MedHyperGraph: EHR-Integrated Multimodal Hyperedges for Clinical VQA

Anonymized Authors

Anonymized Affiliation
anonymous@example.com

Abstract. Many clinical questions about chest X-rays cannot be answered from images alone, as they depend on longitudinal context, including prior studies, earlier reports, diagnoses, and the timing of clinical events. Most medical vision-language models (VLMs) still process isolated image-question pairs. We present MedHyperGraph, a per-patient multimodal hypergraph for electronic health record (EHR)-grounded medical visual question answering (VQA). It integrates structured EHR events, prior-report entities, imaging studies, ontology concepts, and image regions into a single temporally indexed structure; evidence is then retrieved under explicit temporal constraints. On EHRXQA boolean questions ($n = 1{,}900$), the retrieval baselines remain close to a majority-class predictor on the MedGemma backbones; MedHyperGraph improves on all three backbones, reaching 72.63% accuracy (95% CI $[70.6, 74.6]$), without task-specific training, which exceeds the reinforcement-learning-trained HyperWalker's 70.43%. On open-ended questions, it raises weighted F1 by 10.0 points at 27B, driven by multi-study questions ($5.2 \rightarrow 43.4$). Ablations attribute the gain primarily to the structured EHR (-9.1) and prior-report (-6.8) evidence, indicating that EHR-grounded VQA benefits from retrieval that preserves the temporal and n-ary structure of clinical evidence.

Keywords: Medical VQA · Electronic health records · Multimodal retrieval · Hypergraphs · Retrieval-augmented generation · Clinical knowledge graphs

1 Introduction and Related Work

A chest X-ray is rarely interpreted in isolation, because its meaning depends on the patient's history, including prior admissions, studies, reports, diagnoses, labs, and the time between events [2,22]. Whether a finding is new, unchanged, expected after a procedure, or consistent with a known diagnosis is often not visible in a single image. Medical visual question answering (VQA) should reflect this, because answers may depend on when the image was taken, which clinical event it follows, or whether a related condition already appears in the record.

Recent medical vision-language models (VLMs) such as MedGemma [17], LLaVA-Med [11], and Lingshu [18] interpret medical images well, but typically answer isolated image-question pairs with little access to the surrounding record.

This is a poor fit for EHRXQA [2], a benchmark over EHR tables and chest X-rays; its questions can combine structured SQL queries with visual reasoning over MIMIC-CXR images [9]. Therefore, many of its questions cannot be answered from the referenced image alone, but need evidence from across the patient's longitudinal record.

Retrieval-augmented generation (RAG) can provide evidence to a VLM [7], but flat text retrieval does not capture the temporal and multimodal relations that clinical reasoning relies on. Graph-based RAG represents knowledge as entities and relations [6], but a single clinical fact usually combines several elements at once (e.g., a finding, an anatomical region, a study time, a report statement, and an EHR event); pairwise edges cannot represent this grouping unless the fact is represented explicitly. HyperGraphRAG [14] preserves such n-ary facts through hyperedges. In medicine, multimodal knowledge graphs such as MEDMKG [20], VLM-KG [1], and R2GenKG [19] link images to clinical concepts; however, they do not target patient-specific longitudinal retrieval for EHR-grounded VQA, where evidence must be jointly selected from EHR events, prior reports, timestamps, and image regions. Recently, Yang et al. [22] introduced HyperWalker, which builds a dynamic clinical hypergraph and trains a reinforcement-learning agent to traverse diagnostic paths, reaching 70.43% boolean accuracy on a filtered EHRXQA test set. Although HyperWalker shows the value of hypergraph structures on this benchmark, it learns a global navigation policy, does not isolate each patient's own evidence, and does not constrain what the model may read at answer time. This motivates a simpler, patient-scoped formulation that needs no training and keeps its evidence explicit.

Therefore, we present MedHyperGraph, a per-patient multimodal hypergraph for EHR-grounded medical VQA. For each patient, MedHyperGraph integrates structured EHR events, prior-report entities, imaging studies, ontology concepts, and image regions into a single temporally indexed structure; cross-modal links are computed from BioViL-T [4] embeddings; the hypergraph design is informed by RadGraph [8], Chest ImaGenome [21], and CheXRelNet [10]. Given a question, MedHyperGraph retrieves a small evidence subgraph under explicit temporal constraints and passes it, together with the current image, to a VLM. The target study's own report is withheld from retrieval, and future studies are excluded, so a visual answer cannot be copied from the current report. Unlike HyperWalker, MedHyperGraph is patient-scoped rather than global, does not use task-specific training, selects evidence by retrieval rather than a learned policy, and is evaluated on open-ended questions and clinical attribute groups in addition to boolean accuracy.

Our Contributions. We introduce MedHyperGraph, a per-patient multimodal hypergraph that unifies EHR events, prior-report entities, imaging studies, ontology concepts, and image regions for EHR-grounded medical VQA. In addition, we propose temporally aware retrieval with leakage control, ontology-grounded normalization with SNOMED CT [5], and an evaluation on the filtered EHRXQA test set against vanilla inference, text-RAG, and GraphRAG baselines, including temporal and clinical-attribute breakdowns.

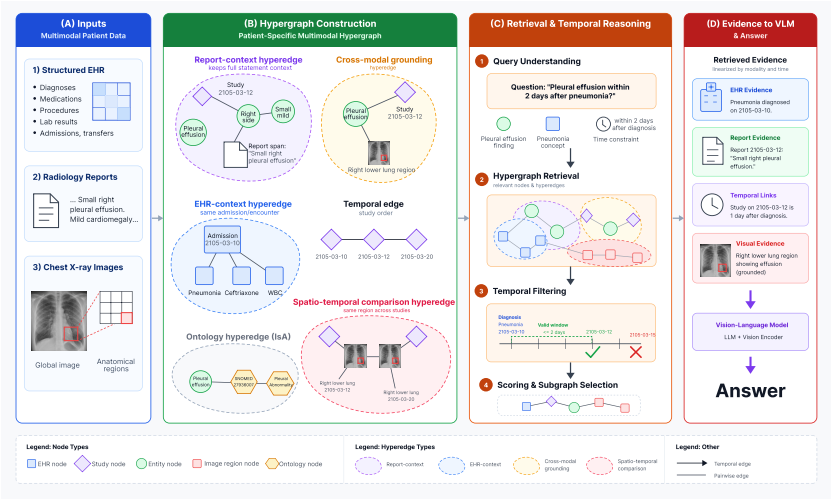

Fig. 1. Overview of MedHyperGraph. Multimodal patient data from EHRs, radiology reports, and chest X-rays form a patient-specific hypergraph, retrieved with temporal constraints, and linearized as evidence for VLM-based answer generation.

2 Methodology

MedHyperGraph represents each patient record as a typed multimodal hypergraph rather than as isolated image-question samples. For patient p, we define $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$, where $\mathcal{V}_p$ is the set of clinical hypernodes and $\mathcal{E}_p$ is the set of typed hyperedges. Each hypernode represents a clinically meaningful unit, such as an EHR event, radiology study, report entity, image region, or ontology concept. Each hyperedge is defined as $e = (S_e, \rho_e, \alpha_e)$ with $S_e \subseteq \mathcal{V}_p$ and $|S_e| \geq 2$, where ρ_e denotes the hyperedge type and α_e stores attributes such as timestamps, source text, grounding scores, or ontology identifiers. Unlike a standard graph edge, a hyperedge can connect more than two clinical elements, allowing MedHyperGraph to preserve higher-order clinical contexts such as report statements, admission events, cross-modal image-text links, and longitudinal comparisons.

2.1 Problem Formulation

For each patient p, let $\mathcal{D}_p = (\mathcal{T}_p, \mathcal{R}_p, \mathcal{I}_p)$ denote the longitudinal record, consisting of structured EHR tables $\mathcal{T}_p$, radiology reports $\mathcal{R}_p$, and chest X-ray studies $\mathcal{I}_p$. Each study $s \in \mathcal{I}_p$ has a timestamp $\tau(s)$, one or more images, and an associated report. Given a question $x = (p, q)$, where q is a natural-language clinical question about patient p, the task is to predict $\hat{y} \in \{\text{yes}, \text{no}\}$.

At inference time, MedHyperGraph retrieves an evidence subgraph $\mathcal{S}_q = (\mathcal{V}_q, \mathcal{E}_q)$, with $\mathcal{V}_q \subseteq \mathcal{V}_p$ and $\mathcal{E}_q \subseteq \mathcal{E}_p$, and conditions a VLM on textual and visual evidence (Figure 1):

$$\hat{y} = F_\theta\big(q,\ \mathcal{L}(\mathcal{S}_q),\ \mathcal{M}(\mathcal{S}_q)\big).$$

Here, $\mathcal{L}(\mathcal{S}_q)$ is the textual linearization of the retrieved evidence, and $\mathcal{M}(\mathcal{S}_q)$ contains retrieved visual evidence such as images or anatomical regions.

2.2 MedHyperGraph Construction

For each patient, MedHyperGraph builds a typed hypernode set $\mathcal{V}_p = \mathcal{V}_p^{\text{ehr}} \cup \mathcal{V}_p^{\text{study}} \cup \mathcal{V}_p^{\text{ent}} \cup \mathcal{V}_p^{\text{img}} \cup \mathcal{V}_p^{\text{ont}}$. **EHR hypernodes** $\mathcal{V}_p^{\text{ehr}}$ are derived from structured tables, including admissions, diagnoses, procedures, prescriptions, laboratory events, and transfers. **Study hypernodes** $\mathcal{V}_p^{\text{study}}$ represent radiology studies and serve as temporal anchors linking images, reports, and nearby EHR evidence. **Report-entity hypernodes** $\mathcal{V}_p^{\text{ent}}$ are extracted from reports and include findings, anatomical regions, severity modifiers, devices, and other clinical concepts; when possible, mentions are normalized to SNOMED CT identifiers [5]. **Image hypernodes** $\mathcal{V}_p^{\text{img}}$ represent chest X-rays at both global-image and anatomical-region levels. **Ontology hypernodes** $\mathcal{V}_p^{\text{ont}}$ encode normalized concepts and their parent concepts to support synonym-aware expansion.

MedHyperGraph defines hyperedges as a union of typed clinical relations:

$$\mathcal{E}_p = \mathcal{E}_p^{\text{str}} \cup \mathcal{E}_p^{\text{rep}} \cup \mathcal{E}_p^{\text{ehr}} \cup \mathcal{E}_p^{\text{cmp}} \cup \mathcal{E}_p^{\text{ont}}.$$

Structural hyperedges $\mathcal{E}_p^{\text{str}}$ link naturally paired elements, such as studies and their reports or images, images and anatomical regions, or consecutive studies. **Report-context hyperedges** $\mathcal{E}_p^{\text{rep}}$ preserve statement-level meaning by linking a study node to all report entities extracted from the same text segment. For a segment c associated with study s, we define $e_c^{\text{rep}} = (\{s\} \cup \text{Ent}(c), \text{ReportContext}, \alpha_c)$, where α_c stores the original text span. **EHR-context hyperedges** $\mathcal{E}_p^{\text{ehr}}$ link an admission node to structured events in the same clinical context, e.g., diagnoses, prescriptions, procedures, and lab results. **Spatio-temporal comparison hyperedges** $\mathcal{E}_p^{\text{cmp}}$ connect corresponding anatomical regions across consecutive studies for longitudinal reasoning. **Ontology hyperedges** $\mathcal{E}_p^{\text{ont}}$ connect entity mentions to normalized concepts and their parents.

Representational motivation. Hyperedges preserve statement-level clinical context that can be lost in a naive pairwise graph: a joint fact over a finding, region, timestamp, and EHR event can collapse to the same clique as several independent pairwise links. By storing these elements as a single hyperedge, MedHyperGraph preserves shared membership and allows incidence expansion to recover the full local context from a partial retrieval hit before ranking and context-budget pruning.

2.3 Temporally Aware Multimodal Retrieval

Given a question q, the retrieval module selects patient-specific evidence from $\mathcal{G}_p$ in four steps (Figure 1). **First**, the question is analyzed to identify clinical concepts, referenced studies, anatomical regions, and temporal constraints. The query is encoded as $z_q = f_{\text{text}}(q)$; when a study or image is explicitly referenced,

the corresponding image representation $z_I = f_{\text{img}}(I)$ is also used. Temporal expressions define a window $W(q)$ relative to an admission, diagnosis, or study.

Second, candidate evidence is retrieved through three complementary channels: (i) entity-level retrieval over EHR, report-entity, and ontology hypernodes; (ii) hyperedge-level retrieval over complete clinical contexts; and (iii) visual retrieval over image or anatomical-region hypernodes when available.

Third, candidates are expanded through hypergraph incidence links. If a retrieved hypernode belongs to a hyperedge, the full hyperedge and its incident hypernodes are added. If a hyperedge is retrieved directly, all of its connected hypernodes are included. Candidates inconsistent with the temporal window $W(q)$ are removed or down-weighted.

Finally, each candidate u is scored using textual and visual relevance, temporal compatibility, and hypergraph proximity:

$$S(u,q) = \lambda_{\text{text}} s_{\text{text}}(u,q) + \lambda_{\text{vis}} s_{\text{vis}}(u,q) + \lambda_{\tau} s_{\tau}(u,q) + \lambda_h s_h(u,q). \quad (1)$$

The top-scoring hypernodes and hyperedges under a fixed context budget form the final evidence subgraph $\mathcal{S}_q$, which is linearized by modality and temporal order into EHR evidence, report evidence, temporal links, and visual evidence. The VLM then generates the final yes/no answer from the question and the retrieved multimodal evidence.

3 Experimental Setup

Dataset and metric. We evaluate on EHRXQA [2], built over MIMIC-IV and MIMIC-CXR [9], using the HyperWalker-filtered test set [22]: 1,900 boolean and 600 open-ended questions. The attribute analysis uses the 719 boolean questions that require reading the image. The boolean answers are imbalanced (56% "no", 44% "yes"), so a constant "no" already scores 56.0% accuracy at 50.0% balanced accuracy. Therefore, we report *balanced accuracy* as the primary metric. Dates in MIMIC are shifted into the future during de-identification, so years such as "2105" are expected rather than errors.

Backbones and baselines. We use three backbones: MedGemma 1.5 (4B), MedGemma 1 (27B), and Qwen 3.6-35B-A3B-AWQ [3]. Under each, we run *Vanilla* (no retrieval), *text-RAG* (dense retrieval over prior-report and EHR text), *GraphRAG* [6] (a naive pairwise entity graph over the same text), and *MedHyperGraph*. All retrieval methods share the same text sources and context budget. Lingshu [18] and HyperWalker [22] use their own backbones and training and serve only as external reference points.

Leakage control. Many EHRXQA visual answers also appear in the study's own radiology report, so we control what the model may read at answer time. Retrieval is restricted to data dated before the query time, so future studies are never used. The target study's own report is withheld from the retrievable pool, so a visual answer cannot be copied from the current report; only prior reports are retrievable. Image regions use Chest ImaGenome [21] bounding-box geometry only, not its finding labels, so report-derived labels do not enter retrieval. Retrieval weights are fixed in advance and are not tuned on the test set.

Table 1. Boolean results on the filtered EHRXQA test set ($n = 1{,}900$; 56% "no"). Balanced accuracy is the primary metric. Best method per column in **bold**. HyperWalker (RL-trained, own backbone) reports 70.43% accuracy.

	MedGemma 1.5 (4B)		**MedGemma 1 (27B)**		**Qwen 3.6**	
Method	Acc.	Bal. Acc.	Acc.	Bal. Acc.	Acc.	Bal. Acc.
Vanilla	59.26	54.99	58.11	56.54	60.00	57.67
text-RAG	58.47	54.84	58.94	58.13	67.89	66.66
GraphRAG [6]	58.74	54.88	58.84	54.92	64.84	62.40
MedHyperGraph	**64.63**	**63.61**	**69.40**	**69.13**	**72.63**	**72.40**

Open-ended scoring. n-gram metrics such as BLEU [16] and ROUGE [12] penalize clinically equivalent paraphrases, while BERTScore [24] does not check factual correctness [15]. Therefore, we use two LLM judges [25,13]: Gemma-4-31B and Qwen 3.6-35B-A3B-AWQ [3]; both run locally with deterministic decoding, a fixed rubric prompt, and structured JSON output; we report their average. Following ReXVal [23] and GREEN [15], each judge identifies matched, missed, and hallucinated findings before scoring recall, precision, correctness, and semantic similarity on a 0–100 scale. We report F1 as the harmonic mean of mean recall and mean precision. We evaluate open-ended questions only with the MedGemma backbones and interpret within-backbone differences, since a judge sharing a model family with a generator may favor its own outputs.

4 Results and Discussion

Boolean accuracy (Table 1). On both MedGemma backbones, the retrieval baselines stay close to the majority-class predictor: 58–59% accuracy, only 2–3 points above the 56% of a constant "no" predictor, and 54.8–58.1% balanced accuracy, only 5–8 points above chance. Flat text retrieval and a binary entity graph add little over no retrieval. This ceiling is not universal: on the stronger Qwen backbone, the baselines recover (66.66% balanced accuracy for text-RAG), which shows the effect depends on backbone capability rather than on a weak baseline implementation. MedHyperGraph performs best on every backbone and is the only method that improves regardless of backbone, reaching 63.61%, 69.13%, and 72.40% balanced accuracy (13.6, 19.1, and 22.4 points above chance, respectively). It also nearly closes the accuracy–balanced-accuracy gap on Qwen (72.63 vs. 72.40), so it discriminates both answers rather than favoring "no". With Qwen, MedHyperGraph reaches 72.63% accuracy (95% Wilson interval $[70.6, 74.6]$, $n = 1{,}900$), its lower bound lies above the 70.43% reported by HyperWalker [22], despite MedHyperGraph requiring no task-specific training. The generalist medical VLM Lingshu [18] sits at 60.77%.

Open-ended VQA (Table 2). MedHyperGraph gives the best overall F1 at both MedGemma backbones and is the only method that improves over the plain model at both: +3.8 at 4B (22.9 → 26.7) and +10.0 at 27B (28.6 → 38.6). GraphRAG lowers F1 (−4.6 at 4B, −7.7 at 27B), while text-RAG helps at 4B

Table 2. Open-ended EHRXQA, weighted F1 by image count ($n = 600$; average of two LLM judges, Gemma-4-31B and Qwen 3.6). Δ is overall F1 vs. the same-backbone plain model. **Bold** is best per column within each backbone.

Configuration	1-img	2-img	N-img	Overall	Δ
MedGemma-4B plain	33.4	26.6	5.2	22.9	–
MedGemma-4B + text-RAG	**38.5**	**28.1**	11.7	26.4	+3.5
MedGemma-4B + GraphRAG	28.6	15.8	10.0	18.3	−4.6
MedGemma-4B + MedHyperGraph	33.5	25.9	**19.0**	**26.7**	+3.8
MedGemma-27B plain	43.4	**32.1**	5.2	28.6	–
MedGemma-27B + text-RAG	26.5	16.1	25.4	22.1	−6.5
MedGemma-27B + GraphRAG	35.5	19.0	4.3	20.9	−7.7
MedGemma-27B + MedHyperGraph	**45.3**	29.7	**43.4**	**38.6**	+10.0

Table 3. Component ablation on MedGemma-27B (boolean accuracy, $n = 1{,}900$).

Configuration	Acc. (%)	Δ
MedHyperGraph (full)	**69.40**	–
w/o EHR evidence	60.35	−9.05
w/o prior reports	62.57	−6.83
w/o temporal filter	68.50	−0.90
w/ grounding	67.70	−1.70

(+3.5) but hurts at 27B (−6.5). The gain comes from multi-study (N-image) questions, where plain models nearly fail (F1 5.2 at both sizes) and MedHyperGraph reaches 19.0 and 43.4. The advantage is not uniform: at 4B text-RAG is stronger on single- and two-image questions, and at 27B the plain model keeps a small edge on two-image questions (32.1 vs. 29.7). MedHyperGraph wins overall through its large multi-study margin, where longitudinal evidence is decisive.

Component ablation (Table 3). Removing evidence sources from the 27B system shows where the accuracy comes from. Structured EHR evidence contributes most (−9.05 when removed) and prior reports second (−6.83), which supports the central claim that longitudinal, multi-source evidence drives the gain. Temporal filtering contributes −0.90 overall; this small effect is expected, since the filter is inactive for questions without a time constraint, and its benefit concentrates on the time-windowed subset rather than the full test set. Unlike the removal ablations, we test visual grounding as an optional addition. The base system already passes the image to the VLM, while grounding adds inferred entity-region links. Adding them lowers accuracy by 1.70 points, suggesting that similarity-based alignment is too noisy and distracts retrieval. Although it may support visual inspection, improving alignment remains future work.

Temporal reasoning. Grouping the 27B questions by their temporal form (each question belongs to one category) shows where the temporal filter helps. On the two time-windowed categories the baselines stay at or below the plain model: global-window, base 57.0 vs. 53.7 (text-RAG) and 54.5 (GraphRAG); relative-window ("within"), 52.3 vs. 57.3 and 48.6; in contrast, MedHyperGraph reaches

71.5 and 66.7 balanced accuracy. The same ordering holds at 4B (53.2 → 67.3 and 52.9 → 60.6). Non-temporal questions also improve (55.8 → 77.0 at 27B), but that gain comes from structural and multimodal retrieval rather than the time window, which is consistent with the small overall ablation effect above. Change-comparison is the hardest category, where MedHyperGraph (60.3) trails text-RAG (63.9); these questions ask about progression across studies and remain the clearest target for future work.

Analysis by clinical attribute. Improvement varies across finding types. On the 719 visual questions, the gain over the plain 27B model is largest for findings that are named explicitly in prior reports and correspond to discrete structures, such as support devices and lines (+24.4 balanced accuracy points) and focal opacities (+21.1). It is smaller for diffuse parenchymal findings, where the plain model already performs comparatively well, for example chronic lung disease (+3.5); this follows from the retrieval mechanism, where report entities that carry a stable name and a consistent anatomical location are normalized to ontology concepts and grounded to image regions more reliably than findings described in varying terms. Two attribute groups depart from this pattern; for pleural and air-related findings, text-RAG reaches 73.0 against 68.1 for MedHyperGraph, which is plausible because such findings are typically stated in plain language in the report and are retrievable without additional structure. The *technical* group contains only 12 questions and does not support interpretation. Group sizes range from 12 to 122 and the comparisons are not corrected for multiple testing, so these differences indicate tendencies rather than established effects.

Results Summary. Across boolean and open-ended questions, and temporal and attribute breakdowns, flat and binary-graph retrieval are inconsistent and in several settings degrade the plain model, while MedHyperGraph improves performance on all three backbones. MedHyperGraph's advantage concentrates on questions whose answer depends on longitudinal evidence, namely multi-study and time-windowed questions; the ablation attributes the gain primarily to the structured EHR and prior-report evidence.

5 Conclusion and Future Work

We presented MedHyperGraph, a patient-specific multimodal hypergraph for EHR-grounded medical VQA that unifies EHR events, prior reports, imaging studies, ontology concepts, and image regions into a temporally indexed retrieval structure. On the EHRXQA test set, MedHyperGraph outperformed vanilla inference, text-RAG, and GraphRAG on all three backbones, reaching 72.63% accuracy and 72.40% balanced accuracy on Qwen, without task-specific training. Ablations attribute most of this gain to structured EHR and prior-report evidence. MedHyperGraph also improved open-ended weighted F1, driven mainly by multi-study questions. Gains were clearest for longitudinal queries, while cross-study change-comparison remained the hardest category, a target for future work such as agentic question decomposition into sub-questions. We also built a simple UI, shown in the supplementary video, for future user studies.

References

1. Abdullah, A., et al.: VLM-KG: Multimodal Radiology Knowledge Graph Generation (May 2025). https://doi.org/10.48550/arXiv.2505.17042
2. Bae, S., et al.: EHRXQA: a multi-modal question answering dataset for electronic health records with chest X-ray images. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 3867–3880. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023), https://openreview.net/forum?id=Pk2x7FPuZ4
3. Bai, S., et al.: Qwen3-VL Technical Report (Nov 2025). https://doi.org/10.48550/arXiv.2511.21631
4. Boecking, B., et al.: Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In: Computer Vision – ECCV 2022. pp. 1–21. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1
5. Donnelly, K.: SNOMED-CT: The Advanced Terminology and Coding System for eHealth. In: Medical and Care Compunetics 3, pp. 279–290. IOS Press (2006), https://ebooks.iospress.nl/publication/9130
6. Edge, D., et al.: From Local to Global: A Graph RAG Approach to Query-Focused Summarization (Feb 2025). https://doi.org/10.48550/arXiv.2404.16130
7. Gao, Y., et al.: Retrieval-Augmented Generation for Large Language Models: A Survey (Mar 2024). https://doi.org/10.48550/arXiv.2312.10997
8. Jain, S., et al.: RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (Jun 2021), https://openreview.net/forum?id=pMWtc5NKd7V
9. Johnson, A.E.W., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (Dec 2019). https://doi.org/10.1038/s41597-019-0322-0
10. Karwande, G., et al.: CheXRelNet: An Anatomy-Aware Model for Tracking Longitudinal Relationships Between Chest X-Rays. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 581–591. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_55
11. Li, C., et al.: LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 28541–28564. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023), https://dl.acm.org/doi/10.5555/3666122.3667362
12. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
13. Liu, Y., et al.: G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.153
14. Luo, H., et al.: HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (Oct 2025), https://openreview.net/forum?id=ravS5h8MNg

15. Ostmeier, S., et al.: GREEN: Generative Radiology Report Evaluation and Error Notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 374–390. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.21
16. Papineni, K., et al.: 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. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135
17. Sellergren, A., et al.: MedGemma Technical Report (Apr 2026). https://doi.org/10.48550/arXiv.2507.05201
18. Team, L., et al.: Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning (Jun 2025). https://doi.org/10.48550/arXiv.2506.07044
19. Wang, F., et al.: R2GenKG: Hierarchical Multi-modal Knowledge Graph for LLM-based Radiology Report Generation (Aug 2025). https://doi.org/10.48550/arXiv.2508.03426
20. Wang, X., et al.: MEDMKG: Benchmarking Medical Knowledge Exploitation with Multimodal Knowledge Graph (May 2025). https://doi.org/10.48550/arXiv.2505.17214
21. Wu, J.T., et al.: Chest ImaGenome Dataset for Clinical Reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Aug 2021), https://openreview.net/forum?id=H-d5634yVi
22. Yang, Y., et al.: HyperWalker: Dynamic Hypergraph-Based Deep Diagnosis for Multi-Hop Clinical Modeling across EHR and X-Ray in Medical VLMs (Jan 2026). https://doi.org/10.48550/arXiv.2601.13919
23. Yu, F., et al.: Evaluating progress in automatic chest X-ray radiology report generation. Patterns **4**(9) (Sep 2023). https://doi.org/10.1016/j.patter.2023.100802
24. Zhang, T., et al.: BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations (Sep 2019), https://openreview.net/forum?id=SkeHuCVFDr
25. Zheng, L., et al.: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 46595–46623. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023), https://openreview.net/forum?id=uccHPGDlao