Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: Applications, Challenges, and Future Directions

Authors: Kai Sun, Siyan Xue, Fuchun Sun, Haoran Sun, Yu Luo, Ling Wang, Siyuan Wang, Na Guo, Lei Liu, Tian Zhao, Xinzhou Wang, Lei Yang, Shuo Jin, Jun Yan, Jiahong Dong

Year: 2025

Venue: Artificial Intelligence in Medicine

1. Classification

- **Domain Category:**
  - Medical VLM / MLLM / MMFM + General FM survey / theory
  - This work is a survey that systematizes datasets, architectures, and clinical applications of medical multimodal foundation models (MMFMs).
- **FM Usage Type:**
  - Multimodal FM or cross-modal integration (survey and taxonomy)
- **Key Modalities:**
  - Medical images (CT, MRI, ultrasound, radiography, surgical video), text (reports, clinical notes), structured clinical data, and in some cases robotic and physiological signals.

---

2. Executive Summary

This survey reviews the rapidly growing field of medical multimodal foundation models (MMFMs), which aim to leverage diverse medical data—images, text, signals, and more—to support clinical diagnosis and treatment. The authors first trace the evolution of foundation models from transformers and vision transformers to multimodal models, then categorize MMFMs into medical multimodal vision foundation models (MMVFMs) and medical multimodal vision–language foundation models (MMVLFMs). They systematically describe available datasets, proxy tasks (segmentation, generation, contrastive learning, hybrid tasks), and model architectures, and then connect these to downstream applications such as radiology report

generation, disease diagnosis, treatment planning, and surgical robotics. The survey emphasizes both the opportunities (better generalization, data efficiency, cross-task transfer) and the significant challenges (data quality, compute cost, fairness, interpretability, deployment, and regulation). For a new grad student, this paper serves as a comprehensive starting point for understanding the MMFM landscape and identifying promising research directions.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - Provide a unified view of **medical multimodal foundation models** that can:
    - Integrate heterogeneous medical data (multi-organ, multi-modality).
    - Support a wide spectrum of clinical tasks from diagnosis to treatment.
    - Move toward generalized medical AI rather than narrow, task-specific models.
- **Why this is hard:**
  - **Data issues:**
    - Medical data are high-dimensional, heterogeneous (images, waveforms, text, lab results), and often noisy or incomplete.
    - Large multimodal datasets with consistent labeling and harmonization are rare; privacy and legal restrictions complicate sharing.
  - **Modeling challenges:**
    - Designing architectures that scale to multiple organs, modalities, and tasks while remaining

efficient.

        - Balancing generality with specialization; avoiding catastrophic forgetting while adapting to new tasks.
    - **Clinical and societal constraints:**
        - Need for interpretability, fairness, robustness, and safe deployment in real clinical environments.
        - Regulatory frameworks and validation protocols are still evolving for foundation-model-based systems.

---

4. Data and Modalities

The survey devotes substantial space to **dataset landscapes** for MMFMs.

- **Plain text datasets:**
    - Large corpora of biomedical literature, clinical guidelines, and clinical notes.
    - Used primarily to train or adapt medical LLM backbones that pair with vision or other modalities.
- **Medical image datasets:**
    - Diverse imaging modalities: CT, MRI, X-ray, ultrasound, endoscopy, digital pathology, ophthalmology, etc.
    - Multi-organ and multi-center datasets that enable cross-domain learning.
- **Image–text pair datasets:**
    - Radiology and pathology image–report pairs enabling CLIP-style or CoCa-style vision–language pretraining.

    - Datasets annotated with segmentation masks, bounding boxes, or keypoints for segmentation-oriented proxy tasks.
- **Other modalities:**
    - Surgical videos, robotics telemetry, physiological signals, and multi-omics data for advanced MMFM applications.
- **Preprocessing / representation themes:**
    - Standardization of image resolutions and voxel spacing; patch extraction and multi-scale crops for large images.
    - Tokenization and normalization of medical text; mapping structured EHR fields to embeddings.
    - Careful balancing of modalities to avoid dominance of one data type in multimodal pretraining.

---

5. Model / Foundation Model

- **Model Type (taxonomy):**
    - The paper distinguishes:
        - **MMVFMs (Medical Multimodal Vision Foundation Models):** multimodal vision encoders focusing on multiple medical image modalities.
        - **MMVLFMs (Medical Multimodal Vision–Language Foundation Models):** models that combine image and text (and sometimes more modalities) in a shared framework.
- **Is it a new FM or an existing one?**
    - This is a **survey**; it does not introduce a specific model but analyzes and categorizes many.

- **Proxy task categories (core contribution):**

Category	Role in MMFMs
Segmentation proxy	Use segmentation tasks to learn detailed anatomical representations
Generative proxy	Use generative tasks (reconstruction, synthesis) for representation learning
Contrastive proxy	Use CLIP-like or contrastive objectives for cross-modal alignment
Hybrid proxy	Combine segmentation, generative, and contrastive objectives to cover multiple skills

- **Architectural themes:**
  - Transformer-based encoders and decoders for both vision and language.
  - Two-tower architectures for CLIP-style alignment vs unified encoders for fully fused multimodal representations.
  - Use of adapters, prompts, and low-rank fine-tuning (LoRA) for efficient adaptation of large backbones.
- **Training setup (generic patterns):**
  - Large-scale pretraining on multi-organ, multi-modality datasets using one or more proxy tasks.

- Fine-tuning or prompting on downstream tasks such as segmentation, detection, classification, report generation, and surgical control.
- Multi-task and multi-stage training regimes to gradually build generalized capabilities.

---

6. Multimodal / Integration Aspects (If Applicable)

Multimodal integration is the central theme of MMFMs.

- **Modalities integrated:**
  - Combinations of images, text, clinical variables, and sometimes signals like robotics trajectories or physiology.
- **Integration strategies:**
  - **Early fusion:** combine modalities at the input or low-level feature stage (e.g., concatenated embeddings).
  - **Intermediate fusion:** fuse modality-specific encoders via cross-attention or shared latent spaces.
  - **Late fusion:** combine modality-specific model outputs via ensembles or simple aggregators.
  - **Vision–language alignment:** CLIP-style or CoCa-style objectives to map image and text into a shared space.
- **What this enables:**
  - More holistic modeling of patient state by combining imaging, text, and structured data.
  - Cross-task and cross-organ transfer: representations learned for one modality or organ can help others.

- Unified models that can support multiple downstream tasks from a shared backbone.

---

7. Experiments and Results

As a survey, the paper does not present new experiments; instead, it **summarizes trends** in published MMFM work.

- **Tasks discussed:**
  - Segmentation (organ, lesion) and detection tasks across various imaging modalities.
  - Disease classification, risk prediction, and prognosis.
  - Radiology report generation and medical image captioning.
  - Surgical planning, navigation, and robotics control.
- **Key observations:**
  - MMFMs often **outperform single-task, single-modality baselines**, especially when downstream labels are scarce.
  - Contrastive and hybrid proxy tasks tend to support better zero-shot and few-shot generalization.
  - Increasing data scale and modality diversity generally improves robustness, but also raises compute and data-governance issues.
  - There is still a gap between offline benchmark performance and the reliability needed for clinical deployment.

---

8. Strengths, Limitations, and Open Questions

Strengths (of MMFMs and the survey):

- Provides a **unified taxonomy** that organizes MMFMs by proxy tasks, modalities, and architectures.
- Connects method choices (segmentation vs contrastive vs hybrid) to clinical application domains.
- Highlights how MMFMs can support **precision medicine**, from early diagnosis to personalized treatment.
- Identifies key datasets and benchmarks, giving readers a practical starting point for experimentation.

Limitations and challenges:

- Many MMFMs rely on **limited or biased datasets**, often from a small number of institutions or populations.
- Compute and data requirements are high, raising concerns about **environmental impact and accessibility**.
- Interpretability and explainability remain limited, especially for complex multimodal reasoning.
- Fairness and generalization to under-represented groups are not yet well studied.
- Real-world deployment faces hurdles in regulation, integration, and clinician acceptance.

Open Questions and Future Directions:

1. How can we develop **data-efficient** MMFMs that retain strong performance without requiring enormous datasets and compute?

2. What are effective strategies for **fair and robust multimodal pretraining**, especially across institutions and populations?
3. How can we integrate **causal and mechanistic knowledge** into MMFMs to move beyond pattern recognition?
4. What evaluation frameworks are needed to measure **trustworthiness, interpretability, and clinical impact** of MMFMs?
5. How should MMFMs interface with **clinical workflows and decision-support systems** to provide value without over-automation?

---

9. Context and Broader Impact

- **Position in the landscape:**
  - This survey sits alongside other HFM and MLLM reviews but focuses specifically on **medical multimodal foundation models** that bridge multiple imaging modalities, text, and clinical tasks.
  - It articulates how MMFMs can underpin next-generation clinical AI systems that span diagnosis, treatment, and surgical assistance.
- **Relation to well-known ideas:**
  - Builds on CLIP-like vision–language alignment, transformer-based FMs (BERT, ViT, GPT), and recent medical FMs like M3FM, Me-LLaMA, and TITAN.
  - Frames MMFMs as a path toward **medical artificial general intelligence**, while emphasizing the importance of safety and governance.

- **Why this paper is a useful reference:**
  - For students and practitioners, it provides a broad yet structured overview of the MMFM space, helping them situate specific models and choose research directions.
  - It also highlights critical non-technical dimensions (fairness, regulation, deployment) that will shape the real-world impact of MMFMs.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Medical AI needs foundation models that can handle **multiple modalities and organs** and support tasks ranging from diagnosis to treatment planning.
- **Method / model (conceptual):**
  - MMFMs are built on transformer-based backbones and trained using segmentation, generative, contrastive, or hybrid proxy tasks on large multimodal datasets.
  - The survey categorizes models into MMVFMs and MMVLFMs and analyzes their architectures, datasets, and applications.
- **Results / insights:**
  - MMFMs generally improve performance and data efficiency over narrow models, and support zero-shot or few-shot adaptation to new tasks.
  - However, they face major challenges around data quality, compute cost, fairness, interpretability, and deployment.

- **Why it matters:**
  - MMFMs are likely to be the **backbone technology** for future clinical AI systems; understanding their design and limitations is crucial.
  - This survey offers a roadmap for advancing MMFMs responsibly toward real-world impact.