Multimodal Large Language Models in Medical Imaging: Current State and Future Directions

Authors: Yoojin Nam, Dong Yeong Kim, Sunggu Kyung, Jinyoung Seo, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Sangah Park, Heeyeon Kwon, Taehee Kwon, Kanghyun Kim, Namkug Kim

Year: 2025

Venue: Korean Journal of Radiology

1. Classification

- **Domain Category:**
  - Medical VLM / MLLM / MMFM **+** General FM survey / theory
  - This is a review article that surveys multimodal large language models (MLLMs) for medical imaging, especially radiology, and analyzes architectures, datasets, capabilities, and challenges.
- **FM Usage Type:**
  - Multimodal FM or cross-modal integration (survey of existing systems)
- **Key Modalities:**
  - Imaging: 2D chest X-ray, CT, MRI, ultrasound, endoscopy, digital pathology, and other clinical photos.
  - Text: radiology reports, clinical notes, question–answer pairs, and structured EHR data.

---

2. Executive Summary

This review paper provides a comprehensive overview of multimodal large language models (MLLMs) in medical imaging, focusing on how they integrate image and text data to support tasks such as radiology report generation, visual question answering (VQA), and interactive diagnostic assistance. The authors first introduce LLMs and vision transformers (ViTs) as the core building blocks, then explain how multimodal connectors and training strategies turn them into MLLMs. They categorize architectures by how images and text are encoded, how connectors project visual features into the LLM's token space, and how multimodal fusion is achieved (contrastive pretraining, instruction-tuned fusion, generative pipelines, etc.). The paper surveys available datasets, clinical applications, and early systems, highlighting both impressive capabilities and serious limitations such as hallucination, poor region grounding, and the scarcity of large-scale medical multimodal datasets. It

closes with a roadmap for future research, emphasizing region-grounded reasoning, robust pretraining on medical data, and safe clinical integration. For a new grad student, this review is an accessible map of the design space and open problems in medical MLLMs.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - Understand how to build and deploy **multimodal LLMs** that can:
    - Interpret medical images together with clinical text.
    - Generate accurate, clinically useful reports and answers.
    - Act as trustworthy assistants in radiology workflows.
- **Why this is hard:**
  - **Data challenges:**
    - Large, high-quality multimodal datasets (images + reports + EHR) are scarce and often siloed by institution.
    - Annotations such as region-level labels and detailed textual descriptions are expensive to obtain.
    - Privacy regulations constrain data sharing and centralized training.
  - **Modeling challenges:**
    - Radiology images (2D/3D, multi-phase) and clinical notes are heterogeneous and high-dimensional.
    - Aligning visual features with language at the right granularity (organ, lesion, pixel) is non-trivial.
    - LLMs trained on web text may hallucinate findings or misuse clinical jargon when connected to images.
  - **Clinical deployment challenges:**
    - Need for region-grounded explanations and robust behavior across scanners, sites, and populations.

        - Integration into PACS/RIS and clinician workflows without increasing cognitive load or risk.

---

4. Data and Modalities

- **Datasets and modalities covered (high level):**
    - **Imaging:**
        - Chest X-ray (e.g., MIMIC-CXR, CheXpert, ChestX-ray14).
        - CT and MRI for various organs.
        - Ultrasound, endoscopy, ophthalmology images, and digital pathology slides.
    - **Text and structured data:**
        - Radiology and pathology reports.
        - Clinical notes and EHR fields (demographics, lab values, vital signs).
        - QA pairs and instruction-style prompts for training medical MLLMs.
- **Pretraining / representation patterns:**
    - Vision encoders (often ViTs or CNNs) map images to dense feature maps or patch tokens.
    - Text encoders/decoders (LLMs) operate on tokenized reports and prompts.
    - Multimodal connectors (projection layers, query transformers, fusion modules, or expert-driven converters) transform image features into token sequences consumable by the LLM.
    - For contrastive pretraining, image and report embeddings are projected into a shared space for CLIP-like alignment.
- **Limitations of current datasets:**
    - Many datasets are single-center, with limited diversity in disease spectrum, scanners, and languages.

- Public multimodal datasets often focus on chest imaging; other organs and modalities are underrepresented.
- Region-level and temporal annotations (for localization, progression tracking) are relatively rare.

---

5. Model / Foundation Model

- **Model Type (surveyed archetypes):**
  - **Contrastive VLMs:** CLIP-like models that learn a shared embedding space for images and reports.
  - **Fusion-based MLLMs:** LLaVA-style architectures where image tokens are injected into an LLM via cross-attention or fusion blocks.
  - **Generative models:** Systems that generate images or segmentations conditioned on text, or generate text conditioned on images (e.g., report generation).
- **New FM vs existing:**
  - The paper does not introduce a single new model; instead, it synthesizes **architectural patterns and design choices** across many existing MLLMs.
- **Key components and innovations (framework level):**

Aspect	Details
Encoders	Pretrained vision encoders (ViTs, CNNs) and LLMs as backbones
Connectors	

Aspect	Details
	Projection-based, query-based (Q-former), fusion-based, and expert-driven language transformers
Training strategies	Contrastive pretraining, instruction tuning, chain-of-thought prompting, RLHF for clinical alignment
Capabilities	Report generation, VQA, retrieval, triage, decision support, image-grounded dialog

- **Training setup (typical):**
  - Pretrain image–text alignment on paired radiology datasets.
  - Adapt a general or medical LLM to accept visual tokens through connectors.
  - Instruction-tune on multimodal tasks (RRG, VQA, captioning, dialog) using curated or synthetic data.
  - Optionally apply RLHF or preference optimization to improve clinical safety and usefulness.

---

6. Multimodal / Integration Aspects (If Applicable)

This review is fundamentally about **multimodal integration** in medical imaging MLLMs.

- **Modalities integrated:**
  - Radiology and other medical images + text (reports, clinical notes, QA) + sometimes structured EHR signals.

- **Integration mechanisms:**
  - **CLIP-style two-tower alignment:**
    - Separate image and text encoders trained with contrastive loss for retrieval and zero-shot classification.
  - **Connector-based fusion:**
    - Projection-based: linear or MLP projections from image features into the LLM token space.
    - Query-based: learnable query tokens attend to visual features and feed condensed information to the LLM.
    - Fusion-based: cross-attention layers inside or around the LLM that jointly process image and text tokens.
    - Expert-driven language transformation: upstream models convert imaging findings into textual descriptions consumed by an LLM.
- **New capabilities enabled:**
  - Image-grounded natural-language interaction (VQA, “chat with your scan”).
  - Zero-shot or few-shot disease classification via text prompts.
  - Automated or draft radiology report generation.
  - Multimodal retrieval (image ? text, patient-level search).

---

7. Experiments and Results

- **Tasks and benchmarks discussed:**
  - Radiology report generation (RRG) from chest X-rays and CT.
  - Visual question answering about imaging findings and clinical context.
  - Image-text retrieval and cross-modal search.
  - Disease detection and classification from multimodal inputs.
  - Early explorations of planning, triage, and longitudinal reasoning.

- **Baselines and comparison themes:**
  - Traditional unimodal CNN/ViT models vs multimodal systems.
  - General LLMs with simple image adapters vs domain-specific medical MLLMs.
  - Trade-offs between model size, task performance, and compute requirements.
- **Key findings (high-level trends):**
  - MLLMs show **promising capabilities** for RRG, VQA, and multimodal reasoning, often outperforming unimodal baselines on complex tasks.
  - However, performance can be unstable across datasets and institutions, and models frequently **hallucinate** or misinterpret subtle findings.
  - Region grounding and localization remain weak; many models reason about images at a coarse, global level.
  - There is a growing shift toward **foundation-model approaches**, leveraging large general or medical LLMs and pre-trained vision encoders.

---

8. Strengths, Limitations, and Open Questions

Strengths of the review and current field:

- Provides a **clear taxonomy** of MLLM architectures, connectors, and training strategies for medical imaging.
- Highlights the importance of **multimodal reasoning** that mirrors how radiologists combine images and clinical context.
- Synthesizes evidence from recent prototypes and studies, giving readers an overview of what is currently feasible.
- Emphasizes practical considerations for clinical adoption (data needs, infrastructure, workflow integration).

Limitations and challenges (field-level):

- Scarcity of large, diverse, high-quality multimodal datasets with region-level labels and outcome annotations.
- High risk of **hallucinated findings** and uncalibrated confidence, especially when MLLMs operate outside their training distribution.
- Limited interpretability and weak region grounding, making it hard to trust model outputs for critical decisions.
- Heavy computational and infrastructure demands, which may be unsuitable for resource-constrained hospitals.
- Regulatory, privacy, and liability questions around deploying MLLMs in clinical care.

Open Questions and Future Directions:

1. How can we design MLLMs with **robust region grounding**, so that textual outputs are tightly coupled to specific image regions?
2. What training strategies and evaluation protocols are needed to **reduce hallucination** and ensure clinically safe behavior?
3. How can we leverage **synthetic data, weak supervision, and federated learning** to overcome data scarcity and privacy constraints?
4. What are effective ways to integrate **EHR data, temporal imaging series, and multi-organ information** into a unified multimodal reasoning system?
5. How should guidelines, benchmarks, and regulations evolve to evaluate and govern MLLMs in real radiology workflows?

9. Context and Broader Impact

- **Position in the landscape:**
  - This paper is one of the first detailed reviews focused specifically on **MLLMs for medical imaging**, complementing broader MMFM and HFM surveys.
  - It connects general LLM and VLM advances (e.g., CLIP, LLaVA-style architectures) to radiology-specific tasks and constraints.
- **Relation to well-known ideas:**
  - Frames medical MLLMs as extensions of **CLIP-like alignment** and **LLM-centric fusion** architectures, adapted to clinical data.
  - Discusses how instruction tuning, chain-of-thought prompting, and RLHF—successful in general AI—might be adapted to medical imaging.
- **Why this review is a useful reference:**
  - For a grad student, it offers a curated tour of design patterns, datasets, and open challenges, making it easier to choose a research direction.
  - It also highlights the **gap between prototype demos and clinically robust systems**, encouraging critical evaluation and responsible innovation.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Radiology practice is inherently multimodal, but most traditional AI systems are unimodal and cannot fully leverage combined image + text + EHR information.

- **Method / model (conceptual):**
  - MLLMs couple powerful LLMs with vision encoders via multimodal connectors (projection, query, fusion, or expert-driven), enabling joint reasoning over images and text.
  - This review categorizes these architectures and training strategies, providing a design space for medical imaging MLLMs.
- **Results / insights:**
  - Early MLLMs show strong potential for RRG, VQA, and multimodal decision support, often surpassing unimodal baselines.
  - However, they are hampered by data scarcity, hallucination, poor region grounding, and deployment challenges.
- **Why it matters:**
  - Understanding MLLMs is crucial for building the next generation of **clinically useful, trustworthy multimodal foundation models** in radiology.
  - This review gives practitioners and researchers a roadmap for tackling open problems and responsibly advancing the field.

Generated via custom pipeline · 2025-11-20