Review

Advancements in Multimodal Foundation Models for Healthcare: An In-Depth Review and Future Outlook

Yajie Zhang *, Zhi-An Huang, Xingyu Wu, Songpan Gao, Rui Liu, Zhen Chen, Jibin Wu, Yao Hu, Kay Chen Tan

Posted Date: 16 June 2026

doi: 10.20944/preprints202606.1256.v1

Keywords: medical; vision-language model; vision model; multimodal large language model; multimodal foundation model

 doi:10.20944/preprints202606.1256.v1

Review

Advancements in Multimodal Foundation Models for Healthcare: An In-Depth Review and Future Outlook

Yajie Zhang [1,*], Zhi-An Huang [2], Xingyu Wu [1], Songpan Gao [2], Rui Liu [1], Zhen Chen [1], Jibin Wu [1], Yao Hu [1] and Kay Chen Tan [1]

[1] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong
[2] Department of Computer Science, City University of Hong Kong (Dongguan), Dongguan 523000, China
* Correspondence: yajie.zhang@connect.polyu.hk

Abstract

Medical multimodal foundation models (MMFMs) have become a central element of medical artificial intelligence, supporting progress in clinical workflows like diagnosis, report generation, and multimodal reasoning. However, existing surveys face issues with quick aging and brief coverage of model types. This paper offers a fine-grained review of MMFMs covering January 2023 to July 2025, filling these needs through a structured framework. We analyze key technical features—including model size, dataset size, and architectural designs—across three main model categories: Universal MMFMs (Uni-MMFMs) with wide use, Modality-specific MMFMs (MS-MMFMs) with single-modality specialization, and Organ-specific MMFMs (OS-MMFMs) with organ-specific tuning. We chart development paths and highlight major challenges: data scarcity and privacy constraints, insufficient cross-modal alignment, limited clinical interpretability, and poor generalization in real-world scenarios. We also propose future directions including data-level growth (multi-source integration, synthetic generation), architecture-level updates (unified image-text frameworks), user-centric features (interpretability, ethical compliance), and developer-focused improvements (continuous learning, multimodal conflict resolution). This survey summarizes the current state of MMFMs and provides a guide for building reliable, interpretable, and useful multimodal medical AI systems.

Keywords: medical; vision-language model; vision model; multimodal large language model; multimodal foundation model

1. Introduction

Medical multimodal foundation models (MMFMs) have gained significant traction in recent years. Trained via self-supervised or semi-supervised learning on extensive multimodal datasets [1], these models capture robust medical knowledge representations. High-performing medical MMFMs can be integrated across healthcare institutions to streamline workflows such as triage, specialist diagnosis, and text analysis [2]. This application offers substantial improvements in diagnostic efficiency and accuracy. Therefore, the development of medical MMFMs has become a pivotal research focus within medical artificial intelligence, drawing widespread interest.

MMFMs show several distinct patterns. First, model performance follows scaling laws [3], where greater data volume and architectural complexity typically lead to improved outcomes. Second, architectural diversity has expanded, incorporating modern methods such as masked autoencoders (MAE) [4], contrastive language-image pre-training (CLIP) [5], and large vision-language models (e.g., LLaVA [6]). These frameworks support various MMFM variants tailored for healthcare, facilitating widespread use in medical image analysis and clinical text understanding. Third, a tiered MMFM ecosystem has developed, matching the complex nature of medical data and application diversity. MMFMs are therefore grouped into three main types [7]:

 doi:10.20944/preprints202606.1256.v1

- Universal MMFMs (Uni-MMFMs): Trained on massive, multi-institutional, multimodal datasets covering varied medical knowledge, these models emphasize reliable generalization and wide utility across diseases and tasks, acting as base platforms for downstream work.
- Modality-specific MMFMs (MS-MMFMs): Focused on a single medical data modality (e.g., radiological imaging, histopathology slides, or ultrasound imaging), these models are tuned for distinct characteristics to deliver better results in specific tasks.
- Organ-specific MMFMs (OS-MMFMs): Targeting specific organ systems (e.g., brain, heart, or eye) and their associated conditions, these models utilize organ-focused multimodal data to enhance organ-targeted diagnostic and prognostic tasks.

However, a thorough and detailed analysis of progress in medical MMFMs is missing from the current literature. Existing surveys in the field often discuss the classifications of architectural designs [8–10], include a selection of key methodologies [8], and offer an overview to the scope of application [11]. Notably, these reviews face several major drawbacks: they quickly age due to the speed of development [8,12,13], which covers core medical MMFMs only up to 2024. Moreover, these works [9,10,14,15] often give general summaries of medical MMFMs, omitting the subtleties across different model categories. As a result, they offer few practical suggestions for the distinct needs and contexts of medical professionals.

This work presents a systematic, fine-grained review of medical MMFMs spanning January 2023 to July 2025. By using a structured classification framework, we analyze key technical features—including model size, release dates, dataset size, and novel architectures—across model types, highlighting development paths and critical advances. We suggest future directions for the advancement of major medical models.

2. Background and Taxonomy

This section briefly outlines the scope of our analysis and introduces fundamental concepts essential for interpreting subsequent sections.

Medical Multimodal Foundation Models. Foundation model is defined as base model trained on large-scale data using self-supervised or semi-supervised learning, adaptable to diverse downstream tasks [16]. Medical MMFMs extend this paradigm by training on medical imaging or multimodal vision-language datasets. Training a medical MMFM typically involves medical dataset curation, model architecture selection, and customized training strategies. Deployment of pretrained MMFMs to downstream tasks employs zero-shot/few-shot learning [17], fine-tuning, or prompt tuning techniques [18].

Literature Search. As shown in Figure 1, we systematically searched for relevant papers published between January 2023 and July 2025 focusing on medical MMFMs. Literature searches were conducted across the following databases: Google Scholar, PubMed, arXiv, Web of Science, IEEE Xplore, Nature, ACM Digital Library, and Scopus. Search strategies were designed to reflect our inclusion criteria:

- Uni-MMFMs: Keywords combined ('medical' or 'healthcare' or 'biomedical') and ('foundation model' or 'vision-language model' or 'multimodal large language model' or 'multimodal foundation model').
- MS-/OS-MMFMs: Keywords combined ('[modality-/organ-specific name]') and ('foundation model' or 'vision-language model' or 'multimodal large language model' or 'multimodal foundation model'), where [modality-specific/organ-specific name] is replaced with the specific modality/organ (e.g., 'CT').

The medical modalities organs covered include pathology, X-ray, CT, and Ultrasound (US), as well as organ-specific areas such as the eye, brain, heart, and skin. Some modalities and organs are excluded as certain modalities (e.g., MRI) typically require specialized preprocessing pipelines that diverge significantly from the unified frameworks adopted for the selected modalities in current MMFM research. Further details on universal MMFMs, MS-MMFMs, and OS-MMFMs are provided in the referenced sections. Initial screening of titles and abstracts yielded 685 potentially relevant papers.

 doi:10.20944/preprints202606.1256.v1

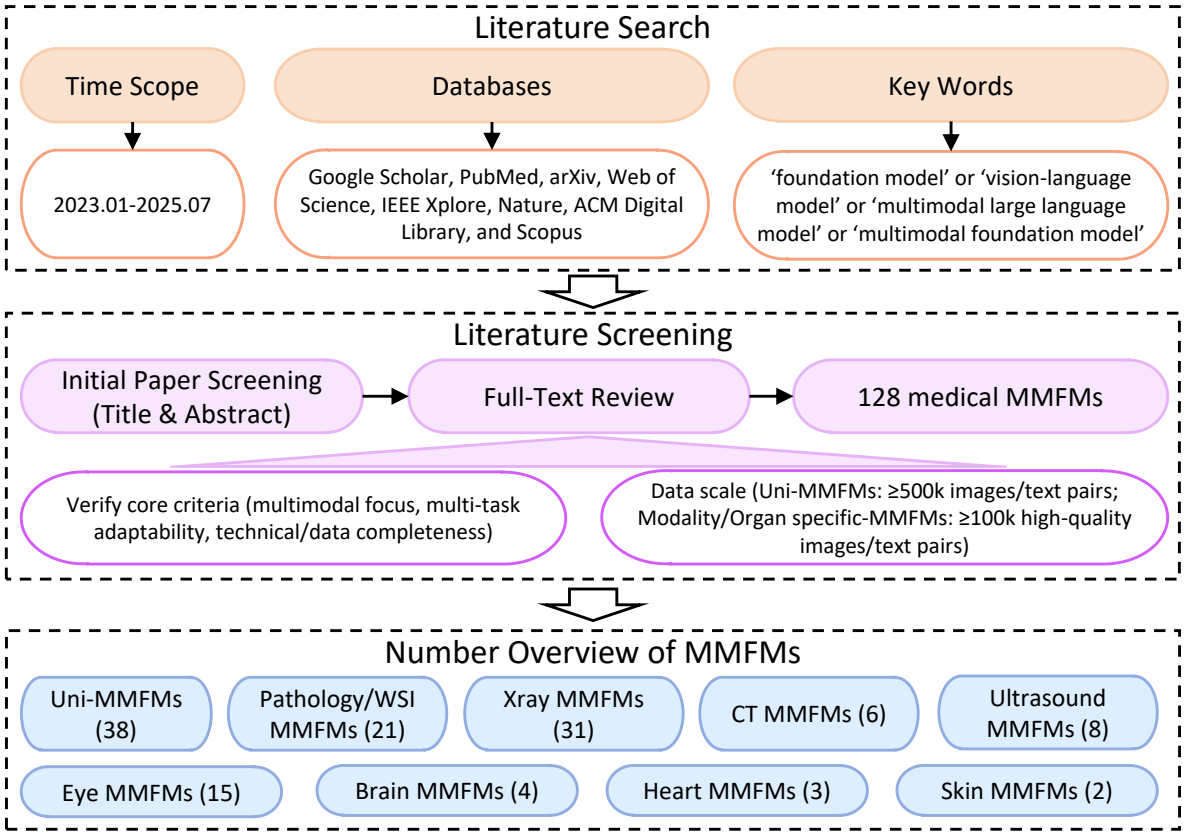

Figure 1. Workflow of literature search, screening, and selection for medical MMFMs (Jan 2023–Jul 2025).

Subsequent full-text review refined the selection to 128 papers included in our analysis. We focus on medical MMFMs adaptable to diverse downstream tasks and single-task models are excluded. Based on empirical research and public dataset benchmarks in medical artificial intelligence, we include medical MMFMs that meet the following training data scale in our statistical analysis:

- Uni-MMFMs trained with at least 500,000 images or image-text pairs are included. These models require integration of heterogeneous data distributions, with public resources (e.g., TCGA [19] (> 33,000 pathology image-report pairs), MIMIC-CXR [20], CheXpert [21], and NIH ChestX-ray [22] collectively providing > 870,000 image-text pairs) suggesting 500,000 samples as a benchmark threshold to ensure disease diversity and cross-modal consistency.
- MS-/OS-MMFMs train at least 100,000 high-quality medical images or image-text pairs are included. This threshold is based on clinical validation: public datasets such as MIMIC-CXR [20] (378,000 image-text pairs) and CheXpert [21] (224,000 image-text pairs) demonstrate that 100,000 image-text pairs is the minimum viable scale for models to achieve radiologist-level diagnostic performance. Below this threshold, the model's generalization ability concerning rare lesions, equipment differences, and complex clinical presentations significantly declines.

Architecture Types. As shown in Figure 2, the architectures of MMFMs encompass six fundamental types:

 doi:10.20944/preprints202606.1256.v1

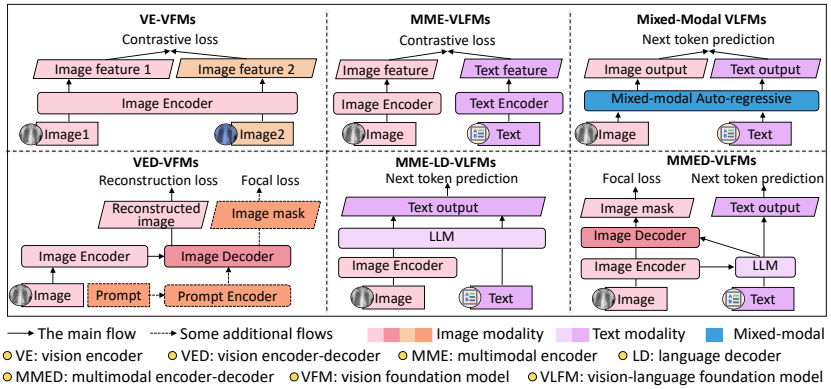

Figure 2. An illustrative diagram summarizing the six primary architectural paradigms of vision and vision-language foundation models. The architecture types include: VE-VFMs, VED-VFMs, MME-VLFMs, MME-LD-VLFMs, MMED-VLFMs, and MM-VLFMs.

- Vision-encoder based vision foundation models (VE-VFMs): they are built on a vision-encoder architecture (e.g., ViT [23] or CNN [24]) that maps raw pixels into dense feature embeddings. These models are trained via self-supervised contrastive learning [25] to maximize similarity between different augmented views of the same image while pushing different images apart.
- Vision encoder-decoder based vision foundation models (VED-VFMs): they are built on a vision encoder-decoder architecture (e.g., MAE [4] or VAE [26]), where the encoder compresses raw pixels into latent representations and the decoder reconstructs the original input. They are trained via self-supervised reconstruction objectives to minimize the discrepancy between original and reconstructed images.
- Multimodal-encoder vision-language foundation models (MME-VLFMs): they are built on a dual-stream encoder architecture (e.g., CLIP [5]) that projects raw pixels and natural language tokens into a shared latent space via separate encoders. They are trained via contrastive learning on massive image-text pairs to maximize the similarity of matched pairs while minimizing that of mismatched ones.
- Multimodal-encoder and language-decoder vision-language foundation models (MME-LD-VLFMs): they are built on an architecture integrating a pre-trained visual encoder with an LLM decoder (e.g., LLaVA [6]), projecting visual features as tokens into the LLM's embedding space. They are trained on instruction-following data using a causal language modeling loss to generate text responses grounded in visual contexts.
- Multimodal encoder-decoder vision-language foundation models (MMED-VLFMs): they are built on a composite architecture integrating a vision encoder, a reasoning LLM, and a visual decoder. This framework processes visual-textual inputs to interpret semantic data and synthesize new images or masks, optimized via a multi-task objective combining language modeling and visual reconstruction/segmentation losses.
- Mixed-modal vision-language foundation models (MM-VLFMs) [27]: they are built on a unified architecture that processes and generates interleaved sequences of discrete text and visual tokens within a single transformer. Trained with an autoregressive next-token prediction objective, they handle diverse cross-modal tasks including visual question answering, captioning, and content creation.

3. Universal Multimodal Foundation Models

3.1. Overview of Uni-MMFMs

Between January 2023 and July 2025, 37 Uni-MMFMs were trained on datasets of at least 500,000 samples, enabling adaptation to diverse downstream tasks. We provide the architectural and scalability overview of Uni-MMFMs below.

 doi:10.20944/preprints202606.1256.v1

Architectural overview. As shown in Figure 3, the ratio of VFMs to VLFMs is 6:31, reflecting a predominant focus on concurrently supporting medical image analysis and NLP tasks. This trend aligns with efforts to develop unified agents capable of interpreting multimodal medical data, including medical images, electronic health records, clinical text, and scientific literature. Specifically, VE-VFMs, VED-VFMs, MME-VLFMs, MME-LD-VLFMs, MMED-VLFMs, MM-VLFMs have a ratio of 1:5:6:21:3:1. The predominance of MME-LD-VLFMs (21 methods) establishes visual instruction tuning [6] as the prevailing paradigm. This approach employs multimodal next-token prediction objectives to enhance visual-textual feature alignment and generate coherent responses, significantly improving zero-shot and few-shot capabilities in complex applications such as open-ended medical QA, report generation, and multimodal reasoning. Notably, four recently developed models emerged during 2024–2025: MMED-VLFMs (Med-Gemini [45], UniBiomed [55], BiomedParse [39]) and the MM-VLFM implementation MedMax [27]. This pattern suggests growing exploration of multimodal generative and analysis architectures as emerging alternatives to the established MME-LD-VLFM framework.

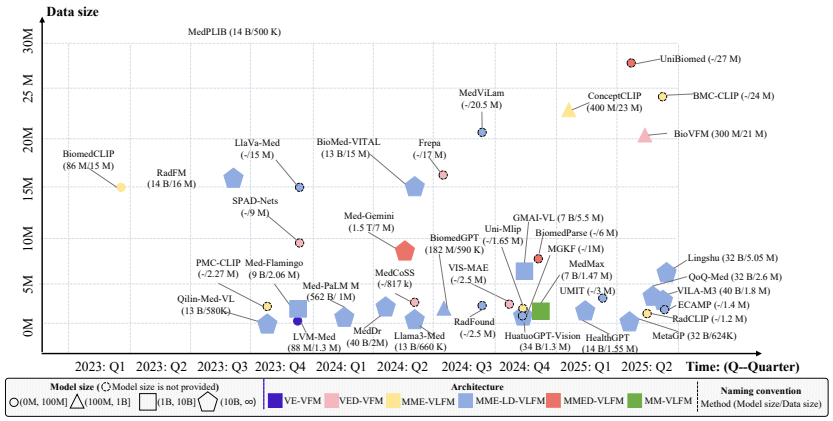

Figure 3. Architectural configurations, model scale, training data size, and developmental timeline of Uni-MMFMs. Uni-MMFMs include: Qilin-Med-VL [28], LVM-Med [29], RadFM [30], LLaVA-Med [31], Med-Flamingo [32], BiomedCLIP [33], PMC-CLIP [34], VIS-MAE [35], BiomedGPT [36], Med-PaLM M [37], RadFound [38], BiomedParse [39], MedViLam [40], GMAI-VL [41], Llama3-Med [42], Frepa [43], MedDr [44], Med-Gemini [45], BioMed-VITAL [46], SPAD-Nets [47], MedCoSS [48], VILA-M3 [49], HuatuoGPT-Vision [50], MedMax [27], Uni-Mlip [51], MetaGP [52], HealthGPT [53], Lingshu [54], UniBiomed [55], MGKF [56], ConceptCLIP [57], BioVFM [58], QoQ-Med [59], ECAMP [60], BMC-CLIP [61], RadCLIP [62], MedPLIB [63], UMIT [64].

Scalability overview. Model parameters show a bimodal distribution, with 70% falling in the 1B–100B range. Med-PaLM M [37] marks the high end at 562B parameters, while smaller models like ConceptCLIP [57] and BiomedCLIP [33] stay below 100M parameters, suggesting potential for edge use. Training dataset sizes for Uni-MMFMs are mostly limited to the scale of millions of samples, usually ranging from 1M to 10M (median 5.5M). Datasets over 20M samples are uncommon (e.g., UniBiomed: 27M [55], MedVilam: 20.5M [40]). This issue comes mainly from the shortage of high-quality, annotated medical data. Collecting such data involves strict privacy rules, and expert clinical knowledge is needed for annotation, especially for multimodal tasks (e.g., text matched with radiology/pathology images).

3.2. Challenges and Methods

The creation of Uni-MMFMs could transform healthcare by offering versatile artificial intelligence able to analyze varied, multimodal clinical data. However, reaching this goal faces major obstacles linked to the specifics of medical data and the needs of clinical application. These issues can be mainly grouped into five key areas: (1) gaps in data quantity and quality, (2) strict data privacy rules, (3) architectural and design difficulties, (4) barriers in representation learning, and (5) lack of model robustness and trust. This section examines these problems, explaining their causes and describing the common strategies used by recent research to address them.

 doi:10.20944/preprints202606.1256.v1

Quantity and quality limitations in medical datasets. The shortage and poor state of massive, high-quality multimodal medical datasets create major barriers. Limited scale, task coverage, modality diversity, and weak image-text alignment together restrict support for complex clinical decision-making and cross-modal learning. To tackle quantity issues, approaches like RadFM [30], Med-MLLM [65], and Med-Gemini [45] utilize public datasets. BiomedCLIP [33] and PMC-CLIP [34] collect multimodal data from scientific literature. Notably, BiRD [66] uses GPT [67] to directly generate image-caption datasets from segmentation annotations, turning masks into descriptive text. Synthetic data generation is improved by MINIM's [68] text-conditioned diffusion model for realistic images. For balancing size and quality, GMAI-VL-5.5M [41] combines 219 datasets across 13 modalities with pixel-level alignment and noise reduction, while Lingshu [54] uses GPT-4o synthesis and strict quality control for 12 modalities.

Data privacy constraints. Medical data is highly private, and its use is bound by tight privacy rules. This poses a major hurdle to pooling data from various institutions for model training, curbing the variety and size of datasets. Federated learning (FL) has become the key approach to this issue. As shown by Lu et al. [69], FL allows joint model training across many institutions without moving raw data. Instead, only encrypted model parameter changes are shared, letting the model learn from a scattered data source while keeping private patient information secure.

Model design challenges. Building a unified architecture that works well across various medical tasks is difficult due to the "single-advantage differentiation" of many models. Some models have strong image/text encoders but weak fusion, while others are tuned for multimodal integration losing flexibility. Also, the spatial variation of medical images and the high computational cost of large models add additional design limits. New architectures are being proposed to build more flexible and efficient models. PTUnifier [70] presents a unified framework using visual and textual prompts to close the functional gap between different model types. To manage spatial variation, SPAD-Nets [47] use spatially adaptive convolutions that change their parameters based on the input. To tackle computational needs, BiomedGPT [36] shows that a small 186M-parameter vision-language model can keep strong results, pointing to a path toward efficiency.

Representation learning hurdles. Learning effective representations from medical data is limited by various issues: poor voxel-level semantics, weak regional analysis, loss of sharp details, poor cross-modal alignment, and conflict between representations needed for different tasks and modalities. These problems come from the detailed, region-specific, and complex reality of medical diagnosis. For spatial and regional understanding: VoCo [71] uses volume contrastive learning to separate regional features, while MedRegA [72] starts region-focused pretraining for combined image-level and detailed analysis. For clear details: Frepa [43] keeps high-frequency preservation using adversarial learning with custom masking. For cross-modal alignment: MPMA [73] adds global-local modules to better mix clinical domain knowledge into the aligned representation. For multi-task conflict: Uni-Med [74] uses a mixture-of-experts module to actively tune cross-task features. VILA-M3 [49] combines task-specific expert models, letting the visual language model trigger relevant experts on demand and merge their outputs. HealthGPT [53] applies heterogeneous low-rank adaptation to split comprehension and generation knowledge. This method helps the model to perform well both on medical visual comprehension tasks and generation tasks, reducing task conflict. QoQ-Med [59] handles imbalanced modality learning with a domain-aware policy that changes training weights.

Robustness and trust deficits. This major issue mixes weak performance on unseen data or distributions, catastrophic forgetting when learning new tasks, and a gap in interpretability for clinical users. The critical nature of medical applications requires models that are not only accurate but also reliable, flexible, and clear. MedDr [44] improves generalization using retrieval-augmented methods to fix modality misalignment. MedCoss [48] uses rehearsal-based continual learning with staged modality mixing to stop catastrophic forgetting during new modality training. For interpretability, UniBiomed [55] and MedPLIB [63] target segmentation and reporting for reliable diagnosis, while ConceptCLIP [57] connects image regions to concepts through dual-alignment pretraining. Medical reasoning

 doi:10.20944/preprints202606.1256.v1

transparency is improved by Med-PaLM M [37] and MetaGP's [52] chain-of-thought methods for step-by-step decision-making, with Lingshu's [54] reinforcement learning with verifiable rewards.

4. Modality-Specific Multimodal Foundation Models

Following the concepts of Uni-MMFMs noted earlier, medical data naturally includes varied modalities—such as radiological imaging, histopathology slides, and ultrasound—each with unique formats and clinical uses. To solve the issues of Uni-MMFMs' wide scope but weak depth, MS-MMFMs are built to focus on single medical modalities. This section covers four main MS-MMFMs: Pathology/WSI (Sec. 4.1), X-ray (Sec. 4.2), CT (Sec. 4.3), and ultrasound (Sec. 4.4), discussing their technical problems and novel methods to show the benefit of specific models in targeted medical AI areas.

4.1. Pathology/WSI Multimodal Foundation Models

Pathology images are microscopic views of stained tissue samples, showing cellular structure, tissue patterns, and biomarker levels, guiding traditional pathological diagnosis. Whole slide images (WSI) are their high-resolution digital equivalents, produced by specialized scanners imaging entire glass slides [75]. Both image types show strong similarities, fitting them for grouping within the same class of foundational models.

4.1.1. Overview of Pathology/WSI MMFMs

Overview Analysis. From January 2023 to July 2025, a total of 21 pathology/WSI MMFMs have been proposed, as illustrated in Figure 4. Among these models, the quantity ratio of the four architectural approaches —VE-VFM, VED-VFM, MME-VLFM, and MME-LD-VLFM—stands at 5:1:5:10. Notably, the two architectures of MMED-VLFM and MM-VLFM have not yet emerged in the field of large pathology models. This phenomenon indicates that although current pathology/WSI MMFMs can accomplish tasks such as image segmentation and image generation through fine-tuning, they still lack the capability to perform image manipulation tasks without fine-tuning. In terms of model parameter scale, the maximum value reaches 15B, achieved by the CPath-Omni [93] model. Regarding data scale, the largest dataset, with a size of 1.3B, is from the Prov-GigaPath [82] model. It is worth noting that models with pre-training data volume exceeding 100M (including Prov-GigaPath [82], RudolfV [87], GPFM [88], and UNI [77]) all belong to the VFM type and do not possess language reasoning capabilities. Among VLFM models, CHIEF [89] has the largest pre-training data scale, utilizing a total of 15M image-text data pairs. The above data demonstrate that there remains a significant gap in image-text data within the field of pathological modalities.

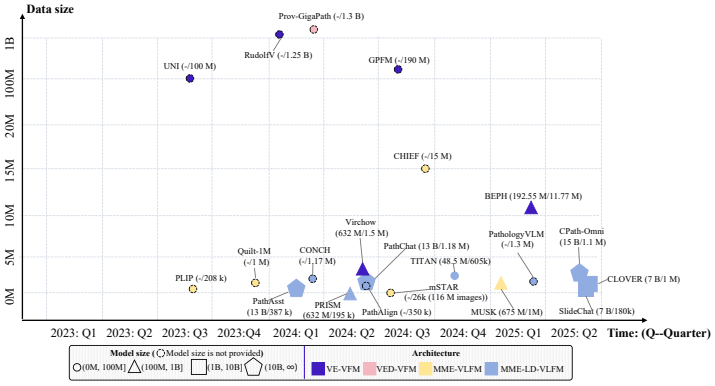

Figure 4. Architectural configurations, model scale, training data size, and developmental timeline of Pathology/WSI MMFMs. Pathology/WSI MMFMs include: PathChat [76], UNI [77], PLIP [78], Quilt-1M [79], Virchow [80], CONCH [81], Prov-GigaPath [82], TITAN [83], PathAlign [84], PathAsst [85], PRISM [86], RudolfV [87], GPFM [88], CHIEF [89], mSTAR [90], BEPH [91], CLOVER [92], CPath-Omni [93], PathologyVLM [94], SlideChat [95], MUSK [96].

 doi:10.20944/preprints202606.1256.v1

4.1.2. Challenges and Methods

Pathology/WSI MMFMs meet several major hurdles that limit their broad clinical use. These issues mainly involve data shortage, struggles in grasping both local and global features, getting robust image-text matching, and poor flexibility across clinical tasks. Here is a summary of these problems and the matching modern methods.

Data scarcity. The critical lack of annotated pathology datasets is a major bottleneck for pathology/WSI MMFMs training. To mitigate this issue, several studies have focused on constructing large-scale annotated datasets from diverse sources. Huang et al. [78] developed a publicly shared image-text dataset using data from medical Twitter and other online sources, which enabled the training of the pathology-specific visual-language foundation model. Ikezogwo et al. [79] proposed Quilt-1M, a method that leverages untapped histopathology educational videos from YouTube through a sophisticated processing pipeline; by combining the resulting QUILT dataset with existing open-source data, the approach achieved 1M paired samples. Additionally, Lu et al. [81] introduced CONCH, which curates image-caption pairs from educational sources and the PMC database using an automated pipeline, generating a total of 1.17M pairs.

Difficulty capturing both local (tile-level) and global (slide-level) patterns. To address this issue, Prov-GigaPath [82] and SlideChat [95] propose to integrates LongNet [97] to enable ultra-long-context modeling. CPath-Omni [93] is proposed as a unified multimodal foundation model that handles both patch and WSI analysis within a single model, thereby facilitating knowledge integration and reducing redundancy. mSTAR [90] resolves both local and global details by the two-stage pretraining. In the first stage, it pretrains a slide aggregator via multimodal contrastive learning to acquire slide-level knowledge; in the second stage, this aggregator acts as a teacher to inject slide-level knowledge into the patch extractor through self-taught training, bridging the gap between patch-level and slide-level analysis.

Image-text alignment. Aligning gigapixel WSIs with slide-level diagnostic text is challenging due to the "many-to-one" relationship between image patches and text, as well as the scarcity of curated pairs. PathAlign [84] addresses this by strategically curating part-level diagnostic text pairs, generating synthetic captions to supplement missing annotations. These modifications enable the model to learn effective slide-level vision-language alignment.

Lack of generalization across clinical tasks. Existing pathology/WSI MMFMs often excel at specific task types but lack generalization across the full range of clinical tasks, with no single model performing optimally across all tasks. GPFM [88] tackles this by implementing a knowledge distillation approach: it distills knowledge from multiple specialized off-the-shelf expert MMFMs into a single student model. This integration of strengths from multiple models does not require access to the original private training data of the expert models, enhancing the student model's generalization across diverse clinical tasks.

4.2. X-ray Multimodal Foundation Models

X-ray images are 2D projections created by X-ray beams passing through human tissues. Different tissues absorb X-rays unevenly, causing visible contrast in the image. These images capture the structure, density variations, and spatial layout of major organs. In this section, we cover the recent progress of X-ray MMFMs.

4.2.1. Overview of X-ray MMFMs

Overall, among X-ray MMFMs, models based on VE-VFM, VED-VFM, MME-VLFM, MME-LD-VLFM, MMED-VLFM, and diffusion models account for a quantity ratio of 1:1:17:8:3:1 as Figure 5. This indicates that current research on X-ray MMFMs mostly focuses on image-text task processing. The vast majority of models emphasize image-text alignment, while another 8 models specialize in image-text reasoning. Notably, there is an additional study RoentGen [109] focusing on diffusion models, exploring the technical approach to generating X-ray chest radiograph images through text. In

 doi:10.20944/preprints202606.1256.v1

terms of training data, the approach using the largest amount of training data currently is CheXagent [111], which utilizes 8.5M image-text pairs. In contrast, the scale of most other pre-training datasets ranges between 100K and 500K. The rapid development of X-ray VLFMs benefits from public datasets such as MIMIC-CXR [20], CheXpert [21], and ChestX-ray 8 [127], which contain large-scale X-ray chest radiograph image-text pairs. However, this has also led to the current focus of X-ray MMFMs research on chest radiograph analysis, and the analysis of X-ray images of other body parts remains insufficient.

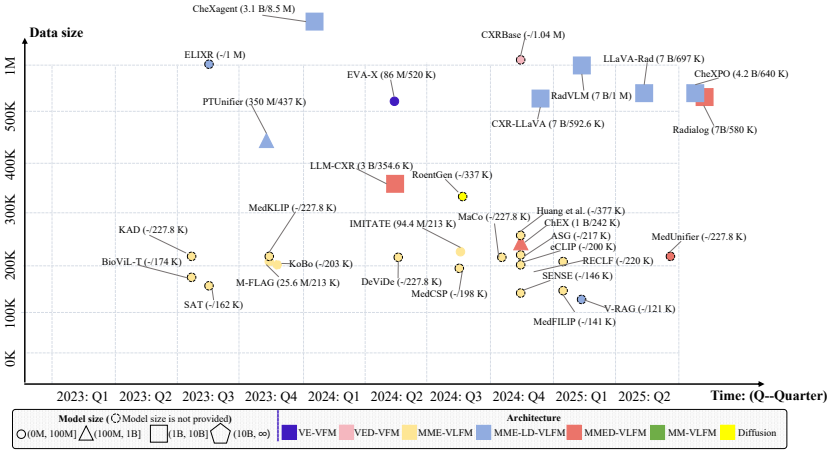

Figure 5. Architectural configurations, model scale, training data size, and developmental timeline of X-ray MMFMs. X-ray MMFMs include: ELIXR [98], KAD [99], LLM-CXR [100], MedKLIP [101], Radialog [102], PTUnifier [70], BioViL-T [103], SAT [104], M-FLAG [105], KoBo [106], EVA-X [107], CXRBase [108], RoentGen [109], ChEX [110], CheXagent [111], MaCo [112], eCLIP [113], LLaVA-Rad [114], MedCSP [115], ASG [116], Huang et al. [112], DeViDe [117], SENSE [118], CheXPO [119], CXR-LLaVA [120], RadVLM [121], RECLF [122], IMITATE [123], MedFILIP [124], V-RAG [125], MedUnifier [126].

4.2.2. Challenges and Methods

X-ray MMFMs have become a key tool in radiological image analysis, improving clinical tasks such as report generation and disease diagnosis. However, the field still faces complex issues—covering data integration, cross-modal alignment, model functionality, and clinical flexibility—that hinder their real-world use. This section outlines these main challenges, explains their technical hurdles, and lists related solutions found in existing studies.

Cross-source and cross-modal data integration and alignment. Clinical data from multiple sources often has temporal inconsistencies, and information from different modalities such as images and reports varies in informational scale and acquisition time, making integration and alignment difficult. To address cross-source integration, MEDCSP [115] employs modality-specific encoders to unify data within individual sources and leverages temporal information and diagnosis history to link patients across sources. For cross-modal alignment, BioViL-T [103] incorporates a multi-image encoder to integrate prior images and longitudinal context, ASG [116] focuses on fine-grained semantic matching between anatomical regions and sentences, and LLM-CXR [100] tokenizes images into the same discrete space as text to enable direct multimodal interaction.

Suboptimal medical semantic representation and data semantic inconsistency. General vision-language models miss subtle medical semantics, confuse similar unpaired image-report pairs in contrastive learning, and overlook the hierarchical structure of radiology reports. Medical datasets also show visual-textual concept mismatch and inconsistent terminology, detail, and syntax in reports. To solve these problems, SAT [104] changes the contrastive learning framework by categorizing samples based on semantic similarity, MaCo [112] improves detailed matching through masked contrastive learning with a correlation weighting method, eCLIP [113] adds radiologist eye-gaze data for guidance, DeViDe [117] gets precise matching using a token-level cross-attention knowledge retrieval module, SENSE [118] and RECLF [122] refine textual feature extraction and reasoning for report structure, MedFILIP [124] uses LLM to pull specific disease information, KoBo [106] applies a clinical knowledge

 doi:10.20944/preprints202606.1256.v1

mixing framework during pre-training, and Huang et al. [112] proposed a report refinement method using clinical dictionaries and knowledge-enhancement metrics to reduce noise.

Architectural and optimization bottlenecks. Dual-encoder and fusion-encoder architectures lack sufficient synergy, and joint training of multiple modules is often risky for issues like latent space collapse. PTUnifier [70] combines pre-training tasks from both designs to support joint learning of single-, cross-, and multi-modal representations. M-FLAG [105] freezes pre-trained language modules and uses latent space geometry tuning mainly on vision processing modules to reduce training problems.

Limited generalization and task-specific overfitting. Models have weak zero-shot generalization to unseen diseases and are prone to catastrophic forgetting when fine-tuned on specific radiology tasks. KAD [99] uses medical knowledge graphs and pulls clinical entities from reports to boost zero-shot results, while MedKLIP [102] [101] encodes disease entities as detailed descriptions. RaDialog adopts a "replay task" method with pseudo-labels from non-fine-tuned models and adds general language tasks during radiology-specific training to stop catastrophic forgetting.

Deficiencies in interaction, interpretability, and generation. Models lack multi-turn dialogue capabilities, localized interpretability, are prone to generating factually incorrect content ("hallucination"), and pay limited attention to image generation. ChEX [110] integrates textual prompts and bounding boxes into a multitask architecture to improve interactivity and interpretability, ASG [116] enhances interpretability through fine-grained anatomical region-sentence alignment, RadVLM [121] is fine-tuned on multi-turn conversation datasets for conversational ability, V-RAG [125] introduces a visual retrieval-augmented generation framework to reduce hallucination, and MedUnifier [126] integrates text-grounded image generation capabilities into a multimodal learning framework.

Challenges in preference alignment. Aligning MMFMs with complex clinical preferences is difficult due to a lack of training data with clinical significance, the long-tailed distribution of real diseases, and the high cost of creating preference datasets. CheXPO [119] addresses these challenges by constructing a large-scale multi-task visual instruction dataset, building counterfactual rationales, and conducting direct preference optimization (DPO).

4.3. CT Multimodal Foundation Models

4.3.1. Overview of CT MMFMs

As shown in Figure 6, from January 2023 to July 2025, a total of six eligible CT MMFMs have emerged. These include CT-FM [132] and FM-CT [130] based on the VE-VFM architecture, M3FM [128] based on the MME-VLFM architecture, OrthoDoc [129] and Med3DVLM [133] based on the MME-LD-VLFM architecture, and LCTfound [131] based on the diffusion architecture. The parameter scale of these CT MMFMs is generally small, with the largest model currently being Med3DVLM [133] at 7B parameters, while the others mostly remain under 1B. In terms of pre-training data, LCTfound [131] has the largest dataset, comprising 28M lung CT images, whereas the other methods use pre-training data ranging from 100K to 400K samples. These results indicate that, compared to other modalities such as X-ray and pathology, CT MMFMs are still in the early stages of development and suffer from a relative scarcity of training data—particularly high-quality paired CT-text datasets.

4.3.2. Challenges and Methods

CT MMFMs face several critical challenges that hinder their widespread adoption and effectiveness. These include issues such as hallucination in generated content, limitations in processing 3D data efficiently, and the inability to perform tasks like image generation and restoration.

Hallucination in generated content. CT MMFMs may produce inaccurate or misleading content—this is particularly risky in medical scenarios as it could affect clinical judgments. To address this issue, OrthoDoc [129] employs a graph-based retrieval-augmented generation module, which combines the model's generative capabilities with information retrieval functionality. By leveraging authoritative medical knowledge to constrain the text generation process, this module effectively reduces the risk of generating unreliable information.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 June 2026 doi:10.20944/preprints202606.1256.v1

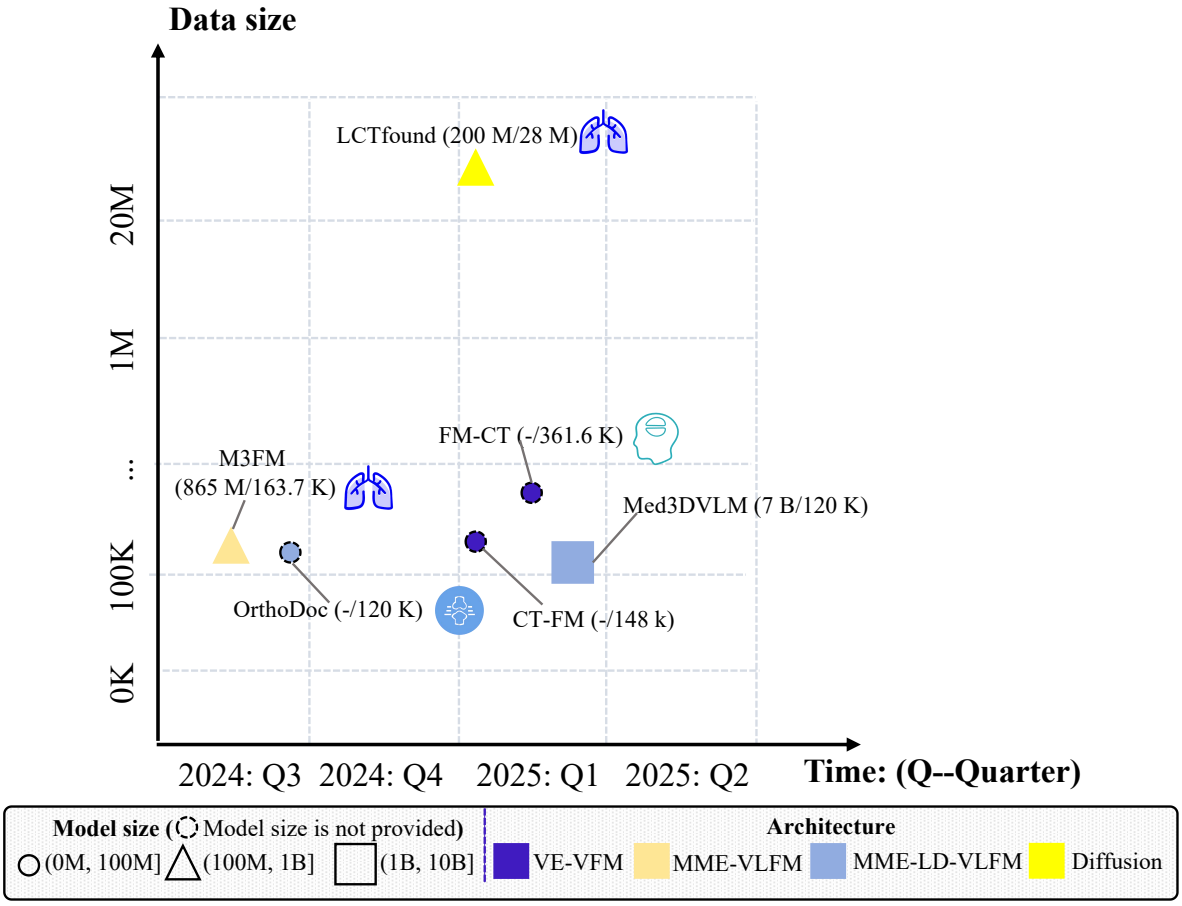

Figure 6. Architectural configurations, model scale, training data size, and developmental timeline of CT MMFMs. CT MMFMs include: M3FM [128], OrthoDoc [129], FM-CT [130], LCTfound [131], CT-FM [132], Med3DVLM [133].

Limitations in image generation and restoration. A further limitation is that current CT MMFMs lack the capability to perform image generation and restoration. To resolve this, LCTfound [131] adopts a diffusion architecture that learns the complex distribution of lung CT images through a process of gradual noise addition and denoising. This enables the model to effectively capture multi-level features, from pixel-level details to semantic representations, facilitating high-quality image generation and restoration tasks.

4.4. Ultrasound Multimodal Foundation Models

Ultrasound images are imaging results generated by transmitting ultrasonic waves into human tissues and receiving the reflected echo signals. These images have unique advantages including real-time imaging capabilities, no ionizing radiation, and high resolution for soft tissue visualization, allowing them to clearly document the morphological structure of target organs, dynamic physiological movements, and pathological changes. This makes ultrasound a core imaging modality in clinical fields such as obstetrics, cardiology, breast imaging, and abdominal diagnosis. In this section, we will systematically introduce the current development status of ultrasound MMFMs, focusing on their architectural designs, training data characteristics, and the critical technical challenges encountered in practical applications, while also summarizing the cutting-edge solutions proposed in recent research.

4.4.1. Overview of Ultrasound MMFMs

As shown in Figure 7, this study includes a total of nine ultrasound MMFMs, covering various architecture types: UltraDINO [139] based on the VE-VFM architecture; URFM [141], USFM [137], and UltraFedFM [135] based on the VED-VFM architecture; UltraSam [136] based on the segment anything model (SAM) [142]; FetalCLIP [138] based on the MME-VLFM architecture; and ThyGPT [140] and

 doi:10.20944/preprints202606.1256.v1

LLaVA-Ultra [134] based on the MME-LD-VLFM architecture. Overall, current research on large ultrasound models primarily focuses on image analysis tasks, while there remains a notable deficiency in image-text reasoning capabilities. It is worth noting that ThyGPT [140] is a multimodal foundation model specifically designed for thyroid diseases, FetalCLIP [138] and UltraDINO [139] specialize in fetal ultrasound analysis, and the other models aim to cover a broader range of ultrasound-related diseases. In terms of model parameter scale, LLaVA-Ultra [134] leads with 13B parameters, whereas the parameter sizes of the other models are not explicitly disclosed. Regarding pre-training data, the scale of ultrasound image data is relatively large—for instance, both UltraDINO [139] and USFM [137] utilize 2B ultrasound images. However, high-quality image-text paired data remain limited, as exemplified by LLaVA-Ultra [134], which uses only 118K image-text pairs.

Figure 7. Architectural configurations, model scale, training data size, and developmental timeline of ultrasound MMFMs. Ultrasound MMFMs include: LLaVA-Ultra [134], UltraFedFM [135], UltraSam [136], USFM [137], FetalCLIP [138], UltraDINO [139], ThyGPT [140], URFM [141].

4.4.2. Challenges and Methods

Ultrasound MMFMs face several critical challenges that impede their clinical application. These include difficulties in capturing fine-grained semantic features from inherently noisy ultrasound images, the "black-box" and "mute-box" nature of traditional AI models which limits interpretability and hinders effective collaboration with radiologists, and a general lack of interactive capabilities for joint diagnosis.

To address these issues, recent research has developed specialized solutions. For enhanced fine-grained feature extraction, LLaVA-Ultra [134] employs a dual-encoder architecture combining SAM and CLIP to fuse local and global features, while URFM [141] leverages self-supervised semantic target reconstruction guided by BiomedCLIP [33] to overcome low signal-to-noise ratios. UltraSam [136] tackles data scarcity by constructing a large-scale segmentation dataset and utilizing the fine-tuned SAM as a backbone for downstream tasks. To improve clinical interpretability, ThyGPT [140] integrates LLMs with computer vision to enable natural language interaction and introduces an AIGC-CAD

 doi:10.20944/preprints202606.1256.v1

framework, thereby enhancing diagnostic transparency, interactivity, and enabling collaborative error detection and correction in ultrasound reports.

5. Organ-Specific Multimodal Foundation Models

While modality-specific models address single-data-type adaptation, clinical diagnosis and treatment often center on specific organ systems. OS-MMFMs emerge as a key branch of MMFMs, integrating multimodal data centered on target organs (e.g., eye, brain, heart, skin) to excel in organ-focused diagnostic and prognostic tasks. This section focuses on OS-MMFMs for critical organs, including the eye, brain, heart, and skin, summarizing their technical routes.

5.1. Eye Multimodal Foundation Models

Medical images of the eye, including fundus images and optical coherence tomography (OCT) images, are core imaging modalities in ophthalmology. Fundus images are captured by specialized fundus cameras to clearly visualize key intraocular structures such as the retina, optic disc, macula, and retinal blood vessels, while OCT images are generated via optical interference technology to obtain high-resolution cross-sectional views of ocular tissues (e.g., retinal layers, choroid). In this section, we will systematically introduce the current development status of eye MMFMs.

5.1.1. Overview of eye MMFMs

As shown in Figure 8, a total of 15 eye MMFMs were documented between January 2023 and July 2025. The distribution of model architectures—VE-VFM, VED-VLFM, MME-VLFM, and MME-LD-VLFM—follows a ratio of 2:7:5:1. This distribution reveals that current research on eye MMFMs is predominantly focused on image analysis tasks. In the domain of text-image analysis, most approaches still rely on the MME-VLFM architecture, i.e., the CLIP framework. As a result, their capabilities remain largely confined to image-text alignment analysis and have not yet fully expanded to deeper image-text reasoning tasks. In terms of pre-training data, the vision foundation model VisionFM [144] utilized the largest image dataset, comprising 3.4M images, while the vision-language model VisionUnite [158] employed the most extensive set of image-text pairs, totaling 1.24M. Overall, the development of eye MMFMs exhibits a distinct characteristic of "focusing on perception over cognition, and on images over text-image integration."

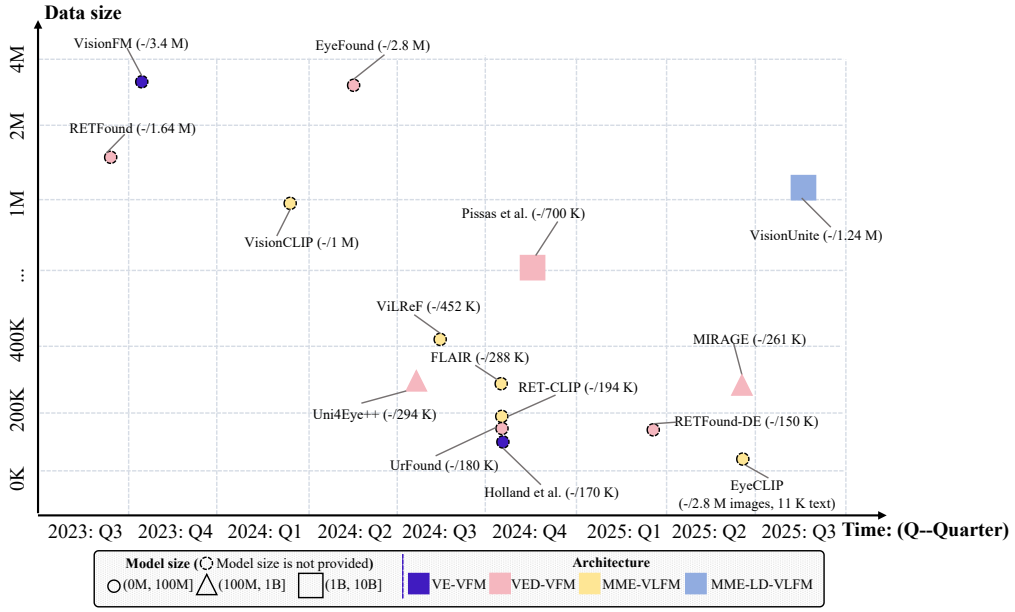

Figure 8. Architectural configurations, model scale, training data size, and developmental timeline of eye MMFMs. Eye MMFMs include: RETFound [143], VisionFM [144], UrFound [145], EyeCLIP [146], EyeFound [147], Pissas et al. [148], Holland et al. [149], FLAIR [150], RET-CLIP [151], ViLReF [152], VisionCLIP [153], VisionUnite [154], Uni4Eye++ [155], MIRAGE [156], RETFound-DE [157].

 doi:10.20944/preprints202606.1256.v1

5.1.2. Challenges and Methods

In the construction and application of eye MMFMs, multiple challenges exist. In terms of data, there is the problem of data scarcity, accompanied by expensive and time-consuming medical data collection as well as high privacy risks. For feature representation learning, existing eye MMFMs fail to fully utilize domain knowledge from expert annotations for effective domain-specific feature representation learning; traditional contrastive learning methods incorrectly regard inherently similar medical images as negative pairs, and false negative samples further interfere with model learning. Besides, fundus datasets lack sufficient text supervision, resulting in a shortage of text descriptions for ophthalmic data. In user interaction, eye MMFMs are deficient in effective interaction capabilities. Additionally, existing eye MMFMs encounter the issue of catastrophic forgetting during data utilization.

To address these challenges, researchers have proposed various methods that collectively aim to improve model performance and applicability. RETFound-DE [157] tackles data scarcity by employing a data-efficient strategy that uses controllable generative AI to generate large-scale synthetic data for pretraining, followed by self-supervised pretraining with a small amount of real data. For feature representation learning, UrFound [145] introduces a knowledge-guided masked modeling pre-training strategy that reconstructs masked image patches and predicts masked text tokens to learn generalizable yet domain-specific features, while Holland et al. [149] redefined inter-image relationships in contrastive learning using patient metadata such as identity, eye laterality, and time-series information to eliminate false negative pairs. RET-CLIP [151] enhances alignment of image and text features through joint contrastive learning at monocular and patient levels, leveraging associations between left and right eye information to improve diagnostic task performance and generalization, and ViLReF [152] designs a weighted similarity coupling loss to dynamically adjust the separation of sample pairs in feature space, boosting zero-shot and transfer learning capabilities. To overcome the scarcity of text descriptions, FLAIR [150] constructs a mapping function from clinical ophthalmology literature to convert categorical labels into descriptive text and combines this with image-text-label contrastive learning pre-training for better generalization without full fine-tuning. Privacy issues are mitigated by VisionCLIP [153], which trains on the synthetic SynFundus-1M dataset to avoid patient data leakage while maintaining performance comparable to models using real data. For improving user interaction, VisionUnite [154] is fine-tuned on the MMFundus dataset, which includes image-text pairs and simulated doctor-patient dialogues, to enable multi-turn dialogue functionality that deepens image understanding and user engagement. Lastly, Uni4Eye++ [155] alleviates catastrophic forgetting through a dynamic head generator module that incorporates image modality information to efficiently utilize multi-dimensional and multi-modal ophthalmic data without losing prior knowledge.

5.2. Other Multimodal Foundation Models

Table 1 further catalogues MMFMs for other organs, which extend beyond those covered in preceding sections. The following discussion delves into the principal challenges and prevailing research approaches within these organs.

Table 1. Summary of other MMFMs.

Method	Organ	Time	Param	Data format	Data quantity	Architecture
CBraMod [159]	Brain	2025.04	4.0 M	EEG	1M samples	MAE
LaBraM [160]	Brain	2024.05	369 M	EEG	2,500 hours	MAE
NeuroLM [161]	Brain	2025.04	1.7 B	EEG-text	25,000 hours	VAE
EEGFormer [162]	Brain	2024.02	-	EEG	1.7T	Encoder-decoder
EchoCLIP [163]	Heart	2024.04	-	US video-text	1M	MME-VLFM
EchoFM [164]	Heart	2025.01	-	US video	286K	VED-VFM
EchoPrime [165]	Heart	2024.01	-	US video-text	12M	MME-VLFM
PanDerm [166]	Skin	2024.01	-	Skin image	2M	VED-VFM
MONET [167]	Skin	2024.01	-	Skin image-text	105K	MME-VLFM

 doi:10.20944/preprints202606.1256.v1

5.2.1. Challenges and Methods

Brain MMFMs. In this survey, we focus on EEG-based brain MMFMs due to their prominent role in recent research on temporal and functional neural representations, while other neuroimaging modalities (e.g., MRI, fMRI) remain less integrated into unified multimodal foundation model frameworks. Existing brain EEG foundation models face challenges: full EEG modeling ignores spatio-temporal dependency heterogeneity of EEG signals, and diverse EEG formats limit downstream task generalization; EEG signals (with complex cognitive/non-cognitive info) are hard to describe via language, causing difficulty in EEG-text alignment; traditional end-to-end models have black-box issues, leading to poor interpretability.

To address these challenges, several methods are proposed. LaBraM [160] first designed a MAE-based EEG basic model, using 2500 hours of EEG data to achieve SOTA in downstream tasks. For spatio-temporal heterogeneity and format-related generalization issues, CBraMod [159] uses a criss-cross transformer (separate spatio-temporal attention) and asymmetric conditional positional encoding (adapts to diverse formats). To solve EEG-text alignment, NeuroLM [161] adds a gradient reverse layer to align EEG and text embeddings in the same feature space. For the black-box problem, EEGFormer [162] uses a vector quantized transformer (encodes EEG into discrete indices) and a learned codebook to provide interpretable results, breaking the traditional black-box paradigm.

Heart MMFMs. Existing echocardiography AI models are single-view, single-task systems that cannot integrate complementary information from multiple views in a comprehensive exam. EchoPrime [165] proposes a multi-view, view-informed, video-based vision-language foundation model, which integrates view classification, an anatomical attention module, and retrieval-augmented interpretation to incorporate all video information from echocardiography for comprehensive clinical interpretation.

Skin MMFMs. Current skin MMFMs are task-specific and cannot handle multimodal clinical workflows. PanDerm [166] is pretrained on a large multimodal dermatology dataset using self-supervised learning to enable general-purpose performance across diverse tasks.

6. Future Directions

In this chapter, we outline key paths to advance MMFMs across four core areas: data-level efforts, architecture-level designs, user demand-focused enhancements, and developer-oriented technical strategies. These directions aim to overcome current limitations and drive MMFMs toward clinical scalability.

6.1. Data Level

Multi-source data expansion and synthetic data generation. Data scarcity and privacy constraints bottleneck model development. Future efforts must expand data scale and diversity by integrating multi-source resources like cross-institutional public datasets, medical literature image-text pairs, and educational videos, utilizing federated learning for collaborative multi-center training without privacy leaks. Concurrently, synthetic data technologies—including text-conditioned diffusion models and controllable generative AI—should be leveraged to generate anatomically plausible 3D medical images and diverse lesions for background fusion. Automated pipelines are also needed to extract paired data from clinical documents and pathological sections, filling data gaps in modalities like ultrasound and CT.

Data quality optimization and noise suppression. Visual noise, semantic inconsistencies, and temporal misalignments necessitate a comprehensive quality control system. Pixel-level alignment, denoising, and domain knowledge-guided feature enhancement can improve lesion recognition. For textual data, clinical dictionary calibration and knowledge-enhanced indicator screening can rectify terminology inconsistencies and ambiguities. To resolve temporal misalignments, multi-image encoders should integrate longitudinal clinical data via timestamp calibration and diagnostic history

 doi:10.20944/preprints202606.1256.v1

correlation. Additionally, contrastive learning should employ semantics-aware classification to refine positive/negative sample definitions and prevent misjudging highly similar unpaired data.

Collaborative adaptation between data formats and model architectures. The disconnect between separate image-text encoding structures and data formats requires driving tighter integration. Building a unified encoding framework that transforms multimodal data into uniform sequence tokens ensures alignment with unified training objectives. Tailored preprocessing methods must match specific modal traits, such as ultra-long context modeling for high-resolution pathological WSIs and spatio-temporally separated attention modules for complex EEG data. Furthermore, a dynamic mapping system should actively tune encoding steps based on data types to maximize data utilization.

Construction of multi-turn conversational data. To shift from single-turn image descriptions to in-depth clinical interactions, a multi-turn conversational data system is required. Multi-turn interactive data encompassing condition inquiry, examination suggestions, diagnostic interpretation, and treatment discussions should be collected from simulated doctor-patient dialogues and annotated for logic and correlation. Integrating longitudinal electronic health records can yield time-series dialogue datasets covering dynamic disease progression. Large models can also assist in expanding single-turn image-text pairs into multi-turn Q&A content, ensuring medical logic and coherence aligned with clinical workflows.

6.2. Architecture Level

Design of unified and collaborative image-text architecture. To resolve split training objectives and poor cross-modal matching in separate architectures, focus should shift toward MM-VLFM and MMED-VLFM frameworks. Using a single transformer to process mixed image-text sequences enables bidirectional cross-modal interaction. Integrating cross-modal attention blocks supports immediate feedback and fine-grained alignment between local visual features and specific medical terms. Additionally, visual reconstruction, language modeling, and modal matching tasks should be unified via a multi-task joint loss function to ensure structural and objectives cohesion.

Modality-adaptive token processing mechanism. A unified, modality-adaptive token processing system is needed to bridge modal differences. Research should explore discrete visual tokenizers to convert medical images into discrete units homologous to text, mapping lesion regions directly to medical terms. A cross-modal token alignment module can match continuous image tokens with discrete text tokens in the semantic space. Sharing a unified loss function for both image token generation and text autoregressive prediction will harmonize conflicting generation logics, while compressing image token dimensions can optimize computational efficiency and balance alignment training costs.

6.3. User Demand Aspect

Enhanced interpretability and clinical transparency. Clinical safety demands a multi-level explanation system. Embedding visualization modules can display key analytical regions via heatmaps and lesion annotations, directly linking conclusions to visual evidence. Chain-of-Thought mechanisms can output diagnostic logic step-by-step, clarifying the derivation from symptoms to conclusions. Decisions should also be mapped to standard medical terminology databases like UMLS to explain underlying medical principles. Finally, an interactive interpretation interface will allow clinicians to trace decision paths and adjust parameters to enhance clinical trust.

Ethical compliance and humanistic care orientation. Standardizing accuracy, fairness, and humanistic care requires whole-process compliance. Data-wise, differential privacy and synthetic generation can minimize reliance on real data; training-wise, clinical expert supervision and validation mechanisms can mitigate hallucination risks. Models must be optimized for fairness across diverse demographic populations to eliminate data bias. Output-wise, models should use empathetic yet professional language to avoid patient panic, offer multi-lingual adaptive options, and establish an ethical review feedback loop for continuous optimization.

 doi:10.20944/preprints202606.1256.v1

Interactive clinical and research support functions. Tailored interactive functions are essential for clinical operations and research. Clinically, real-time image manipulation should support dynamic user annotations and parameter adjustments with synchronous model updates, while multimodal interaction should enable comprehensive cross-modal Q&A across images and text reports. For research, the system should offer batch data processing for WSI feature extraction and statistics, alongside research assistance modules for experimental design and hypothesis generation, all delivered through a lightweight, accessible interface.

6.4. Technical Considerations from a Developer's Perspective

Continuous learning and new data adaptation capabilities. Handling continuous streams of new data, diseases, and modalities requires an efficient continuous learning framework. Incremental training algorithms incorporating rehearsal-based strategies and dynamic heads can prevent catastrophic forgetting. Meta-learning can facilitate rapid adaptation to new modalities like CT and EEG without training from scratch. Additionally, data distribution monitoring modules should detect real-time input shifts to adjust parameters for out-of-distribution (OOD) scenarios, supported by an open modal interface for flexible, unified multi-source data access.

Multimodal collaboration and modal conflict resolution solutions. Optimization solutions are required for two distinct technical paths. Under a unified model architecture, a Mixture-of-Experts (MoE) module with a gating mechanism can assign dedicated expert networks to different modalities, alongside adaptive weight adjustment strategies to alleviate modal conflicts. Under a multi-model cooperative architecture, a collaborative scheduling center can leverage knowledge distillation for complementary expertise, utilizing cross-model communication for real-time feedback between image analysis and text generation. Conflict detection modules should automatically identify and reconcile multi-source inconsistencies.

Precision of cross-modal alignment. Maximizing cross-modal accuracy necessitates multi-scale alignment optimizations. Models must achieve global thematic alignment between whole images and diagnostic reports, alongside local fine-grained matching between lesion regions and medical terms. Incorporating knowledge-enhanced alignment losses can embed anatomical and pathological structures to prevent semantic misalignment. For large-scale images like WSIs, a "patch-sentence" level alignment strategy should be deployed to manage many-to-one associations, monitored by an alignment quality evaluation module to correct deviations.

Medical-grade cross-modal reasoning capabilities. To support complex clinical scenarios, cross-modal reasoning must be strengthened. Integrating medical knowledge graph embeddings can supply essential domain knowledge regarding diseases, symptoms, and drugs. Temporal reasoning mechanisms should ingest longitudinal data to predict disease progression and evaluate treatment efficacy. For critical care, a multimodal joint reasoning framework must synthesize images, text, and physiological signals for real-time critical value judgments and intervention suggestions, refined through case fine-tuning and expert feedback to match clinical standards.

7. Conclusions

This survey provides a systematic and fine-grained review of medical MMFMs spanning January 2023 to July 2025, aiming to address the gaps in existing literature—such as rapid obsolescence and superficial overviews. We first established a hierarchical analytical framework to characterize core technical attributes (parameter scale, training data volume, architectures) of MMFMs, delineating their evolutionary trajectories across three primary tiers: Uni-MMFMs with broad applicability and MS/OS-MMFMs tailored to fields like X-ray, pathology, and brain.

Overall, MMFMs have evolved into a core pillar of medical AI, but their transition to large-scale clinical use requires synergistic advancements in data engineering, architectural innovation, and clinical adaptation. This survey not only consolidates the current state of MMFMs but also provides a roadmap for researchers and clinicians to prioritize efforts, ultimately propelling the development of more reliable, interpretable, and clinically impactful multimodal medical AI systems.

 doi:10.20944/preprints202606.1256.v1

References

1. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
2. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. https://doi.org/https://doi.org/10.1038/s41586-023-05881-4.
3. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* **2020**.
4. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553.
5. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
6. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
7. Zhang, S.; Metaxas, D. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis* **2024**, *91*, 102996. https://doi.org/https://doi.org/10.1016/j.media.2023.102996.
8. Shrestha, P.; Amgain, S.; Khanal, B.; Linte, C.A.; Bhattarai, B. Medical vision language pretraining: A survey. *arXiv preprint arXiv:2312.06224* **2023**.
9. Khan, W.; Leem, S.; See, K.B.; Wong, J.K.; Zhang, S.; Fang, R. A Comprehensive Survey of Foundation Models in Medicine. *IEEE Reviews in Biomedical Engineering* **2025**, pp. 1–22. https://doi.org/10.1109/RBME.2025.3531360.
10. He, Y.; Huang, F.; Jiang, X.; Nie, Y.; Wang, M.; Wang, J.; Chen, H. Foundation Model for Advancing Healthcare: Challenges, Opportunities and Future Directions. *IEEE Reviews in Biomedical Engineering* **2025**, *18*, 172–191. https://doi.org/10.1109/RBME.2024.3496744.
11. Liu, C.; Jin, Y.; Guan, Z.; Li, T.; Qin, Y.; Qian, B.; Jiang, Z.; Wu, Y.; Wang, X.; Zheng, Y.F.; et al. Visual–language foundation models in medicine. *The Visual Computer* **2024**, pp. 1–20.
12. Sun, K.; Xue, S.; Sun, F.; Sun, H.; Luo, Y.; Wang, L.; Wang, S.; Guo, N.; Liu, L.; Zhao, T.; et al. Medical Multimodal Foundation Models in Clinical Diagnosis and Treatment: Applications, Challenges, and Future Directions. *arXiv preprint arXiv:2412.02621* **2024**.
13. Azad, B.; Azad, R.; Eskandari, S.; Bozorgpour, A.; Kazerouni, A.; Rekik, I.; Merhof, D. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689* **2023**.
14. Huang, S.C.; Jensen, M.; Yeung-Levy, S.; Lungren, M.P.; Poon, H.; Chaudhari, A.S. Multimodal Foundation Models for Medical Imaging-A Systematic Review and Implementation Guidelines. *medRxiv* **2024**, pp. 2024–10.
15. AlSaad, R.; Abd-Alrazaq, A.; Boughorbel, S.; Ahmed, A.; Renault, M.A.; Damseh, R.; Sheikh, J. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research* **2024**, *26*, e59505. https://doi.org/doi:10.2196/59505.
16. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundation Models Defining a New Era in Vision: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**, *47*, 2245–2264. https://doi.org/10.1109/TPAMI.2024.3506283.
17. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* **2022**, *35*, 22199–22213.
18. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* **2023**, *5*, 220–235. https://doi.org/https://doi.org/10.1038/s42256-023-00626-4.
19. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **2013**, *45*, 1113–1120.
20. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* **2019**, *6*, 317. https://doi.org/https://doi.org/10.1038/s41597-019-0322-0.
21. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert

 doi:10.20944/preprints202606.1256.v1

comparison. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 590–597. https://doi.org/https://doi.org/10.1609/aaai.v33i01.3301590.

22. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
25. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975.
26. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
27. Bansal, H.; Israel, D.; Zhao, S.; Li, S.; Nguyen, T.; Grover, A. MedMax: Mixed-Modal Instruction Tuning for Training Biomedical Assistants. *arXiv preprint arXiv:2412.12661* **2024**.
28. Liu, J.; Wang, Z.; Ye, Q.; Chong, D.; Zhou, P.; Hua, Y. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956* **2023**.
29. MH Nguyen, D.; Nguyen, H.; Diep, N.; Pham, T.N.; Cao, T.; Nguyen, B.; Swoboda, P.; Ho, N.; Albarqouni, S.; Xie, P.; et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems* **2023**, *36*, 27922–27950.
30. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463* **2023**.
31. 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* **2023**, *36*, 28541–28564.
32. Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E.P.; Rajpurkar, P. Med-flamingo: a multimodal medical few-shot learner. In Proceedings of the Machine Learning for Health (ML4H). PMLR, 2023, pp. 353–367.
33. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* **2023**.
34. Lin, W.; Zhao, Z.; Zhang, X.; Wu, C.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 525–536. https://doi.org/https://doi.org/10.1007/978-3-031-43993-3_51.
35. Liu, Z.; Tieu, A.; Patel, N.; Soultanidis, G.; Deyer, L.; Wang, Y.; Huver, S.; Zhou, A.; Mei, Y.; Fayad, Z.A.; et al. VIS-MAE: An Efficient Self-supervised Learning Approach on Medical Image Segmentation and Classification. In Proceedings of the International Workshop on Machine Learning in Medical Imaging. Springer, 2024, pp. 95–107.
36. Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B.D.; Ren, H.; et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* **2024**, pp. 1–13. https://doi.org/https://doi.org/10.1038/s41591-024-03185-2.
37. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards generalist biomedical AI. *Nejm Ai* **2024**, *1*, AIoa2300138. https://doi.org/10.1056/AIoa2300138.
38. Liu, X.; Yang, G.; Luo, Y.; Mao, J.; Zhang, X.; Gao, M.; Zhang, S.; Shen, J.; Wang, G. Expert-level vision-language foundation model for real-world radiology and comprehensive evaluation. *arXiv preprint arXiv:2409.16183* **2024**.
39. Zhao, T.; Gu, Y.; Yang, J.; Usuyama, N.; Lee, H.H.; Kiblawi, S.; Naumann, T.; Gao, J.; Crabtree, A.; Abel, J.; et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods* **2025**, *22*, 166–176. https://doi.org/https://doi.org/10.1038/s41592-024-02499-w.

 doi:10.20944/preprints202606.1256.v1

40. Xu, L.; Sun, H.; Ni, Z.; Li, H.; Zhang, S. MedViLaM: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation. *arXiv preprint arXiv:2409.19684* **2024**.
41. Li, T.; Su, Y.; Li, W.; Fu, B.; Chen, Z.; Huang, Z.; Wang, G.; Ma, C.; Chen, Y.; Hu, M.; et al. GMAI-VL & GMAI-VL-5.5 M: A Large Vision-Language Model and A Comprehensive Multimodal Dataset Towards General Medical AI. *arXiv preprint arXiv:2411.14522* **2024**.
42. Chen, Z.; Pekis, A.; Brown, K. Advancing High Resolution Vision-Language Models in Biomedicine. *arXiv preprint arXiv:2406.09454* **2024**.
43. Chu, Y.; Zhang, Y.; Han, Z.; Yang, C.; Zhou, L.; Luo, G.; Gao, X. Improving Representation of High-frequency Components for Medical Foundation Models. *arXiv preprint arXiv:2407.14651* **2024**.
44. He, S.; Nie, Y.; Chen, Z.; Cai, Z.; Wang, H.; Yang, S.; Chen, H. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints* **2024**, pp. arXiv–2404.
45. Yang, L.; Xu, S.; Sellergren, A.; Kohlberger, T.; Zhou, Y.; Ktena, I.; Kiraly, A.; Ahmed, F.; Hormozdiari, F.; Jaroensri, T.; et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162* **2024**.
46. Cui, H.; Mao, L.; Liang, X.; Zhang, J.; Ren, H.; Li, Q.; Li, X.; Yang, C. Biomedical visual instruction tuning with clinician preference alignment. *Advances in neural information processing systems* **2024**, *37*, 96449–96467.
47. Luo, L.; Chen, X.; Tang, B.; Chen, X.; Han, R.; Hu, C.; Li, Y.; Chen, T. Building Universal Foundation Models for Medical Image Analysis with Spatially Adaptive Networks. *arXiv preprint arXiv:2312.07630* **2023**.
48. Ye, Y.; Xie, Y.; Zhang, J.; Chen, Z.; Wu, Q.; Xia, Y. Continual Self-Supervised Learning: Towards Universal Multi-Modal Medical Data Representation Learning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11114–11124. https://doi.org/10.1109/CVPR52733.2024.01057.
49. Nath, V.; Li, W.; Yang, D.; Myronenko, A.; Zheng, M.; Lu, Y.; Liu, Z.; Yin, H.; Law, Y.M.; Tang, Y.; et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14788–14798.
50. Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.H.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. Towards injecting medical visual knowledge into multimodal llms at scale. In Proceedings of the Proceedings of the 2024 conference on empirical methods in natural language processing, 2024, pp. 7346–7370.
51. Bawazir, A.; Wu, K.; Li, W. Uni-Mlip: Unified self-supervision for medical vision language pre-training. *arXiv preprint arXiv:2411.15207* **2024**.
52. Liu, F.; Zhou, H.; Wang, K.; Yu, Y.; Gao, Y.; Sun, Z.; Liu, S.; Sun, S.; Zou, Z.; Li, Z.; et al. MetaGP: A generative foundation model integrating electronic health records and multimodal imaging for addressing unmet clinical needs. *Cell Reports Medicine* **2025**, *6*.
53. Lin, T.; Zhang, W.; Li, S.; Yuan, Y.; Yu, B.; Li, H.; He, W.; Jiang, H.; Li, M.; Song, X.; et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838* **2025**.
54. Xu, W.; Chan, H.P.; Li, L.; Aljunied, M.; Yuan, R.; Wang, J.; Xiao, C.; Chen, G.; Liu, C.; Li, Z.; et al. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *arXiv preprint arXiv:2506.07044* **2025**.
55. Wu, L.; Nie, Y.; He, S.; Zhuang, J.; Luo, L.; Mahboobani, N.; Vardhanabhuti, V.; Chan, R.C.K.; Peng, Y.; Rajpurkar, P.; et al. UniBiomed: A Universal Foundation Model for Grounded Biomedical Image Interpretation. *arXiv preprint arXiv:2504.21336* **2025**.
56. Chen, K.; Li, Y.; Zhu, X.; Zhang, W.; Hu, B. A vision-language model with multi-granular knowledge fusion in medical imaging. *World Wide Web* **2025**, *28*, 5.
57. Nie, Y.; He, S.; Bie, Y.; Wang, Y.; Chen, Z.; Yang, S.; Chen, H. ConceptCLIP: Towards Trustworthy Medical AI via Concept-Enhanced Contrastive Langauge-Image Pre-training. *arXiv preprint arXiv:2501.15579* **2025**.
58. Liu, J.; Zhou, H.Y.; Huang, W.; Yang, H.; Song, D.; Tan, T.; Liang, Y.; Wang, S. BioVFM-21M: Benchmarking and Scaling Self-supervised Vision Foundation Models for Biomedical Image Analysis. In Proceedings of the International Workshop on Foundation Models for General Medical AI. Springer, 2025, pp. 23–33.
59. Dai, W.; Chen, P.; Ekbote, C.; Liang, P.P. QoQ-Med: Building Multimodal Clinical Foundation Models with Domain-Aware GRPO Training. *arXiv preprint arXiv:2506.00711* **2025**.
60. Wang, R.; Yao, Q.; Jiang, Z.; Lai, H.; He, Z.; Tao, X.; Zhou, S.K. ECAMP: entity-centered context-aware medical vision language pre-training. *Medical Image Analysis* **2025**, p. 103690.

 doi:10.20944/preprints202606.1256.v1

61. Lozano, A.; Sun, M.W.; Burgess, J.; Chen, L.; Nirschl, J.J.; Gu, J.; Lopez, I.; Aklilu, J.; Rau, A.; Katzer, A.W.; et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 19724–19735.
62. Lu, Z.; Li, H.; Parikh, N.A.; Dillman, J.R.; He, L. RadCLIP: Enhancing Radiologic Image Analysis Through Contrastive Language–Image Pretraining. *IEEE Transactions on Neural Networks and Learning Systems* **2025**.
63. Huang, X.; Shen, L.; Liu, J.; Shang, F.; Li, H.; Huang, H.; Yang, Y. Towards a multimodal large language model with pixel-level insight for biomedicine. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 3779–3787. https://doi.org/https://doi.org/10.1609/aaai.v39i4.32394.
64. Yu, H.; Yi, S.; Niu, K.; Zhuo, M.; Li, B. UMIT: Unifying Medical Imaging Tasks via Vision-Language Models. *arXiv preprint arXiv:2503.15892* **2025**.
65. Liu, F.; Zhu, T.; Wu, X.; Yang, B.; You, C.; Wang, C.; Lu, L.; Liu, Z.; Zheng, Y.; Sun, X.; et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine* **2023**, *6*, 226. https://doi.org/https://doi.org/10.1038/s41746-023-00952-2.
66. Huang, X.; Huang, H.; Shen, L.; Yang, Y.; Shang, F.; Liu, J.; Liu, J. A refer-and-ground multimodal large language model for biomedicine. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 399–409. https://doi.org/https://doi.org/10.1007/978-3-031-72390-2_38.
67. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. Improving language understanding by generative pre-training **2018**.
68. Wang, J.; Wang, K.; Yu, Y.; Lu, Y.; Xiao, W.; Sun, Z.; Liu, F.; Zou, Z.; Gao, Y.; Yang, L.; et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine* **2024**, pp. 1–9. https://doi.org/https://doi.org/10.1038/s41591-024-03359-y.
69. Lu, S.; Liu, Z.; Liu, T.; Zhou, W. Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence* **2023**, *126*, 107037. https://doi.org/https://doi.org/10.1016/j.engappai.2023.107037.
70. Chen, Z.; Diao, S.; Wang, B.; Li, G.; Wan, X. Towards unifying medical vision-and-language pre-training via soft prompts. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 23403–23413.
71. Wu, L.; Zhuang, J.; Chen, H. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22873–22882.
72. Wang, L.; Wang, H.; Yang, H.; Mao, J.; Yang, Z.; Shen, J.; Li, X. Interpretable bilingual multimodal large language model for diverse biomedical tasks. *arXiv preprint arXiv:2410.18387* **2024**.
73. Zhang, K.; Yang, Y.; Yu, J.; Jiang, H.; Fan, J.; Huang, Q.; Han, W. Multi-Task Paired Masking With Alignment Modeling for Medical Vision-Language Pre-Training. *IEEE Transactions on Multimedia* **2024**, *26*, 4706–4721. https://doi.org/10.1109/TMM.2023.3325965.
74. Zhu, X.; Hu, Y.; Mo, F.; Li, M.; Wu, J. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-MoE. *arXiv preprint arXiv:2409.17508* **2024**.
75. Aeffner, F.; Zarella, M.D.; Buchbinder, N.; Bui, M.M.; Goodman, M.R.; Hartman, D.J.; Lujan, G.M.; Molani, M.A.; Parwani, A.V.; Lillard, K.; et al. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. *Journal of pathology informatics* **2019**, *10*, 9. https://doi.org/10.4103/jpi.jpi_82_18.
76. Lu, M.Y.; Chen, B.; Williamson, D.F.; Chen, R.J.; Ikamura, K.; Gerber, G.; Liang, I.; Le, L.P.; Ding, T.; Parwani, A.V.; et al. A foundational multimodal vision language AI assistant for human pathology. *arXiv preprint arXiv:2312.07814* **2023**.
77. Chen, R.J.; Ding, T.; Lu, M.Y.; Williamson, D.F.; Jaume, G.; Chen, B.; Zhang, A.; Shao, D.; Song, A.H.; Shaban, M.; et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474* **2023**.
78. Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T.; Zou, J. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv* **2023**, pp. 2023–03.
79. Ikezogwo, W.; Seyfioglu, S.; Ghezloo, F.; Geva, D.; Sheikh Mohammed, F.; Anand, P.K.; Krishna, R.; Shapiro, L. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems* **2023**, *36*, 37995–38017.

 doi:10.20944/preprints202606.1256.v1

80. Vorontsov, E.; Bozkurt, A.; Casson, A.; Shaikovski, G.; Zelechowski, M.; Severson, K.; Zimmermann, E.; Hall, J.; Tenenholtz, N.; Fusi, N.; et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine* **2024**, *30*, 2924–2935.
81. Lu, M.Y.; Chen, B.; Williamson, D.F.; Chen, R.J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L.P.; Gerber, G.; et al. A visual-language foundation model for computational pathology. *Nature medicine* **2024**, *30*, 863–874. https://doi.org/https://doi.org/10.1038/s41591-024-02856-4.
82. Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **2024**, *630*, 181–188. https://doi.org/https://doi.org/10.1038/s41586-024-07441-w.
83. Ding, T.; Wagner, S.J.; Song, A.H.; Chen, R.J.; Lu, M.Y.; Zhang, A.; Vaidya, A.J.; Jaume, G.; Shaban, M.; Kim, A.; et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine* **2025**, pp. 1–13.
84. Ahmed, F.; Sellergren, A.; Yang, L.; Xu, S.; Babenko, B.; Ward, A.; Olson, N.; Mohtashamian, A.; Matias, Y.; Corrado, G.S.; et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578* **2024**.
85. Sun, Y.; Zhu, C.; Zheng, S.; Zhang, K.; Sun, L.; Shui, Z.; Zhang, Y.; Li, H.; Yang, L. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 5034–5042.
86. Shaikovski, G.; Casson, A.; Severson, K.; Zimmermann, E.; Wang, Y.K.; Kunz, J.D.; Retamero, J.A.; Oakley, G.; Klimstra, D.; Kanan, C.; et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254* **2024**.
87. Dippel, J.; Feulner, B.; Winterhoff, T.; Milbich, T.; Tietz, S.; Schallenberg, S.; Dernbach, G.; Kunft, A.; Heinke, S.; Eich, M.L.; et al. RudolfV: a foundation model by pathologists for pathologists. *arXiv preprint arXiv:2401.04079* **2024**.
88. Ma, J.; Guo, Z.; Zhou, F.; Wang, Y.; Xu, Y.; Li, J.; Yan, F.; Cai, Y.; Zhu, Z.; Jin, C.; et al. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449* **2024**.
89. Wang, X.; Zhao, J.; Marostica, E.; Yuan, W.; Jin, J.; Zhang, J.; Li, R.; Tang, H.; Wang, K.; Li, Y.; et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **2024**, *634*, 970–978. https://doi.org/https://doi.org/10.1038/s41586-024-07894-z.
90. Xu, Y.; Wang, Y.; Zhou, F.; Ma, J.; Jin, C.; Yang, S.; Li, J.; Zhang, Z.; Zhao, C.; Zhou, H.; et al. A multimodal knowledge-enhanced whole-slide pathology foundation model. *arXiv preprint arXiv:2407.15362* **2024**.
91. Yang, Z.; Wei, T.; Liang, Y.; Yuan, X.; Gao, R.; Xia, Y.; Zhou, J.; Zhang, Y.; Yu, Z. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nature Communications* **2025**, *16*, 2366.
92. Chen, K.; Liu, M.; Yan, F.; Ma, L.; Shi, X.; Wang, L.; Wang, X.; Zhu, L.; Wang, Z.; Zhou, M.; et al. Cost-effective instruction learning for pathology vision and language analysis. *Nature Computational Science* **2025**, pp. 1–10.
93. Sun, Y.; Si, Y.; Zhu, C.; Gong, X.; Zhang, K.; Chen, P.; Zhang, Y.; Shui, Z.; Lin, T.; Yang, L. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 10360–10371.
94. Dai, D.; Zhang, Y.; Yang, Q.; Xu, L.; Shen, X.; Xia, S.; Wang, G. Pathologyvlm: a large vision-language model for pathology image understanding. *Artificial Intelligence Review* **2025**, *58*, 1–19.
95. Chen, Y.; Wang, G.; Ji, Y.; Li, Y.; Ye, J.; Li, T.; Hu, M.; Yu, R.; Qiao, Y.; He, J. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 5134–5143.
96. Xiang, J.; Wang, X.; Zhang, X.; Xi, Y.; Eweje, F.; Chen, Y.; Li, Y.; Bergstrom, C.; Gopaulchan, M.; Kim, T.; et al. A vision–language foundation model for precision oncology. *Nature* **2025**, *638*, 769–778.
97. Ding, J.; Ma, S.; Dong, L.; Zhang, X.; Huang, S.; Wang, W.; Zheng, N.; Wei, F. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486* **2023**.
98. Xu, S.; Yang, L.; Kelly, C.; Sieniek, M.; Kohlberger, T.; Ma, M.; Weng, W.H.; Kiraly, A.; Kazemzadeh, S.; Melamed, Z.; et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317* **2023**.
99. Zhang, X.; Wu, C.; Zhang, Y.; Xie, W.; Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications* **2023**, *14*, 4542. https://doi.org/https://doi.org/10.1038/s41467-023-40260-7.

 doi:10.20944/preprints202606.1256.v1

100. Lee, S.; Kim, W.J.; Chang, J.; Ye, J.C. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation. In Proceedings of the The Twelfth International Conference on Learning Representations.
101. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21315–21326. https://doi.org/10.1109/ICCV51070.2023.01954.
102. Pellegrini, C.; Özsoy, E.; Busam, B.; Wiestler, B.; Navab, N.; Keicher, M. Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In Proceedings of the Medical Imaging with Deep Learning, 2025.
103. Bannur, S.; Hyland, S.; Liu, Q.; Pérez-García, F.; Ilse, M.; Castro, D.C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15016–15027. https://doi.org/10.1109/CVPR52729.2023.01442.
104. Liu, B.; Lu, D.; Wei, D.; Wu, X.; Wang, Y.; Zhang, Y.; Zheng, Y. Improving Medical Vision-Language Contrastive Pretraining With Semantics-Aware Triage. *IEEE Transactions on Medical Imaging* **2023**, *42*, 3579–3589. https://doi.org/10.1109/TMI.2023.3294980.
105. Liu, C.; Cheng, S.; Chen, C.; Qiao, M.; Zhang, W.; Shah, A.; Bai, W.; Arcucci, R. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 637–647. https://doi.org/https://doi.org/10.1007/978-3-031-43907-0_61.
106. Chen, X.; He, Y.; Xue, C.; Ge, R.; Li, S.; Yang, G. Knowledge boosting: Rethinking medical contrastive vision-language pre-training. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023, pp. 405–415. https://doi.org/https://doi.org/10.1007/978-3-031-43907-0_39.
107. Yao, J.; Wang, X.; Song, Y.; Zhao, H.; Ma, J.; Chen, Y.; Liu, W.; Wang, B. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. *npj Digital Medicine* **2025**, *8*, 678.
108. Xu, L.; Ni, Z.; Sun, H.; Li, H.; Zhang, S. A foundation model for generalizable disease diagnosis in chest X-ray images. *arXiv preprint arXiv:2410.08861* **2024**.
109. Bluethgen, C.; Chambon, P.; Delbrouck, J.B.; Van Der Sluijs, R.; Połacin, M.; Zambrano Chaves, J.M.; Abraham, T.M.; Purohit, S.; Langlotz, C.P.; Chaudhari, A.S. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering* **2025**, *9*, 494–506. https://doi.org/https://doi.org/10.1038/s41551-024-01246-y.
110. Müller, P.; Kaissis, G.; Rueckert, D. ChEX: Interactive localization and region description in chest X-rays. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 92–111.
111. Chen, Z.; Varma, M.; Xu, J.; Paschali, M.; Van Veen, D.; Johnston, A.; Youssef, A.; Blankemeier, L.; Bluethgen, C.; Altmayer, S.; et al. A Vision-Language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv e-prints* **2024**, pp. arXiv–2401.
112. Huang, W.; Li, C.; Zhou, H.Y.; Yang, H.; Liu, J.; Liang, Y.; Zheng, H.; Zhang, S.; Wang, S. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications* **2024**, *15*, 7620. https://doi.org/https://doi.org/10.1038/s41467-024-51749-0.
113. Kumar, Y.; Marttinen, P. Improving medical multi-modal contrastive learning with expert annotations. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 468–486.
114. Zambrano Chaves, J.M.; Huang, S.C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; et al. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications* **2025**, *16*, 3108.
115. Wang, X.; Luo, J.; Wang, J.; Zhong, Y.; Zhang, X.; Wang, Y.; Bhatia, P.; Xiao, C.; Ma, F. Unity in diversity: Collaborative pre-training across multimodal medical sources. In Proceedings of the Proceedings of the conference. Association for Computational Linguistics. Meeting, 2024, Vol. 2024, p. 3644.
116. Li, Q.; Yan, X.; Xu, J.; Yuan, R.; Zhang, Y.; Feng, R.; Shen, Q.; Zhang, X.; Wang, S. Anatomical structure-guided medical vision-language pre-training. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 80–90. https://doi.org/https://doi.org/10.1007/978-3-031-72120-5_8.
117. Luo, H.; Zhou, Z.; Royer, C.; Sekuboyina, A.; Menze, B. Devide: Faceted medical knowledge for improved medical vision-language pre-training. *arXiv preprint arXiv:2404.03618* **2024**.

 doi:10.20944/preprints202606.1256.v1

118. Liu, B.; Lu, Z.; Wang, Y. Towards medical vision-language contrastive pre-training via study-oriented semantic exploration. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4861–4870.
119. Liang, X.; Hu, J.; Wang, D.; Ma, Z.; Zhao, L.; Li, R.; Wan, B.; Wang, Q. CheXPO: Preference Optimization for Chest X-ray VLMs with Counterfactual Rationale. *arXiv preprint arXiv:2507.06959* **2025**.
120. Lee, S.; Youn, J.; Kim, H.; Kim, M.; Yoon, S.H. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. *European Radiology* **2025**, pp. 1–13.
121. Deperrois, N.; Matsuo, H.; Ruipérez-Campillo, S.; Vandenhirtz, M.; Laguna, S.; Ryser, A.; Fujimoto, K.; Nishio, M.; Sutter, T.M.; Vogt, J.E.; et al. RadVLM: A multitask conversational vision-language model for radiology. *arXiv preprint arXiv:2502.03333* **2025**.
122. Li, M.; Meng, M.; Fulham, M.; Feng, D.D.; Bi, L.; Kim, J. Enhancing Medical Vision-Language Contrastive Learning via Inter-Matching Relation Modeling. *IEEE Transactions on Medical Imaging* **2025**, *44*, 2463–2476. https://doi.org/10.1109/TMI.2025.3534436.
123. Liu, C.; Cheng, S.; Shi, M.; Shah, A.; Bai, W.; Arcucci, R. IMITATE: Clinical Prior Guided Hierarchical Vision-Language Pre-Training. *IEEE Transactions on Medical Imaging* **2025**, *44*, 519–529. https://doi.org/10.1109/TMI.2024.3449690.
124. Liang, X.; Li, X.; Li, F.; Jiang, J.; Dong, Q.; Wang, W.; Wang, K.; Dong, S.; Luo, G.; Li, S. MedFILIP: Medical Fine-Grained Language-Image Pre-Training. *IEEE Journal of Biomedical and Health Informatics* **2025**, *29*, 3587–3597. https://doi.org/10.1109/JBHI.2025.3528196.
125. Chu, Y.W.; Zhang, K.; Malon, C.; Min, M.R. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. *arXiv preprint arXiv:2502.15040* **2025**.
126. Zhang, Z.; Yu, Y.; Chen, Y.; Yang, X.; Yeo, S.Y. Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29744–29755.
127. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106. https://doi.org/10.1109/CVPR.2017.369.
128. Niu, C.; Lyu, Q.; Carothers, C.D.; Kaviani, P.; Tan, J.; Yan, P.; Kalra, M.K.; Whitlow, C.T.; Wang, G. Specialty-oriented generalist medical ai for chest ct screening. *arXiv preprint arXiv:2304.02649* **2023**.
129. Jin, Y.; Zhang, Y. Orthodoc: Multimodal large language model for assisting diagnosis in computed tomography. *arXiv preprint arXiv:2409.09052* **2024**.
130. Zhu, W.; Huang, H.; Tang, H.; Musthyala, R.; Yu, B.; Chen, L.; Vega, E.; O'Donnell, T.; Dehkharghani, S.; Frontera, J.A.; et al. 3D foundation AI model for generalizable disease detection in head computed tomography. *arXiv preprint arXiv:2502.02779* **2025**.
131. Gao, Z.; Zhang, G.; Liang, H.; Liu, J.; Ma, L.; Wang, T.; Guo, Y.; Chen, Y.; Yan, Z.; Chen, X.; et al. A Lung CT Foundation Model Facilitating Disease Diagnosis and Medical Imaging. *medRxiv* **2025**, pp. 2025–01.
132. Pai, S.; Hadzic, I.; Bontempi, D.; Bressem, K.; Kann, B.H.; Fedorov, A.; Mak, R.H.; Aerts, H.J. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001* **2025**.
133. Xin, Y.; Ates, G.C.; Gong, K.; Shao, W. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *arXiv preprint arXiv:2503.20047* **2025**.
134. Guo, X.; Chai, W.; Li, S.Y.; Wang, G. LLaVA-ultra: Large Chinese language and vision assistant for ultrasound. In Proceedings of the Proceedings of the 32nd ACM international conference on multimedia, 2024, pp. 8845–8854. https://doi.org/https://doi.org/10.1145/3664647.3681584.
135. Jiang, Y.; Feng, C.M.; Ren, J.; Wei, J.; Zhang, Z.; Hu, Y.; Liu, Y.; Sun, R.; Tang, X.; Du, J.; et al. Privacy-preserving federated foundation model for generalist ultrasound artificial intelligence. *arXiv preprint arXiv:2411.16380* **2024**.
136. Meyer, A.; Murali, A.; Mutter, D.; Padoy, N. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *arXiv preprint arXiv:2411.16222* **2024**.
137. Jiao, J.; Zhou, J.; Li, X.; Xia, M.; Huang, Y.; Huang, L.; Wang, N.; Zhang, X.; Zhou, S.; Wang, Y.; et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis* **2024**, *96*, 103202. https://doi.org/https://doi.org/10.1016/j.media.2024.103202.
138. Maani, F.; Saeed, N.; Saleem, T.; Farooq, Z.; Alasmawi, H.; Diehl, W.; Mohammad, A.; Waring, G.; Valappi, S.; Bricker, L.; et al. FetalCLIP: A visual-language foundation model for fetal ultrasound image analysis. *arXiv preprint arXiv:2502.14807* **2025**.

 doi:10.20944/preprints202606.1256.v1

139. Ambsdorf, J.; Munk, A.; Llambias, S.; Christensen, A.N.; Mikolaj, K.; Balestriero, R.; Tolsgaard, M.G.; Feragen, A.; Nielsen, M. General methods make great domain-specific foundation models: A case-study on fetal ultrasound. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2025, pp. 271–281. https://doi.org/https://doi.org/10.1007/978-3-032-04981-0_26.
140. Yao, J.; Wang, Y.; Lei, Z.; Wang, K.; Feng, N.; Dong, F.; Zhou, J.; Li, X.; Hao, X.; Shen, J.; et al. Multimodal GPT model for assisting thyroid nodule diagnosis and management. *npj Digital Medicine* **2025**, *8*, 245. https://doi.org/https://doi.org/10.1038/s41746-025-01652-9.
141. Kang, Q.; Lao, Q.; Gao, J.; Bao, W.; He, Z.; Du, C.; Lu, Q.; Li, K. URFM: a general Ultrasound Representation Foundation Model for advancing ultrasound image diagnosis. *iScience* **2025**, *28*.
142. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.
143. Zhou, Y.; Chia, M.A.; Wagner, S.K.; Ayhan, M.S.; Williamson, D.J.; Struyven, R.R.; Liu, T.; Xu, M.; Lozano, M.G.; Woodward-Court, P.; et al. A foundation model for generalizable disease detection from retinal images. *Nature* **2023**, *622*, 156–163.
144. Qiu, J.; Wu, J.; Wei, H.; Shi, P.; Zhang, M.; Sun, Y.; Li, L.; Liu, H.; Liu, H.; Hou, S.; et al. Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. *arXiv preprint arXiv:2310.04992* **2023**.
145. Yu, K.; Zhou, Y.; Bai, Y.; Soh, Z.D.; Xu, X.; Goh, R.S.M.; Cheng, C.Y.; Liu, Y. Urfound: Towards universal retinal foundation models via knowledge-guided masked modeling. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 753–762. https://doi.org/https://doi.org/10.1007/978-3-031-72390-2_70.
146. Shi, D.; Zhang, W.; Yang, J.; Huang, S.; Chen, X.; Xu, P.; Jin, K.; Lin, S.; Wei, J.; Yusufu, M.; et al. A multimodal visual–language foundation model for computational ophthalmology. *npj Digital Medicine* **2025**, *8*, 381.
147. Shi, D.; Zhang, W.; Chen, X.; Liu, Y.; Yang, J.; Huang, S.; Tham, Y.C.; Zheng, Y.; He, M. Eyefound: a multimodal generalist foundation model for ophthalmic imaging. *arXiv preprint arXiv:2405.11338* **2024**.
148. Pissas, T.; Márquez-Neila, P.; Wolf, S.; Zinkernagel, M.; Sznitman, R. Masked image modelling for retinal oct understanding. In Proceedings of the International Workshop on Ophthalmic Medical Image Analysis. Springer, 2024, pp. 115–125.
149. Holland, R.; Leingang, O.; Bogunović, H.; Riedl, S.; Fritsche, L.; Prevost, T.; Scholl, H.P.; Schmidt-Erfurth, U.; Sivaprasad, S.; Lotery, A.J.; et al. Metadata-enhanced contrastive learning from retinal optical coherence tomography images. *Medical Image Analysis* **2024**, *97*, 103296. https://doi.org/https://doi.org/10.1016/j.media.2024.103296.
150. Silva-Rodríguez, J.; Chakor, H.; Dolz, J.; Ayed, I.B.; et al. On the importance of expert knowledge to improve foundation models for retinal fundus images. In Proceedings of the Medical Imaging with Deep Learning, 2024.
151. Du, J.; Guo, J.; Zhang, W.; Yang, S.; Liu, H.; Li, H.; Wang, N. Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In Proceedings of the International conference on medical image computing and computer-assisted intervention. Springer, 2024, pp. 709–719. https://doi.org/https://doi.org/10.1007/978-3-031-72390-2_66.
152. Yang, S.; Du, J.; Guo, J.; Zhang, W.; Liu, H.; Li, H.; Wang, N. ViLReF: an expert knowledge enabled vision-language retinal foundation model. *arXiv preprint arXiv:2408.10894* **2024**.
153. Wei, H.; Liu, B.; Zhang, M.; Shi, P.; Yuan, W. Visionclip: An med-aigc based ethical language-image foundation model for generalizable retina image analysis. *arXiv preprint arXiv:2403.10823* **2024**.
154. Li, Z.; Song, D.; Yang, Z.; Wang, D.; Li, F.; Zhang, X.; Kinahan, P.E.; Qiao, Y. VisionUnite: a Vision-Language Foundation Model for Ophthalmology Enhanced with Clinical Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**, pp. 1–14. https://doi.org/10.1109/TPAMI.2025.3598734.
155. Cai, Z.; Lin, L.; He, H.; Cheng, P.; Tang, X. Uni4Eye++: A General Masked Image Modeling Multi-Modal Pre-Training Framework for Ophthalmic Image Classification and Segmentation. *IEEE Transactions on Medical Imaging* **2024**, *43*, 4419–4429. https://doi.org/10.1109/TMI.2024.3422102.
156. Morano, J.; Fazekas, B.; Sükei, E.; Fecso, R.; Emre, T.; Gumpinger, M.; Faustmann, G.; Oghbaie, M.; Schmidt-Erfurth, U.; Bogunović, H. MIRAGE: Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *arXiv preprint arXiv:2506.08900* **2025**.

 doi:10.20944/preprints202606.1256.v1

157. Sun, Y.; Tan, W.; Gu, Z.; He, R.; Chen, S.; Pang, M.; Yan, B. A data-efficient strategy for building high-performing medical foundation models. *Nature Biomedical Engineering* **2025**, pp. 1–13. https://doi.org/https://doi.org/10.1038/s41551-025-01365-0.
158. Li, Z.; Song, D.; Yang, Z.; Wang, D.; Li, F.; Zhang, X.; Kinahan, P.E.; Qiao, Y. VisionUnite: a Vision-Language Foundation Model for Ophthalmology Enhanced with Clinical Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**, pp. 1–14. https://doi.org/10.1109/TPAMI.2025.3598734.
159. Wang, J.; Zhao, S.; Luo, Z.; Zhou, Y.; Jiang, H.; Li, S.; Li, T.; Pan, G. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. In Proceedings of the The Thirteenth International Conference on Learning Representations.
160. Jiang, W.; Zhao, L.; Lu, B.l. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In Proceedings of the The Twelfth International Conference on Learning Representations.
161. Jiang, W.; Wang, Y.; Lu, B.l.; Li, D. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals. In Proceedings of the The Thirteenth International Conference on Learning Representations.
162. Chen, Y.; Ren, K.; Song, K.; Wang, Y.; Wang, Y.; Li, D.; Qiu, L. EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model. In Proceedings of the AAAI 2024 Spring Symposium on Clinical Foundation Models.
163. Christensen, M.; Vukadinovic, M.; Yuan, N.; Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine* **2024**, *30*, 1481–1488.
164. Kim, S.; Jin, P.; Song, S.; Chen, C.; Li, Y.; Ren, H.; Li, X.; Liu, T.; Li, Q. Echofm: Foundation model for generalizable echocardiogram analysis. *IEEE transactions on medical imaging* **2025**.
165. Vukadinovic, M.; Tang, X.; Yuan, N.; Cheng, P.; Li, D.; Cheng, S.; He, B.; Ouyang, D. EchoPrime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation. *arXiv preprint arXiv:2410.09704* **2024**.
166. Yan, S.; Yu, Z.; Primiero, C.; Vico-Alonso, C.; Wang, Z.; Yang, L.; Tschandl, P.; Hu, M.; Tan, G.; Tang, V.; et al. A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038* **2024**, *2*.
167. Kim, C.; Gadgil, S.U.; DeGrave, A.J.; Omiye, J.A.; Cai, Z.R.; Daneshjou, R.; Lee, S.I. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature medicine* **2024**, *30*, 1154–1165. https://doi.org/https://doi.org/10.1038/s41591-024-02887-x.