LLM-Bootstrapped Targeted Finding Guidance for Factual MLLM-based Medical Report Generation

Cunyuan Yang[1], Dejuan Song[2], Xiaotao Pang[3], Qianqian Shen[1], Wenjie Nie[1], Yifan Huang[1], Wei Han[2], Jiajun Bu[1], Lei Wu[1](✉), and Haishuai Wang[1](✉)

[1] Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science and Technology, Zhejiang University
{cunyuany,shenhai1895,haishuai.wang}@zju.edu.cn
[2] The Second Affiliated Hospital Zhejiang University School of Medicine
[3] Hangzhou Pu Jian Medical Technology Co., Ltd.

Abstract. The automatic generation of medical reports utilizing Multimodal Large Language Models (MLLMs) frequently encounters challenges related to factual instability, which may manifest as the omission of findings or the incorporation of inaccurate information, thereby constraining their applicability in clinical settings. Current methodologies typically produce reports based directly on image features, which inherently lack a definitive factual basis. In response to this limitation, we introduce Fact-Flow, an innovative framework that separates the process of visual fact identification from the generation of reports. This is achieved by initially predicting clinical findings from the image, which subsequently directs the MLLM to produce a report that is factually precise. A pivotal advancement of our approach is a pipeline that leverages a Large Language Model (LLM) to autonomously create a dataset of labeled medical findings, effectively eliminating the need for expensive manual annotation. Extensive experimental evaluations conducted on two disease-focused medical datasets validate the efficacy of our method, demonstrating improved tuberculosis clinical efficacy and most NLG metrics over state-of-the-art models. The code is available at `https://github.com/UCPRER/Fact-Flow`.

Keywords: Medical Report Generation · MLLMs

1 Introduction

The automated generation of medical reports from diagnostic images is a critical task in computational medicine. Early approaches to Medical Report Generation (MRG) relied on encoder-decoder architectures, evolving from CNN-RNN combinations [9,27,29] to Transformer-based models [6,5,19,13]. However, they struggled with capturing complex semantic relationships and generating factually precise narratives [17].

The recent advent of Multimodal Large Language Models (MLLMs) [14] has introduced a new paradigm for MRG. MLLMs such as LLaVA-Med [11] have

demonstrated immense potential, yet a significant challenge emerges when fine-tuned end-to-end: factual instability. These models are prone to hallucinating findings or omitting critical pathological observations, which are clinically unacceptable and remain the primary barrier to real-world deployment [3,28,26].

Inspired by previous work [10,25], we hypothesize that this unreliability stems from coupling two distinct cognitive processes in a single model: visual feature recognition and medical language organization. We propose that by decoupling these tasks—first compelling the model to identify all salient clinical facts before composing the report—we can improve factual grounding.

A primary obstacle is the lack of large-scale datasets pairing medical images with exhaustive key-finding labels. This is especially pronounced in disease-focused settings, where relevant categories are prohibitively expensive to annotate manually. Prior label-guided methods such as TieNet [27] assume a fixed vocabulary tightly coupled to specific datasets, limiting adaptability and rendering them incompatible with modern MLLM architectures.

To overcome this, we introduce Fact-Flow, a novel framework that improves the factual accuracy of MLLM-based report generation through multi-label guidance. The main contributions are as follows.

- We propose Fact-Flow, which improves MLLM-based report generation via explicit multi-label clinical finding conditioning.
- We design a fully automated, LLM-bootstrapped data pipeline that constructs a large-scale (image, multi-label) dataset from existing image-report pairs without manual annotation.
- We validate Fact-Flow on two disease-focused datasets (ophthalmology and tuberculosis), demonstrating consistent improvements over state-of-the-art methods in NLG metrics and, where applicable, clinical efficacy.

2 Method

We present **Fact-Flow**, a framework that enhances MLLM-based medical report generation through multi-label guidance. Given a training set of medical image–report pairs $\{(I_i, R_i)\}_{i=1}^{N}$, our goal is to generate a diagnostic report $\hat{R}$ for a new image I that accurately captures all clinical findings. Rather than mapping $I \to R$ end-to-end, Fact-Flow introduces an intermediate multi-label representation $Y = \{y_1, \ldots, y_k\} \in \{0,1\}^k$, where each y_j denotes the presence or absence of a specific clinical finding. As illustrated in Figure 1, the framework comprises three stages: (1) LLM-bootstrapped label dataset construction, (2) multi-label classification model training, and (3) label-guided report generation, with the predicted labels serving as factual grounding to reduce hallucinations and improve finding recall.

2.1 Stage 1: LLM-Bootstrapped Multi-Label Dataset Construction

Since annotating large-scale reports with fine-grained clinical labels is prohibitively expensive, we construct the label dataset automatically via an LLM-driven pipeline (see Stage 1 in Figure 1). The pipeline proceeds in two steps.

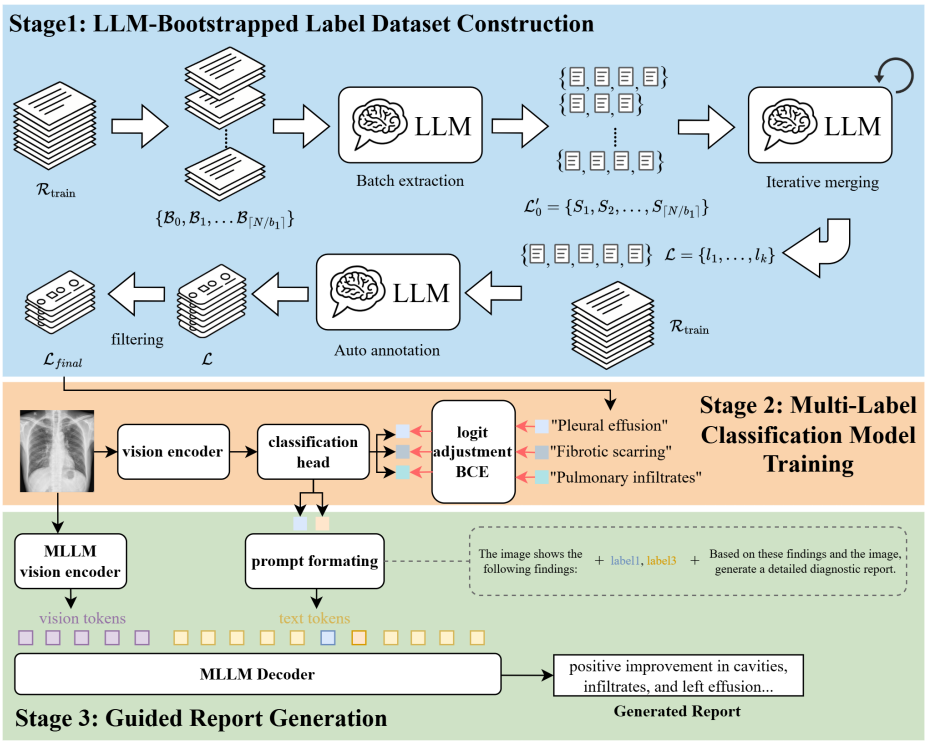


Fig. 1. The overall framework of Fact-Flow

Taxonomy extraction. We first build a unified label taxonomy $\mathcal{L} = \{l_1, \ldots, l_k\}$ from the training reports. Processing all reports in a single LLM call is infeasible due to context-length constraints, so we partition $\{R_i\}_{i=1}^{N}$ into batches of size b and prompt the LLM to extract clinically significant concepts—including diseases, pathological features, anatomical locations, and severity descriptors [4]—from each batch independently. We enforce flat binary, positive labels in canonical terminology, excluding negations, generic anatomy, and acquisition/projection terms. The resulting per-batch label sets inevitably contain synonyms and redundancies, so we consolidate them through iterative hierarchical merging: at each round, sets are grouped into $\leq \kappa$-label groups and the LLM is prompted to perform synonym normalization and deduplication, repeating until a single canonical taxonomy $\mathcal{L}$ is obtained.

Report annotation and filtering. Given $\mathcal{L}$, we annotate each training report R_i by prompting the LLM to identify all labels explicitly or implicitly mentioned, producing a binary vector $Y_i \in \{0,1\}^k$ where $Y_i[j] = 1$ if label l_j is present. To reduce annotation noise and ensure minimum support, we apply frequency-based filtering and retain only labels that appear in at least θ reports.

This yields the final bootstrapped dataset $\mathcal{D}_{\text{MLC}} = \{(I_i, Y_i)\}_{i=1}^{N}$ used to train the guidance model in Stage 2.

2.2 Stage 2: Guidance Model Training

In this stage (see Stage 2 in Figure 1), we train a multi-label classification model f_{MLC}, formally defined as $f_{\text{MLC}}(I) = \sigma(W_{\text{cls}} \cdot \phi(I))$, to predict clinical findings from medical images, where $\phi(\cdot)$ is a pre-trained vision encoder (DINOv3 [22] with a ConvNeXt backbone [15]) and $\sigma(\cdot)$ denotes the sigmoid function applied independently per label.

The retained taxonomy can remain imbalanced. Standard binary cross-entropy tends to overlook low-frequency labels. To address this, we adapt the logit adjustment method [16] to multi-label binary classification, shifting each raw logit z_j by the log-odds of its empirical frequency p_j before computing the loss:

$$\tilde{z}_j = z_j + \tau \log \frac{p_j}{1-p_j} \tag{1}$$

where $p_j = \frac{1}{N}\sum_{i=1}^{N} Y_i[j]$ is the empirical label frequency and τ is a temperature parameter (set to 1). The adjusted logits $\tilde{z}_j$ are then used in the standard binary cross-entropy loss, re-balancing the decision boundary to improve both precision and recall on retained low-frequency labels.

2.3 Stage 3: Guided Report Generation

In the final stage (see Stage 3 in Figure 1), we fine-tune a multimodal large language model (MLLM) to generate diagnostic reports conditioned on both visual features and predicted clinical findings.

During training, the bootstrapped pseudo-labels Y_i are serialized into a natural-language prompt (e.g., *"The image shows the following findings: [label$_a$], [label$_b$], [label$_c$]. Based on these findings and the image, generate a detailed diagnostic report."*), which is prepended to the generation target. The model is optimized with the standard next-token prediction objective, learning to condition its output on both the visual input and the explicit label guidance.

At inference, pseudo-labels are unavailable. Instead, the predicted labels $\hat{Y} = f_{\text{MLC}}(I)$ from Stage 2 are serialized into the same prompt format and used to guide generation, grounding the report in explicitly identified factual findings.

3 Experiments

3.1 Dataset and Evaluation Metrics

We evaluate Fact-Flow on two medical imaging datasets. The first is the public tuberculosis chest X-ray dataset [8], comprising 561/80/160 frontal chest X-rays for training, validation, and testing, each paired with a diagnostic report describing tuberculosis-related findings. The second is a multimodal ophthalmology dataset from a clinical institution, split into 1,854/206/515 training/validation/test cases, with institutional ethics approval (No. 2024-1357) and

Table 1. Comparison with state-of-the-art methods on the Tuberculosis dataset. "+FF" denotes our Fact-Flow approach. For RadFact F1, paired bootstrap resampling against the corresponding non-FF baseline yielded p-values < 0.05 for MedGemma and Qwen2.5-VL, and < 0.1 for LLaVA-Med.

Type	Method	Clinical Efficacy			Natural Language Generation			
		F1	Prec.	Recall	B-4	R-L	CIDEr	METEOR
Traditional	R2Gen	0.1909	0.2273	0.1646	0.0682	0.6423	1.7061	0.2131
	R2GPT	0.2105	0.1310	<u>0.5361</u>	0.1032	0.3101	1.1659	0.1813
	CvT2DistilGPT2	0.2128	0.2333	0.1957	0.1655	0.5752	1.7217	0.1896
	R2GenCSR	0.2041	<u>0.3000</u>	0.1546	0.1478	0.6333	1.8396	0.1953
	R2GenCMN	0.2334	0.2903	0.1952	0.0603	<u>0.6500</u>	1.7216	0.1990
MLLM	MedGemma	0.2266	0.2422	0.2130	0.0795	**0.6566**	1.7307	0.1759
	LLaVA-Med	0.0000	0.0000	0.0000	0.0242	0.4691	1.1719	0.1177
	Qwen2.5-VL	0.0286	1.0000	0.0145	0.0265	0.4779	1.2035	0.1141
Zero-Shot	Gemini-2.5-Flash	0.0677	0.0525	0.0952	0.0014	0.0383	0.0102	0.0295
	Gemini-2.5-Pro	0.0713	0.0457	0.1621	0.0008	0.0301	0.0003	0.0245
Fact-Flow	MedGemma + FF	**0.3055**	0.3070	**0.3041**	**0.2290**	0.6243	**2.0664**	**0.2305**
	LLaVA-Med + FF	0.2016	0.1974	0.2059	0.1287	0.6018	1.5869	0.1869
	Qwen2.5-VL + FF	<u>0.2831</u>	**0.3535**	0.2361	<u>0.2129</u>	0.6107	<u>1.9434</u>	<u>0.2236</u>

ChiCTR registration (ChiCTR2500106060). Each case provides fundus photography, Optical Coherence Tomography (OCT), and OCT angiography (OCTA), paired with an ophthalmologist-authored Chinese diagnostic report describing myopia complications and retinal changes.

We assess report quality using two sets of metrics. For natural language generation (NLG), we employ BLEU-1 through BLEU-4 [20], ROUGE-L [12], CIDEr [24], and METEOR [2]. For clinical efficacy on the tuberculosis dataset, we adopt RadFact [3], which leverages large language models to extract clinical entities from both generated and reference reports and computes Precision, Recall, and F1-score without requiring predefined disease-specific categories; precision reflects unsupported generated entities, recall reflects omitted reference entities, and F1 balances both. No established clinical efficacy metric is currently available for Chinese ophthalmology reports, so only NLG metrics are reported for that dataset.

3.2 Implementation Details

For Stage 1 label bootstrapping, we use GPT-5-mini [23] to extract and merge clinical findings from training reports with $b = 200$, $\kappa = 200$, and a merging threshold $\theta = 15$, yielding 7 labels for tuberculosis and 42 for ophthalmology datasets. For Stage 2, the multi-label classifier f_{MLC} employs DINOv3-ConvNeXt-Base [22,15] as the visual encoder with full fine-tuning and logit adjustment [16] ($\tau = 1.0$) for retained-label imbalance, trained with a learning rate of 2×10^{-5}. For the ophthalmology dataset, we use separate encoders for each modality (fundus, OCT, OCTA) with feature fusion. For Stage 3, we fine-tune

Table 2. Comparison with state-of-the-art methods on the Ophthalmology dataset. "+FF" denotes our Fact-Flow approach.

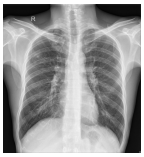
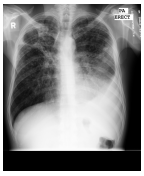
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
Traditional Methods						
R2Gen	0.6004	0.4703	0.3807	0.3138	0.3605	0.0232
R2GenCMN	0.6232	0.5178	0.4420	0.3838	0.7065	0.0650
R2GPT	0.7324	0.6446	0.5777	0.5231	0.8389	0.2589
CvT2DistilGPT2	0.6249	0.5224	0.4485	0.3904	0.4131	0.0268
R2GenCSR	0.7216	0.6372	0.5710	0.5165	0.8386	0.1786
MLLM						
MedGemma	0.6489	0.5562	0.4867	0.4318	0.7921	0.0986
LLaVA-Med	0.7242	0.6266	0.5521	0.4935	0.8345	0.1741
Qwen2.5-VL	0.6873	0.5853	0.5098	0.4506	0.7947	0.1382
Fact-Flow						
MedGemma + FF	0.7079	0.6267	0.5627	0.5099	0.8272	0.2306
LLaVA-Med + FF	**0.7685**	0.6798	0.6103	0.5533	0.8457	0.2350
Qwen2.5-VL + FF	0.7681	**0.6820**	**0.6133**	**0.5567**	**0.8558**	**0.2759**

three MLLMs—Qwen2.5-VL 7B [1], MedGemma 4B [21], and LLaVA-Med-v1.5-Mistral 7B [11]—using LoRA [7] on the decoder, vision encoder, and projector; all models are trained with the AdamW optimizer (learning rate 2×10^{-4}, weight decay 0.1) using BFloat16 mixed precision on four NVIDIA RTX A6000 GPUs. Taxonomy construction, classifier training, and MLLM fine-tuning use only training data; validation is used for model selection and test data only for final evaluation.

3.3 Comparison with State-of-the-Art Methods

We compare Fact-Flow against traditional MRG models (R2Gen [6], R2GenCMN [5], R2GPT [28], R2GenCSR [26], CvT2DistilGPT2 [18]), MLLM baselines (Qwen 2.5-VL, MedGemma, LLaVA-Med fine-tuned directly on image-report pairs without factual guidance), and powerful closed-source VLMs (Gemini-2.5-Flash and Gemini-2.5-Pro) evaluated in a zero-shot setting. On the tuberculosis dataset (Table 1), our approach consistently improves three MLLMs across clinical efficacy and most NLG metrics. **MedGemma + Fact-Flow** achieves the best overall performance, while some vanilla MLLM baselines suffer from mode collapse—Qwen2.5-VL yields perfect precision but near-zero recall, and LLaVA-Med produces zero clinical efficacy scores—a problem our factual guidance directly addresses. Despite their strong general capabilities, Gemini-2.5-Flash and Gemini-2.5-Pro perform poorly in the zero-shot setting, underscoring the necessity of domain-specific training for medical report generation. On the ophthalmology dataset (Table 2), **Qwen2.5-VL + Fact-Flow** achieves the best results on most NLG metrics, demonstrating the effectiveness of our approach across complex multimodal scenarios. Qualitative examples in Table 3 further confirm that factual guidance enables more precise localization of findings (e.g., disease laterality, anatomical region) compared to the vanilla MLLM baseline.

Table 3. Case study of generated reports on the Tuberculosis dataset. Bold text in the reference report denotes key findings. Superscript ✓ indicates findings supported by the reference, while † indicates unsupported or incorrect findings.

Chest X-ray	**Ground Truth**	**MLC Prediction**	**MedGemma + Fact-Flow**	**MedGemma**
	**Right secondary PTB in the upper and middle fields**	Pulmonary tuberculosis$^{\checkmark}$, Secondary (post-primary) pulmonary tuberculosis$^{\checkmark}$	secondary PTB$^{\checkmark}$ in the right upper field$^{\checkmark}$	bilateral† PTB$^{\checkmark}$
RadFactF1			**1**	0
	extensive **infiltrates bilaterally** with large **cavity** in RUL and a moderate **pleural effusion on the left**. AFB smears and RNA probes pos for MTB. **Active TB**, **cavitary**.	Pulmonary tuberculosis$^{\checkmark}$, Pulmonary cavitation$^{\checkmark}$, Pulmonary infiltrates$^{\checkmark}$, Pleural effusion$^{\checkmark}$	bilateral† PTB$^{\checkmark}$ with cavitation$^{\checkmark}$ and left pleural effusion$^{\checkmark}$	bilateral† PTB$^{\checkmark}$
RadFactF1			**0.67**	0

3.4 Analysis of Visual and Factual Guidance

We evaluate five configurations on the tuberculosis dataset using Qwen2.5-VL (Table 4), varying input modalities (image and/or label) and label sources (predicted labels vs. bootstrapped pseudo-labels). The **Image Only** baseline suffers from mode collapse despite high precision, as MLLMs generate overly conservative reports without factual grounding. Introducing predicted labels (**Label Only (Pred)**) substantially improves both clinical efficacy and NLG metrics, and combining image with predicted labels (**Image + Label (Pred)**)—our full Fact-Flow approach—achieves the best practical performance, confirming that visual context and factual guidance are complementary. The performance gap between predicted-label and pseudo-label highlights label quality as the key bottleneck, and the oracle **Image + Label (PL)** consistently outperforms **Label Only (PL)**, indicating that visual information provides spatial details that discrete labels alone cannot capture.

3.5 Intermediate Stage Analysis

We validate intermediate stages (Table 5). **Stage 1:** All metrics are verified by a professional ophthalmologist on the ophthalmology dataset. (1) **Coverage** measures whether each report's key clinical information can be expressed by a subset of labels in $\mathcal{L}$, rated as *fully* / *partially* / *not covered* on 50 randomly sampled reports; (2) **LLM-Match Accuracy** checks whether LLM-Match assigns correct

Table 4. Analysis of visual and factual guidance on the Tuberculosis dataset using Qwen2.5-VL. "Pred" and "PL" denote predicted labels and bootstrapped pseudo-labels, respectively.

Input Modality		Clinical Efficacy			Natural Language Generation					
Image	**Label**	**F1**	**Prec.**	**Recall**	**B-2**	**B-3**	**B-4**	**R-L**	**CIDEr**	**METEOR**
Practical Configurations										
✓	–	0.0286	1.0000	0.0145	0.0265	0.0265	0.0265	0.4779	1.2035	0.1141
–	Pred	0.2115	0.2656	0.1757	0.2895	0.2409	0.2074	0.5943	1.9209	0.2188
✓	Pred	**0.2831**	**0.3535**	**0.2361**	**0.2951**	**0.2455**	**0.2129**	**0.6107**	**1.9434**	**0.2236**
Oracle Upper Bounds										
–	PL	*0.3750*	*0.5000*	*0.3000*	*0.3658*	*0.3026*	*0.2618*	*0.6990*	*2.3198*	*0.2691*
✓	PL	*0.4421*	*0.5116*	*0.3892*	*0.3727*	*0.3095*	*0.2680*	*0.7198*	*2.4420*	*0.2746*

Table 5. Intermediate stage analysis. **Left**: Intrinsic quality evaluation of Stage 1 on the ophthalmology dataset (all metrics verified by a professional ophthalmologist). Report-level metrics (Coverage, LLM-Match Accuracy) are measured on 50 randomly sampled reports; label-level metrics (Redundancy, Clinical Validity) characterize $\mathcal{L}$ itself. **Right**: Stage 2 MLC classification performance on test sets.

Stage 1	**Result**
Report-level (50 sampled reports)	
Coverage	100% (50/50)
LLM-Match Accuracy	100% (50/50)
Label-level ($\mathcal{L}$)	
Redundancy	7.5%
Valid & decidable	80%
Valid, context-dependent	15%
Overly generic	5%
Invalid / pseudo-concept	0%

Stage 2	**Macro-F1**	**Micro-F1**
Tuberculosis	0.5227	0.7328
Ophthalmology	0.7071	0.8426

labels (set-level) on the same 50 reports; (3) **Redundancy** estimates the proportion of clinically synonymous label pairs within $\mathcal{L}$; and (4) **Clinical Validity** categorizes each label as *valid & decidable*, *valid but context-dependent*, *overly generic*, or *invalid/pseudo-concept*. Results show 100% coverage and matching accuracy, low redundancy (7.5%), and 80% of labels valid and decidable with none invalid, providing a sanity check for the bootstrapped vocabulary. **Stage 2:** The MLC achieves Macro-/Micro-F1 of 0.52/0.73 on tuberculosis and 0.71/0.84 on ophthalmology, providing practical grounding for Stage 3.

4 Conclusion

To address the critical challenge of factual instability in Multimodal Large Language Models (MLLMs) when applied to medical report generation, we propose Fact-Flow, a novel framework that enhances factual accuracy by decoupling visual feature recognition from textual composition. Our core contributions include an LLM-bootstrapped pipeline for automatic multi-label dataset

creation and a multi-label classification model that provides an explicit "Fact-Flow" to guide the MLLM. Fact-Flow is a plug-and-play framework compatible with any MLLM architecture, particularly suited to clinical scenarios where reports revolve around targeted and enumerable finding categories. Experiments on two disease-focused datasets with three MLLMs (Qwen2.5-VL, MedGemma, and LLaVA-Med) demonstrate that Fact-Flow improves tuberculosis clinical efficacy and most NLG metrics over state-of-the-art methods and direct fine-tuning, without compromising textual quality. Limitations include LLM-based pseudo-labeling/evaluation, Stage-2 error propagation without rejection, and untested hyperparameter sensitivity.

Acknowledgments. This work was supported by the Key Research and Development Program of Zhejiang Province (2024C03205), Zhejiang Province Major Science and Technology Plan Project (2026C01021), the Provincial Key Laboratory Program (2025E10048), Zhejiang Province Science and Technology Plan Project (2025C01128).

Disclosure of Interests. The authors have no competing interests in the paper.

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. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
4. Chen, H., Jia, J., Li, R., Yang, C., Li, W., Pang, X., Chen, Y., Wang, H., Bu, J., Wu, L.: Directed ordinal diffusion regularization for progression-aware diabetic retinopathy grading. arXiv preprint arXiv:2602.21942 (2026)
5. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). pp. 5904–5914 (2021)
6. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
8. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
9. 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 (2018)

10. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6666–6673 (2019)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
13. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
16. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
17. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)
18. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
19. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
23. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
24. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
25. Wang, S., Tang, L., Lin, M., Shih, G., Ding, Y., Peng, Y.: Prior knowledge enhances radiology report generation. AMIA Summits on Translational Science Proceedings **2022**, 486 (2022)
26. Wang, X., Li, Y., Wang, F., Wang, S., Li, C., Jiang, B.: R2gencsr: Retrieving context samples for large language model based x-ray medical report generation. arXiv preprint arXiv:2408.09743 (2024)

27. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9049–9058 (2018)
28. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
29. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–466. Springer (2018)