); Figure 1 summarizes the pipeline.

3.1 Report-to-Label Canonicalization (RLC)

Free-text reports exhibit heterogeneous phrasing, negation, and uncertainty [14, 24]. RLC converts each report R into a structured, closed-vocabulary sequence $\mathbf{y} = (y_1, \ldots, y_T)$ comprising (i) a liver-background token $y^{\text{liv}} \in \mathcal{Y}_{\text{liv}}$ and (ii) up to $K = 8$ lesion triplets $\{(y_k^{\text{type}}, y_k^{\text{pos}}, y_k^{\text{qty}})\}$, serialized as `[BOS, liver, type`$_1$`, pos`$_1$`, qty`$_1$`, ..., EOS]`; each triplet is paired with an evidence sentence $e_k \subset R$ for auditability [16].

Concretely, RLC is an LI-RADS-guided LLM prompting protocol, not a black-box call: an LLM (Claude-3.5; GPT-5 and DeepSeek-R1 also evaluated, Table 1) receives the LI-RADS major-feature decision logic with chain-of-thought

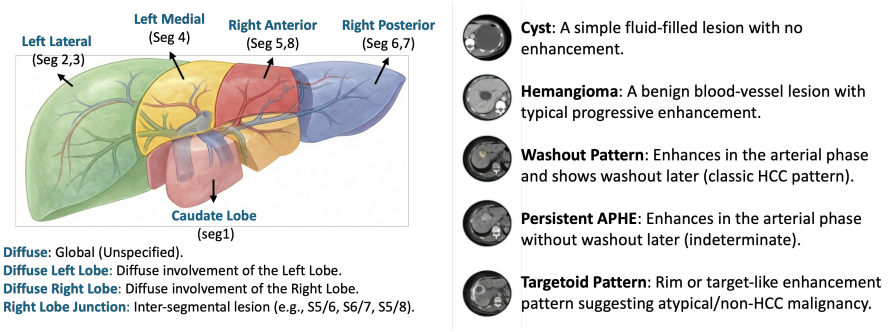

Fig. 2. Output vocabulary: location sectors $\mathcal{Y}_{\text{pos}}$ (left, Couinaud-derived) and lesion types $\mathcal{Y}_{\text{type}}$ with representative MRI examples (right).

instructions [19, 30] and, per candidate observation, emits a $(y_k^{\text{type}}, y_k^{\text{pos}}, y_k^{\text{qty}})$ triplet *with* the verbatim evidence sentence $e_k \subset R$ licensing it; triplets are assembled deterministically into $\mathbf{y}$ and abstained when evidence is absent, making every label reproducible and auditable. Following the LI-RADS algorithm [25], the prompt queries sequentially for definitive benign features, targetoid features, and washout dynamics to assign lesion categories. The four token subsets are: $\mathcal{Y}_{\text{liv}}$ (2 tokens: normal, fibrotic/cirrhotic); $\mathcal{Y}_{\text{type}}$ (5 LI-RADS categories), $\mathcal{Y}_{\text{pos}}$ (9 Couinaud-derived surgical sectors [10]), and $\mathcal{Y}_{\text{qty}}$ (4 quantity buckets: single, 2, $\geq$3, diffuse); representative examples of $\mathcal{Y}_{\text{type}}$ and $\mathcal{Y}_{\text{pos}}$ are shown in Fig. 2. The shared token dictionary $\mathcal{V}$ contains special tokens, liver-background tokens, lesion-type tokens, location tokens, quantity-bucket tokens, and `NO_LESION`, giving $|\mathcal{V}| = 24$. Although the LI-RADS categories (LR-1–5) formally apply only to patients at risk for HCC, our vocabulary additionally includes non-HCC tokens (cyst, hemangioma, targetoid) and normal/cirrhotic liver-background tokens, extending coverage beyond the at-risk population.

3.2 Encoder–Decoder Architecture

Following 3D radiology-captioning practice [5,12], we model report generation as an autoregressive conditional distribution $p(\mathbf{y} \mid \mathbf{V}) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, \mathbf{V})$, where $\mathbf{V} \in \mathbb{R}^{3\times H\times W\times D}$ stacks the arterial-phase (ART), portal-venous-phase (PV), and a binary liver segmentation mask (each independently intensity-normalised).

The visual encoder f_θ is a 3D CNN with four Conv3d-IN-GELU blocks and stride-2 downsampling, producing visual memory tokens $\mathbf{E}^{\text{vis}} = f_\theta(\mathbf{V}) \in \mathbb{R}^{S_{\text{vis}}\times d}$. A visual Transformer applies self-attention for global 3D context; the autoregressive decoder g_ϕ [27] then cross-attends to $\mathbf{E}^{\text{vis}}$ and projects to logits over $\mathcal{V}$.

The backbone is pre-trained to reconstruct the PV phase from the ART input (MSE and SSIM loss within the liver mask) before joint fine-tuning with the Transformer. To make the decoder attend to visual features rather than text co-occurrence, we add two mechanisms: (i) an auxiliary binary head—global average pooling over $\mathbf{E}^{\text{vis}}$ feeds a linear classifier predicting lesion presence $\hat{z} \in \{0, 1\}$,

encouraging the backbone to retain lesion-discriminative features; and (ii) word dropout [15], randomly replacing a fraction of decoder input tokens with `PAD` to reduce reliance on the text prefix. The training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda\, \mathcal{L}_{\mathrm{BCE}}(z, \hat{z}), \tag{1}$$

where $\mathcal{L}_{\mathrm{CE}}$ uses per-class weights [9] and label smoothing.

Implementation. Volumes are resized to $192 \times 192 \times 96$; ART/PV normalised independently. CNN: four Conv3d-IN-GELU blocks (channels 32/64/128/256; strides 2/2/2/1), yielding 256-d visual tokens. Decoder: 6 layers, 8 heads, FFN 512, dropout 0.3. AdamW (lr $5{\times}10^{-5}$), cosine schedule, batch 12, early stopping on val Lesion-F1. NVIDIA A6000/RTX 4090.

3.3 Sequence-to-Report Rendering

The predicted token sequence is deterministically mapped to a natural-language report via hand-crafted sentence templates: the liver-background token selects a parenchymal description and each lesion triplet yields a finding sentence (e.g., "A single observation consistent with definite HCC is identified in the right posterior sector"). Stylistic variation is introduced by sampling from a small pool of paraphrase alternatives per template, produced by prompting Claude-3.5 with 2,000 randomly sampled training-set reports to capture institutional writing conventions; this prevents fixed lexical artefacts from unblinding readers in the reader study (§4.3).

4 Results

Dataset & evaluation setup. The cohort comprises 3,929 multi-phase liver-MRI report–image pairs from a 10-year retrospective collection at a single tertiary academic hospital, acquired on Siemens scanners (predominantly 3T) with extracellular gadolinium-based contrast; reports were written by board-certified abdominal radiologists. After RLC canonicalization (§3.1), the 3,830 studies with complete arterial and portal-venous phases are split 80/10/10% (train/val/test) by a stratified set-cover on $(\mathrm{type}, \mathrm{pos}, \mathrm{qty})$, yielding 383 held-out test cases that cover rare LI-RADS categories. k-fold cross-validation is impractical for this 3D autoregressive pipeline; the set-cover split mitigates split variance, and the $\geq$21-point Lesion-F1 margin over the best baseline (n=383) lies well outside any plausible bootstrap interval.

4.1 RLC Validation

We assess RLC reliability via a radiologist retention test on each category's reports: two radiologists (junior, senior) independently mark each AI-produced label as retained if supported by the original report (and evidence sentence when available), averaged across annotators. Table 1 compares direct taxonomy mapping (RLC disabled) against evidence-backed canonicalization (RLC enabled)

Table 1. Radiologist retention rate (% Retained = Rad/AI) across LLM backends for labels extracted from liver MRI radiology reports. Gray parentheses report the underlying counts (Retained / Produced).

	RLC	LLM backend		
		Claude-3.5 [1]	GPT-5 [23]	DeepSeek-R1 [11]
Organ (Liver)	×	76.1 (1619/2128)	84.6 (1801/2128)	80.1 (1705/2128)
	✓	97.1 (1652/1701)	**97.7** (1662/1701)	95.2 (1620/1701)
Lesion (Non-HCC)	×	82.3 (1455/1769)	84.2 (1490/1769)	80.7 (1428/1769)
	✓	**98.0** (1498/1528)	96.3 (1472/1528)	92.3 (1411/1528)
Lesion (HCC)	×	77.9 (392/503)	82.7 (416/503)	76.7 (386/503)
	✓	**98.1** (418/426)	97.0 (413/426)	91.8 (391/426)

across three LLM backends: RLC consistently raises retention from ∼77–84% to ≥92% across all categories and backends, with produced labels slightly reduced as RLC abstains when evidence is insufficient.

4.2 Ablation Study

Predicted lesions are scored on the held-out test set by matching *type* and *location* (and optionally *quantity*) to the ground truth. We report: case-level **Sen**/**Spe** (whether a case contains any lesion); **Lesion-F1**, set-F1 over matched (type, pos) pairs on lesion-positive cases; **Triplet-F1**, strict set-F1 over (type, pos, qty); **Liver-Acc** for the liver-background token; and **LLM-Eval**, the fraction of rendered reports rated acceptable (score 1–2) by Claude-3.5 [1] on a 3-point scale, given rendered and reference reports in randomised order.

Table 2 compares MRI2Rep with 3D vision–language baselines and ablates key choices. For fair comparison, baseline free-text outputs are mapped to the same closed (type, pos, qty) taxonomy and liver-background tokens via Claude-3.5 [1].

MRI2Rep achieves the highest sensitivity (76.0%) and Lesion-F1 (29.4%); label quality is independently validated by the 98.1% retention rate (Table 1). Specificity (68.4%) reflects a deliberate sensitivity–specificity trade-off (oversampling, class-reweighted loss) tunable at inference; prioritising sensitivity aligns with guidelines favouring minimal missed malignancies. The *Multi-label cls.* baseline's lower Lesion-F1 (14.8%) confirms autoregressive decoding better captures inter-lesion dependencies, and the narrow Lesion-F1/Triplet-F1 gap (29.4 vs. 28.1) indicates reliable quantity prediction once a lesion is localised.

Removing ART-to-PV pre-training (*w/o Pretrain*) drops Lesion-F1 to 22.1 (Sen 57.3), confirming pre-training as essential initialisation. Dropping the portal-venous phase (*w/o PV*) causes the largest fall (Lesion-F1 13.5; Sen 46.4): PV washout dynamics are essential for HCC and unrecoverable from ART alone.

Table 2. Ablation and comparison study. We report lesion *detection* (Sen/Spe), lesion *characterisation* as set-F1 over (type, pos) pairs (Lesion-F1) and strict (type, pos, qty) triplets (Triplet-F1), organ-context classification (Liver-Acc), and Claude-3.5 [1] judged clinical consistency (LLM-Eval; see §4 for details). All values are in %; **bold** indicates the best result.

Method	Sen↑	Spe↑	Lesion-F1↑	Triplet-F1↑	Liver-Acc↑	LLM-Eval↑
Comparison with prior work						
M3D [2]	10.5	86.3	6.4	4.1	58.2	17.8
CT2Rep [12]	5.2	91.8	7.2	4.8	61.5	14.6
Merlin [5]	25.1	84.7	5.8	3.9	55.6	21.3
RadFM [31]	21.6	**92.4**	8.3	5.6	66.8	25.2
Architecture baseline						
Multi-label cls.	43.2	74.1	14.8	9.3	79.6	33.4
Ablation						
w/o Vis. Enc	42.7	60.9	10.4	5.9	76.8	29.7
w/o Pretrain	57.3	50.6	22.1	21.3	79.1	47.9
w/o PV	46.4	77.8	13.5	9.9	77.8	37.1
MRI2Rep (Ours)	**76.0**	68.4	**29.4**	**28.1**	**82.4**	**61.8**

Removing visual self-attention (*w/o Vis. Enc.*) lowers Lesion-F1 to 10.4 and Spe to 60.9%, as the decoder over-generates lesions without global 3D context.

4.3 Clinical Reader Study

To validate LLM-Eval, two radiologists (junior/senior) independently scored 100 randomly sampled, randomly ordered test cases (40 AI-generated, 60 original controls) given only the MRI images, blinded to report origin and to the AI:control ratio, on a 3-point scale (acceptance = score 1–2). Radiologists reported no systematic strategy for identifying AI reports, supporting the integrity of the blind.

Radiologists accepted 75%/70% of AI reports vs. 95%/100% for originals, with almost-perfect agreement (Cohen's $\kappa = 0.97$, weighted 0.96). LLM-Eval's lower rate (61.8%) shows Claude-3.5 applies a *stricter* standard than human readers—a desirable conservative bias—and, together with its ≥98% label-extraction accuracy (Table 1), validates it as a reliable, high-throughput automated judge.

4.4 Case Study

Three representative cases (Fig. 3) span a full success (a), a partial success with one hallucinated LR-M (b), and a failure where post-ablation changes are misread as two active lesions (c), illustrating the difficulty of complex multi-finding sequences.

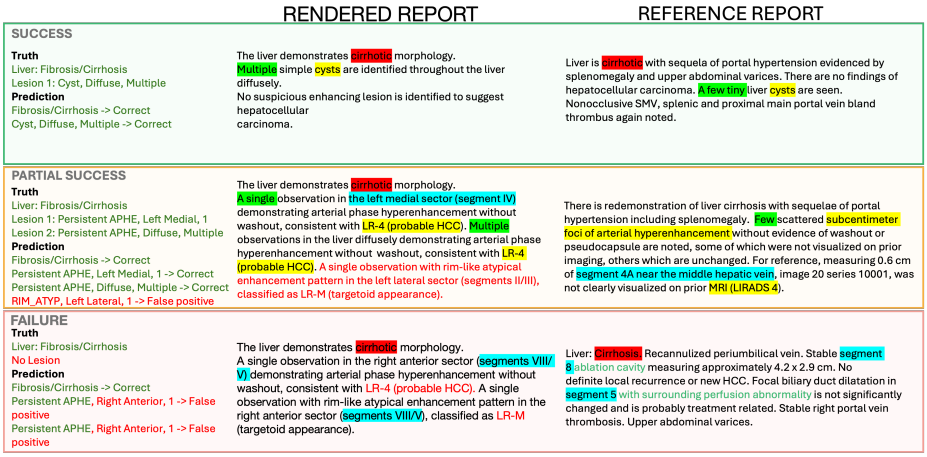

Fig. 3. Qualitative examples from the held-out test set. Each row shows the ground-truth and predicted structured labels (left), the rendered report (centre), and the reference report (right), with key clinical terms highlighted. (a) Full success: all predicted labels match the ground truth; the rendered report captures the essential diagnostic content. (b) Partial success: liver background and two lesions are correctly predicted, but one false-positive LR-M is hallucinated. (c) Failure: the model predicts two active lesions where the reference describes only post-ablation changes, with no evidence of HCC recurrence.

Systematic error analysis reveals three patterns. Per-class Lesion-F1 ranges from 51.7% (cysts) to 11.3% (targetoid, subtle rim enhancement), with washout-pattern lesions (LR-4/5) at 24.6%, reflecting hard cross-phase reasoning. Lesion-F1 drops from 38.4% (single-lesion) to 21.7% (multi-lesion, ≥ 2), as longer sequences accumulate inter-triplet errors. Of all false positives, 34.1% occur in prior-ablation cases where residual enhancement mimics active lesions, motivating treatment history as a future input.

5 Conclusion

MRI2Rep is an autoregressive framework for structured liver MRI report generation: its LI-RADS-guided RLC module yields clean, closed-vocabulary supervision without lesion-level annotations, enabling an encoder–decoder model that outperforms medical vision–language baselines and reaches 75%/70% blinded clinical acceptability. Key limitations: report-only supervision without spatial annotations limits lesion-level F1 and multi-lesion/post-treatment cases (a dominant false-positive source, §4); the closed RLC vocabulary bounds expressible semantics; and single-centre data leaves cross-centre generalisation open. Future work includes treatment-history integration, lesion-level spatial supervision, and external multi-site validation.

Acknowledgments. This work was supported by the National Institutes of Health (NIH) under Award Number R01EB037669.

Disclosure of Interests. The authors declare no competing interests.

References

1. Anthropic: Claude 3.5 sonnet (2024), https://www.anthropic.com/news/claude-3-5-sonnet, announcements, Jun 21, 2024
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
3. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., et al.: Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23720–23730 (2025)
4. Bhargavan, M., Kaye, A.H., Forman, H.P., Sunshine, J.H.: Workload of radiologists in united states in 2006–2007 and trends since 1991–1992. Radiology **252**(2), 458–467 (2009)
5. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., Bluethgen, C., Jensen, M.E.K., Ostmeier, S., Varma, M., Valanarasu, J.M.J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., Weng, W.H., Amaro Junior, E., Ahuja, N., Fries, J., Shah, N.H., Johnston, A., Boutin, R.D., Wentland, A., Langlotz, C.P., Hom, J., Gatidis, S., Chaudhari, A.S.: Merlin: A vision language foundation model for 3d computed tomography (2024). https://doi.org/10.48550/arXiv.2406.06512
6. Castro, D.C., Bustos, A., Bannur, S., et al.: Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. arXiv preprint arXiv:2411.05085 (2024)
7. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 1439–1449 (2020)
8. Chetlen, A.L., Chan, T.L., Ballard, D.H., Frigini, L.A., Hildebrand, A., Kim, S., et al.: Addressing burnout in radiologists. Academic Radiology **26**(4), 526–533 (2019)
9. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
10. Germain, T., Favelier, S., Cercueil, J.P., Denys, A., Krause, D., Guiu, B.: Liver segmentation: practical tips. Diagnostic and Interventional Imaging **95**(11), 1003–1016 (2014). https://doi.org/10.1016/j.diii.2013.12.005
11. Guo, D., Yang, D., Zhang, H., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**, 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z
12. Hamamci, I.E., et al.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer (2024)
13. Hou, W., Xu, K., Cheng, Y., Li, W., Liu, J.: Organ: Observation-guided radiology report generation via tree reasoning. arXiv preprint arXiv:2306.06466 (2023)

14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
15. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) (2015)
16. Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P.: RadGraph: Extracting clinical entities and relations from radiology reports. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks). vol. 1 (2021)
17. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1240
18. Johnson, A.E., 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)
19. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems. vol. 35, pp. 22199–22213 (2022)
20. Liang, J.: Multi-task learning for radiology report generation with structured findings consistency. Computer Science Bulletin **8**(1), 477–489 (2025). https://doi.org/10.71465/csb178
21. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. Neural Networks p. 107228 (2025)
22. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
23. OpenAI: GPT-5 system card. OpenAI technical report (2025), https://openai.com/index/gpt-5-system-card/
24. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings **2018**, 188 (2018)
25. Santillan, C., Fowler, K., Kono, Y., Chernyak, V.: Li-rads major features: Ct, mri with extracellular agents, and mri with hepatobiliary agents. Abdominal Radiology **43**(1), 75–81 (2018). https://doi.org/10.1007/s00261-017-1291-4
26. Song, J., Hu, Y., Wang, H., Chen, Y.W.: Liver-vlm: Enhancing focal liver lesion classification with self-supervised vision-language pretraining. Applied Sciences **15**(23), 12578 (2025). https://doi.org/10.3390/app152312578
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
28. Wang, L., et al.: A generative vision-language model for holistic pathological assessment using preoperative imaging in hepatocellular carcinoma. eBioMedicine **122**, 106060 (2025). https://doi.org/10.1016/j.ebiom.2025.106060

29. Wang, S., Safari, M., Li, Q., Chang, C.W., Qiu, R.L.J., Roper, J., Yu, D.S., Yang, X.: Vision foundation model for 3d magnetic resonance imaging segmentation, classification, and registration. Medical Image Analysis **110**, 103992 (2026). https://doi.org/10.1016/j.media.2026.103992
30. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
31. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025). https://doi.org/10.1038/s41467-025-62385-7
32. Yang, Z., DSouza, N., Megyeri, I., et al.: Decipher-mr: A vision-language foundation model for 3d mri representations. arXiv preprint arXiv:2509.21249 (2025)