M3FM: A Multimodal, Multidomain, Multilingual Medical Foundation Model for Zero-Shot Clinical Diagnosis

Authors: Fenglin Liu, Zheng Li, Qingyu Yin, Jinfa Huang, Jiebo Luo, Anshul Thakur, Kim Branson, Patrick Schwab, Bing Yin, Xian Wu, Yefeng Zheng, David A. Clifton

Year: 2025

Venue: npj Digital Medicine

1. Classification

- **Domain Category:**
  - Medical VLM / MLLM / MMFM

- This paper proposes a medical multimodal vision–language foundation model that jointly handles radiology images and multilingual text for diagnosis and report generation.

- **FM Usage Type:**
  - Core FM development **+** Multimodal FM or cross-modal integration
- **Key Modalities:**
  - 2D chest X-rays (CXR)
  - 3D CT images
  - English radiology reports
  - Chinese radiology reports (machine-translated for training; human reports for evaluation on some datasets)

---

2. Executive Summary

M3FM (Multimodal Multidomain Multilingual Foundation Model) is a medical foundation model designed to perform zero-shot radiology report generation and disease diagnosis across imaging domains (CXR and CT) and languages (English and Chinese). The core idea is to first learn a shared vision–language embedding space (MultiMedCLIP) using large English-centric image–report and English–Chinese text pairs, and then train a multilingual medical language model (MultiMedLM) on top of this space. By aligning visual features from different imaging modalities and textual features from multiple languages to English, M3FM can generate reports and perform diagnosis in languages and domains where little or no labeled data exist. The model is pretrained on hundreds of thousands of CXR and CT images with English reports, plus translated Chinese corpora, and is evaluated on nine downstream datasets that cover report generation and disease classification for infectious (COVID-19, TB) and noninfectious diseases. Across zero-shot, few-shot, and fully supervised settings, M3FM often matches or outperforms strong supervised baselines that have full access to labeled data, especially for cross-

language and cross-domain generalization. For a new grad student, this paper is a canonical example of how to build a medical CLIP-style + medical LLM stack and use it for multilingual, low-label clinical scenarios.

---

3. Problem Setup and Motivation

- **Scientific / practical problem:**
  - Build a single foundation model that can:
    - Generate radiology reports from CXR and CT images in both English and Chinese.
    - Diagnose diseases from images and/or generated reports, including rare and emerging conditions.
  - Crucially, the model should work in **zero-shot** or **few-shot** regimes where labeled data in the target language or domain are scarce or absent.
- **Why this is hard:**
  - **Data scarcity and imbalance:**
    - High-quality labeled data for rare diseases and new pathogens (e.g., early COVID-19 waves) are limited exactly when they are most needed.
    - Non-English radiology reports, especially in languages like Chinese, are much less abundant and standardized than English reports.
  - **Multidomain heterogeneity:**
    - CXR and CT have very different appearance, resolution, and information content, yet clinicians want unified reporting and diagnosis workflows.
  - **Multilingual alignment:**
    - Reports in different languages describe similar findings but with different vocabularies, structures, and clinical conventions.

- **Label efficiency:**
  - Training supervised models separately for each disease, modality, and language combination is infeasible; a foundation model should generalize with minimal task-specific labeling.
- **Safety and fairness:**
  - Failure to support low-resource languages and rare diseases risks widening disparities in access to AI-assisted care.

---

4. Data and Modalities

- **Pretraining data (English-centric corpora):**
  - **MIMIC-CXR:**
    - ≈377k chest X-ray images with ≈228k English radiology reports.
  - **COVID-19-CT-CXR:**
    - ≈1k CT/CXR images with English reports focused on COVID-19.
  - **Chinese–English parallel corpora:**
    - Constructed by machine-translating portions of the English reports into Chinese to form Chinese–English text pairs for multilingual training.
- **Downstream datasets (disease reporting and diagnosis):**
  - **Report generation:**
    - IU-Xray (CXR–English reports).
    - COVID-19-CT (CT–Chinese reports).
    - COV-CTR (CT images with English & Chinese reports).
    - Additional qualitative examples for CXR-to-Chinese where human-annotated datasets are unavailable.
  - **Disease classification:**
    - Shenzhen Tuberculosis (CXR, TB vs normal).
    - COVID-CXR (CXR, COVID-19 vs non-COVID).
    - NIH ChestX-ray14 (14-label multi-label classification).

        - CheXpert (multi-label classification).
        - RSNA Pneumonia (pneumonia detection).
        - SIIM-ACR Pneumothorax (pneumothorax detection).
- **Modalities and languages:**
    - Imaging: CXR, CT.
    - Text: radiology reports in English and Chinese.
- **Preprocessing / representation:**
    - Images are encoded by a vision backbone into dense features.
    - Text reports (English and Chinese) are tokenized and embedded for contrastive learning and language modeling.
    - English reports from different datasets are treated as separate corpora to avoid leakage between training and evaluation institutions.

---

5. Model / Foundation Model

- **Model Type:**
    - A **two-stage medical foundation model** composed of:
        - **MultiMedCLIP:** CLIP-style vision–language encoder that aligns images and text.
        - **MultiMedLM:** multilingual medical language model (LLM) trained for report generation and text understanding.
- **Is it a new FM or an existing one?**
    - M3FM is a **new medical foundation model stack** that builds on standard transformer backbones but introduces a specific pretraining strategy for multidomain, multilingual radiology.
- **Key components and innovations:**

Aspect	Details
Vision encoder	CNN/ViT-style backbone encoding CXR and CT into visual embeddings
Text encoder / decoder	Transformer-based encoders/decoders for English and Chinese reports
Alignment module	CLIP-like contrastive loss aligning images and English text, and English–Chinese text pairs
Language model	MultiMedLM trained on multilingual medical corpora to reconstruct and generate reports
Inference strategy	Zero-shot and few-shot report generation and diagnosis using aligned embeddings

- **Training setup (high level):**
  - **Stage 1 – MultiMedCLIP (vision–language alignment):**
    - Pretrain on English-centric corpora: CXR–English, CT–English, and Chinese–English pairs.
    - Contrastive objective encourages matched image–text pairs (or bilingual pairs) to be close in a shared latent space and mismatched ones to be far apart.
    - Enables alignment of new non-English reports by mapping them into the same space as English text and images.
  - **Stage 2 – MultiMedLM (multilingual medical LLM):**
    - Train an autoregressive language model over the aligned text embeddings, reconstructing inputs across languages.
    - Leverages large text corpora (including machine-translated Chinese) to learn biomedical vocabulary and style.

    - **Inference:**
        - For zero-shot report generation, visual embeddings are fed through the aligned text space into MultiMedLM to decode reports in English or Chinese without downstream supervised training.
        - For diagnosis, image features and generated reports can be combined to drive disease classifiers.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Modalities integrated:**
    - Radiology images (CXR, CT) and textual reports in English and Chinese.
- **How integration works:**
    - **CLIP-style two-tower alignment (MultiMedCLIP):**
        - Separate image and text encoders are trained with a contrastive loss so that images and their corresponding reports are nearby in the joint embedding space.
        - English text is the anchor; Chinese reports are aligned via English–Chinese text pairs, effectively “pivoting” through English.
    - **Text-only language modeling (MultiMedLM):**
        - Trained on multilingual text (including translated reports) to generate fluent medical text conditioned on embeddings from MultiMedCLIP.
    - **Zero-shot transfer:**
        - Once everything is aligned, CXR or CT images from unseen datasets can be fed through the image encoder and decoded into English or Chinese reports, even when no labeled image–text pairs exist for that specific institution or language.

- **Why this integration is useful / new capabilities:**
  - Enables **zero-shot multilingual report generation** from images, including generating Chinese reports without any human-labeled CXR-Chinese or CT-Chinese training pairs.
  - Supports **zero-shot and few-shot disease diagnosis** by combining image features and generated reports across diseases and datasets.
  - Provides a flexible template for extending to additional modalities (e.g., other imaging types or languages) by adding new aligned corpora.

---

7. Experiments and Results

- **Tasks / benchmarks:**
  - **Multilingual, multidomain report generation:**
    - CXR-to-English, CT-to-English, CXR-to-Chinese (qualitative), CT-to-Chinese.
  - **Disease diagnosis (classification):**
    - Binary tasks (COVID vs non-COVID, TB vs normal) and multi-label disease prediction (ChestX-ray14, CheXpert) using image + generated reports.
  - **Settings:**
    - Zero-shot report generation and diagnosis (no downstream labels).
    - Few-shot and fully supervised diagnosis using limited or full labels on top of M3FM embeddings.
- **Baselines:**
  - Supervised report generation models such as R2Gen and other encoder–decoder methods trained on specific datasets.
  - Supervised disease classifiers trained directly on images (e.g., CNNs) with full labels.

    - Few-shot and fully supervised methods that do not use a unified foundation model.
- **Key findings (trends):**
    - M3FM achieves **strong zero-shot report generation performance**, often matching or surpassing supervised baselines on CXR-to-English and CT-to-Chinese tasks despite using no downstream labels.
    - In **zero-shot and few-shot diagnosis**, combining M3FM representations and generated reports yields performance close to fully supervised baselines that use large labeled training sets.
    - M3FM is particularly strong in **cross-language generalization**, outperforming baselines on Chinese report generation and diagnosis tasks built from machine-translated training corpora.
    - Across nine benchmark datasets spanning infectious and noninfectious diseases, M3FM provides consistently competitive or superior results, especially in low-label regimes.

---

8. Strengths, Limitations, and Open Questions

Strengths:

- Provides a **single foundation model** that unifies multimodal (image + text), multidomain (CXR + CT), and multilingual (English + Chinese) clinical diagnosis and report generation.
- Demonstrates **genuine zero-shot capabilities**, generating reports and diagnoses without labeled downstream data in the target domain or language.
- Uses existing English-centric corpora and machine translation to bootstrap multilingual capabilities, which is practical for many health systems.
- Offers a concrete, reproducible blueprint for combining CLIP-style alignment with a domain-specific medical LLM.

Limitations:

- Relies heavily on **machine-translated Chinese text**, which may introduce translation artifacts and biases into the model's understanding of Chinese medical language.
- Focuses on **CXR and CT** only; other imaging modalities and richer EHR data are not modeled.
- Zero-shot and few-shot performance, while strong, may still fall short of what clinicians require for fully autonomous deployment.
- Training such a model still requires substantial compute and careful data curation; scaling to more modalities and languages further raises costs.
- Evaluation emphasizes standard generation and classification metrics, which may not fully capture clinical safety and decision impact.

Open Questions and Future Directions:

1. How well does the M3FM paradigm extend to additional languages (e.g., Spanish, Arabic) and imaging modalities (MRI, ultrasound, pathology) when only weak supervision is available?
2. Can higher-quality human-translated corpora or bilingual clinical notes substantially improve multilingual performance and safety relative to machine translation alone?
3. How should we design clinical trials and human-in-the-loop workflows to safely deploy zero-shot report generators in real hospitals?
4. What is the best way to combine M3FM with structured EHR data and other signals (labs, medications) for end-to-end diagnostic support?
5. Can similar multimodal, multilingual foundations be built in a more compute-efficient way, for example via parameter-efficient fine-tuning or distillation?

9. Context and Broader Impact

- **Position in the landscape:**
  - M3FM is to **radiology report generation and diagnosis** what CLIP-style models and GPT-style LLMs are to generic vision–language tasks: a unified backbone that supports multiple tasks, domains, and languages from a single pretraining run.
  - It sits alongside other medical multimodal FMs (e.g., radiology VLMs, medical CLIP variants) but is distinctive in explicitly targeting **multilingual, multidomain zero-shot diagnosis**.
- **Relation to well-known ideas:**
  - Architecturally, MultiMedCLIP is a **CLIP-like two-tower model** for medical images and text, while MultiMedLM is a **medical LLM** akin to Me-LLaMA but optimized for radiology report style and multilinguality.
  - The system follows the modern pattern of using a strong vision encoder + language encoder/decoder, aligned via contrastive learning and then instruction- or task-tuned for downstream use.
- **Why this paper is a useful reference:**
  - It provides a clear recipe for building **multilingual, multidomain medical VLMs** and demonstrating their advantages in low-label regimes.
  - For a grad student, it serves as a practical starting point for designing new medical MLLMs, extending them to other modalities, or exploring safer, more equitable deployment strategies in global health.

---

10. Key Takeaways (Bullet Summary)

- **Problem:**
  - Clinical radiology needs AI systems that can generate reports and support diagnosis across imaging domains and languages,

especially when labeled data for rare diseases or non-English populations are scarce.

- **Method / model:**
  - M3FM combines **MultiMedCLIP**, a CLIP-style vision–language encoder trained on English-centric image–text corpora and English–Chinese text pairs, with **MultiMedLM**, a multilingual medical language model trained on large text corpora.
  - The model aligns CXR and CT images with English and Chinese reports in a shared latent space and then uses an LLM to generate reports and support downstream classification.
- **Results:**
  - Achieves strong **zero-shot** and **few-shot** performance on nine datasets covering report generation and disease diagnosis, often matching or exceeding supervised baselines that rely on labeled data.
  - Particularly strong for cross-language tasks such as CT-to-Chinese report generation and COVID-19 diagnosis with minimal labeled data.
- **Why it matters:**
  - Demonstrates that a single medical multimodal foundation model can reduce dependence on large labeled datasets and extend AI benefits to **low-resource languages and rare diseases**.
  - Provides a concrete blueprint for future medical VLMs and MLLMs targeting multilingual, label-efficient clinical decision support.