Biomedical Engineering Letters (2025) 15:809–830
https://doi.org/10.1007/s13534-025-00484-6

REVIEW ARTICLE

Vision-language foundation models for medical imaging: a review of current practices and innovations

Ji Seung Ryu[1] · Hyunyoung Kang[2] · Yuseong Chu[1] · Sejung Yang[1,2]

Received: 15 January 2025 / Revised: 30 April 2025 / Accepted: 20 May 2025 / Published online: 6 June 2025

Abstract
Foundation models, including large language models and vision-language models (VLMs), have revolutionized artificial intelligence by enabling efficient, scalable, and multimodal learning across diverse applications. By leveraging advancements in self-supervised and semi-supervised learning, these models integrate computer vision and natural language processing to address complex tasks, such as disease classification, segmentation, cross-modal retrieval, and automated report generation. Their ability to pretrain on vast, uncurated datasets minimizes reliance on annotated data while improving generalization and adaptability for a wide range of downstream tasks. In the medical domain, foundation models address critical challenges by combining the information from various medical imaging modalities with textual data from radiology reports and clinical notes. This integration has enabled the development of tools that streamline diagnostic workflows, enhance accuracy (ACC), and enable robust decision-making. This review provides a systematic examination of the recent advancements in medical VLMs from 2022 to 2024, focusing on modality-specific approaches and tailored applications in medical imaging. The key contributions include the creation of a structured taxonomy to categorize existing models, an in-depth analysis of datasets essential for training and evaluation, and a review of practical applications. This review also addresses ongoing challenges and proposes future directions for enhancing the accessibility and impact of foundation models in healthcare.

Keywords Foundation model · Vision-language model · Medical imaging · Deep learning

1 Introduction

1.1 History of foundation models and recent trends

Over the past decade, artificial intelligence (AI) and machine learning (ML) have experienced groundbreaking advancements spurred by the development of foundation models, large language models (LLMs), and vision-language models (VLMs). These cutting-edge innovations have transformed the fields of computer vision and natural language processing (NLP), introducing versatile and efficient methodologies to address a wide array of visual understanding tasks. AI/ML systems have become indispensable in achieving significant progress across various domains, including object detection, image segmentation, or multimodal applications such as visual question answering (VQA) and cross-modal retrieval. Foundation models represent a fundamental shift in AI/ML approaches. Unlike traditional deep learning models, which rely heavily on task-specific annotated datasets, these models utilize extensive and diverse datasets during pre-training [1]. This pre-training spans multiple data modalities, including images, text, and their multimodal combinations, which allows foundation models to develop generalized representations that require minimal additional training for downstream tasks. Models such as CLIP [2] and DINO [3] epitomize

Ji Seung Ryu and Hyunyoung Kang equally contributed to this work.

✉ Sejung Yang
syang@yonsei.ac.kr

Ji Seung Ryu
ryujissss@yonsei.ac.kr

Hyunyoung Kang
sonya23@yonsei.ac.kr

[1] Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea

[2] Department of Medical Informatics and Biostatistics, Yonsei University Wonju College of Medicine, Wonju, Republic of Korea

this paradigm by employing large-scale, self-supervised learning to align visual and textual data, thereby performing efficiently across a variety of applications.

Critical differentiators between foundation models and earlier deep learning architectures are their scalability, adaptability, and efficiency. Traditional models often require large, labeled datasets and significant computational resources for task-specific training. In contrast, foundation models leverage self-supervised or unsupervised learning techniques, drawing on large, uncurated datasets such as web-crawled image-text pairs [4]. This approach minimizes the reliance on annotated data while enabling the extraction of rich, transferable representations. Consequently, foundation models not only reduce the computational overhead but also address a broader spectrum of vision-related tasks. Moreover, these models demonstrate a remarkable ability to generalize visual features across different domains and tasks. The modular architecture of foundation models further enhances their utility by supporting incremental fine-tuning, thereby enabling seamless adaptation to new domains or tasks with minimal computational effort.

In computer vision and NLP, foundation models have driven revolutionary advancements in complex multimodal applications. Tasks such as cross-modal retrieval, action recognition, and high-level semantic understanding benefit from the robustness of these models. LLMs, including GPT-3 [5], PaLM [6], Galactica [7], and LLaMA [8] are pretrained on vast text corpora using self-supervised learning techniques. These models are particularly adept at zero-shot and few-shot learning, allowing them to perform a wide range of tasks with minimal fine-tuning. Unlike LLMs, VLMs focus on integrating visual and language modalities. By leveraging paired datasets during pre-training, models such as CLIP [2] align images with text, making them highly effective for tasks requiring multimodal reasoning.

Healthcare is a field that naturally demands diverse data types—medical imaging, clinical records, and laboratory results, to name a few. Foundation models such as LLMs and VLMs are well-suited for addressing this complexity. For instance, LLMs reduce the reliance on task-specific training by efficiently extracting critical insights from unstructured textual data. They enable the seamless analysis of electronic health records (EHR) and support natural language-driven decision-making [9]. Simultaneously, VLMs excel in bridging textual and visual data and tackling tasks such as cross-modal retrieval, disease diagnosis, and automated medical report generation. Their extensive pretraining allows them to generalize across applications, thereby minimizing manual efforts and enhancing accuracy (ACC) [10].

These foundation models contribute to the transformative reshaping of healthcare workflows. GatorTron [9] optimizes EHR analysis, improves clinical documentation, and enables faster access to critical patient data. In interactive settings, ChatDoctor enhances patient-provider communication with conversational AI capabilities, bridging gaps in understanding [11]. Visually, VLMs have proven to be indispensable for multimodal applications. BioViL [10] combines imaging and textual data to support disease classification and reporting, which are critical requirements for modern diagnostics. By enhancing diagnostic ACC, reducing manual workloads, and delivering comprehensive insights across multiple modalities, these foundation models can redefine the future of healthcare. Their ability to integrate and analyze diverse datasets not only improves efficiency but also paves the way for more personalized and effective patient care.

This article presents a comprehensive review of foundation models, emphasizing their applications in medical imaging and the recent advancements in VLMs within the medical domain. We have organized and evaluated existing studies to provide a structured and insightful overview. Our analysis highlights the applications and strengths of these models, focusing specifically on research published between 2022 and 2024, to capture the latest developments in this dynamic field. A key highlight of this review is the meticulously curated summary of the datasets used for training and evaluation, which provide a valuable resource for researchers. Additionally, by categorizing models based on medical imaging modalities, we offer in-depth insights into the unique challenges and tailored solutions associated with each imaging modality.

Specifically, this paper focuses on the application of VLMs in the medical imaging domain, offering a structured analysis of studies in this field. To complement these findings, Fig. 1 presents a four-part visual taxonomy that classifies the reviewed studies by imaging modality (Fig. 1a), anatomical target (Fig. 1b), task type (Fig. 1c), and data source (Fig. 1d).

This review is intended to serve as a guiding framework for researchers, to foster deeper exploration and collaboration between the vision and medical communities. The major contributions of this study are as follows:

- This review presents a structured taxonomy and thorough analysis of vision-language foundation models in medical imaging, with a focus on groundbreaking research conducted between 2022 and 2024 (Fig. 1).
- By categorizing the models according to their medical imaging modalities, we provide detailed insights into the modality-specific challenges and innovative solutions designed to address them.
- Furthermore, through a comparative evaluation of model performance across tasks and modalities, we emphasize the clinical applicability and practical implications of VLMs.

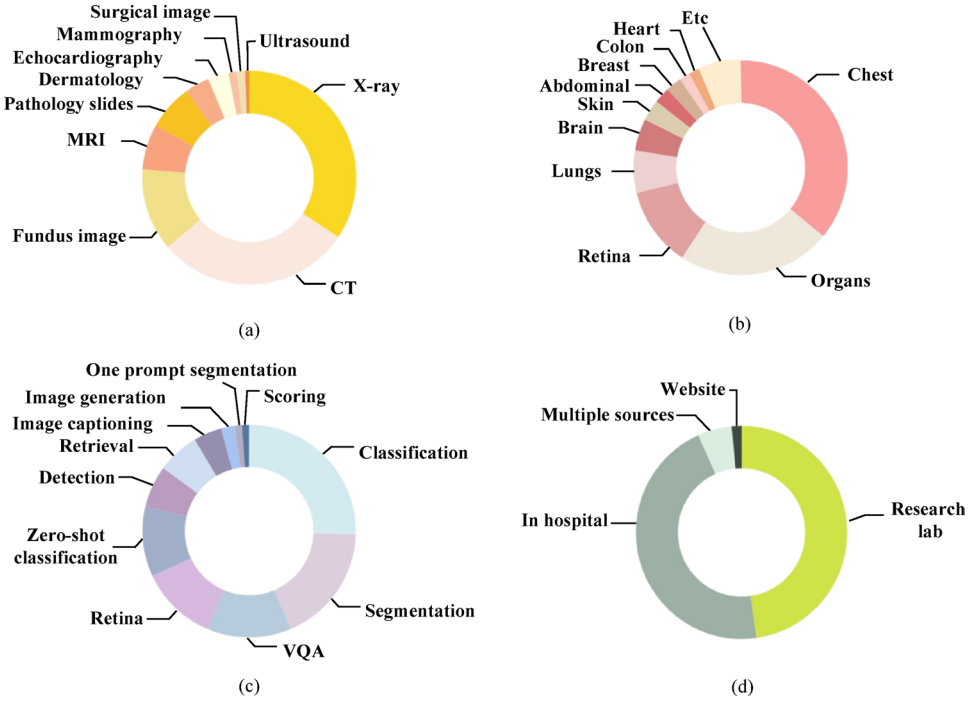

Fig. 1 Distribution of foundation model in medical field. The diagrams provide an analysis of the training datasets utilized in the reviewed studies. Each subfigure illustrates the distribution of key aspects: **a** imaging modalities, **b** target classifications, **c** organs of focus, and **d** data sources. The total number of papers included in the analysis is 61

- This review highlights the key applications and strengths of existing methodologies and proposes directions for future research.

1.2 Prior reviews on foundation models and the medical domain

Wang et al. [12] explored the impact of deep learning methodologies on medical image analysis, with a particular focus on advances in convolutional neural networks (CNNs). Their review delved into applications such as disease detection, image segmentation, and classification, while addressing critical challenges such as data scarcity, model interpretability, and the integration of these techniques into clinical workflows. Suganyadevi et al. [13] provided a broad analysis of deep learning approaches across various medical imaging modalities, including magnetic resonance imaging (MRI), CT, and X-rays. By covering the entire technical pipeline—from preprocessing and model development to evaluation—their work also highlighted practical implementation barriers. To address these challenges, they offered actionable recommendations aimed at facilitating real-world adoption. Azad et al. [14] focused on the emerging role of foundation models in medical imaging, emphasizing their scalability and adaptability to downstream tasks. Their review categorized the existing foundation models based on the architectural design and pretraining strategies, offering a critical assessment of their limitations, and proposed future research directions to enhance the efficacy of these models and broaden their applicability in medical contexts. Hartsock et al. [15] examined the application of VLMs to tasks such as medical report generation and VQA. By investigating the advancements in aligning visual and textual data, their review highlighted commonly used datasets and evaluation metrics. They also discussed the potential of these models in streamlining healthcare workflows by improving clinical documentation and decision support. Zhang et al. [16] addressed the challenges of deploying foundation models for medical image analysis, particularly those related to data availability, bias, and clinical validation. These issues often hinder the transition from research to practical application. Their review emphasized the need for model interpretability and robust evaluation frameworks to ensure clinical relevance, offering a forward-looking perspective on bridging the gap between innovation and implementation.

2 Research approach

We conducted an extensive search using Google Scholar and Arxiv, utilizing the advanced search tools available on these platforms. Custom queries were developed to compile a diverse and comprehensive collection of academic studies. This process encompassed multiple types of publications, including peer-reviewed journal articles, conference papers, workshop materials, preprints, and other non-peer-reviewed work. To ensure this breadth and diversity, our search criteria were carefully tailored to capture the full scope of relevant research.

The queries were carefully crafted to include the following keywords: (foundation*| generalist*| medical*| [Task]), (med-[FM]| medical vision language), and (foundation*| biomedical*| image*| model*). Here, [FM] denotes well-known foundation models such as PaLM and CLIP and [Task] denotes specific tasks, such as segmentation and question answering, within the context of medical imaging.

To provide a structured overview of this emerging field, this review adopts a narrative synthesis approach, focusing specifically on VLMs applied in medical imaging. The objective is to analyze recent advances in architectures, data modalities, and clinical applications, with emphasis on interpretability, scalability, and domain-specific challenges. Studies were selected for inclusion based on the following criteria: (1) publication between 2020 and 2024, (2) application of VLMs to medical domains including radiology, pathology, and ophthalmology, and (3) the presence of experimental results or evaluations conducted on clinical datasets. Studies that were purely theoretical or not directly related to medical tasks were excluded. Reflecting the rapid and ongoing developments in vision-language models within the medical imaging domain, the literature search strategy incorporated preprints published between 2022 and 2024. Scientific rigor and reliability were preserved by applying a critical appraisal process, through which only preprints demonstrating sound methodological quality and adequate experimental validation were retained. In addition, non-English studies were excluded in order to minimize the risk of misinterpretation stemming from linguistic ambiguity or inconsistencies in translation, thereby ensuring coherence and clarity in the synthesis of findings. Ultimately, this review aims to serve as a resource for both researchers and clinicians by offering a comprehensive understanding of the current state of medical VLMs, identifying prevailing limitations, and outlining potential directions for future research.

2.1 Review organization

The remainder of this review is organized, as follows. Section 2 provides an overview of the foundational principles underlying foundation models and their significance in the healthcare domain. It also summarizes the major tasks in medical imaging and classifies the primary frameworks of the foundation models used in this field. Section 3 focuses on VLMs used in medical imaging. It distinguishes between *Specialist* VLMs tailored for specific imaging modalities, such as CT, X-ray, and fundus; and *Generalist* VLMs designed to handle multiple imaging modalities for diverse applications (Fig. 2). Section 4 addresses the challenges in medical VLMs, including dataset bias, inadequate multilingual representation, and the limitations of evaluation metrics, along with an analysis of the overall trends in methodology adoption. It highlights the widespread use of cross-modal alignment for scalability, whereas multimodal attention and encoder–decoder integration face computational challenges.

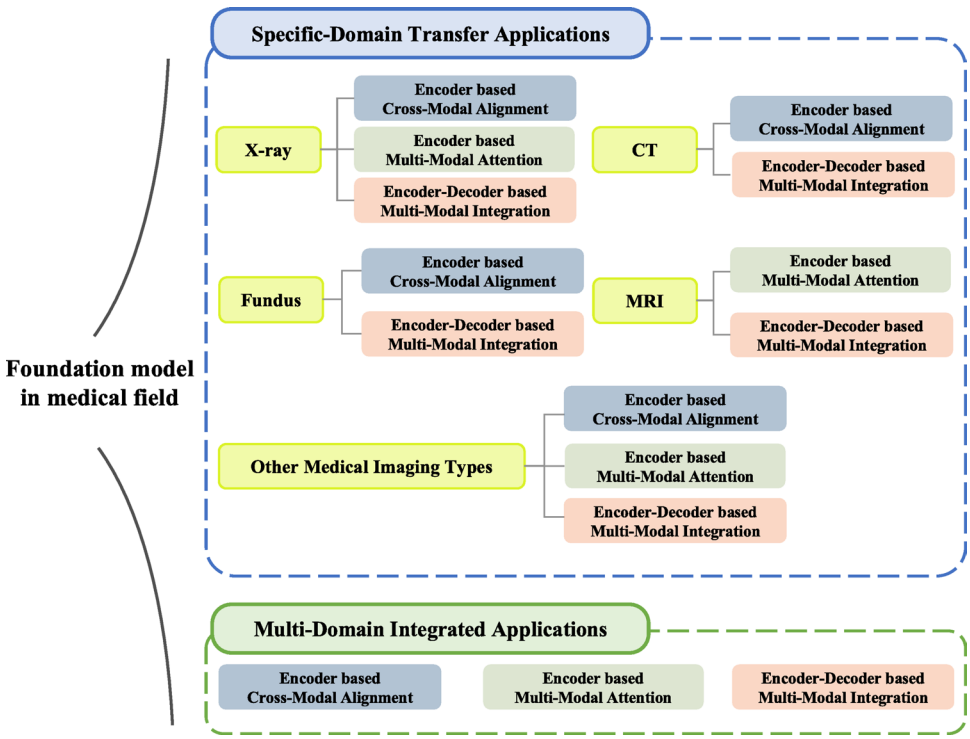

Fig. 2 Organization of the review paper. The proposed taxonomy organizes foundational models in medical field into two broad categories. Specific-domain transfer applications, which include X-ray, CT, fundus imaging, MRI, and other medical imaging types. Multi-domain integrated applications, which combine insights across multiple imaging modalities

3 Preliminary information

The concept of "foundation models" was first introduced by the Stanford Institute for Human-Centered AI, which defined them as "base models trained on large-scale data in a self-supervised or semi-supervised manner, adaptable for various downstream tasks" [1]. These models are built on the principles of deep learning, such as deep neural networks and self-supervised learning, and are influenced by the development of LLMs. Their growth has been driven by the scaling up of both data and model sizes, thereby enabling their use across many fields. This section discusses the main tasks that foundation models address in the medical field, their underlying architectures, and the factors that make them effective for medical applications.

3.1 Primary tasks in the medical field

3.1.1 Classification and zero-shot classification

Classification is a cornerstone task in medical imaging, in which models predict categories such as disease types or imaging conditions. Zero-shot classification, a more advanced approach, utilizes pre-trained VLMs to classify images without requiring fine-tuning using task-specific data. This capability is particularly valuable in scenarios where labeled datasets are scarce. One notable example is CheXNet [17] which achieves a radiologist-level performance in detecting pneumonia from chest radiographs. By leveraging DenseNet architecture and the large-scale labeled dataset ChestX-ray14, the study highlights the critical role of extensive datasets in achieving high diagnostic accuracy.

3.1.2 Segmentation

Segmentation focuses on identifying and delineating specific anatomical structures or regions of interest such as tumors, organs, or lesions. This task is crucial for applications such as treatment planning and surgical procedures. The U-Net architecture introduced by Ronneberger et al. [18] has become the gold standard for biomedical image segmentation. U-Net features an encoder–decoder design enhanced with skip connections and excels in precise boundary delineation, even with limited training data. Its adaptability makes it an indispensable tool for a wide range of medical imaging tasks.

3.1.3 Detection

Detection tasks are centered on identifying and localizing abnormalities, such as nodules, fractures, or tumors, within medical images. These tasks are essential for early diagnosis and treatment planning. MedYOLO [19], a 3D object detection framework based on the YOLO family, was introduced and specifically tailored for medical imaging applications. The model has demonstrated an exceptional performance in detecting various medical structures, highlighting its potential for use in clinical workflows.

3.1.4 Retrieval

Retrieval tasks focus on identifying visually or semantically similar images from medical datasets and play a critical role in comparative diagnosis and research. This capability is particularly valuable in fields such as radiology, pathology, and dermatology, where historical cases often guide diagnostic decisions. Lehmann et al. [20] proposed a comprehensive framework for content-based image retrieval in medical applications. By incorporating feature extraction and relevance feedback mechanisms, their system can significantly improve the retrieval ACC across multimodal datasets.

3.1.5 VQA

VQA integrates visual understanding with clinical reasoning to address natural language questions regarding medical images. This task is particularly vital in domains such as radiology and pathology, where clinicians require targeted insights from imaging data. Ben Abacha et al. [21] achieved significant strides in this area by developing the VQA-Med benchmark dataset. Designed to evaluate the VQA models in medical imaging, the dataset features clinically relevant questions related to imaging findings and diagnostic tasks. By providing a standardized resource, VQA-Med has become instrumental in advancing VQA systems for medical applications.

3.1.6 Image captioning

Image captioning automates the generation of textual descriptions for medical images, enhancing documentation and communication among healthcare professionals. Wang et al. [22] introduced TieNet, a model that embeds radiological images and reports into a shared representation space to produce descriptive captions for chest radiograph. By aligning visual data with textual representations, TieNet can improve the efficiency of automated reporting systems and support streamlined radiological workflows.

3.1.7 Image and report generation

Image generation focuses on synthesizing realistic medical images to augment datasets, particularly in cases involving rare conditions or limited training data. Hou et al. [23] developed a hybrid synthesis pipeline for histopathology image segmentation that combines real histopathology textures with generative adversarial networks (GANs). This innovative approach generates diverse training image patches across various tissue types, enhancing generalization performance. By improving the heterogeneity of synthetic datasets, this method is especially valuable for cancer types lacking annotated training data.

Report generation automates the creation of structured diagnostic reports by summarizing the key imaging findings. Jing et al. [24] designed a model that learns the joint representations of imaging data and textual information and produces radiology reports. By bridging the gap between image analysis and textual synthesis, this approach contributes to more accurate and efficient reporting in clinical radiology.

3.2 Model architecture

VLMs represent a groundbreaking category of AI systems designed to process and reason across both visual and textual modalities. These models support a wide range of tasks including image captioning, cross-modal retrieval, VQA, and text-conditioned image generation. Methodologically, VLMs can be divided into three main approaches: encoder-based cross-modal alignment, encoder-based multimodal attention, and encoder–decoder based multimodal integration (Fig. 3). This section explores each approach in detail, focusing on the architecture, learning strategies, and expected effects in the medical domain.

3.2.1 Encoder based cross-modal alignment

Encoder based cross-modal alignment employs separate encoders for visual and textual inputs and aligns their representations in a shared embedding space, shown in Fig. 3a. This alignment enables the model to compute semantic similarity between modalities—such as visual features in an X-ray and corresponding medical terms in a diagnostic report—without requiring pixel-level annotations. By comparing the similarity between encoded features, the model learns to associate paired inputs and distinguish them from unpaired examples. This methodology relies primarily on contrastive learning in which paired inputs are brought closer together in the embedding space and mismatched pairs are pushed apart. CLIP [2] by OpenAI is a seminal model in this category that has achieved zero-shot capabilities across various tasks by pre-training on 400 million image-text pairs. CLIP utilizes a vision transformer (ViT) or ResNet as its image encoder and a transformer for text encoding, jointly optimizing them using contrastive loss. ALIGN [25] has extended this approach using a larger dataset, demonstrating state-of-the-art results in image-text retrieval. Subsequent advancements, such as CLOOB [26] and DeCLIP [27] have focused on improving robustness and efficiency by integrating self-supervised learning objectives and better sampling strategies for contrastive pairs. In medical imaging, encoder based alignment models facilitate the development of robust retrieval systems that match medical images with their corresponding textual annotations or reports. This capability can significantly enhance the efficiency of case-based reasoning and diagnostic support in radiology.

3.2.2 Encoder based multimodal attention

Encoder based multi-modal attention combines the visual and textual inputs within a unified encoder architecture (Fig. 3b). By embedding both modalities into a single encoder, the model learns joint representations that capture their contextual relationships through layer-wise interaction. Unlike cross-modal alignment, which processes the modalities separately, this approach uses self-attention mechanisms to model cross-modal interactions directly within an encoder, thereby enabling joint representation learning. An example of this methodology is SimVLM [28], which treats image patches and text tokens as inputs to a shared encoder, using attention layers to capture the dependencies between the two modalities. Similarly, VisualBERT [29] employs a transformer encoder to jointly encode image regions and text tokens, allowing it to excel in tasks such as VQA and visual entailment. By fully integrating the information in each layer, these models perform exceptionally well in tasks requiring complex cross-modal reasoning. In the medical context, encoder based multimodal attention models are highly effective for tasks such as medical VQA, in which nuanced interactions between clinical images and associated textual data are critical. This approach is particularly useful for tasks that require contextual understanding, such as combining diagnostic imaging with clinical notes to provide comprehensive insights.

3.2.3 Encoder–decoder based multimodal integration

Encoder–decoder based multi-modal integration models adopt a generative approach, making them highly effective for tasks such as image captioning, report generation, and text-conditioned image creation. Unlike models that simply align or jointly embed inputs, this architecture is

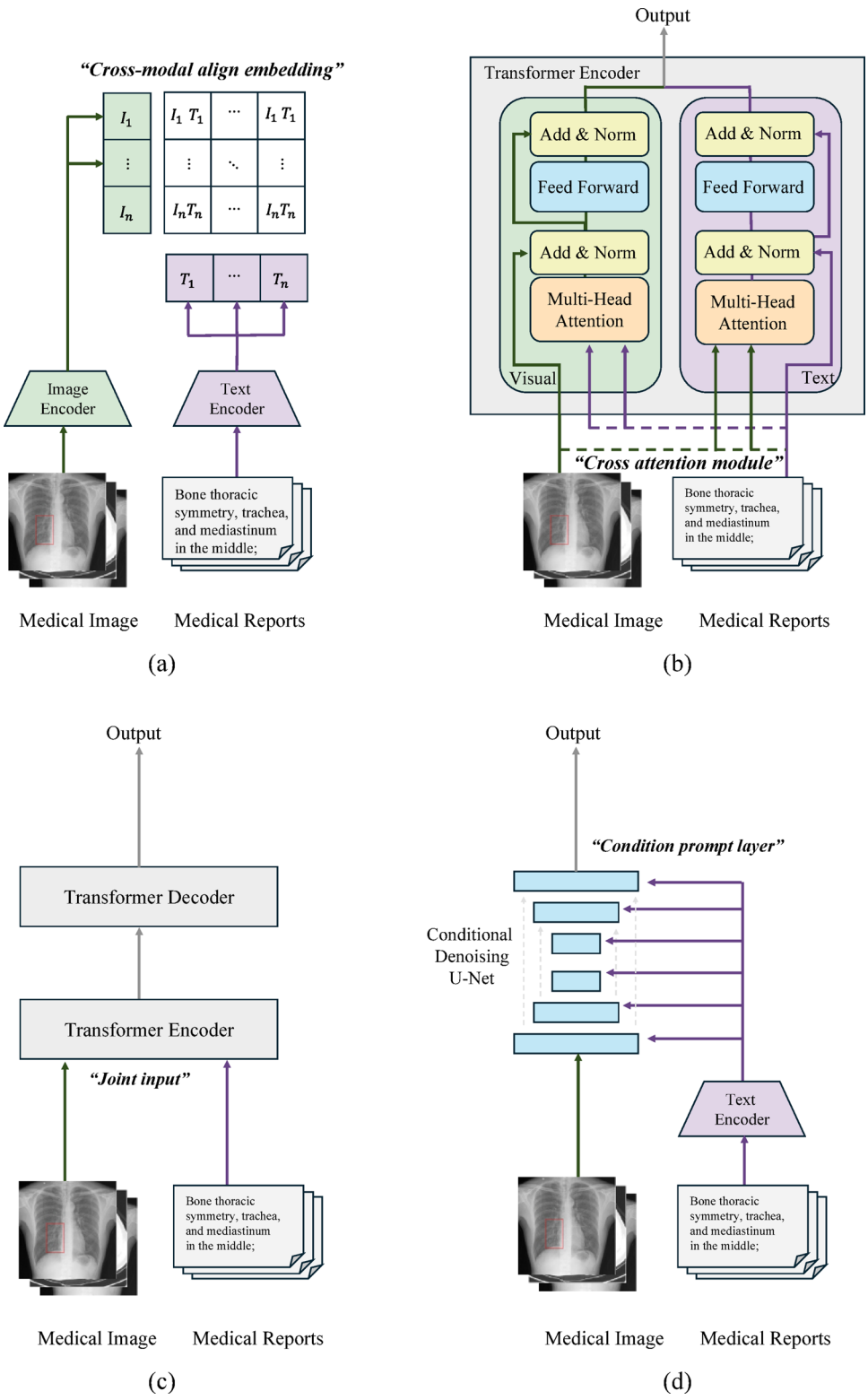

Fig. 3 Detailed illustration of model architecture. **a** Encoder-based cross-modal alignment method employs separate encoders for images and text, aligning their embeddings across modalities to facilitate integration. **b** In encoder-based multi-modal attention, both image and text inputs are processed within a unified model, using the encoder alone to execute tasks. **c** Encoder–decoder-based multi-modal integration combines images and text as simultaneous joint inputs to the encoder, adopting a generative approach for decoding outputs. **d** In another encoder–decoder-based multi-modal integration approach, text serves as a conditional prompt, directing the generation process by attention-based mechanisms

designed to actively generate outputs, allowing the model to produce natural language or synthesized images conditioned on mult-imodal input. These models typically process visual and textual inputs within a shared encoder and utilize a decoder to generate outputs based on the encoded representations. Some implementations also allow one modality, such as text, to conditionally influence another, such as images, during the intermediate stages of processing the intermediate stages of processing, as illustrated as Fig. 3c. In its encoder–decoder configuration, SimVLM

[28] treats image patches as pseudo-text tokens and integrates them seamlessly into prefixed language modeling for tasks such as conditional text generation. Expanding on this concept, VisualGPT [30] conditions pre-trained language models on visual inputs, enabling the generation of detailed captions or answers. Similarly, DeepMind's Flamingo [31] leverages cross-attention modules to fuse images and text modalities dynamically, achieving impressive few-shot performance across a variety of vision-language tasks. A representative architectural structure of these models is shown in Fig. 3d. In medical applications, encoder–decoder models have significant potential for automating diagnostic report generation, thereby reducing the workload of radiologists. For instance, given a chest radiograph, such models can produce comprehensive findings and impressions, improve workflow efficiency, and minimize human error. Furthermore, text-conditioned image generation can be used to simulate rare pathological cases, thereby enhancing the diversity of training datasets for medical education and model development.

4 Foundation models in medical imaging

4.1 Specific domain transfer applications

4.1.1 X-ray imaging

In the domain of X-ray imaging using encoder based cross-modal alignment (Table 1), Phan et al. [32] proposed a novel medical foundation model that breaks down disease descriptions into fundamental visual components. This model, which is primarily trained on X-ray images, aligns visual data with key pathological features, thereby significantly improving its ability to detect and interpret pathological findings. Similarly, Luo et al. [33] introduced DeViDe, a transformer-based approach that enhanced the performance of medical foundation models. The integration of diverse medical knowledge sources, such as radiographic descriptions, enables this model to establish a stronger connection between visual data and textual representations. Focusing on clinical knowledge, Liu et al. [34] developed a hierarchical foundation model, IMITATE. With a structure that relies on X-ray images, the model uses the findings and impressions sections of medical reports to align multilevel visual features with descriptive and conclusive text, thereby achieving effective integration of clinical insights. Finally, Wang et al. [35] presented multi-modal collaborative prompt learning (MCPL), a framework aimed at refining the relationship between medical texts and image representations. By employing collaborative prompt learning, this model demonstrated enhanced precision and interpretability, making it suitable for a wide range of medical tasks.

In the domain of encoder based multi-modal attention, Moon et al. [36] introduced the Medical Vision Language Learner (MedViLL), a framework that bridges the understanding and generation of medical images and text. Through an innovative self-attention mechanism, MedViLL effectively captures joint representations and achieves superior performance across a variety of medical tasks. Wang et al. [37] proposed ECAMP, a model designed to enhance the interpretation of medical data by emphasizing entity-specific contexts within radiology reports. By leveraging advanced language models, ECAMP extracts and refines entity-centered information from medical reports, thereby strengthening the interaction between the textual and visual modalities to improve diagnostic insights. Yan et al. [38] adapted the bidirectional encoder representations from transformers (BERT) architecture for clinical text by pretraining it on extensive medical corpora, including the medical information mart for intensive care III (MIMIC-III) clinical notes. The resulting ClinicalBERT model excels in understanding the unique language patterns and specialized terminology of the medical domain, making it highly effective for various clinical text-processing applications.

In the domain of encoder–decoder based multi-modal integration, Chambon et al. [39] presented RoentGen, a vision-language foundation model specifically designed to produce clinically accurate and descriptive chest X-ray reports. This model bridges the gap between imaging and text by generating detailed radiological insights, making it a robust tool for automated report generation. Huemann et al. [40] developed ConTEXTual Net, a multi-modal vision-language foundation model that integrates radiology reports into the segmentation process for chest radiographs. By incorporating a free-form textual context, the model can enhance pneumothorax segmentation, surpassing the performance of vision-only models, and demonstrating the value of combining visual and textual modalities. Li et al. [41] introduced an Anatomical Structure-Guided (ASG) framework that integrates anatomical knowledge into a medical vision-language foundation model. This innovative approach aligns the anatomical regions in images with the corresponding textual descriptions, enabling superior performance in classification and segmentation tasks across multiple datasets. Liu et al. [42] proposed M-FLAG, which focuses on improving training stability and efficiency. By freezing the language models and optimizing the latent space geometry with a novel orthogonality loss, the model achieves significant advancements in medical tasks. Thawakar et al. [43] introduced XrayGPT, which was tailored for radiology applications. By combining the MedClip visual encoder with a fine-tuned Vicuna language model,

Table 1 Summary of foundation models in X-ray imaging

Modality	Model	Dataset	Prompt type	Task	Metrics	Mean (evaluation dataset)
Encoder based cross-modal alignment	MAVL	MIMIC-CXR v2	Text	Zero-shot classification Detection	AUROC, F1, ACC IoU, Dice, ACC	0.735, 26.25, 82.77 (ChestX-ray14) 21.97, 34.11, 84.29 (COVID Rural)
	DeViDe	MIMIC-CXRv2	Text	Zero-shot classification Segmentation	AUROC, F1, ACC Dice	0.777, 31.5, 82.3 (ChestX-ray14) 70.27 (ChexDet)
	IMITATE	MIMIC-CXR, CheXpert, RSNA, SIIM, COVIDx, ChestX-ray14	Text	Classi-fication, Segmentation Detection, Retrieval	AUROC, Dice mAP, Precision@5	0.897 (CheXpert), 64.5 (SIIM) 26.4 (RSNA), 71.83 (CheXpert 5 × 200)
	MCPL	MIMIC-CXR	Report, Hand-craft	Classification Detection	ACC, AUROC mAP, mIoU	83.3, 0.843 (CheXpert) 20.1, 27.5 (Object-CXR)
Encoder based multi-modal attention	MedViLL	MIMIC-CXR, Open-I	Report	Classification Retrieval	Avg AUROC, F1 MRR, H@5, R@5	0.980, 0.839 (MIMIC-CXR) 56.5, 77.0, 47.4 (MIMIC-CXR)
	ECAMP	MIMIC-CXR	Text gener-ated by ChatGPT	Classification Segmentation	AUROC Dice	0.867, 0.851 (ChestX-ray14) 84.5 (SIIM-ACR Pneumothorax)
	Clinical-BERT	MIMIC-CXR, IU X-Ray, COV-CTR, NIH ChestXray14	Report	Image Captioning Classification	BLUE1,CIDEr AUROC	0.383,0.151 (MIMIC-CXR) 0.845 (NIH ChestXray14)
Encoder–decoder based multi-modal integration	RoentGen	MIMIC-CXR	Text	Image generation Classification	FID AUROC	3.6 (MIMIC-CXR) 0.824 (CheXpert)
	ConTEXTual Net	CANDID-PTX	Report	Segmentation	Dice	0.716 (CANDID-PTX)
	ASG	MIMIC-CXR	Report	Classification Segmentation	AUROC Dice	0.836 (NIH Chest X-ray) 73 (RSNA Pneumonia)
	M-FLAG	MIMIC-CXR	Report	Classification Segmentation	AUC Dice	69.50 (MIMIC-CXR) 64.80 (SIIM-ACR)
	XrayGPT	MIMIC-CXR, Open-I	Report	Image captioning Classification	BLEU AUROC	17.8 (OpenI) 0.832 (CheXpert)
	Libra	MIMIC-CXR, Medical-Diff-VQA, MIMIC-Ext-MIMIC-CXR-VQA	Report	Report generation	BLEU-1, BLEU-4,	51.3, 24.5 (MIMIC-CXR)

AUROC, area under receiver operating characteristic curve; ACC, accuracy; IoU, intersection over union; mAP, mean average precision; mIoU, mean intersection over union; MRR, mean reciprocal rank; H, Hit Rate; R, Recall; CIDEr, consensus-based image description evaluation; FID, fréchet inception distance; BLEU, bilingual evaluation understudy

their approach excels in radiology report generation and interactive reasoning, offering state-of-the-art performance in these areas. Zhang et al. [44] designed Libra, a temporally aware multi-modal LLM aimed at improving radiology report generation. Libra effectively captures temporal changes in radiological data, achieving good performance with the MIMIC-CXR dataset across lexical and clinical evaluation metrics.

4.1.2 Computed tomography imaging

Chen et al. [45] presented 3D-CT-GPT, a cutting-edge VQA-based medical VLM developed to generate radiology reports from 3D CT scans, with a specific focus on chest computed tomography (CT) using encoder-based cross-modal alignment in CT imaging (Table 2). By employing advanced VQA techniques, this model improves the interpretability and ACC of automated radiological assessments, thereby providing a significant step forward in generating detailed and clinically meaningful reports. Building on the need for

Table 2 Summary of foundation models in CT imaging

Modality	Model	Dataset	Prompt type	Task	Metrics	Mean (evaluation dataset)
Encoder based cross-modal alignment	3D-CT-GPT	CT-RATE, Dataset-XY	Text	Report generation	BLEU, ROUGE-1	13.27, 25.94 (CT-RATE)
	CT-CLIP, CT-CHAT	CT-RATE	Text	Detection Zero-shot classification	MAP@1 MAP@1	0.886 (CT-RATE) 0.886 (CT-RATE)
	E3D-GPT	BIMCV-R, CT-RATE Unlabeled 3D CT	Text	Report generation VQA	BLEU ACC	18.19 (BIMCV-R) 42.24 (BIMCV-R-VQA)
Encoder–decoder based multi-modal integration	Merlin	Abdominal CT	Report	Zero-shot classification	F1	0.741 (Abdominal CT)
	ProMISe	Medical Segmentation Decathlon (MSD)	Point	Segmentation	Dice, NSD	66.81, 81.24 (MSD)
	Med-2E3	M3D-Cap, M3D-VQA	Report	Report generation VQA	BLEU-1, ROUGE-1 BLEU-1, ROUGE-1	51.51, 54.48 (M3D-Cap) 58.55, 62.04 (M3D-VQA)
	Proposed Methods	MIMIC-CXR, Open-I, CT-KIDNEY	Text	Detection	AUROC, Precision	0.96,0.95 (MIMIC-CXR)

BLEU, bilingual evaluation understudy; ROUGE, recall-oriented understudy for gisting evaluation; MAP, mean average precision; ACC, accuracy; NSD, normalized surface dice; AUROC, area under receiver operating characteristic curve

robust datasets, Hamamci et al. [46] introduced CT-RATE, which is the first open-source multi-modal dataset that pairs 3D CT scans with the corresponding textual reports. By leveraging this dataset, the authors also developed CT-CLIP and CT-CHAT, two innovative foundation models that excel in tasks such as zero-shot multi-abnormality detection and multi-modal AI assistance for 3D medical imaging. To address the challenges of extracting high-quality 3D visual features, Lai et al. [47] proposed E3D-GPT, an enhanced 3D visual foundation model tailored for medical vision-language applications. The model is built on a substantial corpus of unlabeled 3D CT data utilized in a self-supervised learning framework to extract robust 3D visual features. By incorporating 3D spatial convolutions, E3D-GPT efficiently aggregates and projects high-level image features while reducing computational complexity.

In the domain of encoder–decoder based multi-modal integration, Blankemeier et al. [48] presented Merlin, a computationally efficient 3D vision-language foundation model specifically designed for interpreting abdominal CT scans. Merlin achieves exceptional performance across a wide range of downstream tasks by integrating supervision from both structured EHR and unstructured radiology reports. Notably, Merlin achieves state-of-the-art results while maintaining minimal computational resource requirements, making it a practical and scalable solution. To address the challenges of 3D medical image segmentation, Li et al. [49] introduced ProMISe, a framework driven by prompt engineering that adapts general VLMs for domain-specific applications. By leveraging the flexibility of prompts, this method demonstrates both high effectiveness and versatility, thereby establishing a new standard for segmentation in complex medical imaging. Focusing on multi-modal integration, Shi et al. [50] developed Med-2E3, a vision-language foundation model that combines 3D and 2D encoders to enhance medical-image analysis. The model incorporates an innovative text-guided interslice (TG-IS) scoring module that mimics the attention mechanisms used by radiologists when analyzing CT images. This approach enables Med-2E3 to excel in tasks such as report generation and VQA using large-scale multi-modal benchmarks. Zhou et al. [51] proposed a sophisticated vision-language framework that merges LLMs with hierarchical attention mechanisms. By effectively integrating multi-modal inputs, the model excels in fine-grained abnormality detection and the generation of natural language descriptions for medical CT images. This approach significantly improves the clinical relevance and detection ACC, establishing a new benchmark for precision in medical imaging tasks.

4.1.3 Fundus imaging

In the domain of fundus imaging using encoder based cross-modal alignment (Table 3), Cherukuri et al. [52] employed a guided context self-attention mechanism to integrate visual and textual features within a vision-language foundation model designed for retinal image captioning. The GCS-M3VLT architecture effectively captures intricate visual details and a broader clinical context, even with limited data. Evaluations of the DeepEyeNet dataset have demonstrated improvements in BLEU-4 scores, indicating its capability to generate accurate and comprehensive medical captions. Du et al. [53] developed RET-CLIP, a vision-language foundation model pre-trained on a large dataset of color fundus photographs paired with clinical diagnostic reports. The model employs a tripartite optimization strategy to extract features at three levels: the left eye, right eye, and report data. This multilevel approach facilitates effective representation learning, leading to enhanced diagnostic performance in diseases such as diabetic retinopathy and glaucoma. Luo et al. [54] addressed demographic biases in VLMs by introducing FairCLIP, a framework designed to promote fairness

Table 3 Summary of foundation models in Fundus imaging

Modality	Model	Dataset	Prompt type	Task	Metrics	Mean (evaluation dataset)
Encoder based cross-modal alignment	GCS-M3VLT	DeepEyeNet	Text	Report generation	BLEU-1, BLEU-2	0.430, 0.345 (DeepEyeNet)
	RET-CLIP	Private Dataset	Report	Classification	AUROC, AUPR	0.856 0.616 (IDRID)
	FairCLIP	Harvard-FairVLMed	Report generated by ChatGPT	Classification	AUROC, ES-AUC	0.702, 0.655 (Harvard-FairVLMed)
	FLAIR	37 Combined datasets	Text	Detection Segmentation	ACA/κ AUROC	0.604/0.772 (MESSIDOR) 0.92 (FIVES)
	VisionCLIP	SynFundus-1 M	Text	Zero-shot classification	ACC	43.1 (MESSIDOR)
	ViLReF	Private Dataset	Report	Classification Segmentation	AUROC, mAP DSC, IoU	94.29, 63.62 (RFMiD) 52.65, 38.38 (IDRiD)
Encoder–decoder based multi-modal integration	VisionUnite	MMFundus	Text	Classification	ACC, Diagnostic Relevance	77.8, 2.937 (MMFundus)

BLEU, bilingual evaluation understudy; AUROC, area under receiver operating characteristic curve; AUPR, area under the precision-recall curve; ES-AUC, early stopping area under the curve; ACA, average classification accuracy; κ, Cohen's Kappa; ACC, accuracy; mAP, mean average precision; DSC, dice similarity coefficient; IoU, intersection over union

Table 4 Summary of foundation models in MRI imaging

Modality	Model	Dataset	Prompt type	Task	Metrics	Mean (evaluation dataset)
Encoder based multi-modal attention	MedBLIP	ADNI, NACC, OASIS	Text generated by EHRs	Classification Zero-shot classification	ACC ACC	78.7 (ADNI) 80.8 (AIBL)
Encoder–decoder based multi-modal integration	Med-UniC	MIMIC-CXR, PadChest	Report	Image captioning Classification	BLEU AUROC	18.25 (MIMIC-CXR) 0.832 (CheXpert)
	FM-ABS	Left Atrium, Brain Tumor	Bbox generated by MobileSAM	Segmentation	Dice, Jaccard	86.14, 75.85 (Left Atrium)

Bbox, bounding box; ACC, accuracy; BLEU, bilingual evaluation understudy

across diverse data distributions. Using optimal transport methods, the model mitigates performance disparities between demographic groups, ensuring more equitable outcomes in medical image analysis while maintaining robust diagnostic capabilities. Silva-Rodriguez et al. [55] incorporated domain-specific retinal knowledge into the training process of FLAIR, a vision-language foundation model for medical image analysis. The model embeds expert clinical insights into text supervision and demonstrates improved interpretative abilities, resulting in an enhanced performance in disease classification and anomaly detection tasks. Wei et al. [56] utilized synthetic fundus images paired with natural language descriptions to develop VisionCLIP, a vision-language foundation model for retinal image analysis. This strategy enabled the model to effectively generalize to real-world datasets while preserving patient confidentiality. Yang et al. [57] designed ViLReF, a vision-language foundation model optimized for detecting fine-grained abnormalities in retinal images. By leveraging expert-driven label extraction and implementing weighted similarity coupling loss, the model effectively captures subtle yet clinically significant patterns. This approach improves the ACC of lesion detection and segmentation tasks and highlights its utility in precision diagnostics.

In the domain of encoder–decoder based multi-modal integration, Li et al. [58] introduced VisionUnite, which is designed specifically for ophthalmology, to address critical challenges in multi-disease diagnosis, user interaction, and interpretability. The model is trained on MMFundus, the largest multi-modal fundus dataset to date that contains more than 1.24 million image-text pairs, including high-resolution fundus images and simulated doctor-patient dialogues.

4.1.4 MRI imaging

In the domain of MRI imaging using encoder based multi-modal attention (Table 4), Chen et al. [59] introduced MedBLIP, a vision-language foundation model aimed at seamlessly integrating 3D medical imaging with textual data derived from EHRs. By leveraging vision language pre-training, this model effectively captures the intricate relationships between volumetric medical images and the associated textual information. Consequently, MedBLIP has achieved significant breakthroughs in applications such as

automated radiology report generation and clinical decision making.

In response to the biases often present in multilingual medical datasets, Wan et al. [60] developed Med-UniC, a vision-language foundation model that employs cross-lingual text alignment regularization. This innovative framework aligns textual representations across languages, thereby enhancing inclusivity and optimizing performance in a variety of vision-language tasks. In particular, Med-UniC excels in multilingual diagnostic reporting and image-text retrieval, underscoring its adaptability to diverse clinical contexts. Xu et al. [61] proposed foundation model-driven active barely supervised (FM-ABS), a vision-language foundation model designed to address the complexities of 3D medical image segmentation under minimal supervision. By incorporating a prompt-driven architecture alongside active learning methodologies, FM-ABS significantly reduces the reliance on large, annotated datasets while maintaining high segmentation precision.

4.1.5 Other medical imaging

In the domain of other medical imaging using encoder based cross-modal alignment (Table 5), Ferber et al. [62] explored the potential of in-context learning within multimodal LLMs to classify cancer pathology images without the need for task-specific fine-tuning. By harnessing the contextual information embedded in both visual and textual data, the model demonstrates its capability to analyze complex pathology slides with adaptability and efficiency. This innovative approach shows promise in supporting flexible diagnostic workflows that align seamlessly with clinical requirements. Vo et al. [63] investigated the utilization of frozen, large-scale, pretrained vision-language foundation models as foundational backbones for multi-modal breast cancer prediction. Rather than retraining the models, this method preserves the pretrained parameters while incorporating domain-specific mammography data, leading to improved predictive ACC for breast cancer diagnosis. This study highlights the practical advantages of repurposing large-scale VLMs for medical imaging, showcasing their effectiveness in addressing domain-specific diagnostic challenges. Building on the EchoCLIP model, Christensen et al. [64] introduced EchoCLIP-R, a vision-language foundation model specifically designed for echocardiographic analysis. This updated model features a customized echocardiography report text tokenizer, enabling a more precise alignment of multi-modal data. EchoCLIP-R achieves impressive results across various tasks, including identifying individual patients across multiple videos, detecting clinical transitions, and delivering robust image-to-text retrieval with top-tier cross-modal ranking. These advancements underscore its versatility and reliability in echocardiographic interpretation and report generation.

In the domain of encoder–decoder based multi-modal integration, Yin et al. [62] investigated the use of prompt engineering to customize vision foundation models for analyzing pathology images. Task-specific prompts are incorporated within the QAP framework, enabling the model to excel in pathology-oriented tasks such as tissue

Table 5 Summary of foundation models in other medical imaging

Modality	Model	Dataset	Image type	Prompt type	Task	Metrics	Mean (evaluation dataset)
Encoder based cross-modal alignment	GPT-4 V	Private dataset attributed to company	Pathology Slides	Text	Zero-shot classification	ACC	32.5 (CRC-VAL-HE-7 K)
	Proposed Methods	CBIS-DDSM, EMBED	Mammography	Text generated by Tab2Text	Classification	ACC, AUROC	79.6, 0.907 (CBIS-DDSM)
	EchoCLIP-R	Cedars-Sinai Medical Center	Echocardiography	Report	Retrieval Regression	MCMRR MAE	206.1 (Cedars-Sinai Medical Center) 16.9 (Cedars-Sinai Medical Center)
Encoder based multi-modal attention	QAP	NAFLD-Anomaly	Pathology Slides	Morphological Attributes	Classification Scoring	F1 Avg F1	99.58 (NAFLD-Anomaly) 83.37 (NAFLD-Anomaly)
	LLaVA-Ultra	US-Hospital	Ultrasound	Text	VQA	F1, Precision	76.85, 81.88 (SLAKE)
	GP-VLS	11 Combined datasets	Surgical Imaging	Text	VQA	ACC	46.1 (MedQA)
Encoder–decoder based multi-modal integration	SkinGEN	Fitzpatrick17k, SCIN	Clinical image	Text	Image generation Classification	CLIP, DINOV2 score	0.76,0.82 (Fitzpatrick17k)

ACC, accuracy; AUROC, area under receiver operating characteristic curve; MCMRR, mean cumulative mean reciprocal rank; MAE, mean absolute error

classification and anomaly detection without the need for extensive fine-tuning. This innovative approach emphasizes the adaptability and efficiency of prompt-based techniques for streamlining medical imaging workflows for pathological slides. Guo et al. [65] introduced LLaVA-Ultra, a vision-language foundation model specifically designed for ultrasound imaging in Chinese healthcare. This model integrates sophisticated vision and language functionalities to address critical challenges unique to ultrasound, including the variability in interpretation and the demands of real-time interaction. Optimized for tasks such as image interpretation, diagnostic decision-making, and interactive querying, LLaVA-Ultra is effective in advancing clinical ultrasound practices. In surgical applications, Schmidgall et al. [66] developed GP-VLS, a versatile vision-language foundation model that combines domain-specific medical and surgical knowledge with advanced visual scene comprehension. This model supports key tasks such as surgical phase recognition, instrument detection, and intraoperative decision-making. GP-VLS offers real-time, context-sensitive assistance and can enhance surgical workflows, improve clinical efficiency, and support more informed decision-making in surgical environments.

Lin et al. [67] introduced SkinGEN, a vision-language foundation model augmented with stable diffusion, to advance dermatological diagnostics through interactive and explainable visualizations. The model generates lifelike depictions of potential skin conditions, enhancing the diagnostic ACC in tasks such as classification and anomaly detection. By embedding explainability into its design, SkinGEN not only improves clinical outcomes, but also strengthens communication between clinicians and patients, fostering greater trust and understanding in medical consultations.

4.2 Multi-domain integrated applications

4.2.1 Encoder based cross-modal alignment

In the domain of foundation models with encoder based cross-modal alignment (Table 6), Ghosh et al. [68] introduced Mammo-CLIP, a pioneering vision-language foundation model pre-trained on an extensive dataset of mammogram-report pairs. By capitalizing on the inherent alignment between the visual and textual data in mammography, the model achieves improvements in data efficiency and robustness. Its enhanced performance in tasks such as abnormality detection and image-text alignment underscores its potential for integration into breast cancer screening workflows. Liu et al. [69] developed T3D, which is a vision-language framework tailored for high-resolution 3D medical imaging. This model uses text-informed contrastive learning and advanced image restoration techniques to capture intricate visual details without down sampling. Consequently, T3D excels in representation learning for volumetric datasets, making it particularly effective for classification and segmentation

Table 6 Summary of foundation models with encoder-based cross-modal alignment

Model	Dataset	Image type	Prompt type	Task	Metrics	Mean (evaluation dataset)
Mammo-CLIP	UPMC, VinDr	X-ray, CT	Report	Zero-shot classification	ACC	62.0, 76.0, 15.0 (RSNA)
T3D	BIMCV-VLP	X-ray, CT, MRI	Text	Segmentation Classification	avgDice macro-avg AUROC	79.5 (BTCV) 58.1 (MDLT)
BLIP	PubMed Image-Text	Xray, CT, MRI, Microscopy, Fundus Imaging	Caption	Retrieval	i2t@1 i2t@10	36.52 72.62 (PubMed Image-Text)
PM2	BACH, Figshare MRI Brain Tumor, DR	MRI, Fundus Imaging, Pathology Slides	Text generated by CoOp	Zero-shot classification	ACC	47.5 (BACH)
Medclip	MIMIC-CXR, CheXpert, Unpaired Text, COVID, RSNA Pneumodia	X-ray, CT	Text	Zero-shot classification Retrieval	ACC P@1,P@2	59.4 (MIMIC-CXR) 45,49 (CheXpert5 × 200)
UniDCP	ROCO, MIMIC-CXR	X-ray, CT, MRI, Ultrasound, Pathology Slides	Text	VQA Report generation	ACC BLEU-1, BLEU-2	74.5 (VQA-RAD) 0.527, 0.349 (IU X-Ray)
MPMA	ROCO, MIMIC-CXR	X-ray, CT, MRI, Ultrasound, Pathology Slides	Text	Classification Report generation	AUROC BLEU-1, BLEU-2	0.906 (CheXpert) 0.518, 0.337 (IU X-Ray)
BiomedCLIP	PMC-15 M	X-ray, CT, MRI, Ultrasound, PET, Microscopy, Pathology Slides	Text	Retrieval VQA	R@1, R@5 ACC	56.0, 77.9 (PMC-15 M) 72.7 (VQA-RAD)

ACC, accuracy; AUROC, area under receiver operating characteristic curve; i2t, image-to-text; P, precision; R, Recall

tasks involving 3D modalities, such as CT scans. Monajatipoor et al. [70] proposed BLIP, a pipeline designed to align medical images with textual data through subfigure-caption matching and multi-modal pretraining. Particularly adept at analyzing brain abnormalities, this model enhances tasks such as image-text retrieval and multi-modal understanding. Its architecture emphasizes precise alignment between visual inputs and textual descriptions, enabling superior analysis of complex brain imaging datasets. Wang et al. [71] introduced PM2, a multi-modal prompting paradigm that addresses the challenges of few-shot medical image classification. By integrating cross-modal information, PM2 demonstrates flexibility and robust performance, particularly in scenarios with limited labeled data. This versatility makes it a valuable tool for various medical imaging modalities. Wang et al. [72] presented MedCLIP, a vision-language foundation model designed to learn from unpaired medical images and text. Employing a semantic similarity matrix for contrastive learning, MedCLIP bypasses the need for paired datasets, achieving notable success in zero-shot image-text retrieval and classification across modalities such as X-rays and pathology slides. Zhan et al. [73] introduced UniDCP, a VLM that utilizes dynamic cross-modal learnable prompts. This approach harmonizes inputs from diverse pretraining tasks, enabling the model to adapt to a wide range of vision-language tasks in medical imaging without requiring task-specific fine-tuning. UniDCP performs exceptionally well in tasks such as report generation and cross-modal retrieval. Zhang et al. [74] proposed MPMA, a vision-language foundation model that integrates cross-modal alignment into joint image-text reconstruction. By fostering enhanced interactions between modalities, this method improves the performance in tasks such as classification and report generation, particularly when applied to multi-modal datasets. Finally, Zhang et al. [75] introduced BiomedCLIP, a multi-modal biomedical foundation model pre-trained on PMC-15 M [75], a comprehensive dataset containing 15 million image-text pairs sourced from PubMed Central. The model benefits from extensive pretraining and excels in biomedical tasks such as image-text retrieval and zero-shot classification. Its ability to address complex medical queries with remarkable precision highlights its potential for advancing biomedical research and applications.

4.2.2 Encoder based multi-modal attention

In the domain of foundation models with encoder based multi-modal attention (Table 7), Chen et al. [76] devised an approach that integrates domain-specific knowledge. Their method refines the alignment between the visual and textual data, enabling more accurate reasoning for complex tasks. This advancement has proven to be particularly effective in

Table 7 Summary of foundation models with encoder based multi-modal attention

Model	Dataset	Image type	Prompt type	Task	Metrics	Mean (evaluation dataset)
Proposed Methods	ROCO, MedICaT, MIMIC-CXR	X-ray, CT, MRI, Ultrasound	Text, Graph	VQA Classification	ACC ACC	67.60 (VQA-RAD) 80.51 (MELINDA)
Llama3-Med	Claude 3 Opu, LLaMA 3 70B	X-ray, CT, MRI, Ultrasound, PET	Text	VQA	Recall	31.20 (VQA-RAD)
PPE	COCO	X-ray, Microscopy, Pathology Slides, RGB image	Text generated by BLIP, Handcraft, Mask label generated by LViT	Segmentation	Dice, mIoU	80.59, 67.59 (MoNuSeg)
LLaVA-Med	PMC-15 M	X-ray, CT, MRI, Ultrasound, PET	Text generated by GPT-4	VQA	Recall	64.75 (VQA-RAD)
TFA-LT	ISIC2018, APTOS2019	Dermoscopy, Fundus Imaging	Text	Classification	ACC	70.48 (ISIC2018)
LViT	Private dataset attributed to company	X-ray, CT	Report	Segmentation	Dice, mIoU	83.66, 75.11 (MosMed Data+)
One-Prompt Segmentation	78 Combined datasets	X-ray, CT, MRI, Fundus Imaging, CBCT	Click, Bbox, Doodles, Mask label	Segmentation	Avg Dice	67.30 (KiTS23)
Med-VLFM	ROCOv2	X-ray, CT	Text	Report generation	BERT Score, ROUGE-1	0.638, 0.304 (ROCOv2)
BiomedGPT-B	IU X-ray, MIMIC-CXR, Peir Gross, SLAKE, VQA-RAD, PathVQA	X-ray, CT, MRI, Pathology Slides	Report	Image captioning VQA	ROUGE-L, METEOR ACC	28.50, 12.90 (IU X-ray) 88.7 (SLAKE)

Bbox, bounding box; ACC, accuracy; mIoU, mean intersection over union; ROUGE, recall-oriented understudy for gisting evaluation; METEOR, metric for evaluation of translation with explicit ordering

Table 8 Summary of foundation models with encoder–decoder based multi-modal integration

Model	Dataset	Image type	Prompt type	Task	Metrics	Mean (evaluation dataset)
TV-SAM	Private dataset attributed to company	X-ray, CT, MRI, Ultrasound, Microscopy, Dermoscopy	Text generated by GPT-4, Bbox generated by GLIP	Segmentation	Avg Dice	0.831 (Polyp benchmark)
SERPENT-VLM	IU X-Ray, ROCO	X-ray, CT	Text	Report generation	BLEU4, ROUGE-L	0.190,0.326 (IU X-Ray)
BiomedCoOp	CTKidney, DermaMNIST, Kvasir, RETINA, LC25000	CT, Dermoscopy, Endoscopy, Fundus Imaging, Pathology Slides	Text	Classification	ACC, Harmonic Mean	86.93, 82.74 (CTKidney)
MS-VLM	CT-RATE, In-house Rectal MRI	CT, MRI	Report	Report generation VQA	BLEU-4, ROUGE-L Precision, Recall	0.232, 0.438 (CT-RATE) 0.222, 0.329 (CT-RATE)
VILA-M3	MIMIC-CXR, SLAKE, PathVQA, CheXpert	X-ray, CT, MRI, Pathology Slides	Text	Segmentation VQA	Dice ACC	0.95 (RSNA Pneumonia) 84.20 (SLAKE)
MAKEN	ImageCLEFmedical 2023	X-ray, CT, MRI, Ultrasound, PET, Endoscopy	Text	Report generation	BLEU-1, ROUGE-1	0.189, 0.275 (ImageCLEFmedical 2023)
Proposed Methods	TN3K, Kvasir-SEG, QaTa-COV19	Ultrasound, Endoscopic, CT	Bbox	Segmentation	mDice, mIoU	93.67, 89.44 (TN3K)

Bbox, bounding box; BLEU, bilingual evaluation understudy; ROUGE, recall-oriented understudy for gisting evaluation; ACC, accuracy; mDice, mean dice similarity coefficient; mIoU, mean intersection over union

medical applications such as diagnostic support and anomaly detection. Llama3-Med, a vision-language foundation model crafted by Chen et al. [77], is designed for biomedical tasks. The model utilizes a hierarchical image-encoding strategy and an enriched biomedical image-text dataset, significantly enhancing its capacity to analyze intricate biomedical imagery. Its strong performance in generating diagnostic reports and supporting clinical decisions highlights its potential. Focusing on the segmentation ACC and adaptability across imaging modalities, Han et al. [78] created prior prompt encoder (PPE), a VLM guided by textual prompts at multiple scales. The integration of contextually relevant guidance has been invaluable for tasks involving X-rays, CT scans, and MRIs. Li et al. [79] streamlined the training of LLaVA-Med, a foundation model optimized for multi-modal biomedical conversations. The model was trained in less than one day by using an efficient pipeline that combines biomedical figure-caption pairs and GPT-4-generated instruction data. This efficiency, paired with conversational fluency, has made it stand out in biomedical contexts. Li et al. [80] addressed the challenge of long-tailed medical image classification using text-guided foundation model adaptation for long-tailed medical (TFA-LT), which is a text-guided framework. Their system employs lightweight adapters and a two-stage training strategy and excels in handling imbalanced datasets while maintaining computational efficiency. Li et al. [81] introduced LViT, which advances medical image segmentation. By fusing vision transformers with language guidance, the model achieves precise, context-aware segmentation. Its success demonstrates the benefits of combining multi-modal understanding with advanced techniques. Wu et al. [82] innovated a single-prompt framework that simplifies medical image segmentation across diverse imaging modalities. Its versatility and straightforward design make it a promising choice for tasks such as organ segmentation and lesion identification. Yang et al. [83] achieved recognition with Med-VLFM (also known as Pclmed), a vision-language foundation model that triumphed in the ImageCLEFmedical 2024 Caption Prediction Challenge. The model improves both interpretability and clinician-patient communication by generating detailed, context-aware captions for medical images. Finally, Zhang et al. [84] introduced BiomedGPT-B, a multi-modal foundation model designed for biomedical applications. The model uses extensive pretraining to excel in tasks such as VQA and multi-modal analysis, thus solidifying its role as a robust tool for biomedical research.

4.2.3 Encoder–decoder based multi-modal integration

In the domain of foundation models with encoder–decoder based multi-modal integration (Table 8), Jiang et al. [85] improved the zero-shot segmentation capabilities for multi-modal medical images by integrating GPT-4-generated descriptive prompts into the text-visual-prompt segment anything model (TV-SAM) framework. This innovation

eliminated the reliance on human annotations, making segmentation workflows more efficient while maintaining high ACC across imaging modalities such as X-rays, CT scans, and MRIs. Kapadnis et al. [86] introduced SERPENT-VLM, a self-refining framework designed for generating radiology reports. Employing a novel self-supervised loss function, the model aligned generated text with the corresponding input images, thereby effectively minimizing hallucinations and bolstering robustness. Even when handling noisy or incomplete inputs, SERPENT-VLM delivered consistent results across multiple radiology benchmarks. Koleilat et al. [87] addressed the challenges of biomedical image classification using BiomedCoOp, a vision-language foundation model. By blending BiomedCLIP with prompt ensembles derived from LLMs and employing selective knowledge distillation, the framework excelled in few-shot classification tasks. Its effectiveness has been demonstrated using diverse imaging modalities, including pathology slides and mammograms. For 3D medical imaging interpretation, Lee et al. [88] introduced MS-VLM, a model optimized using a slice-by-slice embedding strategy powered by Z-former. This innovative design seamlessly integrated multi-view and multi-phase data to overcome the computational challenges typically encountered by traditional 3D vision encoders. MS-VLM has also achieved impressive performance in generating clinically relevant radiology reports. Nath et al. [89] expanded the potential of vision-language foundation models with VILA-M3, which incorporated domain-specific medical knowledge. Task-specific optimization allowed the model to excel in VQA, report generation, and medical image classification, particularly when used with complex multi-modal datasets. Wu et al. [90] participated in the ImageCLEFmedical 2023 challenge and utilized the MAKEN framework to focus on internal validation because of the absence of ground truth labels for external test datasets. By prioritizing reliable internal benchmarking, their approach ensured robust performance even with data limitations. Zheng et al. [91] explored the segmentation challenges in medical imaging through a curriculum-prompting strategy for vision-language foundation models. This framework gradually increased the task complexity during training, leading to superior segmentation results across imaging modalities, such as CT and ultrasound. This systematic approach offered an effective pathway to enhance the segmentation performance.

5 Discussion

In the medical field, the use of VLMs is closely tied to both imaging modalities and the underlying model architectures. The choice of imaging modality is shaped by factors such as data availability, clinical relevance, and the technical feasibility of integrating these modalities into VLM frameworks. Similarly, the architecture of the model, including encoder–decoder designs, attention mechanisms, and multimodal fusion techniques, significantly affects its ability to process and analyze diverse medical data effectively. This discussion explores the key trends in modern healthcare VLMs, focusing on advancements in their applications and strategies. Additionally, the ongoing challenges in applying VLMs to the medical domain are also addressed, highlighting areas that require further development.

5.1 Frequently used medical image modalities

X-rays are the most widely used imaging modalities in research, serving as a foundation for numerous medical applications. This can be attributed to several factors. First, The availability of large-scale datasets, such as MIMIC-CXR [92], CheXpert [93], and NIH ChestX-ray14 [94], provides millions of X-ray images paired with radiology reports. These datasets are instrumental for VLM training, facilitating robust cross-modal alignment, and supporting tasks such as automated report generation. In addition, the structured nature of radiology reports aligns well with the requirements of cross-modal tasks, further enhancing their utility. Second, the simplicity and consistency of X-ray imaging make it particularly well-suited for scalable model development. Unlike CT or MRI, which produce complex 3D volumetric data, X-rays are 2D single-view images. This lower dimensionality significantly reduces computational demands and helps mitigate the risk of overfitting, especially when working with limited data.

While CT and MRI are indispensable for diagnosing complex conditions, such as cancer staging and neurological disorders, their use in VLM research remains relatively limited compared to X-rays. A major barrier is the computational demands of the modalities. CT and MRI generate high-resolution volumetric data, requiring extensive processing power and sophisticated algorithms, which increases the complexity of training VLMs. Thus, despite their clinical significance, CT and MRI are seldom used in large-scale VLM studies. Efforts to incorporate 3D imaging into vision-language pretraining have faced scalability issues due to GPU memory limitations and the lack of standardized radiology report formats across institutions [45]. These challenges hinder model generalization and underscore a key limitation, that clinically valuable imaging modalities cannot be fully leveraged without adequate computational resources and standardized datasets.

Fundus imaging is a specialized niche in research. Its clinical applications, such as the diagnosis of diabetic retinopathy and glaucoma, highlight its importance. Paired

image-text datasets, such as IDRiD [95] and MMFundus [96, 97] support research in this area by enabling vision-language applications. However, fundus imaging is confined to ophthalmology, which restricts its broad applicability in diverse clinical contexts. Pathology and ultrasound imaging are less researched because of the unique challenges they pose. Pathology datasets require detailed expert annotations, such as cell types or cancer grades, making them time-consuming and costly. Additionally, the visual complexity of pathology images complicates data preparation and model training. Particularly, the extremely high resolution of whole-slide pathology images, gigapixel scale, imposes significant memory demands. Although tiling strategies are often used to manage this, they frequently lead to the loss of spatial context that is essential for accurate diagnosis. In contrast, ultrasound imaging faces challenges related to variability in image quality. Operator skills significantly affect the consistency of the ultrasound data, creating inconsistencies that make model training more difficult. Furthermore, the lack of large-scale paired datasets limits the use of VLMs.

5.2 Frequently used methodologies

Encoder based cross-modal alignment is the most widely used VLM methodology in the medical domain. Its popularity arises from its simplicity, scalability, and efficiency in addressing tasks such as classification and retrieval, particularly when large paired datasets such as X-rays and radiology reports are available. By separating the image and text encoders, this approach reduces the computational overhead, making it an attractive choice for resource-constrained settings. The strength of this methodology in zero-shot learning has revolutionized case-based reasoning and diagnostics. For example, DeViDe [33] excels in both segmentation and classification tasks, whereas RET-CLIP [53] demonstrates high performance in fundus imaging classification. Its effectiveness is primarily due to the fact that many widely used medical datasets, such as MIMIC-CXR [92] and CheXpert [93], contain loosely aligned image-text pairs rather than fully annotated or structured reports. Despite these advantages, the independent processing of visual and textual modalities remains a notable limitation. This separation hinders the model's ability to capture complex interactions between modalities, making it less effective for tasks that demand deep semantic understanding, such as those involving nuanced cross-modal reasoning.

In the medical domain, encoder based multi-modal attention is moderately used, primarily in tasks that demand nuanced reasoning and rich contextual understanding. In contrast to cross-modal alignment, which processes modalities independently, multi-modal attention fosters deeper interactions between visual and textual data. Although this approach increases computational costs, it excels in scenarios that require simultaneous reasoning over both modalities. A notable example is MedViLL [36], which demonstrates strong classification performance by combining X-ray images with clinical notes. This method performs well in tasks that require understanding both image and text together such as matching clinical findings with corresponding visual patterns because it directly models interactions between the two. However, this comes at a cost. The internal workings of the attention mechanism are hard to interpret, making it difficult for clinicians to understand why the model made a certain prediction. This lack of transparency can be a major drawback in medical settings where trust and accountability are essential. In addition, these models often need large amounts of training data to perform well. When trained on smaller datasets, their performance tends to plateau early, limiting their usefulness in low-resource domains like rare diseases or specialized imaging modalities.

Encoder–decoder based multi-modal integration is among the least commonly applied methodologies in VLMs within the medical domain, despite its significant potential for generative tasks. Its limited adoption can be attributed to the considerable computational power and large-scale paired datasets required for effective training. Generative tasks, such as radiology report generation, often depend on structured text outputs; however, such datasets are scarce, particularly for modalities such as MRI and pathology. Even in the case of widely available modalities, such as X-rays, datasets such as MIMIC-CXR [92] offer only partially structured text, further complicating the training process. The high computational demands of encoder–decoder models present another major challenge, particularly for institutions that lack a robust infrastructure. Consequently, such models are often limited to niche applications in resource-rich environments. However, their capabilities are limited to tasks for which structured and coherent outputs are indispensable. For instance, RoentGen [39] demonstrates strong performance in radiology report generation by producing clinically relevant and coherent text. Similarly, XrayGPT [43] has demonstrated its potential for automating diagnostic reporting workflows, thereby reducing the manual effort required for such labor-intensive processes. While the promise of encoder–decoder based integration for generative applications is evident, its current reliance on extensive paired datasets and computationally intensive training limits its broader adoption. Addressing these challenges is essential for making this methodology more accessible and applicable across diverse medical contexts.

5.3 Bias and variance in VLMs

The bias and variance issues in VLMs for medical imaging remains a significant challenge. Bias arises from training datasets that do not adequately represent diverse populations, leading to an imbalanced model performance across different groups. For example, biases related to race, ethnicity, sex, socioeconomic factors, and language can result in unreliable predictions regarding underrepresented communities. Variance, on the other hand, refers to the sensitivity of the model to variations in training data, which limits its ability to effectively generalize across different patient populations or healthcare settings. In VLM datasets, English continues to dominate, despite the fact that most of the world's population does not speak English as their primary language. This dominance restricts the performance of monolingual VLMs in multilingual tasks and introduces community bias, which disproportionately affects non-English speakers. This bias is particularly concerning in medical applications and can have serious consequences [60].

Recent developments in VLMs have indicated a shift towards emphasizing the diversity and representativeness of datasets to address these challenges. For example, datasets such as FairCLIP [54], PadChest [98], PMC-15 M [75], and Mammo-CLIP [68] include racially and demographically diverse data to reduce bias and ensure fairness. Specifically, PadChest [98] can construct reports that incorporate non-English languages, such as Spanish, to integrate cross-lingual representations and improve performance on non-English tasks. The VLMs applied to these datasets include MAVL [32], Medunic [60], BioMedCLIP [75], DeViDe [33], IMITATE [34], LLaVA-Med [79], and Mammo-CLIP [68]. These models demonstrate the potential to address biases, improve multilingual capabilities, and enhance real-world performance.

Despite ongoing efforts to mitigate bias in recent models, existing datasets and methodologies remain inadequate for fully addressing this issue. A lack of diversity in training data, such as under representation of different racial groups, languages, or clinical settings, can lead to uneven model performance, thereby increasing the risk of inaccurate or biased outcomes for marginalized populations. This limitation is particularly concerning in clinical contexts, where fairness, reliability, and generalizability are critical for safe deployment. To overcome this challenge, it is essential to develop more representative and inclusive datasets that accurately reflect the heterogeneity of real-world patient populations. Additionally, robust evaluation frameworks are needed to assess model performance across diverse demographic and linguistic subgroups.

5.4 Lack of standardized evaluation metrics

Evaluation metrics, such as BLEU and ROUGE, are widely used to assess the generative performance of medical VLMs. These metrics serve as evaluation benchmarks for most models [38, 50, 90]. However, these metrics often fail to reflect clinically important findings. BLEU and ROUGE focus on surface-level matching by evaluating n-grams (words or phrases) based on their overlap with reference texts. This approach is limited because clinical reports often describe the same conditions or findings using various terminologies or expressions. As a result, clinically accurate texts may still receive poor evaluations. Moreover, clinical reports frequently emphasize specific disease names or findings that carry greater clinical significance compared to other words. Because BLEU and ROUGE treat all n-grams equally, they cannot assign appropriate weights to clinically critical terms or phrases. For instance, if a clinical report states "no malignancy found" but rephrases it as "malignancy not detected," the two sentences convey identical clinical meaning. However, BLEU and ROUGE may assign low scores because of differences in word choice or phrasing. Consequently, metrics should prioritize ACC, relevance, and interpretability, which reflect the clinical importance of findings, over simple textual similarity. To address these limitations, alternative metrics such as the CLIP and Dinov2 scores have been proposed, focusing on the similarity between medical text and images [67]. Although these metrics represent an improvement, they still fail to fully guarantee ACC for clinical significance and lack sufficient evaluation of specific details or key terms in medical texts. Therefore, future studies should consider developing evaluation metrics that better reflect the way medical professionals understand clinical reports. For example, using medical term databases such as Unified Medical Language System (UMLS) or RadLex could help give more weight to important disease related terms during evaluation. It is also important to recognize that different expressions can mean the same thing in clinical language. In addition, involving clinicians or radiologists in the evaluation process could help judge whether a generated report is truly useful and accurate in a medical context. Finally, creating benchmark datasets that include multiple correct versions of a report for the same image would allow for more fair and realistic scoring, since there is often more than one way to describe the same medical finding.

6 Conclusion

This review of VLMs in the medical domain provides a vital synthesis of the rapidly evolving landscape of foundation models in healthcare. Exploring the diverse applications of

VLMs across key medical imaging tasks, such as segmentation, classification, and report generation, highlights their transformative potential in enhancing diagnostic ACC and clinical workflows. The modality-specific categorization of VLMs, coupled with a detailed analysis of their strengths, and a systematic mapping of their clinical applications offer a structured and comprehensive perspective on the current state of the art. This review aims to serve as both a resource and roadmap to guide researchers and practitioners in advancing the development and application of VLMs to address the complex challenges of modern medicine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13534-025-00484-6.

Acknowledgements We would like to thank Editage (www.editage.co.kr) for the English language editing.

Author contributions Conceptualization was carried out by Sejung Yang. Literature search was conducted by Ji Seung Ryu, Hyunyoung Kang. Formal analysis was performed by Ji Seung Ryu, Hyunyoung Kang. Funding acquisition was secured by Sejung Yang. Investigation was undertaken by Ji Seung Ryu, Yuseong Chu. Project administration was managed by Sejung Yang. Supervision was provided by Sejung Yang. The original draft of the manuscript was written by Ji Seung Ryu. All authors contributed to the methodology, validation, and visualization. The review and editing of manuscript were collaboratively performed by all authors.

Funding This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2022R1A2C2091160). Additional support was provided through a grant from the Information and Communications Promotion Fund via the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT), Republic of Korea. Furthermore, this research received funding from the Bio & Medical Technology Development Program of the National Research Foundation (NRF), supported by the Korean government (MSIT) (Grant No. RS-2024-00440802).

Declarations

Conflict of of interest The authors declare that they have no relevant financial or non-financial interests to disclose.

Ethical approval This article is a review and does not contain any studies with human participants or animals performed by any of the authors. Therefore, ethics approval was not required.

Consent to publish Not applicable. This article does not contain any individual person's data in any form.

Consent to participate Not applicable. This article does not report on studies involving human participants performed by the authors.

References

1. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. 2021. arXiv:2108.07258.
2. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. Int Conf Mach Learn. 2021;139:8748–63.
3. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision; 2021; 9650–60.
4. Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. 2020;1. arXiv:2005.14165.
5. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
6. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: scaling language modeling with pathways. J Mach Learn Res. 2023;24(240):1–113.
7. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A large language model for science. arXiv preprint arXiv:2211.09085. 2022. arXiv:2211.09085.
8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023. arXiv:2302.13971.
9. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. arXiv preprint arXiv:2203.03540. 2022. arXiv:2203.03540.
10. Boecking B, Usuyama N, Bannur S, Castro DC, Schwaighofer A, Hyland S, et al., editors. Making the most of text semantics to improve biomedical vision–language processing. European conference on computer vision; 2022: Springer.
11. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus. 2023;15(6):e40895. https://doi.org/10.7759/cureus.40895.
12. Wang J, Zhu H, Wang SH, Zhang YD. A review of deep learning on medical image analysis. Mobile Netw Appl. 2021;26(1):351–80. https://doi.org/10.1007/s11036-020-01672-7.
13. Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. Int J Multimed Inf Retr. 2022;11(1):19–38. https://doi.org/10.1007/s13735-021-00218-1.
14. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, et al. Foundational models in medical imaging: a comprehensive survey and future vision. arXiv. arXiv preprint arXiv:2310.18689. 2023;10. arXiv:2310.18689.
15. Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: a review. Front Artif

Intell. 2024;7:1430984. https://doi.org/10.3389/frai.2024.1430984.
16. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal. 2024;91:102996. https://doi.org/10.1016/j.media.2023.102996.
17. Rajpurkar P. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv abs/1711. 2017.5225. arXiv:1711.05225.
18. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015. 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015. pp. 234-41. https://doi.org/10.1007/978-3-319-24574-4_28.
19. Sobek J, Medina Inojosa JR, Medina Inojosa BJ, Rassoulinejad-Mousavi S, Conte GM, Lopez-Jimenez F, et al. MedYOLO: a medical image object detection framework. J Imaging Inform Med. 2024;37(6):3208–16. https://doi.org/10.1007/s10278-024-01138-2.
20. Lehmann TM, Güld MO, Thies C, Fischer B, Spitzer K, Keysers D, et al. Content-based image retrieval in medical applications. Methods Inf Med. 2004;43(4):354–61.
21. Ben Abacha A, Hasan SA, Datla VV, Demner-Fushman D, Müller H. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes; 2019:9–12.
22. Wang X, Peng Y, Lu L, Lu Z, Summers RM. Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:9049–58.
23. Hou L, Agarwal A, Samaras D, Kurc TM, Gupta RR, Saltz JH. Robust histopathology image analysis: To label or to synthesize?. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019;2019:8533-42. https://doi.org/10.1109/CVPR.2019.00873.
24. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195. 2017. 10.48550/arXiv:1711.08195.
25. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, PMLR. 2021.
26. Fürst A, Rumetshofer E, Lehner J, Tran VT, Tang F, Ramsauer H, et al. Cloob: Modern hopfield networks with infoloob outperform clip. Adv Neural Inf Process Syst. 2022;35:20450–68.
27. Li Y, Liang F, Zhao L, Cui Y, Ouyang W, Shao J, et al. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208. 2021. arXiv:2110.05208.
28. Wang Z, Yu J, Yu AW, Dai Z, Tsvetkov Y, Cao Y. Simvlm: simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904. 2021. 10.48550/arXiv:2108.10904.
29. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W. Visualbert: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557. 2019. 10.48550/arXiv:1908.03557.
30. Chen J, Guo H, Yi K, Li B, Elhoseiny M. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
31. Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst. 2022;35:23716–36.
32. Phan VMH, Xie Y, Qi Y, Liu L, Liu L, Zhang B, et al. Decomposing disease descriptions for enhanced pathology detection: a multi-aspect vision-language pre-training framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
33. 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. https:/doi.org/10.48550/arXiv:2404.03618.
34. Liu C, Cheng S, Shi M, Shah A, Bai W, Arcucci R. IMITATE: clinical prior guided hierarchical vision-language pre-training. IEEE Trans Med Imaging. 2024. https://doi.org/10.1109/TMI.2024.3449690.
35. Wang P, Zhang H, Yuan Y. MCPL: multi-modal collaborative prompt learning for medical vision-language model. IEEE Trans Med Imaging. 2024. https://doi.org/10.1109/TMI.2024.3418408.
36. Moon JH, Lee H, Shin W, Kim Y-H, Choi E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE J Biomed Health Inform. 2022;26(12):6070–80. https://doi.org/10.1109/JBHI.2022.3207502.
37. Wang R, Yao Q, Lai H, He Z, Tao X, Jiang Z, et al. ECAMP: entity-centered context-aware medical vision language pre-training. arXiv preprint arXiv:2312.13316. 2023. 10.48550/arXiv:2312.13316.
38. Yan B, Pei M. Clinical-bert: vision-language pre-training for radiograph diagnosis and reports generation. Proc AAAI Conf Artif Intell. 2022. https://doi.org/10.1609/aaai.v36i3.20204.
39. Chambon P, Bluethgen C, Delbrouck J-B, Van der Sluijs R, Połacin M, Chaves JMZ, et al. Roentgen: vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737. 2022. 10.48550/arXiv:2211.12737.
40. Huemann Z, Tie X, Hu J, Bradshaw TJ. ConTEXTual net: a multimodal vision-language model for segmentation of pneumothorax. J Imaging Inf Med. 2024. https://doi.org/10.1007/s10278-024-01051-8.
41. Li Q, Yan X, Xu J, Yuan R, Zhang Y, Feng R, et al. Anatomical structure-guided medical vision-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention; 2024: Springer. https://doi.org/10.1007/978-3-031-72120-5_8.
42. Liu C, Cheng S, Chen C, Qiao M, Zhang W, Shah A, et al. M-flag: medical vision-language pre-training with frozen language models and latent space geometry optimization. In: International conference on medical image computing and computer-assisted intervention; 2023: Springer. https://doi.org/10.1007/978-3-031-43907-0_61.
43. Thawakar OC, Shaker AM, Mullappilly SS, Cholakkal H, Anwer RM, Khan S, et al., editors. XrayGPT: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd workshop on biomedical natural language processing. 2024. https://doi.org/10.18653/v1/2024.bionlp-1.35.
44. Zhang X, Meng Z, Lever J, Ho ES. Libra: Leveraging temporal images for biomedical radiology analysis. arXiv preprint arXiv:2411.19378. 2024. arXiv:2411.19378.
45. Chen H, Zhao W, Li Y, Zhong T, Wang Y, Shang Y, et al. 3d-ct-gpt: generating 3d radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330. 2024. arXiv:2409.19330.
46. Hamamci IE, Er S, Almas F, Simsek AG, Esirgun SN, Dogan I, et al. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. 2024; https://doi.org/10.21203/rs.3.rs-5271327/v1.
47. Lai H, Jiang Z, Yao Q, Wang R, He Z, Tao X, et al. E3D-GPT: enhanced 3D visual foundation for medical vision-language model. arXiv preprint arXiv:2410.14200. 2024. arXiv:2410.14200.
48. Blankemeier L, Cohen JP, Kumar A, Van Veen D, Gardezi SJS, Paschali M, et al. Merlin: a vision language foundation model for

3d computed tomography. Res Square. 2024. https://doi.org/10.21203/rs.3.rs-4546309/v1.
49. Li H, Liu H, Hu D, Wang J, Oguz I, editors. Promise: prompt-driven 3D medical image segmentation using pretrained image foundation models. In: 2024 IEEE international symposium on biomedical imaging (ISBI); 2024: IEEE. https://doi.org/10.1109/ISBI56570.2024.10635207.
50. Shi Y, Zhu X, Hu Y, Guo C, Li M, Wu J. Med-2E3: A 2D-enhanced 3D medical multimodal large language model. arXiv preprint arXiv:2411.12783. 2024. arXiv:2411.12783.
51. Zhou Z, Xia S, Shu M, Zhou H. Fine-grained abnormality detection and natural language description of medical CT images using large language models. Int J Innov Res Comput Sci Technol. 2024;12(6):52–62. https://doi.org/10.1109/ICHI61247.2024.00080.
52. Cherukuri TK, Shaik NS, Bodapati JD, Ye DH. GCS-M3VLT: guided context self-attention based multi-modal medical vision language transformer for retinal image captioning. arXiv preprint arXiv:2412.17251. 2024. arXiv:2412.17251.
53. Du J, Guo J, Zhang W, Yang S, Liu H, Li H, et al., editors. Retclip: a retinal image foundation model pre-trained with clinical diagnostic reports. In: International conference on medical image computing and computer-assisted intervention; 2024: Springer.
54. Luo Y, Shi M, Khan MO, Afzal MM, Huang H, Yuan S, et al. Fairclip: harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024.
55. Silva-Rodriguez J, Chakor H, Kobbi R, Dolz J, Ayed IB. A foundation language-image model of the retina (flair): encoding expert knowledge in text supervision. Med Image Anal. 2025;99:103357. https://doi.org/10.1016/j.media.2024.103357.
56. Wei H, Liu B, Zhang M, Shi P, Yuan W. VisionCLIP: an MedAIGC based ethical language-image foundation model for generalizable retina image analysis. arXiv preprint arXiv:2403.10823. 2024. arXiv:2403.10823.
57. Yang S, Du J, Guo J, Zhang W, Liu H, Li H, et al. ViLReF: an expert knowledge enabled vision-language retinal foundation model. arXiv preprint arXiv:2408.10894. 2024. arXiv:2408.10894.
58. Li Z, Song D, Yang Z, Wang D, Li F, Zhang X, et al. VisionUnite: a vision-language foundation model for ophthalmology enhanced with clinical knowledge. arXiv preprint arXiv:2408.02865. 2024. arXiv:2408.02865.
59. Chen Q, Hong Y. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In: Proceedings of the Asian conference on computer vision; 2024.
60. Wan Z, Liu C, Zhang M, Fu J, Wang B, Cheng S, Ma L, Quilodrán-Casas C, Arcucci R. Med-unic: unifying cross-lingual medical vision-language pre-training by diminishing bias. Adv Neural Inf Process Syst. 2023;36:56186–97.
61. Xu Z, Chen C, Lu D, Sun J, Wei D, Zheng Y, et al. FM-ABS: promptable foundation model drives active barely supervised learning for 3D medical image segmentation. In: International conference on medical image computing and computer-assisted intervention; 2024: Springer. https://doi.org/10.1007/978-3-031-72111-3_28.
62. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun. 2024;15(1):10104. https://doi.org/10.1038/s41467-024-51465-9.
63. Vo HQ, Wang L, Wong KK, Ezeana CF, Yu X, Yang W, et al. Frozen large-scale pretrained vision-language models are the effective foundational backbone for multimodal breast cancer prediction. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3507638.
64. Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision–language foundation model for echocardiogram interpretation. Nat Med. 2024. https://doi.org/10.1038/s41591-024-02959-y.
65. Guo X, Chai W, Li S-Y, Wang G, editors. LLaVA-ultra: large Chinese language and vision assistant for ultrasound. In: Proceedings of the 32nd ACM international conference on multimedia; 2024. https://doi.org/10.1145/3664647.3681584.
66. Schmidgall S, Cho J, Zakka C, Hiesinger W. GP-VLS: a general-purpose vision language model for surgery. arXiv preprint arXiv:2407.19305. 2024. 10.48550/arXiv:2407.19305.
67. Lin B, Xu Y, Bao X, Zhao Z, Zhang Z, Wang Z, et al. SkinGEN: an explainable dermatology diagnosis-to-generation framework with interactive vision-language models. arXiv preprint arXiv:2404.14755. 2024. 10.48550/arXiv:2404.14755.
68. Ghosh S, Poynton CB, Visweswaran S, Batmanghelich K. Mammo-clip: a vision language foundation model to enhance data efficiency and robustness in mammography. In: International conference on medical image computing and computer-assisted intervention; 2024: Springer.
69. Liu C, Ouyang C, Chen Y, Quilodrán-Casas CC, Ma L, Fu J, et al. T3d: towards 3d medical image understanding through vision-language pre-training. arXiv preprint arXiv:2312.01529. 2023. 10.48550/arXiv:2312.01529.
70. Monajatipoor M, Dou Z-Y, Chien A, Peng N, Chang K-W. Medical vision-language pre-training for brain abnormalities. arXiv preprint arXiv:2404.17779. 2024. https://doi.org/10.48550/arXiv.2404.17779.
71. Wang Z, Sun Q, Zhang B, Wang P, Zhang J, Zhang Q. PM2: A new prompting multi-modal model paradigm for few-shot medical image classification. arXiv preprint arXiv:2404.08915. 2024. https://doi.org/10.48550/arXiv.2404.08915.
72. Wang Z, Wu Z, Agarwal D, Sun J. Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163. 2022. https://doi.org/10.48550/arXiv.2210.10163.
73. Zhan C, Zhang Y, Lin Y, Wang G, Wang H. Unidcp: unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts. IEEE Trans Multimed. 2024. https://doi.org/10.1109/TMM.2024.3397191.
74. Zhang K, Yang Y, Yu J, Jiang H, Fan J, Huang Q, et al. Multi-task paired masking with alignment modeling for medical vision-language pre-training. IEEE Trans Multimed. 2023. https://doi.org/10.1109/TMM.2023.3325965.
75. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. BiomedCLIP: a multimodal biomedical foundation model pre-trained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915. 2023. https://doi.org/10.48550/arXiv.2303.00915.
76. Chen Z, Li G, Wan X, editors. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the 30th ACM international conference on multimedia. 2022.
77. Chen Z, Pekis A, Brown K. Advancing high resolution vision-language models in biomedicine. arXiv preprint arXiv:2406.09454. 2024. https://doi.org/10.48550/arXiv.2406.09454.
78. Han X, Chen Q, Xie Z, Li X, Yang H. Multiscale progressive text prompt network for medical image segmentation. Comput Graph. 2023;116:262–74. https://doi.org/10.1016/j.cag.2023.08.030.
79. 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. Adv Neural Inf Process Syst. 2023;36:28541–64.
80. Li S, Lin L, Huang Y, Cheng P, Tang X. Text-guided foundation model adaptation for long-tailed medical image classification. In: 2024 IEEE international symposium on biomedical imaging

(ISBI); 2024: IEEE. https://doi.org/10.1109/ISBI56570.2024.10635462.
81. Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, et al. Lvit: language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging. 2023. https://doi.org/10.1109/TMI.2023.3291719.
82. Wu J, Xu M. One-prompt to segment all medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024.
83. Yang B, Yu Y, Zou Y, Zhang T. Pclmed: champion solution for imageclefmedical 2024 caption prediction challenge via medical vision-language foundation models. CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS org, Grenoble, France; 2024.
84. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024. https://doi.org/10.1038/s41591-024-03185-2.
85. Jiang Z, Cheng D, Qin Z, Gao J, Lao Q, Ismoilovich AB, et al. TV-SAM: increasing zero-shot segmentation performance on multimodal medical images using GPT-4 generated descriptive prompts without human annotation. Big Data Min Analyt. 2024;7(4):1199–211.
86. Kapadnis MN, Patnaik S, Nandy A, Ray S, Goyal P, Sheet D. SERPENT-VLM: self-refining radiology report generation using vision language models. arXiv preprint arXiv:2404.17912. 2024. https://doi.org/10.48550/arXiv.2404.17912.
87. Koleilat T, Asgariandehkordi H, Rivaz H, Xiao Y. BiomedCoOp: learning to prompt for biomedical vision-language models. arXiv preprint arXiv:2411.15232. 2024. https://doi.org/10.48550/arXiv.2411.15232.
88. Lee C, Park S, Shin C-I, Choi WH, Park HJ, Lee JE, et al. Read like a radiologist: efficient vision-language model for 3D medical imaging interpretation. arXiv preprint arXiv:2412.13558. 2024. https://doi.org/10.48550/arXiv.2412.13558.
89. Nath V, Li W, Yang D, Myronenko A, Zheng M, Lu Y, et al. VILA-M3: enhancing vision-language models with medical expert knowledge. arXiv preprint arXiv:2411.12915. 2024. https://doi.org/10.48550/arXiv.2411.12915.
90. Wu S, Yang B, Ye Z, Wang H, Zheng H, Zhang T. MAKEN: improving medical report generation with adapter tuning and knowledge enhancement in vision-language foundation models. In: 2024 IEEE international symposium on biomedical imaging (ISBI); 2024: IEEE. https://doi.org/10.1109/ISBI56570.2024.10635421.
91. Zheng X, Zhang Y, Zhang H, Liang H, Bao X, Jiang Z, et al. Curriculum prompting foundation models for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention; 2024: Springer.
92. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0.
93. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI conference on artificial intelligence; 2019.
94. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
95. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. Idrid: diabetic retinopathy–segmentation and grading challenge. Med Image Anal. 2020;59:101561. https://doi.org/10.1016/j.media.2019.101561.
96. Liu R, Wang X, Wu Q, Dai L, Fang X, Yan T, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns. 2022. https://doi.org/10.1016/j.patter.2022.100512.
97. Orlando JI, Fu H, Breda JB, van Keer K, Bathula DR, Diaz-Pinto A, et al. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal. 2020;59:101570. https://doi.org/10.1016/j.media.2019.101570.
98. Bustos A, Pertusa A, Salinas J-M, De La Iglesia-Vaya M. Padchest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal. 2020;66:101797. https://doi.org/10.1016/j.media.2020.101797.