Knowledge-Based Systems 352 (2026) 116901

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Vision-Language Models for medical imaging: A survey on anomaly detection and multimodal applications

Gul E Arzu [a,b], Muhammad Umar [c], Muhammad Fayaz [a], Usman Ali [a], L. Minh Dang [d], Hyeonjoon Moon [a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[b] Department of Digital Healthcare, Graduate School of JABA, Wonkwang University, Iksan, Republic of Korea
[c] Department of Software Engineering, Sejong University, Seoul 05006, Republic of Korea
[d] Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam

ARTICLE INFO

Keywords:
Vision-Language Models
Medical imaging
Visual questions answer
Radiology report generation
Medical dataset scarcity

ABSTRACT

Vision-Language Models (VLMs) have revolutionized medical imaging by integrating visual and textual modalities, enabling advanced tasks such as anomaly detection, diagnostic classification, and multimodal analysis. These models facilitate critical applications, including radiology report generation (RG), visual question answering (VQA), and image-text retrieval, thereby enhancing diagnostic precision and clinical workflows. Leveraging architectures like CLIP, BiomedCLIP, and Med-GPT, VLMs achieve fine-grained alignment between radiology images and textual descriptions, significantly improving interpretability and decision support in medical diagnostics. Despite their potential, adapting VLMs for healthcare presents challenges, including the scarcity of large-scale annotated datasets, the complexity of medical terminology, and the stringent requirements for patient privacy and data security. This survey comprehensively reviews state-of-the-art VLMs in radiology, analyzing their architectures, training methodologies, datasets, and real-world applications. Models such as BioViL and XrayGPT exemplify how multimodal learning enhances pathology detection and automated report generation. Additionally, emerging techniques like contrastive learning, self-supervised approaches, and federated learning (FL) offer promising solutions to address data limitations, privacy concerns, and model robustness. However, several barriers remain, including issues of model generalizability, dataset biases, and the high computational demands of training and deployment in clinical environments. This study highlights the transformative impact of VLMs in medical diagnostics and identifies key research gaps that must be addressed to ensure their practical implementation. This work outlines a roadmap for future advancements and underscores the need for scalable, privacy-preserving, and clinically effective AI-driven solutions to enhance radiological analysis and healthcare decision-making.

1. Introduction

One of the most critical and rapidly evolving research areas in medical informatics is the study of Vision-Language Models (VLMs) using radiographic images and their corresponding free-text descriptions, such as radiology reports (e.g., chest X-rays) [1–4]. Medical images and reports jointly encode rich semantic information that supports clinical interpretation and decision-making. VLMs have become pivotal in enabling multimodal learning within the medical domain by integrating visual and textual modalities. These models provide automated assistance for various clinical tasks, including report generation, diagnosis classification, and cross-modal reasoning [2,4]. Advances in vision-language research have demonstrated strong potential for improving the quality and efficiency of medical care [4,5].

Despite significant progress, medical image classification remains a challenging and persistent problem in healthcare. Although automated classification techniques have achieved promising results, they are often limited to a narrow range of predefined pathologies, restricting their applicability in real-world clinical settings where generalization to unseen conditions is essential [6]. This limitation is particularly pronounced due to the vast diversity of medical conditions and the scarcity of comprehensive annotated training data.

* Corresponding author.
E-mail addresses: arzurabani@sju.ac.kr (G.E. Arzu), muhammadumar@sju.ac.kr (M. Umar), Muhammadfayaz@sju.ac.kr (M. Fayaz), usman.ali@sejong.ac.kr (U. Ali), hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.knosys.2026.116901
Received 25 February 2025; Received in revised form 8 April 2026; Accepted 16 August 2026
Available online 22 August 2026

To tackle this challenge, cutting-edge methods like CLIP [7] contrastively train on various image-text pairs. To provide a more detailed understanding, it is essential to concentrate on the area of interest even though CLIP captures the content of the complete image. This is particularly associated with radiology because pathological structures are frequently fine and tiny in radiology. Moreover, the radiologist frequently observes an anomaly but wants to help in categorizing it [8].

An intuitive tactic to allow the VLM to concentrate on an exact area would be to crop the image. Nevertheless, doing so would remove the pathology's global context, which could impair classification accuracy. Some methods actively model these concentrating abilities [9], but their applicability is restricted because there are not many datasets in the medical field. However, drawing markers directly into the image is a straightforward technique that does not need any extra training. Current research in natural images analyzes this strategy, producing unconventional findings in zero-shot works [6,10,11]. They assume, that the model has encountered the chosen visual features throughout training and, therefore, comprehends the meaning behind those concepts. However, given the lack of such visual characteristics in the training process, they also show that this performance has a greater potential to be learned from large datasets and high-capacity models.

Medical image analysis and pre-training for natural language processing (NLP) [12] has advanced substantially due to the integration of pre-trained language models like BERT, enabling novel insights in the field [13]. Large-scale VLP models like CLIP [12] and BLIP [14], which certainly apply paired image-text contrastive learning, have currently advanced meaningfully in visual tasks. However, it is still difficult to interpret elementary VLP models into real-time healthcare applications. Large language models (LLMs) have demonstrated exceptional proficiency in understanding and generating natural language, providing foundational solutions for specific domains such as law education [15–17], and public healthcare [18–20]. However, each domain has unique challenges, and directly using pre-trained LLMs may not yield ideal results. Considering their inherent complexity, fine-tuning these models enables them to better adapt to downstream tasks, which is a key approach to leveraging large models [21–23]. Usually, VLMs are pre-trained on a huge collection of text image datasets by incorporating their knowledge of visual and textual modalities, but LLMs struggle to process visual information because they can only comprehend discrete text, despite their remarkable proficiency in zero/few-shot reasoning in the majority of NLP tasks. CLIP [24] is a notable example that optimizes language and vision models to align textual and visual images. Computational pathology is not an exception.

QUILT-Net [6] and PLIP [25] are two VLMs that were pre-trained on big histopathology datasets using pairs of histopathology images and descriptions employing the contrastive learning paradigm similar to CLIP. Such VLMs build a solid foundation for various types of downstream tasks. In particular, these models have been successfully applied to several image classification tasks such as lymph-node metastasis detection, tissue phenotyping, and Gleason grading [6,25] without further training or fine-tuning, i.e., zero-shot image classification. We have observed that the existing works on VLMs have mainly focused on the pre-training process and straightforward, direct application to downstream tasks [26–29]. Some have sought to utilize VLMs with transfer learning and knowledge distillation [30]. These approaches are, by and large, similar to how the pre-trained vision and language models are used, which do not fully explore the potential of VLMs. Herein, we suppose that the pre-trained VLMs per se are capable of conducting quantitative histopathology image analysis. In other words, VLMs can quantify histopathology images, and the resultant quantitative features can be used for downstream tasks [10].

However, studying the potential of these models to combine and analyze visual data, specifically in highly dedicated fields such as biomedicine, denotes a new frontier in the use of AI. The field of AI has seen a successful track in model development during the past ten years. We categorize them into four groups, as shown in Fig. 1. The

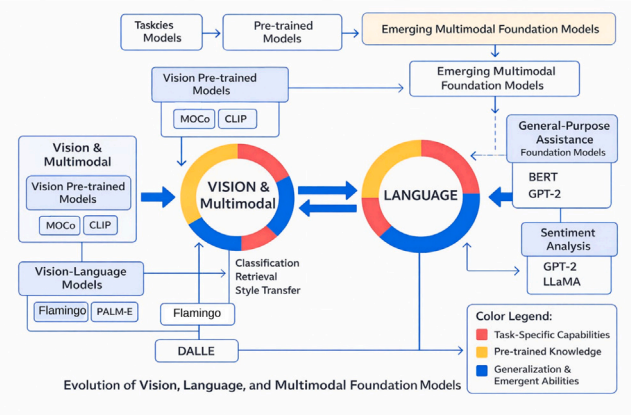

Fig. 1. An illustrative example of the developmental trajectory of foundation models integrating vision and language within a multimodal framework. The first category represents task-specific models, while the remaining three correspond to foundation models. Language and vision foundation models are depicted using green and blue blocks, respectively. Key characteristics of representative models are highlighted for each category. We hypothesize that multimodal foundation models will evolve along a trajectory similar to vision and language models, progressing from pre-trained models designed for specialized tasks toward unified models and general-purpose assistants. However, as multimodal models such as GPT-4 and Gemini remain largely proprietary, the optimal design paradigm is still unclear and is therefore indicated by a question mark.

GPT-4 series and Gemini Vision series models (referred to as GPT-4 and Gemini for the remainder of this article) [31] are noteworthy examples of these developments, as they seamlessly combine the processing of linguistic and visual information.

Following the success of LLMs like ChatGPT and GPT-4 in NLP, researchers aim to develop advanced models for vision, VL, and multimodal tasks, allowing machines to combine and interpret visual and textual data for more versatile applications. Since the introduction of BERT, vision pre-training and VLP have gained significant attention [32]. These methods have become the foremost framework for vision learning, allowing models to develop universal visual and VL representations or to produce incredibly realistic images. Similar to BERT/GPT-2 in the language sector, they could be regarded as the first generation of multimodal foundation models [33,34].

Many demanding tests are part of the assessment process, which proposes to evaluate the models' precision, effectiveness, and versatility in deciphering and using visual data for biomedical applications. This study demonstrates the pros and cons of each model by comparing the performance of GPT-4 and Gemini in tasks like data synthesis [35], anomaly detection, and medical image classification. In addition to demonstrating the revolutionary possibilities of incorporating cutting-edge AI models like GPT-4 and Gemini into biomedical analysis, this study establishes a standard for further investigation in the field. By contrasting these models, the study adds insightful information to the continuing discussion about improving AI's multimodal abilities [34], particularly in fields where joining textual and visual data is important. To our knowledge, this is the first comprehensive study to examine the image analysis of VLMs in the medical domain. Overall, contributors highlighted critical design requirements for the usability and acceptance of VLM concepts, even as they recognized their value. The function of VLMs in healthcare is thoroughly reviewed in this paper, as shown in Fig. 2. The main contributions of this paper are as follows.

- Overview of VLM use cases to assist with healthcare workflows and offer information on how valuable these ideas are thought to be.

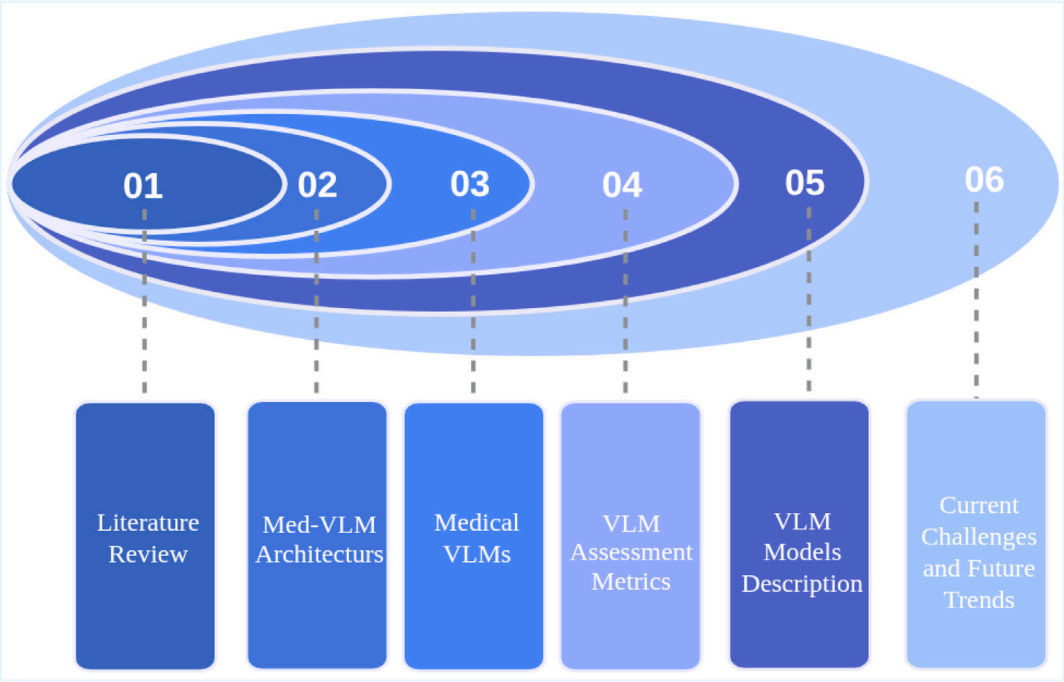

Fig. 2. Overall structure of the paper.

- We exhibit our method's effectiveness by analyzing various downstream tasks, such as medical visual question answering (QA), RG, and image-report retrieval.
- Concentrate on the challenges and design principles associated with VL models in healthcare environments.
- Discuss design implications and future research directions for integrating VLM capabilities into healthcare.

2. Literature review

2.1. Background of VLMs

Large-scale pre-trained VLMs are widely used across many fields, refining tasks like image captioning, VQA, and multimodal search [13, 36]. For instance, CLIP [24] leverages a vast dataset of 400 million image-text pairs sourced from the internet, effectively aligning visual and textual encoders. Pre-trained object detectors (ODs) are often used to obtain visual characteristics in past work. Models like ViLBERT [37] and LXMERT [38] pay attention to multimodal fusion, using transformers to process text and region features separately before combining them. In contrast, integrated attention fusion is used in VisualBERT [39], VL-BERT [40], and UNITER [41], where both text and region features are input into a single transformer. Enhanced with extra image tags, models like OSCAR [42], and VinVL [43] display leading-edge performance on various VL tasks. However, region feature extraction is time-consuming, and pre-trained ODs remain static during pretraining, limiting the potential of VLP models.

2.2. Medical Visual Language Models (Med-VLMs)

Med-VLMs are crucial for automating visual-language tasks like VQA and IRG in the medical field. These models [44–46] first use CNNs and RNNs to extract linguistic and visual characteristics individually. However, because of their structural restrictions, such methods often fail to be generalizable and transferable to other tasks [47]. The transformer architecture, which utilizes a pretraining-finetuning paradigm, is mainly accepted by current Med-VLMs [48]. These models undergo pre-training on huge and diverse medical image-text datasets before being fine-tuned on specific task-oriented datasets for more precise applications. Exploiting the ALBEF [49] methodology, MISS [50], for instance, starts training on 38,800 cautiously chosen image-text pairs from the MedICaT dataset [51] before being fine-tuned for VQA tasks. Comparable to this, LLaVA-Med Fig. 3 uses a two-phase pretraining tactic [52], concluding in full-scale fine-tuning for VQA jobs after first positioning image-text features on two million pairs from PubMed and then improving conversational abilities using instruction format data. For task adaptation, these methods always use full-model fine-tuning, which is an effective technique but requires a lot of resources, especially for large-scale models like LLaVA-Med, and the model's generalizability is further threatened by the limited dataset sizes accessible for downstream task training, which could result in catastrophic forgetting and decrease its extensive applicability in medical contexts.

Several recent survey papers have explored the landscape of VLMs in the medical domain. For example, Li et al. [53] provide a comprehensive overview of VLM architectures and training paradigms across seven core tasks, such as classification, segmentation, and report generation. Yang et al. [54] focus on the integration of multimodal clinical data with large language models (LLMs), highlighting their potential in decision support systems. Xu et al. [55] categorize visual-language datasets and present a taxonomy of challenges in VQA and captioning for medical images. Meanwhile, Zhang et al. [29] emphasize the role of contrastive learning and foundation models in aligning medical images with radiology reports. A comparative summary of these reviews is presented in Table 1, highlighting their core tasks, focus areas, dataset evaluation strategies, and ethical considerations. Compared to these works, our review centers on the real-world clinical applicability of medical VLMs, particularly in downstream tasks like VQA, RG, and image-report retrieval. We also offer a focused discussion on model explainability, ethical implications, and evaluation frameworks areas often underrepresented in earlier reviews. This human-centered perspective aims to bridge the gap between algorithmic advances and practical integration into clinical workflows.

Table 1
Comparison of survey papers on Medical VL Models (VLMs).

Paper/Focus	Core tasks covered	Key focus areas	Dataset evaluation	Explainability & Ethics	Unique contributions
Li et al. (2025) [53]	Classification, Segmentation, Report Generation, etc.	VLM architectures and training paradigms	Yes	Limited	Comprehensive technical overview, but less emphasis on clinical use
Yang et al. (2024) [54]	Multimodal clinical data integration	Integration of multimodal clinical data with LLMs for decision support	Moderate	Some focus	Focused on clinical decision support using LLMs
Xu et al. (2023) [55]	VQA, Captioning	Dataset categorization and challenge taxonomy	Yes	Minimal	Dataset- and challenge-centric taxonomy of VLM applications
Zhang et al. (2023) [29]	Image-Report Alignment	Contrastive learning and foundation models	Yes	Limited	Emphasis on image-report alignment using contrastive pretraining
This Paper (2025)	VQA, Report Generation, Image-Report Retrieval	Real-world clinical applicability, explainability, ethics, evaluation frameworks	Yes	Strong focus	Human-centered perspective bridging algorithmic advances with practical clinical integration

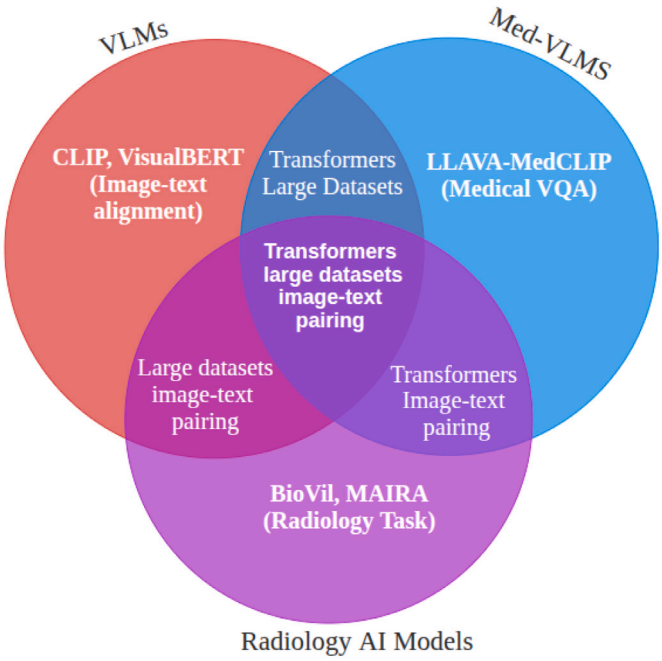

Fig. 3. Illustrating the relationships between VLMs, Med-VLMs, and Radiology AI Models.

2.3. Limitations and advances in medical VLMs

Pre-trained visual language models in the medical domain are constrained by the complexity of medical reports and the scarcity of large-scale medical image-text datasets. ConVIRT, for instance, aligns medical image and text data for incompatible learning [56]. MedCLIP [57] reduces false negatives by decoupling the strict connection between texts and images, expanding the training data pool. Recent strategies, including Med-UniC [58], MedKLIP [59], and DiViDe [60], integrate medical knowledge during model training, but they still lack fine-grained discriminative knowledge necessary for diagnostic tasks. Before the emergence of advanced language models, materials scientists and biologists used simpler machine-learning techniques for tasks like microstructural classification identification [61,62], material property forecasting [63,64], and segmentation [65,66], notable examples include few-shot transfer learning for electron backscatter diffraction (EBSD) pattern classification [67] and semi-supervised few-shot learning for the segmentation of oxide material systems [68].

2.4. Multimodal AI models in radiology

LLMs generate statistically probable continuations of text, demonstrating their ability to handle medical concepts. Advanced LLMs like PaLM 2 [69], FLAN-T5 [70], and LLaMA 2 [71], especially those fine-tuned on medical data (e.g., Med-PaLM 2 [72]), excel in healthcare tasks like medical question-answering, clinical note summarization, and chatbot applications. LLMs have been integrated with medical images to apply multimodal healthcare data better [73]. Examples include BioViL(-T) and MAIRA, which associate radiology images with textual reports. These models enable tasks including text-image reacquisition, VQA [74], report fault detection, and automatic report generation, significantly improving diagnostic accuracy and efficiency

2.5. Prompt design

In knowledge-intensive medical fields, training area-specific language models (LM) on professional knowledge-augmented corpora is crucial. Recent trends involve querying LM in few-shot or zero-shot fashions to extract information (e.g., from WordNet). Techniques like auto-prompt generation [75], and context optimization [76] facilitate the production of knowledge-rich prompts, enhancing the effectiveness of VLMs for subsequent tasks.

3. Mainstream Med-VLM architectures

Modern generative Med-VLMs, whether large- or small-scale, typically follow a similar architectural framework, as illustrated in Fig. 5. This structure consists of a language model (LM) head, a connector module for integrating visual and textual representations, a visual encoder, and a text encoder. Most Med-VLMs employ Vision Transformer (ViT) [77] as the visual feature extractor and adopt text encoders based on architectures such as BERT [78] or GPT [33]. Despite implementation differences, these models share a common transformer-based foundation, including attention mechanisms, layer normalization, and feed-forward neural networks (FFN). A generic architectural overview is presented in Fig. 4.

3.1. Multimodal AI in healthcare

The increasing demand for high-quality healthcare services and limited medical resources highlight the need for intelligent AI-driven solutions. Recent advancements in medical AI have focused on developing foundation models [29,79,80]. Medical VLMs integrate multiple modalities, such as language, images, and clinical data, enabling improved understanding of complex medical conditions compared to unimodal approaches [79,81]. These models support various healthcare applications, including diagnosis, treatment planning, and patient

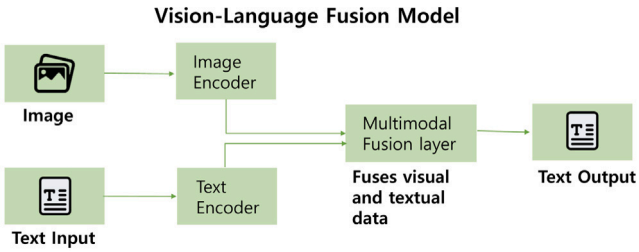

Fig. 4. General architecture of mainstream Medical Vision-Language Models (Med-VLMs), highlighting the visual encoder, text encoder, connector module, and language model head.

management, by leveraging complementary information from different data sources [82]. With the rapid development of large-scale models, advanced Med-VLMs such as Med-Gemini [83] have demonstrated strong potential in multimodal medical analysis. Medical VLMs can generally be categorized into multimodal fusion models and multimodal generation models [84].

3.1.1. Multimodal fusion models (MFM)

CLIP [6] is the foundational work upon which multimodal alignment models are built. Exploiting contrastive learning methods, CLIP is a model designed for image-text matching, generating a unified depiction for both textual and visual inputs. Several studies have attempted to apply this contrastive training formulation to align various modalities in the medical field [57,85]. Furthermore, many studies additionally develop multimodal models for segmentation [86–88], explainable diagnosis [89,90], structured report generation [91,92], etc., based on these well-aligned medical CLIP models. Despite their strong performance in numerous downstream works, these multimodal models (MM) are not adaptable enough to be employed in clinical settings due to a deficiency of medical knowledge and collaboration.

3.1.2. Multimodal generation

Conversely, the generative multimodal foundation models can fully utilize the promising area knowledge from their LLM component and offer unstructured text reports. These models were established particularly for the medical field and are primarily based on MM [93]. These models provide enhanced interaction, along with better few-shot and zero-shot learning capabilities compared to MFM. They can autonomously make medical image reports that include appropriate normal results and abnormalities, and consider the patient's medical history. These models can aid clinicians even more by combining interactive visualizations, like underscoring the region each phrase describes, with text reports. This differentiation between multimodal generation and alignment models provides a basis for understanding their usefulness in various healthcare areas. Regarding medical modalities, the aforementioned multimodal VLMs demonstrate capabilities in the following areas of healthcare:

Imaging analysis and diagnosis: In medical imaging diagnostics, VLMs effectively associate imaging and textual data. For example, by learning from extensive medical imaging datasets and related diagnostic reports, VLMs can accelerate the recognition [94] of diseases like pneumonia and cancer. These models automate the interpretation of CT and MRI scans, incorporating imaging data with patient medical histories and symptoms to yield more precise diagnostic results [95], as illustrated in Fig. 6. These models automate the examination of MRI and CT scans and combine imaging analyses with patient medical histories and symptom information to provide diagnostic insights.

Healthcare Literature and Records Mining: VLMs can analyze and interpret large volumes of research papers, electronic health records, and medical literature, improving medical knowledge mining and decision-making. These models quickly sort and evaluate the most current research findings in medical research, assisting researchers and healthcare professionals in understanding innovative [95] therapeutic approaches and diagnostic methods. The automatic extraction and analysis of important data from patient records by VLMs enhances clinical decision-making.

Advanced Medical Devices: VLMs contribute to the advancement of smart robotic systems and medical tools. For instance, by combining cross-modal information, VLMs improve the accuracy and safety of robotic surgeries in complex procedures [96,97]. Intelligent devices analyze real-time data and images during surgery [98], providing accurate assistance and supervision to minimize surgical risks.

Drug Development Assistance: By analyzing vast amounts of biomedical data and streamlining the drug design process, VLMs in drug growth forecast the effectiveness and potential side effects of novel medications. To enhance the efficacy and success rate of new drug development [99], these models integrate information from drug compounds, protein structures, genetics, and other modalities.

Remote Healthcare and Diagnostics: VLMs hold significant potential in remote healthcare. These models offer high-quality healthcare services to remote areas with limited medical resources by combining textual, audio [95], and video data to enable remote diagnosis and treatment. VLMs can merge imaging and medical record data to provide accurate diagnostic recommendations and analyze real-time communication between physicians and patients. However, most popular medical VLMs are still trained on specific medical conditions or fields, limiting their generalizability and preventing them from functioning as multimodal models that is used for several conditions or tasks [100]. Additionally, they often face challenges due to the small scale of training or fine-tuning datasets, and they must also address general issues like privacy risks and inefficient computational power. Thus, a careful and measured approach is required when considering the growth potential of VLMs in healthcare.

3.2. Training methodology

3.2.1. Transfer Learning (TL)

Transfer Learning (TL) is a widely adopted machine learning (ML) paradigm where a model trained on a source domain is adapted to perform a different but related task on a target domain. This approach is especially useful when the target task has limited data availability. TL typically involves fine-tuning the parameters of a pre-trained model using a smaller dataset corresponding to the new task [101]. Mathematically, given a source domain $\mathcal{D}_S$ with a corresponding task $\mathcal{T}_S$, and a target domain $\mathcal{D}_T$ with its task $\mathcal{T}_T$, transfer learning aims to improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ using the knowledge from $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

$$\text{Objective:} \quad \min_{\theta_T} \mathcal{L}_T\big(f_T(x_T;\theta_T)\big), \quad \text{where } \theta_T \leftarrow \theta_S \tag{1}$$

Here, θ_S represents the parameters learned from the source task, and these are used as initializations (rather than random values) for the target task model parameters θ_T. The loss function $\mathcal{L}_T$ is specific to the target task. This enables the model to converge faster and generalize better, especially with limited target data [102]. The architecture of the original model may also be altered to fit the target task, such as by adding task-specific classification or regression layers. The core idea is to retain the generalized knowledge embedded in the pre-trained model and adapt it efficiently to the new task requirements.

3.2.2. Curriculum Learning (CcL)

CcL is an advanced training strategy inspired by the human learning process, where simpler concepts are mastered before moving on to more difficult ones. It introduces a structured progression in training by ordering the data or tasks based on difficulty, enabling the model to learn more effectively and generalize better [103].

Formally, let $D = (x_i, y_i)_{i=1}^{N}$ be the training dataset. In traditional training, the data is presented in a random order. However, in CcL,

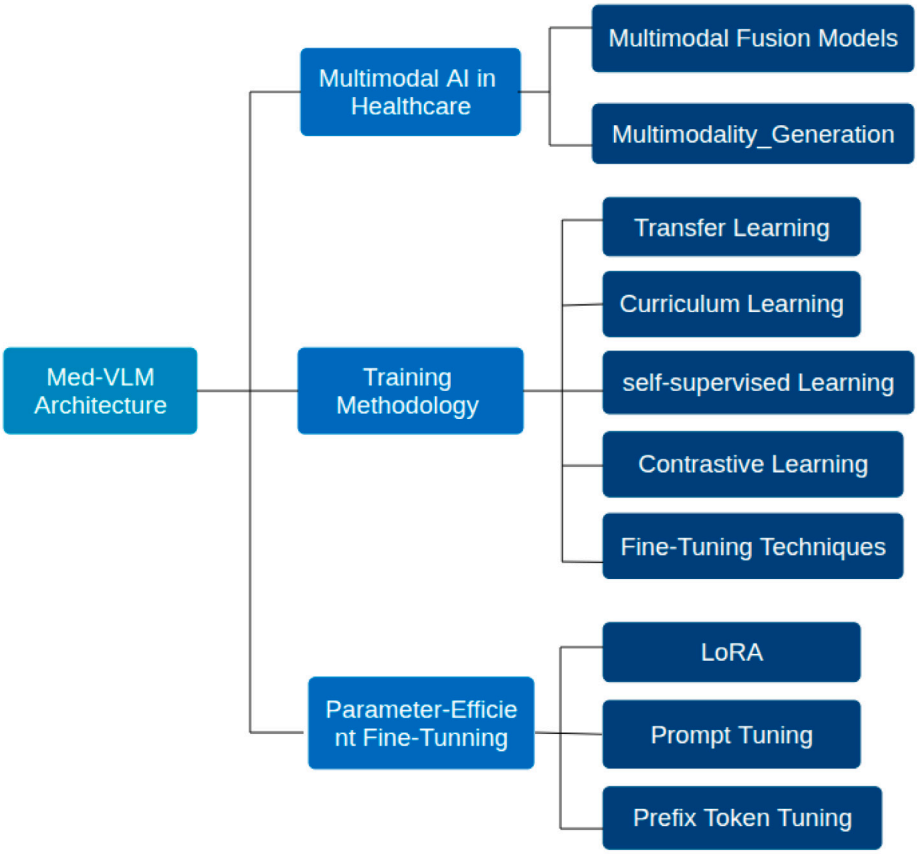

Fig. 5. General architectures of Med-VLM.

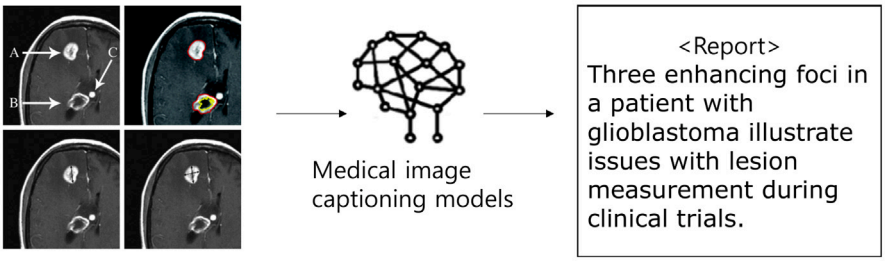

Fig. 6. A medical image captioning model drafts a diagnostic report from the image.

we define a difficulty scoring function $s(x_i, y_i)$ that assigns a difficulty level to each sample. The training proceeds in stages by gradually introducing more difficult examples:

$$D_1 \subset D_2 \subset \cdots \subset D_k = D, \quad \text{where } D_j = (x_i, y_i) \mid s(x_i, y_i) \leq \tau_j \tag{2}$$

Here, τ_j is a threshold that increases at each stage j, and D_j contains all samples with difficulty scores below that threshold.

This incremental training approach allows the model to first establish a solid understanding of easy examples before being exposed to harder ones. For instance, LLaVa-Med [52], a state-of-the-art medical VLM, utilizes CcL by initially training on simple image-text pairs and gradually introducing more complex clinical scenarios. This structured learning process improves model convergence, robustness, and adaptability to complex medical tasks.

3.2.3. Self-Supervised Learning (SSL)

SSL is a crucial paradigm in VLM training, providing an efficient substitute for supervised learning. It enables the models to learn useful representations by directly generating labels from the data [104]. This method is especially helpful when gaining labeled data is either complex, time-consuming, or costly. The models in SSL for VLMs produce staff that take advantage of the data's natural structures, which allows them to acquire remarkable illustrations over modalities without the need for direct external classifications. SSL tasks contain contrastive learning, masked image and language modeling.

A common SSL objective used in multimodal models is masked modeling, where random tokens in the input (either visual patches or text tokens) are masked and the model must predict the masked content. The loss function is defined as:

$$\mathcal{L}_{\text{Mask}} = \frac{1}{|M|} \sum_{i \in M} \mathcal{L}_{\text{recon}}(f(x)_i, x_i) \tag{3}$$

Here, M is the set of masked positions, x_i is the ground truth value, and $f(x)_i$ is the model prediction at that position. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is typically cross-entropy for language and mean squared error for image patches. This objective enables the model to learn contextual representations by reconstructing missing elements from the input. This objective encourages the model to learn embeddings where paired modalities (e.g., aligned image and caption) are close together,

Table 2
Comparison of parameter-efficient fine-tuning vs. Fine-tuning approaches for VL models.

Aspect	Parameter-efficient fine-tuning	Fine-tuning
Goal	Enhance the efficiency of a pre-trained VLM for specific VL tasks using minimal data and computational resources.	Improve a pre-trained VLM's performance on a specific task with extensive data and compute resources.
Training data	Works well with small, task-specific datasets (e.g., vision-text pairs).	Requires large-scale datasets to achieve fine-tuning objectives.
Training time	Faster due to selective parameter updates (adapters or LoRA layers).	Slower because all model parameters are updated during training.
Computational resources	Requires fewer resources; most layers of the VLM remain frozen.	Demands higher computational power for full-layer training.
Model parameters	Modifies only a small fraction of parameters, such as adapters or prefixes.	Modifies each parameter across the visual and textual feature extraction and cross-attention modules.
Pretrained knowledge retention	Better retention of general VL understanding due to minimal changes.	May risk overfitting or losing generalized knowledge from pretraining.
Flexibility across tasks	Easily adaptable to multiple tasks by adding task-specific adapters.	Task-specific fine-tuning requires retraining the entire model for new tasks.
Training performance	Achieves competitive results; ideal for resource-constrained environments.	Generally, yields superior task performance but at a higher cost.

while unpaired (non-matching) samples are pushed apart in the latent space.

3.2.4. Contrastive Learning (CL)

CL enables models to learn meaningful representations by contrasting positive and negative pairs of multimodal data, such as images and corresponding textual descriptions [105]. The objective is to map semantically aligned pairs closer together in a shared embedding space while pushing unaligned (negative) pairs farther apart, thereby facilitating more effective representation learning. Positive pairs typically consist of an image and its correct caption, whereas negative pairs are composed of mismatched image-text combinations [106].

To achieve this, contrastive loss functions are employed, most notably the InfoNCE (Noise-Contrastive Estimation) loss [107]. InfoNCE is designed to increase the similarity between positive pairs while decreasing the similarity between all other (negative) pairs in the batch. This technique is used in models such as CLIP [6], which applies InfoNCE loss with cosine similarity to align vision and language modalities. ALIGN [28] further leverages a normalized SoftMax variant of this loss to enhance training efficiency and scalability. Both approaches aim to maximize the similarity of matching image-text embeddings while minimizing that of non-matching pairs. The InfoNCE loss function with cosine similarity is mathematically defined as:

$$\mathcal{L}\text{InfoNCE} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum k = 1^N \exp(\text{sim}(z_i, z_k)/\tau)} \tag{4}$$

where z_i and z_j are the normalized embeddings of the image and its corresponding text (positive pair), $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter controlling distribution sharpness, and the denominator sums over all image-text pairs in a batch of size N, including negative samples.

3.2.5. Fine-tuning techniques

After pretraining, VLMs are usually adjusted on specialized datasets to improve their capability for particular applications Table 2.

Supervised Fine-Tuning (SFT) The VLM is initially trained on a large-scale image-caption dataset, enabling it to learn rich and generalizable associations between visual and textual modalities. Following this pretraining, SFT is performed on a more narrowly focused, domain-specific dataset tailored to the intended downstream task. This two-stage training approach, broad pretraining followed by targeted fine-tuning, allows the model to leverage broad prior knowledge while adapting efficiently to the specific requirements of particular applications [108].

Mathematically, given pretrained model parameters θ_P, fine-tuning updates these parameters to θ_F by minimizing a task-specific supervised loss $\mathcal{L}_F$ over the fine-tuning dataset $\{(x_i, y_i)\}_{i=1}^M$:

$$\theta_F = \arg\min_\theta \frac{1}{M} \sum_{i=1}^{M} \mathcal{L}_F\big(f(x_i;\theta), y_i\big) \quad \text{with} \quad \theta \leftarrow \theta_P \tag{5}$$

Here, $f(x_i;\theta)$ denotes the model prediction for input x_i with parameters θ, and y_i is the ground-truth label for supervised learning. The initialization from pretrained parameters θ_P allows the fine-tuning to converge faster and generalize better on the target task.

Human-supervised Reinforcement Learning (HSRL): A distinctive approach for the task-specific training of VLM is Reinforcement Learning with Human Feedback (RLHF), which integrates human feedback into the fine-tuning process to enhance model performance [108, 109]. RLHF begins with an initial pretrained model and builds a reward model based on human-generated rankings of its outputs. Unlike conventional reinforcement learning (RL) [110], which relies solely on interactions with an environment, RLHF incorporates human preferences directly into the learning loop. This human-in-the-loop strategy provides a more nuanced and informed framework, enabling VLMs to be better aligned with human values and ultimately producing optimized results that are more useful and trustworthy.

The RLHF training objective can be formalized as maximizing expected reward under the policy π_θ, parameterized by model parameters θ:

$$\max_\theta \ \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[r_\phi(s,a)\right] \tag{6}$$

where s is the state (e.g., input prompt), a is the model output (action), and r_ϕ is the learned reward function trained on human feedback. The reward model r_ϕ guides the model toward outputs preferred by humans, thus aligning the model's behavior with human intent.

Fine-Tuning Instruction (IFT): IFT involves improving a pre-existing language model by providing it with specific directives or guidelines tailored to the target task or application [111]. Typically, this process exposes the model to prompts or examples aligned with the intended instructions and adjusts its parameters based on feedback during fine-tuning. An example in the medical VLM domain is RaDialog [112], which utilizes this fine-tuning approach to enhance model performance and outcomes.

The fine-tuning objective can be formulated as minimizing the expected loss over the instruction-following dataset:

$$\min_\theta \ \mathcal{L}_{IFT}(\theta) = \mathbb{E}_{(x,y)\sim D_{IFT}}\left[\ell\big(f(x;\theta), y\big)\right] \tag{7}$$

where θ are the model parameters, D_{IFT} is the instruction dataset of input–output pairs (x, y), f is the model, and ℓ is the loss function

(e.g., cross-entropy). This objective ensures the model learns to generate outputs consistent with given instructions, improving task-specific performance.

3.2.6. Parameter-Efficient Fine-Tuning (PEFT)

This section reviews methods that adapt Vision-Language Models (VLMs) by updating only a small subset of parameters rather than fully fine-tuning the entire network. PEFT encompasses various strategies designed to optimize parameter usage during fine-tuning, especially when labeled data for the target task is limited, as summarized in Table 2. The core idea is to insert task-specific modules, known as adapters [113], into a pretrained model without altering its original parameters. These adapter modules typically use a bottleneck architecture, projecting features into a lower-dimensional space, applying a non-linear transformation, and then projecting back to the original dimensionality. This design ensures parameter efficiency by adding only a small number of trainable parameters per task. Adapters are placed after each layer of the pretrained model, allowing the model to learn task-specific representations while preserving previously acquired knowledge [114].

LoRA (Low-Rank Adaptation): This is a widely used module-based approach for adjustment of pre-trained models [115]. Two smaller matrices, a breakdown version of the larger weight matrix in a previously trained model must be fine-tuned as part of the adaptation process. These smaller matrices form the LoRA adapter, and the approach focuses on applying low-rank modifications to the model, allowing it to adapt to a specific task without altering the original model's structure. Lora is usually employed to fine-tune pre-existing LLMs, which play a key role in medical VLMs frameworks, such as RaDialog [112] and Visual Med-Alpaca [116]. LoRA introduces low-rank updates to the pretrained weight matrix W_0, such that the adapted weight is

$$W = W_0 + \Delta W = W_0 + AB, \tag{8}$$

where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ are trainable matrices with $r \ll \min(d, k)$, representing the low-rank update. This allows fine-tuning only the small matrices A and B while keeping W_0 fixed.

Prompt Optimization: Continuous feature illustrations are created as feed cues for Prompt Optimization [117]. This permits the model to produce efficient prompts during training. The potential of models to create correct and contextually relevant responses is greatly upgraded by ongoing prompt improvement. The model can adjust its behaviors based on the increasing understanding of the task through an iterative process. Prompt tuning was employed by VLMs such as InstructBLIP and Qwen-VL [93,118]. In prompt tuning, continuous prompt embeddings P of length m are prepended to the input tokens, and only P is optimized during training. The input sequence embedding becomes:

$$X' = [P; X], \tag{9}$$

where X is the original token embedding sequence, and $[;]$ denotes concatenation.

Prefix Token Tuning: Prefix token tuning sets the behavior of models for a task by adding targeted task features to the data, namely to the primary identifier known as Leading tokens [119]. For example, questions from numerous datasets were given different prefixes in VL-T5 [120]. While the other parameters in the pre-trained model remain constant, these features can be independently trained and updated. Task-specific adaptation is made possible through prefix token tuning, which preserves the previously learned information encoded in most of the model's parameters. Prefix tuning prepends learnable prefix tokens P to the input and modifies the key and value vectors inside the transformer layers. Formally, for each transformer layer, the prefix tokens generate key–value pairs (K_p, V_p) added to the original attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q[K_p; K]^T}{\sqrt{d_k}}\right)[V_p; V], \tag{10}$$

Beyond conventional PEFT approaches such as LoRA, prompt tuning, and prefix tuning, several emerging strategies have been proposed to further enhance the adaptability, robustness, and domain-awareness of medical VLMs. These methods go beyond simple parameter reduction and focus on domain alignment, cross-task generalization, and stability under distribution shifts.

BiomedCoOp: Cooperative prompts interact and share information across multiple medical downstream tasks instead of operating independently [121]. Unlike conventional prompt tuning, where a separate prompt must be trained for each task, BiomedCoOp enables shared semantic representation, allowing prompts trained for classification to support report generation or captioning without extensive retraining. This leads to improved multi-task reasoning, particularly useful in radiology pipelines where disease identification, description, and report synthesis occur sequentially. It is especially advantageous when labeled data is scarce, as knowledge transfer between tasks reduces annotation demand. However, cooperative prompts require careful balancing; over-sharing across unrelated tasks may dilute task-specific knowledge and cause interference.

Generalized Domain Prompt Learning (GDPL): GDPL [122] develops domain-level prompts instead of dataset- or task-specific prompts. This contrasts with standard prompt tuning, which is highly sensitive to domain variations and often needs retraining for every new dataset. GDPL encodes biomedical domain priors into a unified prompt space, enabling improved cross-hospital generalization and efficient scaling to new imaging modalities (X-ray → MRI → CT). This significantly reduces retraining costs and enhances performance when deployed across heterogeneous clinical environments. Its limitation lies in domain granularity, highly specialized disease types may still require adaptation, as a single domain prompt cannot fully capture all sub-pathological nuances.

PromptSmooth: PromptSmooth [123] tackles robustness issues present in conventional prompt tuning, which tends to overfit to the training distribution. Instead of learning a single deterministic prompt, it generates a smoothed distribution of prompts that encourages the model to explore diverse representations, making it more stable against noise, scanner variability, and demographic shifts. This method is particularly beneficial for real-world clinical deployment where domain shift is unavoidable. Experiments show improved anomaly detection and diagnostic consistency on unseen hospital datasets. A noted limitation is increased training complexity; sampling prompt distributions requires additional computation compared to fixed prompts.

Med-VLM: Med-VLM [124] uses multimodal prompts that combine textual medical knowledge with visual priors extracted from the image. While conventional prompting treats language and vision independently, Med-VLM fuses anatomical cues (e.g., organ boundaries, lesion shape) with textual clinical semantics to improve segmentation and localization. This makes it effective for detecting subtle or small anomalies such as micro-lesions or polyps, where purely textual prompts fail to guide the model. It excels in medical image segmentation and report-grounded lesion highlighting. The limitation is dependency on high-quality visual priors, poor segmentation initialization or noisy images may weaken multimodal alignment.

Quaternion Prompt Learning: Unlike Euclidean prompt embeddings, Quaternion prompts [125] encode four-dimensional feature interactions, enabling richer spatial-semantic coupling. This is particularly useful in complex 3D medical imaging (MRI, CT) or multi-sequence radiology workflows where anatomical orientation and volumetric relationships matter. Quaternion prompts enhance feature fusion, improving lesion localization across slices and reducing modality mismatch. They have shown superior transferability in cross-modal VLM adaptation. However, quaternion operations introduce higher implementation complexity and require careful parameter initialization to avoid instability during training.

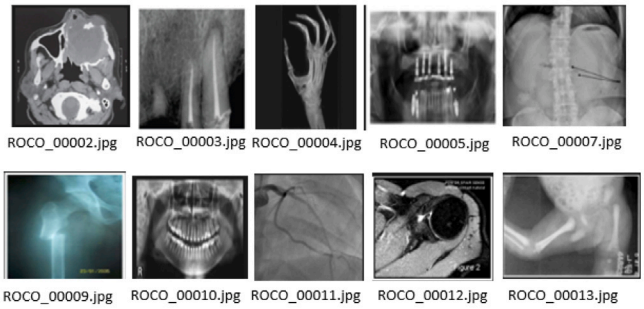

Fig. 7. Radiology images from the ROCO dataset for medical VL tasks [128].

4. Medical VLMs

Our main focus in this work is to apply the knowledge from vision-language (VL) models to the medical domain [126]. To achieve this, we investigate a range of detection tasks in medical settings and explore strategies to effectively transfer knowledge from VLMs trained on natural images. In particular, we focus on medical images and related datasets to enable lesion detection under zero-shot and fine-tuning scenarios, integrating expert-level knowledge with image-specific information. The considered datasets are summarized in Table 4 and illustrated in Fig. 13.

4.1. Medical datasets for VLMs

VLMs are pre-trained and fine-tuned using task-specific datasets to support various medical applications. Table 1 presents publicly available VL datasets containing medical VQA pairs, which serve as key resources for pretraining and evaluation in report generation (RG) and VQA tasks. Further details of widely used medical VL datasets are summarized in Table 3.

4.1.1. Dataset summary

Radiology Objects in Context (ROCO): The ROCO dataset consists of image–caption pairs collected from PubMed Central (PMC), an open-access biomedical literature repository [127]. It is divided into two subsets: radiology and out-of-class. The radiology subset contains over 80,000 images spanning multiple imaging modalities, including CT, X-ray, MRI, ultrasound, PET, and angiography, as illustrated in Fig. 7. The out-of-class subset includes a smaller collection of diverse images such as clinical photographs, synthetic images, and non-radiology content, as shown in Fig. 8.

Each image in the dataset is accompanied by textual metadata, including captions, keywords, and Unified Medical Language System (UMLS) concept identifiers and semantic types, enabling rich multimodal representation learning. Example samples with associated descriptions are presented in Fig. 9.

The dataset is typically divided into training, validation, and test sets following an 80/10/10 split. Due to its large scale, modality diversity, and structured annotations, ROCO serves as a valuable resource for training and evaluating vision-language models in medical image understanding tasks, particularly in scenarios involving multimodal learning and semantic representation.

Indiana University Chest X-ray (IU X-ray):

The IU X-ray dataset is a publicly available benchmark widely used in medical imaging research. It consists of 7470 chest X-ray images paired with 3955 corresponding radiology reports. Each report is structured into multiple sections, including impression, findings, tags, comparison, and indication, providing comprehensive clinical context for each image. The structured nature of these reports enables effective alignment between visual and textual modalities, making the dataset particularly suitable for tasks such as medical report generation, image-text retrieval, and vision-language modeling. Its detailed annotations allow for fine-grained supervision in multimodal learning settings. However, the dataset is relatively limited in scale and diversity compared to larger datasets such as MIMIC-CXR, which may restrict its generalization capability in large-scale applications. Despite this limitation, IU X-ray remains a valuable resource for benchmarking and developing multimodal models in healthcare.

Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR): The MIMIC-CXR dataset comprises 227,835 free-text radiology reports associated with 377,110 chest X-ray images [129]. The data is derived from de-identified radiographic studies conducted at Beth Israel Deaconess Medical Center. It includes imaging studies with one or more images, typically in Digital Imaging and Communications in Medicine (DICOM) format, and provides both lateral and posteroanterior views. Due to its large scale and rich clinical content, MIMIC-CXR is widely used for training and evaluating vision-language models in medical applications. However, challenges remain due to the unstructured nature of reports, variability in medical terminology, and class imbalance across pathologies, which may require appropriate preprocessing for effective model development.

MIMIC-CXR-JPG: The pre-processed MIMIC-CXR dataset [129] is known as MIMIC-CXR-JPG [132]. This version consists of compressed JPG files derived from the original 377,110 images. Labels for multiple pathologies are added to the associated 227,827 reports. These labels are generated using NegBio [144] and CheXpert [145] by analyzing radiology reports. Its main advantage lies in ease of use for deep learning applications due to the standardized image format. However, conversion from DICOM to JPG results in loss of resolution and removal of important metadata, while reliance on rule-based NLP tools may introduce label noise.

MIMIC-NLE: The MIMIC-NLE dataset is designed to generate natural language explanations (NLEs) for medical image predictions, particularly for chest X-ray analysis and thoracic pathologies [146]. It includes 38,003 image–NLE pairs and 44,935 image–diagnosis–NLE combinations. The NLEs are extracted from MIMIC-CXR [129] radiology reports and focus on anteroposterior (AP) and posteroanterior (PA) X-ray views. Each NLE is associated with diagnosis and evidence labels. The dataset is divided into training, validation, and test sets. However, its relatively smaller size and the subjective nature of natural language explanations remain key limitations.

MIMIC-III: MIMIC-III is a large, publicly available database containing anonymized health records of over 40,000 patients admitted to critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012 [147]. It includes diverse clinical data such as demographics, laboratory results, procedures, medications, imaging reports, clinical notes, mortality records, and vital signs [148]. The dataset supports a wide range of research applications, including clinical decision-making, epidemiology, and healthcare analytics. However, it lacks direct linkage with image data, which limits its standalone use in vision-language tasks.

MIMIC-IV: MIMIC-IV is a relational database that encompasses detailed records of hospital stays for patients admitted to a tertiary academic medical center in Boston, USA [149]. It includes comprehensive information such as laboratory results, prescriptions, vital signs, and other clinical data. Designed to support a broad range of healthcare studies, MIMIC-IV builds upon MIMIC-III and introduces several enhancements [150]. However, similar to MIMIC-III, it does not natively include chest X-ray images, which limits its direct applicability in vision-language modeling. Linking it with imaging datasets such as MIMIC-CXR can enhance its utility for multimodal analysis.

CXR with Previous Citations Removed (CXR-PRO): The CXR-PRO dataset is derived from MIMIC-CXR [129] and consists of 374,139 impression-only radiology reports with associated chest radiographs [151]. These reports exclude references to prior imaging studies to reduce hallucinated citations in medical report generation. This design

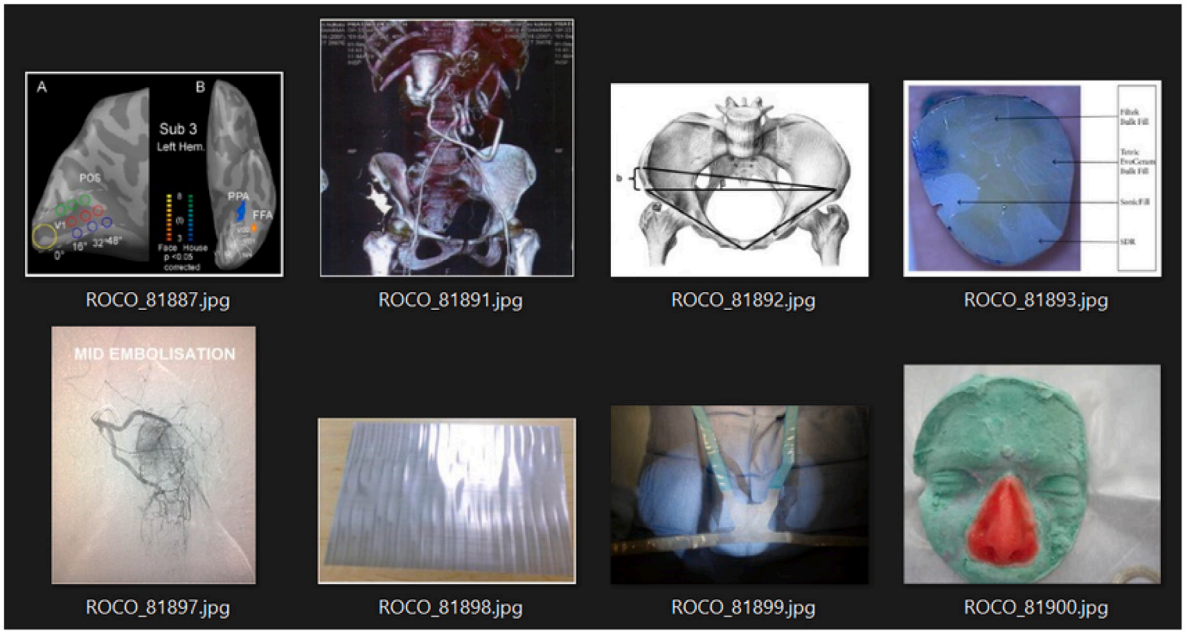

Fig. 8. Non-radiology images from the ROCO dataset for VL research [128].

Table 3
This table provides a detailed description of widely used medical VL datasets, including information on image count, size, annotation types, class counts, and data availability. Class counts are approximate and have been inferred from related publications, dataset repositories, and other sources due to the absence of explicitly defined class labels in some datasets.

Dataset	No. of images	Size	Format	Classes	Annotations	Split	Storage	Availability
ROCO [130]	79,789	Varying	JPG, PNG	12	Captions	70/15/15	Not Specified	Public
IU X-ray [131]	7470	1024 × 1024	JPG	14	Image + Reports	Not Specified	2 GB	Public
MIMIC-CXR-JPG [132]	377,110	256 × 256	JPG	14	None (Images Only)	Predefined	150 GB	Public
MIMIC-NLE [133]	500k	Text Only	–	17* (UMLS entities)	NER labels	Not Specified	Not Specified	Public
MIMIC-II [134]	26,870	Mixed	CSV, TXT	75 (ICD-9 codes)	Structured + Text	Not Specified	32 GB	Public
MIMIC-IV [135]	524k	Mixed	CSV, TXT	82 (ICD-10 codes)	Structured + Notes	Not Specified	100 GB	Public
CXR-PRO (No. Verified)	20,000+	Not Specified	PNG/JPG	14	BBoxes + Labels	Not Specified	Not Specified	Not Public
PadChest [136]	160,000	Varying	PNG	174	Labels + BBoxes	Not Specified	300 GB	Public
MedICap [137]	1,000+	Varying	Mixed	18 (modalities	terms)	Captions + Labels	Not Specified	Public
MS-CXR [138]	377,110	256 × 256	JPG	14	Image-text pairs	Predefined	150 GB	Public
SLAKE [139]	6000	Varying	JPG	8	QA pairs	Not Specified	Not Specified	Public
VQA-RAD [140]	315	Varying	JPG	11	QA pairs	70/10/20	500 MB	Public
PMC [141]	227k pairs	Varying	PNG	20+ (terms from captions)	Captions + Labels	Not Specified	Not Specified	Public
VQA-Med [142]	4200+	Varying	JPG	15	QA pairs	Predefined	Not Specified	Public
PathVQA [143]	4998	Varying	JPG	22	QA pairs	Predefined	Not Specified	Public

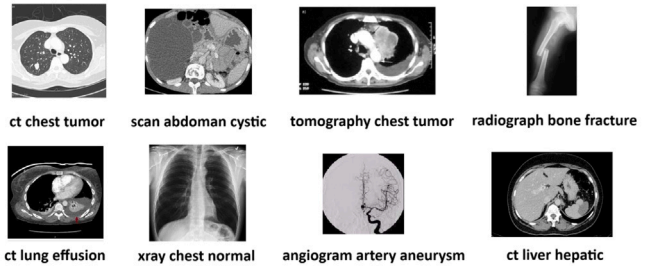

Fig. 9. Sample photos from the data collection with their equivalent three-word descriptions [128].

provides cleaner inputs for training models to generate concise diagnostic summaries. However, the absence of detailed findings sections and incomplete pairing between reports and images may limit effective vision-language alignment.

PadChest: Chest X-ray Dataset with Multi-label Annotations: The PadChest dataset [136] comprises over 160,000 high-resolution chest X-ray images from approximately 67,000 patients. The dataset includes multiple projection views and is accompanied by demographic and acquisition metadata [152]. It provides extensive annotations, including anatomical locations, differential diagnoses, and radiographic findings mapped to a standardized taxonomy using UMLS. A portion of the data is manually annotated by radiologists, while the remainder is labeled using automated methods. PadChest is widely used for multi-label classification, disease localization, and segmentation tasks. However, class imbalance and the presence of automated annotations may introduce label noise.

MedICap: Medical Image and Caption Dataset: The MedICap dataset consists of 217,061 images extracted from open-access PMC articles, covering diverse medical domains including radiology [51]. It integrates data from the S2ORC corpus [153] and ROCO figures [127], resulting in figure-caption pairs derived from compound figures, as illustrated in Fig. 10. MedICap supports tasks such as image-text alignment and caption generation due to its large scale and diversity. However, variability in caption quality and the weakly supervised nature of the dataset may limit performance in fine-grained clinical applications.

Table 4
Medical VL model training datasets.

Dataset	Year	Image-Text pairs	Language
ROCO [130]	2018	87,952 (81,825 radiology, 6127 out-of-class)	English
IU X-ray [131]	2017	7470 chest X-ray images with 3955 reports	English
MIMIC-CXR-JPG [132]	2019	377,110 images with 227,827 reports	English
MIMIC-NLE [133]	2022	38,003 image-NLE combinations	English
MIMIC-III [134]	2016	–	English
MIMIC-IV [135]	2023	–	English
CXR-PRO (No. Verified)	2022	374,139 impression sections	English
PadChestt [136]	2020	160,000+ chest X-ray images	Spanish
MedICap [137]	2020	217,060 images from 131,410 PMC articles	English
MS-CXRR [138]	2022	1162 image-text pairs	English
SLAKE [139]	2021	642 images with 14,028 QA pairs	English, Chinese
VQA-RAD [140]	2018	3515 QA pairs	English
PMC [141]	2022	15,282,336 image-caption pairs	English
VQA-Med (2019–2021) [142]	2019–2021	15,292 QA pairs, 5000 images, 5500 pairs	English
Pathology VQA (PathVQA) [143]	2019	4998 images with 32,799 QA pairs	English

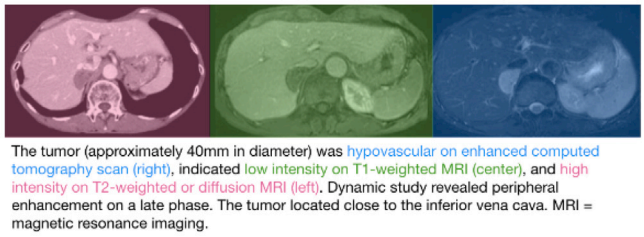

Fig. 10. Because the subfigures in this figure are referenced by spatial position in a right-to-left order, subfigure-to-subcaption alignment is difficult. Color designates corresponding subfigures and subcaptions.
Original source: [154]; from where the figure and caption are adapted.

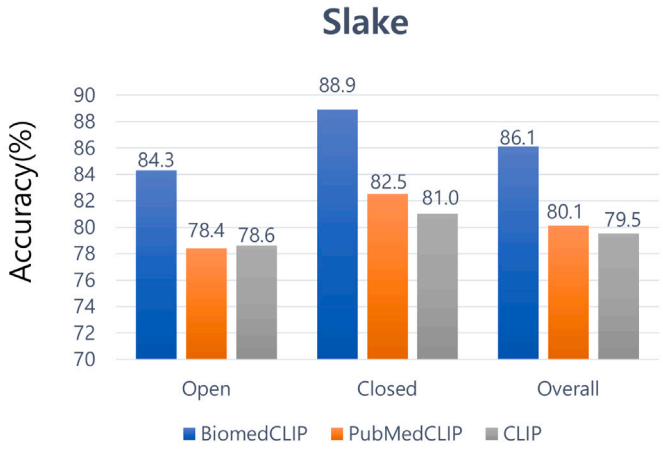

Fig. 11. Accuracy comparison of BiomedCLIP, PubMedCLIP, and CLIP across different tasks (Open, Closed, and Overall) on the SLAKE dataset.
Original source: [118]; from where the figure and caption are adapted.

Multi-Scale Chest X-ray (MS-CXR): The MS-CXR dataset includes 1162 image–report pairs, each annotated with bounding boxes indicating relevant pathological findings verified by board-certified radiologists [155]. It focuses on eight cardiopulmonary conditions, including atelectasis, consolidation, cardiomegaly, pneumonia, edema, lung opacity, pneumothorax, and pleural effusion. The dataset integrates data from MIMIC-CXR [129] and REFLACX [156]. MS-CXR is well-suited for tasks involving fine-grained visual grounding and report generation due to its high annotation precision. However, its relatively small size limits its standalone applicability in large-scale deep learning models [157], making it more suitable for fine-tuning and evaluation purposes.

Pathology Visual Question Answering (PathVQA): The PathVQA dataset is a domain-specific visual question answering (VQA) benchmark designed for pathology applications. It contains 4998 pathology images and 32,799 question–answer pairs curated to simulate realistic diagnostic queries. The images are collected from pathology textbooks and the PEIR Digital Library [143]. Each image is annotated with multiple QA pairs covering a range of clinical reasoning tasks, including abnormality detection, object presence, attribute recognition, counting, positional relationships, tissue identification, and clinical inference [158]. This diverse categorization enables comprehensive evaluation of a model's ability to interpret and reason over medical visual data. By incorporating both simple and complex question types, PathVQA supports the development of multimodal models that combine visual understanding with medical knowledge [159].

Semantically-Labeled Knowledge-Enhanced (SLAKE): The SLAKE dataset [139] is a bilingual (English and Chinese) medical dataset comprising 642 annotated images covering multiple diseases and anatomical regions. Each image is enriched with bounding boxes for object detection and segmentation masks for detailed visual understanding [160]. The dataset includes 14,028 question–answer pairs, consisting of both open-ended and closed-ended questions, as illustrated in Fig. 11. In addition, SLAKE provides structured knowledge in the form of triplets (head, relation, tail), capturing relationships between medical entities such as organs and diseases. These annotations support tasks involving knowledge-based reasoning and visual grounding. While SLAKE offers rich semantic annotations and multilingual support, its relatively small size and limited modality diversity restrict its applicability for large-scale training.

VQA-RAD: The VQA-RAD dataset [161] consists of medical images, including chest X-rays, abdominal CT scans, and head CT/MRI slices, selected to represent unique patients. Each image is accompanied by expert-verified captions describing modality, anatomical plane, and diagnostic findings. The dataset contains 3515 question–answer pairs, with multiple questions per image, including both structured and unstructured formats as well as linguistically varied rephrasings. The questions are categorized into open-ended and closed-ended types, covering a broad range of clinical reasoning tasks, as shown in Table 5. VQA-RAD serves as an important benchmark for evaluating multimodal models in medical VQA, supporting research in image understanding and diagnostic reasoning. However, its relatively small size and limited diversity in imaging modalities and disease coverage may restrict generalizability in large-scale applications (see Fig. 12).

PubMed Central (PMC): PMC-15M is a large-scale dataset of paired biomedical images and corresponding text, extracted from full-text scientific articles available in PubMed Central (PMC), a publicly accessible biomedical repository. As of June 2022, over 4.4 million full-text articles were available, from which more than 15 million image–caption pairs were curated. Each article package typically includes XML files, PDFs, and associated media, enabling the extraction of diverse visual

Table 5
Analysis of medical VL Datasets: Advantages, limitations, and proposed solutions.

Dataset	Advantages	Limitations	Proposed Solutions (Augmentation/Preprocessing)
ROCO [130]	Diverse biomedical content including radiology and out-of-domain data	Imbalance between radiology and other domains; limited QA pairs	Use modality-specific sampling; generate QA pairs using LLMs; augment out-of-domain samples using generative models.
IU X-ray [131]	Balanced image-report pairs; used widely in research	Small size; outdated modality spectrum	Relabel with modern lexicons; use text summarization for concise inputs.
MIMIC-CXR	Large-scale paired dataset; real-world clinical reports	Unstructured and noisy reports; possible image redundancy	Normalize report structures using section labeling; remove duplicates; use weak supervision for image filtering.
MIMIC-CXR-JPG [132]	Easy-to-use JPG format; mirrors MIMIC-CXR content	Loss of detail due to JPG compression; same limitations as MIMIC-CXR	Apply contrast enhancement and noise robustness training; align with DICOM versions if available.
MIMIC-NLE [133]	Introduces natural language explanations (NLEs)	Small scale; limited clinical diversity	Generate synthetic NLEs via prompt-tuned GPT; expand to multi-institutional datasets.
MIMIC-II [134]	Rich EHR data for multi-modal research	No image-text pairing; purely textual	Pair with public radiographs using diagnosis codes; apply table-to-text generation for synthetic captions.
MIMIC-IV [135]	More recent and richer than MIMIC-III	Still lacks native image-text pairs	Link with image banks using timestamps/ICD codes; use structured-to-natural language techniques.
CXR-PRO (No. Verified)	Large impression-only dataset	Lacks full reports and images	Combine with other CXR image sets; generate full reports from impressions via text expansion.
PadChest [136]	Multilingual dataset; annotated with pathologies	Reports in Spanish; limited cross-lingual usability	Translate reports with clinical-domain MT; normalize to UMLS terms.
MedICap [137]	Large captioned dataset from scientific articles	Captions may be noisy or unrelated to image content	Filter captions using NLP similarity measures; group similar captions using clustering.
MS-CXR [138]	Carefully curated image-text pairs	Very small size; limited variability	Use GANs to generate synthetic X-rays; oversample underrepresented pathology classes.
SLAKE [139]	Multilingual QA pairs; visual grounding elements	Small number of images; uneven QA quality	Expand QA with GPT-based multilingual prompts; augment images via transformations.
VQA-RAD [140]	First medical VQA dataset; clinically relevant QA	Limited QA volume; lacks diversity	Use VQA template generation; integrate additional X-rays from MIMIC-CXR.
PMC [141]	Massive scale; covers a wide range of biomedical topics	High noise in captions; varied image quality	Clean captions with QA-filtering; resample from underrepresented journals.
VQA-Med [142]	Annual VQA challenges; benchmarks QA performance	Small image base; inconsistent QA styles across years	Normalize QA format; synthesize diverse questions using LLMs.
PathVQA [162]	Specialized in pathology; extensive QA per image	Image complexity high; QA pairs manually limited	Apply image enhancement; generate QA with few-shot prompting.

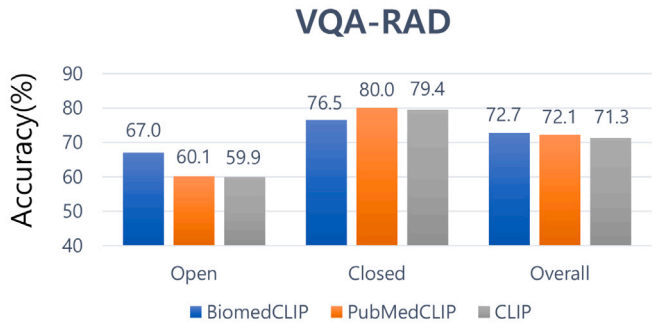

Fig. 12. Accuracy comparison of BiomedCLIP, PubMedCLIP, and CLIP across different tasks (Open, Closed, and Overall) on the VQA-RAD dataset.
Original source: [118]; from where the figure and caption are adapted.

and textual content. While PMC resources have been widely used to pretrain biomedical language models such as PubMedBERT [163] and BioGPT [164], this dataset extends their applicability to the vision-language domain by integrating textual information with biomedical imagery [165]. It includes a wide range of image types, such as digital pathology, microscopy images, radiographs (e.g., X-rays, MRIs, and CT scans), and schematic figures like charts and graphs [166]. Notably, PMC images often exceed standard resolutions used in conventional vision models, and their captions can be significantly longer than typical input limits in models such as CLIP [167], posing challenges for model design. To address this, images are linked with semantic concepts using curated biomedical taxonomies within embedding spaces such as BiomedCLIP [29]. Although variability in data quality and modality exists, the scale and diversity of PMC-15M make it a valuable resource for large-scale pretraining of biomedical vision-language models.

VQA-Med (2019–2021): The VQA-Med dataset, developed between 2019 and 2021, provides a structured benchmark for Visual Question Answering (VQA) in the medical domain, particularly radiology. In 2019, the dataset included 15,292 QA pairs associated with 4200 radiology images sourced from the publicly available MedPix database [142]. The training set comprised 12,792 QA pairs and 3200 images, with each image paired with 3–4 questions, while the validation and test sets each included 500 images and corresponding QA sets. The questions focused on key aspects such as imaging modality, anatomical system, imaging plane, and pathological abnormalities. In 2020, the dataset expanded to 5000 radiology images and QA pairs [168], maintaining a similar structure with 4000 QA pairs for training and 500 each for validation and testing. This version emphasized defect identification and introduced a subset dedicated to Visual

Table 6
Summary of VQA-Med Datasets (2019–2021).

Year	Total Images	QA Pairs	Training (Images, QA)	Validation (Images, QA)	Test (Images, QA)
2019	4200	15,292	3200, 12,792	500, –	500, –
2020	5000	5000	4000, 4000	141, –	80, –
2021	5500	5500	4500, 4500	85, 200	100, –

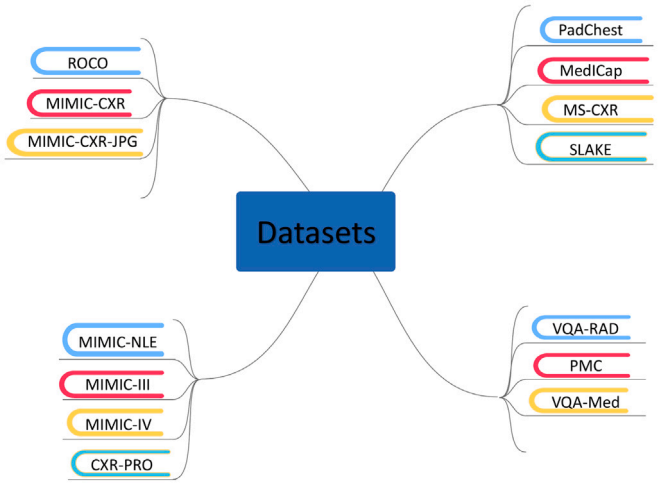

Fig. 13. Comprehensive list of datasets utilized in this survey.

Question Generation (VQG), containing 2156 questions and 780 training images. In 2021, the dataset was further enriched with 5500 QA pairs and corresponding images [169]. The training portion contained 4500 image–QA pairs, with an emphasis on 17 distinct abnormalities. Like its predecessors, the dataset supported both VQA and VQG tasks.

The detailed dataset splits and statistics across all versions are summarized in Table 6. The progressive development of VQA-Med highlights its growing utility in training and evaluating models for medical image understanding and clinical reasoning. Its consistency in structure and increasing complexity across years make it valuable for longitudinal research. However, limitations such as the restricted image sources (primarily MedPix), relatively small dataset size, and domain-specific question formulation may constrain its generalizability to broader medical imaging applications.

4.2. Clinical deployment and real-world considerations

Even though vision-language models have evolved very fast in their use in the medical field, they are not implemented in actual clinical practice. The majority of the systems available are now being tested with retrospective studies and controlled benchmarks, and few have advanced to pilot clinical tests. Among the major challenges are the requirements of regulatory approvals, including the adherence to standards implemented by agencies such as the FDA, and the issues of data privacy, data security and patient confidentiality. Besides that, the implementation of VLMs into the current hospital workflow has practical challenges, such as the interoperability with electronic health records (EHR) systems, inconsistency in imaging protocols, and the necessity to perform effective generalization among heterogeneous patient groups [170]. Model predictions are also important issues because the interpretability and reliability of the prediction must be transparent and trustworthy to healthcare professionals to have them adopt the models. Even though the latest research has shown positive findings in medical report generation, visual question answering, and diagnostic assistance, more research is required to close the gap between research and implementation. The future implementation should be directed at prospective clinical validation, compliance with regulations, and the creation of user-centric systems that could be easily introduced into clinical practice.

5. VLM assessment metrics

This section explores the evaluation of medical VLMs, focusing on the critical early steps of selecting appropriate standard datasets and identifying suitable assessment metrics tailored to specific VL tasks. Choosing the right evaluation metrics is essential for accurately measuring model performance, interpreting results, and guiding further improvements, as shown in Fig. 14. Different tasks, such as image captioning, VQA, and retrieval, require distinct metrics that balance importance and limitations based on clinical relevance and computational feasibility. The detailed overview of these evaluation metrics, their significance, constraints, and applicability to various medical VLM tasks is summarized in Table 8.

5.1. Tasks and calculations

The conventional VLM is made to receive an incorporated image-text prompt as input and created a text response as output. The prompt, [171], for instance, “What is the image’s modality?” may cause a reaction. This image exhibits a brain MRI. The key core of any assessment is the model’s capability to react to diverse question prompts, which in turn shows the range of tasks the model has the potential to complete. We measure the VLM on 6 tasks: visual QA, report summarization [172], QA, image classification, RG, and natural language inference (NLI) to offer a complete calculation of its abilities. We explain in detail the prompt design, datasets, and performance metrics that MultiMedEval applies for assessment in individual tasks. The evaluation criteria and associated metrics for these tasks are summarized in Table 9. For each dataset, with the exclusion of MIMIC-III and Pad-UFES-20, which limits an official divide, we utilize the split. We suggest employing 20% of the Pad-UFES 20 dataset for testing, and we utilize the split suggested by [173] for MIMIC-III. Except for MIMIC-CXR, the model forecasts the answer for all datasets by assessing the BLEU score among every class and the model’s reaction, then opting for the category that scored the highest points. The answer is labeled CheXBert [174] for the MIMIC-CXR dataset. CheXBert is a report-labeling technology that derives 14 environments, of which we keep 5 settings (Edema, pleural effusion, consolidation, cardiomegaly, and atelectasis) for calculating the metrics. We describe the classification performances applying macro AUROC, macro F1, and macro accuracy, once the classes are obtained. The evaluation criteria and associated metrics for these tasks are summarized in Table 7.

5.1.1. Bilingual Evaluation Understudy (BLEU)

The BLEU score is an automatic evaluation metric originally developed for machine translation by quantifying the overlap of n-grams between a system-generated text and one or more reference texts [175]. The fundamental component of BLEU is the n-gram precision, defined as the ratio of matched n-grams in the generated text to the total n-grams present in that output:

$$\text{Precision}(n) = \frac{\text{Number of overlapping } n\text{-grams}}{\text{Total number of } n\text{-grams in generated text}} \tag{11}$$

where overlapping n-grams are those that match with at least one n-gram in the reference text. To avoid inflating precision due to repeated

Table 7
Evaluation criteria and metrics for Vision-Language Model (VLM) tasks.

Task	Dataset(s)	Prompt/Input type	Evaluation metrics
Visual Question Answering (VQA)	VQA-RAD [161], VQA-Med [142]	Image + Question prompt [171]	BLEU [175], ROUGE [176], Accuracy
Report Summarization	MIMIC-CXR [132], ROCO [177]	Radiology report	ROUGE [176], BLEU [175], Perplexity [178]
General QA	MedQA [179], PubMedQA [141]	Text question	Accuracy, F1-score
Image Classification	MIMIC-CXR (CheXbert labels) [174]	Image only	Macro AUROC [180], Macro F1-score, Macro Accuracy
Report Generation (RG)	MIMIC-CXR [132], IU X-ray [181]	Image	BLEU [175], ROUGE [176], BERTScore [182], CheXbertSim [183]
Natural Language Inference (NLI)	MedNLI [184]	Sentence pairs	Accuracy, F1-score

Table 8
Evaluation metrics: Importance, limitations, and suitability for medical VL tasks.

Metric	Importance in medical imaging	Limitations	Useful for medical imaging?
BLEU [175]	Common metric for language generation; measures n-gram overlap between generated and reference reports.	Ignores clinical correctness; sensitive to wording variations; poor correlation with medical relevance.	Limited
ROUGE [176]	Measures recall-oriented overlap, often for summarization tasks; useful to gauge coverage of key phrases.	Fails to capture clinical accuracy or importance of critical findings; insensitive to semantic equivalence.	Limited
METEOR [185]	Includes synonym and paraphrase matching, improving over BLEU/ROUGE for semantic similarity.	Struggles with clinical nuances; cannot evaluate correctness of clinical facts.	Limited
Perplexity [186]	Originates from speech recognition and language modeling; it measures how well a probabilistic model predicts a sample.	Does not evaluate clinical relevance or correctness; low perplexity indicates fluent text but not medically accurate content.	No
BERTScore [182]	Uses contextual embeddings to capture semantic similarity; better semantic matching than traditional metrics.	Does not guarantee clinical correctness or capture diagnostic accuracy.	Moderate
CheXbert Vector Similarity [145]	Designed for chest X-ray reports; compares disease mentions and medical concepts between texts.	May miss subtle contextual or non-disease errors; dataset-specific limitations.	Yes
RadGraph F1 [187]	Evaluates extraction of clinical entities and relations; reflects correctness of structured medical information.	Requires detailed annotation; limited generalizability outside annotated datasets.	Yes
Human Evaluation [188]	Direct assessment of clinical correctness and usefulness by medical experts.	Resource-intensive, subjective, lacks standardization, often limited sample size.	Yes
RadCliQ [189]	Combines clinical correctness and linguistic quality; designed for clinical report evaluation.	Emerging metric requires more validation across datasets and scenarios.	Promising

n-grams, a clipping technique is applied: restrict the count of any n-gram in the generated text to its maximum occurrence in a reference sentence. This prevents redundancy from skewing the evaluation. The final BLEU score incorporates a geometric mean of these clipped precisions across up to N n-gram lengths, commonly set to 4, combined with a brevity penalty (BP) to penalize overly short outputs: Table 10 presents QA tasks which are used in image classification and report summarization, and achieved a value of 40%. Here is the equation for BLEU-n:

$$\text{BLEU-n} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \cdot \log\left[\text{precision}(n)\right]\right) \tag{12}$$

$$\text{BP} = \begin{cases} 1 & \text{if } c \geq r, \\ e^{(1-\frac{r}{c})} & \text{if } c < r, \end{cases} \tag{13}$$

where c is the length of the generated text and r is the reference length. BLEU scores range from 0 to 1, with higher values indicating closer alignment with the reference text. BLEU is widely used across natural language generation (NLG) tasks such as QA [190], image captioning, and medical report summarization [191] due to its simplicity, computational efficiency, and consistency as a benchmarking tool. It reliably measures surface-level lexical similarity and discourages excessively short outputs. Nonetheless, its reliance on exact n-gram matches limits its ability to recognize paraphrased or semantically equivalent expressions. Additionally, its effectiveness diminishes with fewer or less diverse reference texts, potentially leading to misleading evaluations. Hence, while BLEU remains a foundational metric for text generation evaluation, it is advisable to use it alongside metrics that better capture semantic correctness and contextual understanding, especially in domains that require nuanced language, such as medical applications.

Clinical Relevance and Limitations: Although BLEU is widely used in report generation, its lexical matching nature poses limitations in medical text evaluation. A model may generate clinically incorrect findings while still achieving a high BLEU score due to surface-level n-gram overlap. In radiology or medical reporting, factual correctness (e.g., "no pneumothorax" vs. "mild pneumothorax present") is significantly more important than wording similarity. BLEU also fails to evaluate whether the generated report identifies key abnormalities or supports clinical decision-making. Therefore, in medical settings, BLEU should not be used alone; it must be complemented with clinically grounded metrics such as RadGraph-F1 or CheXbert similarity alongside expert radiologist review [175].

5.1.2. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE is a widely adopted set of evaluation metrics designed to compare automatically generated text, especially summaries, against high-quality human-authored reference summaries [176]. It measures the overlap of lexical units such as unigrams, bigrams, and sequences between the generated and reference texts. ROUGE metrics are particularly recall-oriented, emphasizing how much of the reference information is captured in the system output [176]. These metrics are extensively used across various NLP applications, including question answering, machine translation, headline generation, and abstractive text summarization.

Key ROUGE metrics:

ROUGE-N: This metric evaluates the overlap of n-grams between the generated and reference summaries. It is typically used with $n = 1$ (ROUGE-1) for unigram overlap and $n = 2$ (ROUGE-2) for bigram overlap. ROUGE-N measures how many of the reference n-grams are captured in the generated text:

$$\text{ROUGE-N} = \frac{\text{Number of matching } n\text{-grams}}{\text{Total } n\text{-grams in reference summary}} \tag{14}$$

A higher ROUGE-N score indicates better lexical overlap, which correlates with content preservation.

ROUGE-L: This metric is based on the Longest Common Subsequence (LCS) between the generated and reference summaries. Unlike ROUGE-N, ROUGE-L accounts for word order without requiring strict contiguity, capturing fluency and grammatical coherence:

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{\text{Length of reference summary}} \tag{15}$$

Here, $\text{LCS}(X, Y)$ denotes the length of the longest common subsequence between the generated summary X and the reference summary Y. ROUGE-L is especially useful for evaluating summary structure and sequence fidelity.

ROUGE-S (Skip-Bigram): This variant evaluates the co-occurrence of skip-bigrams, any pair of words that appear in the same order in both summaries, not necessarily adjacent. It captures non-contiguous phrasal matches that can reflect more flexible paraphrasing:

$$\text{ROUGE-S} = \frac{\text{Number of skip-bigram matches}}{\text{Total skip-bigrams in reference}} \tag{16}$$

The choice of the ROUGE metric should be aligned with the specific objectives of the task at hand. ROUGE-1 is particularly effective for evaluating content adequacy, making it suitable for extractive summarization or scenarios where capturing the surface-level content of the reference is paramount. ROUGE-2, by measuring bigram overlaps, offers a more refined evaluation by capturing the preservation of short dependencies and key phrases, which is especially useful in tasks such as machine translation and dialogue generation. In contrast, ROUGE-L, which is based on the longest common subsequence, is more appropriate for assessing fluency, grammatical structure, and sentence-level coherence, making it well-suited for abstractive summarization. Meanwhile, ROUGE-S, which measures skip-bigram matches, is advantageous for detecting semantic overlap in paraphrased or loosely structured outputs where flexibility in word positioning is permitted.

Although ROUGE provides an interpretable and standardized framework for text evaluation, it has limitations. It is inherently lexical and does not capture deeper semantic meaning, leading to an underestimation of quality in paraphrased or conceptually accurate outputs with divergent wording. Moreover, performance can vary significantly depending on the number and diversity of reference summaries available [191]. Thus, ROUGE should be complemented with semantic evaluation metrics such as METEOR, BERTScore, or human judgment, especially in complex domains like medical report generation, legal summaries, or conversational AI [192].

Clinical Relevance and Limitations: While ROUGE effectively measures recall and content overlap, it does not capture diagnostic correctness or clinical reasoning. In medical reports, two sentences may share many overlapping words but contradict each other clinically. For instance, "no consolidation observed" and "consolidation present" may have comparable ROUGE overlap yet lead to opposite diagnoses. Due to this limitation, ROUGE alone is insufficient for evaluating medical text generation. Consequently, clinical validation through physician assessment, entity-based metrics, and real-case deployment trials is critical to ensure trustworthiness in healthcare applications [176].

5.1.3. Metric for Evaluation of Translation with Explicit ORdering (METEOR)

The METEOR metric, introduced by Banerjee and Lavie in 2005 [185], was developed to address some of the shortcomings of earlier evaluation metrics like BLEU, particularly by placing a greater emphasis on meaning preservation and fluency in the generated text. Unlike BLEU, which relies primarily on precision of n-gram overlaps, METEOR incorporates both precision and recall, and introduces a penalty term to account for fragmented matches, resulting in a more balanced and interpretable evaluation.

The METEOR score is calculated using the harmonic mean of unigram precision P and recall R, with recall weighted higher, and is further adjusted by a fragmentation penalty. The formula is:

$$\text{METEOR} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \cdot (1 - \text{Penalty}), \tag{17}$$

where precision and recall are defined as:

$$R = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in the reference}}, \tag{18}$$

$$P = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in the candidate}}. \tag{19}$$

The penalty term penalizes disordered or scattered matches and is computed as:

$$\text{Penalty} = \frac{1}{2}\left(\frac{\text{Number of chunks}}{\text{Number of matched unigrams}}\right)^3, \tag{20}$$

where a *chunk* refers to a sequence of unigrams that appear in both the generated and reference texts in the same order and without interruption. This penalization ensures that more fluent and coherent outputs are rewarded. The final METEOR score ranges from 0 to 1, where higher values reflect better alignment in both content and structure between the generated and reference texts. METEOR is specifically valued in tasks like machine translation and summarization because of its ability to accommodate synonyms, stemmed variations, and flexible word ordering, offering a more semantically sensitive evaluation than purely lexical methods. **Clinical Relevance and Limitations:** Even though METEOR considers synonyms and generates more meaning-aware evaluation than BLEU/ROUGE, it still lacks mechanisms to validate medical findings, pathology correctness, or relationships between anatomical structures. A report may receive a high METEOR score despite missing life-critical observations like hemorrhage or misidentifying anatomical locations. Therefore, METEOR should be applied with caution in medical report generation and ideally combined with clinically-oriented metrics and human evaluation by radiologists or medical experts [193].

5.1.4. Perplexity

Perplexity is a widely adopted evaluation metric in NLP, particularly for assessing the performance of language models. Quantifies the degree of uncertainty a model has in predicting the next word in a sequence: Lower perplexity values indicate greater confidence and, generally, better predictive accuracy [186]. This makes it a fundamental tool in evaluating models for tasks such as text generation, machine translation, and speech recognition.

Formally, given a sequence of words $\omega_1, \omega_2, \ldots, \omega_n$ and a language model that assigns a probability distribution P, perplexity is defined as:

$$\text{Perplexity} = 2^{H(P)}, \quad \text{where } H(P) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 p(\omega_i) \tag{21}$$

Here, $H(P)$ denotes the entropy of the probability distribution, and $p(\omega_i)$ is the predicted probability of the word ω_i. Perplexity essentially corresponds to the exponential of the average negative log-likelihood per word. A lower value implies that the model is more confident in its predictions, indicating stronger language modeling capabilities.

While perplexity is a strong indicator of syntactic fluency and internal consistency in a language model, it does not directly account for semantic adequacy or alignment with human judgment.

Therefore, its applicability is most appropriate during model development phases, such as pretraining and fine-tuning. For downstream evaluation, particularly in tasks like summarization or translation, semantic-aware metrics (e.g., ROUGE, METEOR) are often used in conjunction with perplexity to offer a more comprehensive assessment.
Clinical Relevance and Limitations: A low perplexity score indicates strong language fluency but does not guarantee medical correctness or diagnostic accuracy. A model could generate grammatically perfect but clinically dangerous statements. Perplexity evaluates probability distribution quality, not whether the model identifies diseases correctly or avoids hallucinating conditions. For real-world deployment, perplexity must be complemented with clinical evaluation frameworks, factual consistency checks, and domain-specific safety assessments.

5.1.5. BERTScore

BERTScore is a semantic similarity metric for evaluating NLG tasks such as report summarization and VQA. Unlike lexical overlap-based metrics, BERTScore leverages contextual embeddings from pretrained language models (e.g., BERT) to compute similarity at the token level [182]. In this study, the BERTScore attained a value of 45%, reflecting its ability to measure semantic alignment between generated and reference text.

Embedding representation. Each token in the candidate and reference sentences is embedded using contextual representations from a pretrained BERT model. Let C_i denote the embedding of the ith token in the candidate sentence and R_j the embedding of the jth token in the reference sentence.

Cosine similarity. To measure semantic similarity between tokens, cosine similarity is computed for all pairs of candidate and reference embeddings:

$$\text{sim}(C_i, R_j) = \frac{C_i \cdot R_j}{\|C_i\| \|R_j\|} \tag{22}$$

where $\cdot$ denotes the dot product and $\|\cdot\|$ is the L_2-norm of the embedding vector.

Precision. For each token in the candidate sentence, the similarity with every token in the reference sentence is computed. Precision is defined as the average of the maximum similarity scores for each candidate token:

$$P = \frac{1}{|C|} \sum_{i=1}^{|C|} \max_j \text{sim}(C_i, R_j) \tag{23}$$

Recall. Similarly, recall is calculated by taking the average of the maximum similarity scores for each reference token when compared to all candidate tokens:

$$R = \frac{1}{|R|} \sum_{j=1}^{|R|} \max_i \text{sim}(C_i, R_j) \tag{24}$$

F1-based BERTScore. The final BERTScore is computed as the harmonic mean of precision and recall, providing a balanced measure of semantic overlap from both candidate and reference perspectives:

$$\text{BERTScore} = \frac{2 \times P \times R}{P + R} \tag{25}$$

BERTScore has become increasingly important due to its ability to capture deep semantic relationships between texts, overcoming the limitations of traditional lexical metrics that rely solely on exact word matching. This makes it especially valuable in evaluating tasks where paraphrasing, synonymy, and context play significant roles, such as abstractive summarization and dialogue generation. Moreover, BERTScore correlates better with human semantic similarity judgments, providing a more meaningful assessment of language generation quality. However, despite its advantages, BERTScore has some limitations. It requires substantial computational resources due to large pretrained models and can be sensitive to the choice of the underlying language model and tokenization. Additionally, while it captures semantic similarity well, it may not fully account for factual correctness or logical consistency in generated text, which are critical in certain applications. Thus, BERTScore should ideally be complemented with other task-specific metrics or human evaluation for a comprehensive assessment.

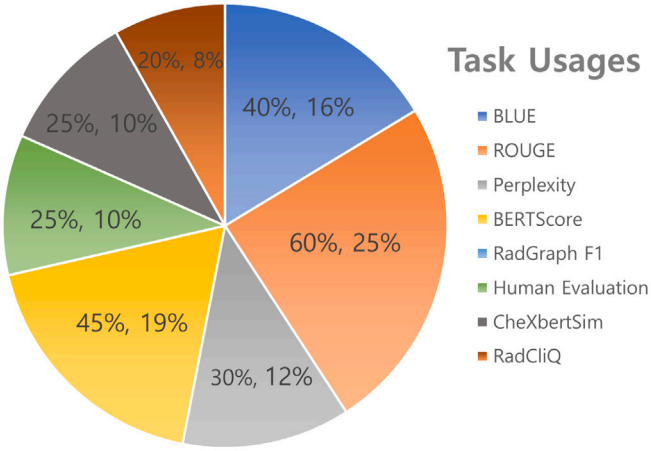

Fig. 14. Distribution of evaluation metrics used across tasks. Note: Each metric is represented by two values (x%, y%), where x% indicates the proportion of task usage (i.e., how frequently the metric is employed across evaluated tasks), and y% reflects its visual proportion in the pie chart (i.e., area occupied in the figure). For example, "60%, 25%" denotes that the metric is used in 60% of tasks, while it occupies 25% of the chart space for clarity and readability.

Clinical Relevance and Limitations: BERTScore captures semantic similarity better than lexical metrics, yet it remains limited in detecting factual mistakes, contradictions, and missed findings. In medical reports, semantic closeness does not ensure clinical safety if key abnormalities are omitted. Additionally, BERTScore depends heavily on the quality of the pretrained language model, which may not be trained on medical terminology. Hence, for clinical applications, BERTScore should be paired with RadGraph-F1, CheXbert, or human expert validation [182].

5.2. CheXbert Vector Similarity

CheXbert Vector Similarity is a targeted metric designed to assess whether generated radiology reports accurately reflect the presence or absence of key clinical findings by comparing binary pathology label vectors. Using the CheXbert labeled [194], a BERT-based system trained on expert-annotated and rule-based labels, each report is mapped to a 14-dimensional binary vector v indicating the mention (1) or absence (0) of common chest pathologies. The similarity between a generated report's vector v_{gen} and the reference report's vector v_{ref} is computed via cosine similarity:

$$\text{CheXbertSim} = \frac{v_{\text{gen}} \cdot v_{\text{ref}}}{\|v_{\text{gen}}\| \|v_{\text{ref}}\|}, \tag{26}$$

yielding a score in $[0, 1]$, where 1 denotes perfect agreement in pathology mentions [195,196]. This metric is important because it abstracts away linguistic variability and directly measures diagnostic accuracy, allowing clear interpretability since low scores immediately flag missing or spurious findings. Its advantages include strong clinical relevance and straightforward error analysis for specific conditions. However, by focusing solely on a fixed set of fourteen pathologies, CheXbertSim may overlook other critical observations and cannot capture the severity or contextual nuances of mentions. Furthermore, its reliance on a specialized labeler trained specifically in chest X-ray data limits its

Table 9
Evaluation Metrics for Various Medical VLM Applications.

Model type	Application	Evaluation metrics
VL Models (VLMs)	Medical image segmentation	Dice Score, Mean Intersection over Union (mIoU)
	Medical report generation	BLEU, ROUGE, METEOR, Perplexity, BERTScore, RadGraph F1, Human Evaluation
	Clinical efficacy metrics	Accuracy, Precision, Recall, F1 Score
	Question answering	Accuracy, Exact Match, Human Evaluation
	Relevance evaluation	CheXbert Vector Similarity, F1-BERTScore, RadCliQ

Table 10
Task usage of evaluation metrics in medical vision-language models.

Metric	Category	Purpose	Usage (%)	Used In
BLEU	N-gram Overlap	Measures surface-level similarity with reference reports	40%	XrayGPT, MedViLL
ROUGE	N-gram Overlap	Captures recall-based similarity of phrases	60%	Med-Flamingo, PubMedCLIP
Perplexity	Language model fluency	Assesses how fluent the generated text is under a language model	30%	RepsNet, BioGPT
BERTScore	Semantic similarity	Measures contextual embedding similarity between outputs and references	45%	LLaVA-Med, BioViL
RadGraph F1	Clinical entity matching	Extracts and compares medical entities in generated vs. ground-truth reports	35%	LLaVA-Med, Med-Flamingo
Human evaluation	Expert review	A Radiologist or clinician assesses clinical correctness and usefulness	35%	Visual Med-Alpaca, RepsNet
CheXbertSim	Clinical label similarity	Measures clinical label overlap using CheXbert's predictions	25%	CheXzero, MedVLP, BioGPT
RadCliQ	Composite clinical quality	Combines BLEU, BERTScore, and CheXbertSim into a factuality-aware quality score	20%	VLP-BioMed, Med-Flamingo, LLaVA-Med

applicability to other imaging modalities or pathology domains without retraining the underlying model [194].

Clinical Trust and Limitations: CheXbert focuses on disease-label correctness, making it highly useful for verifying whether generated reports reflect true clinical findings. Yet it only assesses predefined chest pathologies (14 classes) and may miss rare or nuanced clinical information. For deployment in practice, CheXbert should be combined with human evaluation, multi-pathology reasoning metrics, and prospective clinical testing to assess model reliability [145].

5.3. RadGraph F1

RadGraph F1 is a novel evaluation metric specifically designed for the task of medical report generation, where the primary goal is to produce clinically accurate and meaningful radiology reports from medical images [196]. Traditional automatic evaluation metrics such as BLEU, ROUGE, and METEOR, which were originally developed for general NLG tasks, primarily measure surface-level textual similarity. These metrics compare generated and reference texts based on word or n-gram overlap without considering the underlying clinical content. While useful for gauging fluency and syntactic similarity, such metrics are often insufficient in medical contexts where semantic correctness and factual accuracy are paramount. For instance, a report could have high BLEU or ROUGE scores but still omit or misrepresent critical clinical findings, leading to potentially harmful diagnostic errors.

RadGraph F1 addresses these limitations by shifting the focus from text surface similarity to the accurate extraction and comparison of clinical entities and their relationships. The metric works by first converting both the reference (ground truth) and model-generated reports into graph representations. In these graphs [197], nodes correspond to clinical concepts such as observations, findings, anatomical sites, and medical conditions, while edges represent the semantic or relational links between these concepts, for example, specifying that a finding is located in a particular anatomical region. This graph-based representation aligns closely with the way radiologists think about reports, as a structured network of clinical information rather than just free-form text.

To calculate the RadGraph F1 score, the system performs entity and relation extraction on both reference and generated reports. It then compares the graphs by matching nodes (entities) and edges (relations) based on their textual content and labels. The F1 score is computed separately for clinical entities and relationships, capturing both the correctness of identified medical concepts and the accuracy of their connections. The final RadGraph F1 score for a report pair is the average of these two F1 scores. Averaging over all report pairs in the test set yields an overall performance metric for the generation model.

The key advantage of RadGraph F1 lies in its clinical grounding. By focusing on entities and their relations, it better reflects the clinical validity of generated reports. This ensures that the model is not only producing grammatically correct sentences but also conveying medically important information accurately. Furthermore, RadGraph F1 provides a more interpretable diagnostic tool for model developers by identifying whether errors arise from missing or incorrect entities or flawed relations. However, the metric also has notable limitations. Its accuracy depends on the performance of the underlying entity and relation extraction model, which itself can introduce errors or inconsistencies. In addition, the complexity of graph construction and comparison requires more computational resources than traditional metrics, which can be a bottleneck for large-scale evaluations. Finally, RadGraph F1 focuses on structured information and may not fully capture nuances such as clinical reasoning, implied meanings, or subtle linguistic variations that do not directly map to graph structures [198].

Despite these limitations, RadGraph F1 has gained wide adoption as a complementary metric alongside traditional text similarity measures in medical report generation research. Current state-of-the-art systems typically achieve scores around 35%, indicating the significant challenge of generating radiology reports that are both linguistically fluent and clinically accurate. By providing a clinically meaningful evaluation, RadGraph F1 helps guide the development of more reliable and trustworthy medical AI systems.

Clinical Deployment Impact: RadGraph-F1 directly evaluates clinical correctness by comparing entities and relations, making it suitable for real diagnostic scenarios. It has been used in multiple radiology report generation studies and demonstrates a stronger correlation with radiologist judgment compared to BLEU/ROUGE. However, because this metric relies on extraction models, errors in entity detection can propagate. Clinical trials and expert-in-the-loop evaluations remain necessary for deployment in hospitals [187].

5.4. Human evaluation

Human evaluation is equally, if not more, essential in assessing the quality of VLMs designed for medical report generation. Unlike automated metrics, human assessment leverages the expert knowledge and clinical judgment of trained radiologists to evaluate reports in terms of accuracy, coherence, and clinical usefulness. This qualitative evaluation is indispensable because automatic metrics can only approximate clinical validity and often fail to detect subtle but critical errors that impact patient care. For example, a generated report might describe an abnormality using clinically inaccurate terminology, omit relevant findings, or contradict earlier statements, issues that can be detected reliably only through expert review.

In typical human evaluation protocols, expert radiologists review generated reports line by line, classifying errors such as factual inaccuracies, missing observations, contradictory statements, and linguistic mistakes [199]. Each identified error is assigned a severity score reflecting its clinical impact, allowing evaluators to quantify how harmful or misleading an error might be. Two summary statistics are then computed: Maximum Error Severity (MES), which captures the worst error severity score present anywhere in the report, and Average Error Severity (AES), which averages the severity scores across all report lines to reflect overall report quality.

The use of MES is particularly valuable because even a single severe error can undermine the clinical reliability of a report and affect downstream decision-making. AES, on the other hand, provides a more holistic view of report quality, including minor but cumulative errors. Radiologists' evaluations often reveal that a relatively small proportion of generated reports are error-free, with typical studies reporting that around 18% of model outputs have MES scores of zero (indicating no severe errors), while about 24% have AES scores of zero (indicating an overall error-free impression).

Human evaluation brings multiple advantages. It incorporates clinical expertise and contextual understanding that cannot be easily replicated by automated systems. It can assess pragmatic aspects of report quality such as clarity, interpretability, and appropriateness for clinical workflows. Additionally, it enables the identification of specific error types, informing targeted improvements in model design and training data. However, human evaluation also has limitations. It is resource-intensive, requiring significant time from busy clinical experts. The process is inherently subjective, and inter-rater variability can affect consistency. Establishing standardized protocols and training evaluators is necessary to improve reliability, but it remains challenging.

Despite these challenges, human evaluation remains the gold standard for validating medical report generation systems before clinical deployment. It ensures that models meet the stringent requirements of safety, accuracy, and utility demanded in healthcare settings. Moreover, human evaluation is not limited to report generation but extends to related tasks such as VQA in radiology, where interpretability and clinical reasoning are vital.

5.5. RadCliQ: Radiology Report Clinical Quality metric

RadCliQ (Radiology Report Clinical Quality) is a composite evaluation metric developed to quantify the clinical accuracy of AI-generated radiology reports by combining multiple complementary scores into a single radiologist-aligned value. Traditional text-overlap measures such as BLEU often fail to detect missing or incorrect clinical findings, while semantic metrics like BERTScore may overlook factual errors in entity relationships. RadCliQ addresses these gaps by integrating BLEU, BERTScore, CheXbert vector similarity, and RadGraph F1 through a linear regression model trained on expert error annotations [196,200]. Formally, if $m_1, \ldots, m_4$ denote the normalized scores of BLEU, BERTScore, CheXbertSim, and RadGraphF1, respectively, then RadCliQ is defined as:

$$\text{RadCliQ} = \alpha\, m_{\text{BLEU}} + \beta\, m_{\text{BERTScore}} + \gamma\, m_{\text{CheXbertSim}} + \delta\, m_{\text{RadGraphF1}} + b \tag{27}$$

where $\alpha, \beta, \gamma, \delta$, and b are learned coefficients [201] that minimize prediction error against radiologists' total error counts. RadCliQ is important because it captures both linguistic fidelity and clinical correctness, providing a more accurate and clinically relevant evaluation of radiology reports, achieving a Kendall's Tau-b correlation of 0.522 with human judgments and outperforming individual component metrics. Its advantages include combining multiple complementary metrics that better reflect clinical accuracy, aligning closely with radiologist evaluations by leveraging expert annotations, and capturing nuanced clinical relationships beyond simple text overlap. However, RadCliQ's reliance on multiple underlying metrics and the need to fit a regression model introduce computational overhead, and its design, based on chest X-ray reporting conventions and specific error annotations, may limit its generalizability to other medical domains or reporting styles [195,196].

5.6. Comparative evaluation of medical VLMs

To holistically evaluate the capabilities of cutting-edge medical VLMs and VQA systems, Table 11 presents a unified performance comparison across a diverse set of benchmark datasets, including MIMIC-CXR [132], VQA-RAD [140], ROCO [130], PathVQA [143], SLAKE [139], and PadChest [136]. The models are evaluated using conventional linguistic metrics, BLEU-4, ROUGE-L, METEOR, and BERTScore, as well as clinical relevance metrics like RadGraph F1 [187], CheXbert similarity [194], and the composite RadCliQ score [20], which combines multiple criteria to reflect the factual correctness and alignment with clinical knowledge, as illustrated in Fig. 15. Additionally, classification metrics such as accuracy, precision, recall, and F1-score are included for QA-based tasks to provide a broader view of diagnostic performance. This multifaceted evaluation framework helps assess the models not only on their NLG quality but also on their clinical utility and reliability, thereby facilitating informed choices for downstream medical applications.

5.7. Comparative analysis

Table 12 presents a detailed comparative evaluation of several state-of-the-art medical VLMs across multiple benchmark datasets, including MIMIC-CXR, PadChest, ROCO, and VQA-RAD. These models, such as **LLaVA-Med** [209], **Med-Flamingo** [202], **MedViLL** [204], **PubMedCLIP** [203], **RepsNet** [206], and others, have been assessed using a variety of linguistic and clinical metrics to comprehensively capture their performance in medical image understanding and report generation tasks.

For NLG and alignment with ground-truth radiology reports, metrics such as BLEU-4, ROUGE-L, METEOR, and BERTScore are considered. For instance, MedViLL demonstrates relatively strong performance in BLEU-4 (7.2), ROUGE-L (25.1), and METEOR (11.3), suggesting effective report generation capability. LLaVA-Med and Med-Flamingo, evaluated predominantly on the MIMIC-CXR dataset, show moderate BLEU-4 scores but vary in ROUGE-L and BERTScore, highlighting different fluency and semantic similarity strengths.

Metrics such as RadGraph F1, CheXbert similarity, and RadCliQ measure clinical relevance and factual correctness. LLaVA-Med shows

Table 11
Performance Comparison of Medical VL and QA Models on Various Medical Tasks.

Dataset/Task	Metric	LLaVA-Med [52]	Med-Flamingo [202]	PubMedCLIP [203]	MedViLL [204]	XrayGPT [205]	RepsNet [206]
Visual Question Answering							
Path-VQA [143]	Accuracy ↑	0.230	0.241	0.697	–	–	–
Slake-VQA [139]	Accuracy ↑	0.288	0.309	0.626	–	–	–
VQA-RAD (Closed) [207]	Accuracy ↑	0.545	0.577	0.680	0.658	0.715	0.690
VQA-RAD (Open) [207]	Accuracy ↑	0.140	0.335	0.442	0.430	0.485	0.460
VQA-RAD [207]	F1 Score ↑	0.069	0.442	0.510	0.493	0.550	0.533
Medical QA (Text-only)							
MedQA [208]	Accuracy ↑	0.230	0.241	0.697	–	–	–
MedMCQA [208]	Accuracy ↑	0.288	0.309	0.626	–	–	–
PubMedQA [208]	Accuracy ↑	0.336	0.488	0.800	–	–	–
Report Generation (MIMIC-CXR)							
BLEU-4 ↑	0.006	0.115	0.192	0.175	0.210	0.198	–
ROUGE-L ↑	0.125	0.275	0.354	0.330	0.368	0.345	–
F1-RadGraph ↑	0.048	0.267	0.310	0.295	0.322	0.315	–
RadCliQ ↓	5.094	3.100	2.810	3.010	2.740	2.860	–
CheXbert Sim. ↑	0.150	0.440	0.470	0.460	0.495	0.480	–
Cross-Modal Retrieval (ROCO)							
Image→ Text R@1 ↑	0.381	0.412	0.560	0.532	0.595	0.580	–
Text→ Image R@1 ↑	0.365	0.388	0.545	0.510	0.570	0.548	–
Mean Rank ↓	22.1	19.4	11.2	13.5	10.4	11.0	–
Zero-Shot Classification (PadChest)							
Accuracy ↑	0.365	0.398	0.528	0.510	0.540	0.525	–
AUC ↑	0.621	0.640	0.705	0.689	0.730	0.712	–
F1 Score ↑	0.534	0.556	0.620	0.605	0.638	0.625	–
Report Summarization (MIMIC-III)							
ROUGE-L ↑	0.185	0.320	0.402	0.388	0.415	0.405	–
BLEU-1 ↑	0.192	0.154	0.260	0.242	0.278	0.264	–
METEOR ↑	0.303	0.350	0.410	0.398	0.428	0.415	–

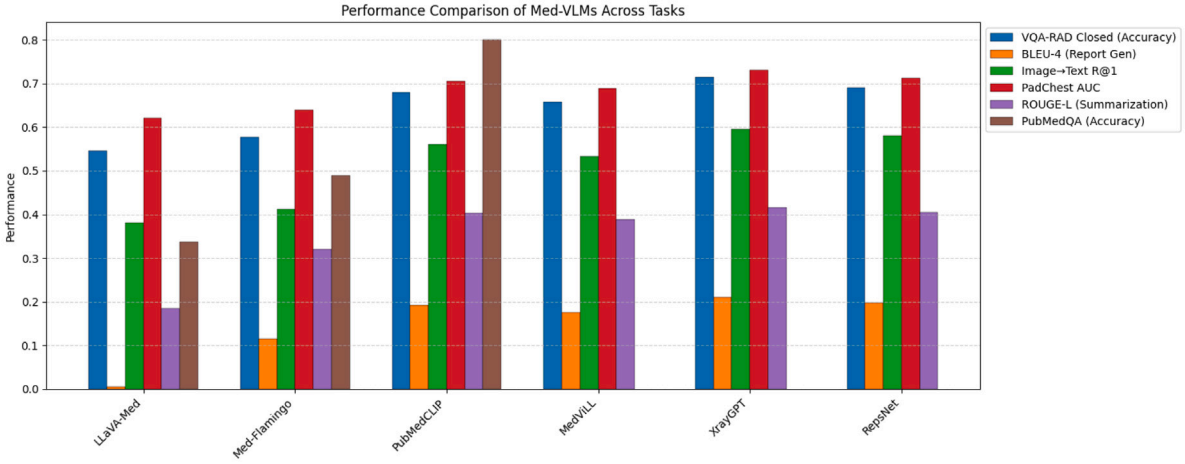

Fig. 15. Performance comparison of state-of-the-art Medical Vision-Language Models (Med-VLMs) across multiple tasks including Visual Question Answering (VQA-RAD), Biomedical QA (PubMedQA), Report Generation (BLEU-4), Cross-modal Retrieval (Image → Text R@1), Zero-shot Classification (PadChest AUC), and Report Summarization (ROUGE-L).

Table 12
Comprehensive evaluation of Medical VLMs Across multiple datasets.

Model	Dataset	Accuracy	Precision	Recall	F1-score	BLEU-4	ROUGE-L	METEOR	BERTScore	RadGraph F1	CheXbert Sim.	RadCliQ
LLaVA-Med [209]	MIMIC-CXR	–	–	–	–	5.05	19.13	–	47.51	15.77	23.06	–
Med-Flamingo [202]	MIMIC-CXR	–	–	–	–	5.17	3.54	3.46	59.12	–	–	–
MedViLL [204]	MIMIC-CXR	84.1	0.698	0.559	0.621	0.066	25.1	11.3	52.4	20.0	25.5	7.5
RepsNet [206]	MIMIC-CXR	–	–	–	–	0.27	–	–	–	–	–	–
XrayGPT [210]	PadChest	–	–	–	–	–	0.1997	–	–	–	–	–
Visual Med-Alpaca [211]	MIMIC-CXR	–	–	–	–	–	–	–	–	–	–	–
LLaVA-Med [209]	VQA-RAD	78.3	75.2	74.5	74.8	–	–	–	–	–	–	–
Med-Flamingo [202]	VQA-RAD	79.1	77.0	75.5	76.2	–	–	–	–	–	–	–

Table 13
Summary of Medical VL Models.

Model	Text Encoder	Framework	Datasets
MedViLL [204]	BERT [78]	ResNet-50, BERT	MIMIC-CXR [129], Open-I [181], VQA-RAD [140]
PubMedCLIP [29]	CLIP Text Encoder [6]	ViT-B/32, ResNet-50	ROCO [127], VQA-RAD [140], SLAKE [139]
BiomedCLIP [29]	PubMedBERT [212]	ViT-B/16	VQA-RAD [161], SLAKE [139]
RepsNet [213]	BERT [78]	ResNeXt-101	VQA-RAD [161], IU-Xray [181]
Visual Med-Alpaca [116]	LLaMA-7B [214]	DePlot, Med-GIT	PubMedQA [141], MedDialog [215], MEDIQA RQE [216], MedQA [179]
LLaVA-Med [52]	LLaMA-7B [214]	ViT-L/14, BiomedCLIP [29]	PathVQA [143], VQA-RAD [161], SLAKE [139]
XrayGPT [205]	Vicuna-7B [217]	MedCLIP [57]	MIMIC-CXR [129], Open-I [181]
Med-Flamingo [218]	LLaMA-7B [85]	CLIP ViT-L/14	MTB [218], PMC-OA [85], PathVQA [143], VQA-RAD [161]
BioViL [219]	Transformer	ViT, Transformer	MIMIC-CXR [129], ROCO [127]
Masked VL Modeling/Contrastive Losses [220]	Transformer	ViT, Transformer	MIMIC-CXR [129], ROCO [127]

competitive performance in RadGraph F1 (15.77) and CheXbert similarity (23.06), indicating better alignment with clinical findings in the reports. However, many models have missing clinical metric values, reflecting the current gap in standardized clinical evaluation across all models.

Accuracy, precision, recall, and F1-score are predominantly reported for VQA tasks on datasets such as VQA-RAD, with LLaVA-Med and Med-Flamingo achieving accuracy scores of approximately 78%–79%, showcasing their strong ability to interpret medical images in a QA context.

Notably, several models lack comprehensive reporting across all metrics and datasets, which points to the need for more systematic, standardized benchmarking frameworks in medical VL modeling. This table underscores that while significant progress has been made in individual linguistic and clinical metrics, further work is needed to unify evaluation protocols to reliably compare model generalizability and clinical applicability across diverse medical imaging datasets.

6. Medical models

This review paper section provides an overview of medical VLM models specifically designed for medical images, as shown in Fig. 18. The information is arranged sequentially according to each model's original introduction in Table 13, and Table 16 presents detailed descriptions of VL models, highlighting their architectural components, pretraining strategies, and evaluation tasks within the medical domain. We primarily concentrate on newly released, free, open-source, or publicly accessible models.

6.1. Medical Vision Language Learner (MedViLL)

MedViLL processes medical pictures to produce associated details and reports. The model uses ResNet-50 [221], trained on ImageNet [222], to extract visual features V. It also employs a base BERT [78] embedding layer to textual features T captured from diagnostic reports, utilizing a WordPiece tokenizer. The input is segmented into a series of tokens. The visual and textual characteristics combine locational data to record the spatial associations and the series of factor arrangements within the input data. MedViLL applied multimodal representation input into the BERT model and pretrained it over two main goals, including masked language modeling (MLM) and image-text matching (ITM). MLM uses a bidirectional autoregressive self-attention mask to merge image and text features, and it uses the negative and incorrect log-likelihood loss function, while the ITM facilitates the training by forecasting whether pairs match, utilizing a loss function for matching and non-matching pairs that is based on predictions. Pre-training happens on 89,395 MIMIC-CXR [129] picture-report pairs and fine-tuning on 3547 Open-I pairs (focused on AP view X-rays). The VQA and RG tasks are assessed on VQA-RAD [161], MIMIC-CXR [129], and Open-I [181] datasets. This model accomplishes a perplexity of 4.185 [145], a BLEU-4 score of 0.066, and an accuracy of 84.1%, with precision, recall, and F1 scores of 0.698, 0.559, and 0.621, respectively.

MedViLL is among the earliest medical VLMs emphasizing radiology report generation rather than open-ended instruction following. Compared to later models such as PubMedCLIP and BioViL, MedViLL was trained on a relatively smaller paired dataset (≈89k image-text pairs), which limits its generalization ability when transferred to broader radiology domains beyond chest X-rays. Despite these constraints, MedViLL represents an important early milestone; its BERT-based MLM and ITM training pipeline inspired subsequent CL approaches. Its strength mainly lies in foundational cross-modal alignment rather than high-level reasoning or retrieval scalability, which later models improved substantially (see Table 14).

6.2. RepsNet

RepsNet is a model tailored for VQA tasks, with the capability to understand medical images and produce computerized medical results. The model exploits a pre-trained BERT [78] as an enhanced type of the previously trained ResNeXt-101 [223] as the visual and the textual encoder. The WordPiece is used for text tokenization [224]. The integration of question and image attributes is accomplished through the BAN framework [225]. The ResNet [226] uses bidirectional CL to make textual captions and visual data reliable. The model is refined and calculated on VQA-RAD for VQA tasks [161]. Meanwhile, the IU-Xray [181] collection is utilized for RG's refinement and assessment. For closed-ended questions, the model classifies responses through classification, and it makes responses applying an improved GPT-2 decoder that incorporates image data and prior circumstances. Performance metrics for RepsNet on the IU-Xray dataset include a BLEU-2 score of 0.44 and a BLEU-4 score of 0.27, highlighting its effectiveness in medical RG and image interpretation, as shown in Table 15.

RepsNet is effective for medical VQA due to BAN integration and GPT-style decoding, enabling question-conditioned response generation. However, unlike PubMedCLIP or BiomedCLIP, which rely on large-scale contrastive pretraining, RepsNet does not leverage extensive multimodal pretraining corpora, resulting in weaker clinical robustness when dealing with unseen modalities or rare findings. It is more suited for structured report generation tasks than broad zero-shot retrieval, making it functionally different rather than directly competing with contrastive VLMs.

6.3. PubMedCLIP (PMC)

PMC [227] is a CLIP-based model, inspired by the architecture of CLIP [6] and pre-trained on the ROCO [127] dataset, which includes over 80,000 Image-label pairs extracted from PubmedClip papers. This model is split up into three different visual encoders, such as ViT-B/32 [8], ResNet-50 and ResNet-50 × 4 [221], in addition to a CLIP text encoder employing the Transformer architecture [228]. PubMedCLIP follows a CL methodology during its pre-training phase, creating unified illustrations by measuring cosine similarity among visual and textual attributes. Calculating and averaging cross-entropy loss for both modalities is the aim of pre-training. For VQA, PMC is adapted as a previously trained visual conversion. This method concatenates the

Table 14
Strengths, limitations, and clinical implications of vision-language models in medical models.

Model	Strengths	Limitations	Deployment feasibility	Clinical applicability
LLaVA-Med [52]	Strong visual grounding and report generation	High computational cost	Moderate (requires powerful GPU)	Suitable for diagnostic support and teaching tools
Med-Flamingo [218]	Flexible multi-task handling, strong language reasoning	Large model size and dependency on few-shot prompts	Low (requires specialized infrastructure)	Good for complex question answering
PubMedCLIP [29]	Robust image–text matching	Limited generative ability	High (lightweight architecture)	Effective for image retrieval and annotation
MedViLL [204]	Well-aligned vision–language features for classification	Limited scalability beyond chest X-rays	High (compact model)	Effective in radiograph classification tasks
XrayGPT [205]	Accurate prompt-based report generation	Sensitive to prompt phrasing; large model footprint	Moderate (dependent on prompt quality)	Useful for automated reporting and summarization
RepsNet [213]	Combines retrieval with generation; boosts relevance	Dependence on external retrieval corpus	Moderate (needs retrieval setup)	Suitable for QA and image captioning
Visual Med-Alpaca [116]	Instruction-following capability; fine-tuned on medical reports	Limited evaluations and real-world testing	Moderate (custom-tuned for medical use)	Promising for interactive diagnosis assistants
BiomedCLIP	Large-scale contrastive training; strong generalization	Less effective in free-form generation	High (scalable and modular)	Ideal for retrieval and image classification
BioViL [219]	Learns robust multimodal representations; handles missing modalities via masking	Requires careful masking strategy; training complexity is higher	Moderate (needs custom pretraining setup)	Promising for joint classification and retrieval tasks
Masked VL Modeling w/ Contrastive Losses [220]	Effective at aligning image patches with clinical text; benefits from contrastive objectives	May underperform on free-form generation; sensitive to mask ratios	Moderate (custom pretraining and fine-tuning required)	Useful for disease localization and multi-label classification

result of the method with the results from a convolutional denoising autoencoder (CDAE) [229], which functions to serve as a module for image denoising as illustrated in Fig. 16. The query is first encoded with GloVe embeddings [230] and processed via an LSTM [231]. Bilinear attention networks (BAN) are utilized to integrate the characteristics of the image and question [225]. The output data is then input into a two-layer FFN to predict the answer. The classification [232] and reconstruction image losses are combined in the VQA loss. Fine-tuning is done on the SLAKE (English) [139] and VQA-RAD [161] datasets, including open- and closed-ended questions. PMC's performance is evaluated relative to two established Medical VQA (MedVQA) methods: Medical Enhanced Visual Features (MEVF) [233], which makes use of Question-Conditioned Reasoning (QCR) [234], and improved visual and graphical attributes, which focus on question-conditioned reasoning. When replacing the visual encoder components in both frameworks with PMC, the model demonstrates improved correctness on the SLAKE and VQA-RAD collections within the QCR background, achieving better results than in the MEVF architecture. PMC consistently achieves the topmost precision within the QCR structure across these datasets.

PMC improves retrieval and VQA performance over earlier systems like MedViLL due to dual-encoder contrastive learning. However, its pretraining dataset (80k ROCO pairs) remains significantly smaller than BiomedCLIP (≈15M), limiting its capability for rare disease reasoning and domain transfer. PMC performs well in multimodal understanding but lacks the generative ability found in LLaVA-Med and XrayGPT, meaning it is more aligned with representation learning than interactive clinical dialogue or free-form reporting [235].

6.4. Visual Med-Alpaca

Visual Med-Alpaca is a specialized medical foundation model developed to handle diverse multidimensional biomedical tasks, including VQA [116]. Its architecture is structured as follows: Primarily, a classifier is utilized to identify the module that suitable component converts visual data into an intermediary text representation from image inputs. DePlot [236], which specializes in analyzing charts and graphs, and Med-GIT [237], fine-tuned using the ROCO dataset [127] to efficiently interpret radiology images, are two of the supported modules. A prompt manager then integrates the text extracted from the visual modules with user-provided textual inputs to formulate a comprehensive prompt for the LLaMA-7B model [214]. Prior to generating outputs, LLaMA-7B undergoes rigorous fine-tuning, including both traditional methods and LoRA [115] adjustment, using 54,000 medical QA pairs from a carefully selected dataset. This set's questions are obtained from QA datasets including PubMedQA [141], MedDialog [215], MEDIQA QA, MEDIQA RQE [216], and MedQA [179]. The corresponding answers are generated using GPT-3.5-Turbo through a self-instruction approach [238]. To ensure accuracy and relevance, human experts carefully review and refine the question-answer pairs. Although the model's evaluation is still underway, its creation shows an important advancement in biomedical artificial intelligence [116].

Unlike encoder-centric models such as PMC or BiomedCLIP, Med-Alpaca emphasizes instruction-tuned interactive reasoning. It is capable of multi-step medical question answering and conversational interaction, yet its visual grounding heavily relies on models like Med-GIT/DePlot. As a result, Med-Alpaca is more patient-chat-oriented and less image-understanding-centric compared to BioViL. It is useful for clinical assistant-style communication but not as strong for pure visual feature alignment.

6.5. BiomedCLIP

BiomedCLIP is a specialized model pre-processed on the 15 PMC dataset, comprising 15M figure-label pairs extracted from Pubmedclip research papers [29]. However, the model is not openly accessible. Its framework closely mirrors that of CLIP [6], with its pre-learned PubMedBERT model text encoder helping as a key distinction [212]

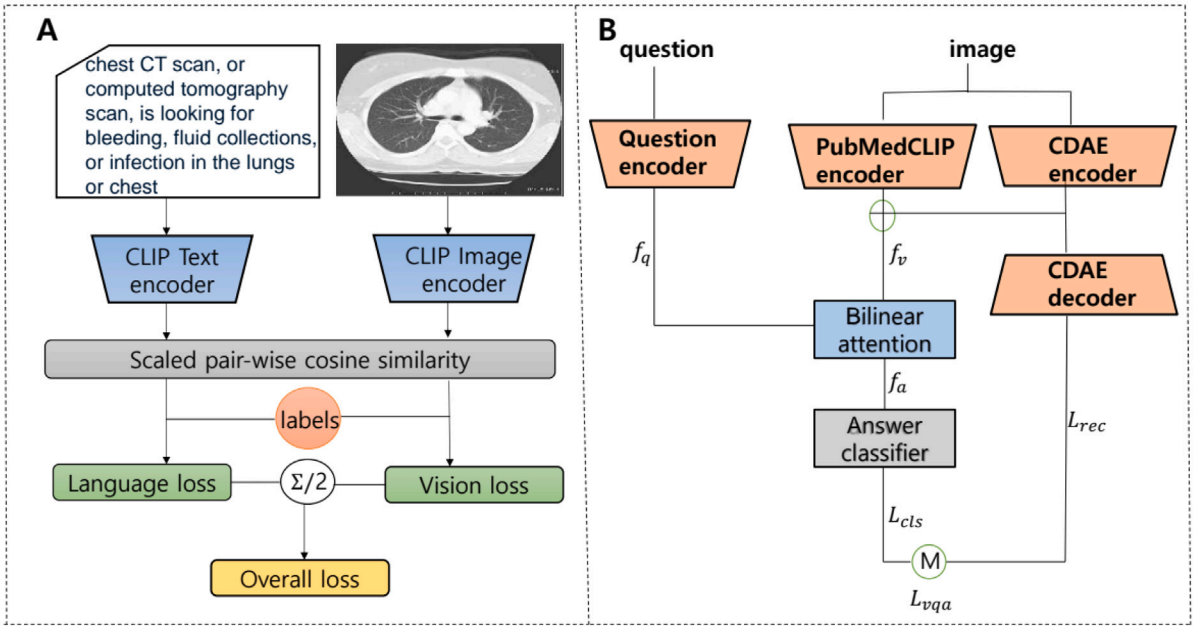

Fig. 16. (A) Summary of the PubMedCLIP pre-training method. (B) Diagram of the MedVQA backbone with the PubMedCLIP encoder.

that employs the Word Piece tokenizer [224]. For visual data encoding, the model uses ViT-B/16 [8]. During pre-training, BiomedCLIP leverages a CL approach while optimizing memory usage through the use of particle contrastive loss [239]. To adapt the model for VQA tasks, integrate the METER framework [240], which employs a co-attention multimodal fusion component based on Transformers. This module generates multimodal embeddings that are subsequently passed to a classifier for answer prediction. The model's performance is evaluated on two datasets: VQA-RAD [161] and SLAKE (English) [139], demonstrating its utility in biomedical VQA tasks, as shown in Fig. 17. The top-1 results in BiomedCLIP are correct more than 56% of the time, while the top-5 results contain the correct ones more than 77%. Despite the biomedical modification, PubMedCLIP [241] performs even worse than CLIP, which surprised us. This is a result of the training data that PubMedCLIP utilized. Even though PubMedCLIP was named after PubMed, it only utilized a small subset of image-text pairs from radiology, which comprise a small percentage of images in the biomedical literature, for continuous pretraining of CLIP. Furthermore, catastrophic forgetting may occur if continuous pretraining on a small dataset is done carelessly (e.g., by augmenting with a pretraining aim) [242]. The significance of using a large and diverse dataset for domain-specific VLM is further demonstrated by the excellent performance of BiomedCLIP, which pre-trains on large-scale data from PMC-15 M.

BiomedCLIP leverages one of the largest biomedical image-text corpora (15M pairs), providing a substantial advantage in zero-shot retrieval and cross-institution generalization. It consistently outperforms earlier models like PMC in semantic alignment and downstream transfer tasks. However, limited public release and partial training accessibility hinder full reproducibility and clinical deployment. While excellent for medical representation learning, it does not natively generate reports like XrayGPT-based LLM hybrids.

6.6. XrayGPT

XrayGPT is an interactive medical and clinical VLM designed for interpreting chest X-ray images [205]. The model applies the vision encoder MedCLIP [57] to derive visual characteristics. These characteristics pass through a comprehensive conversion: using a linear transformation layer, they are primarily reduced in dimension through

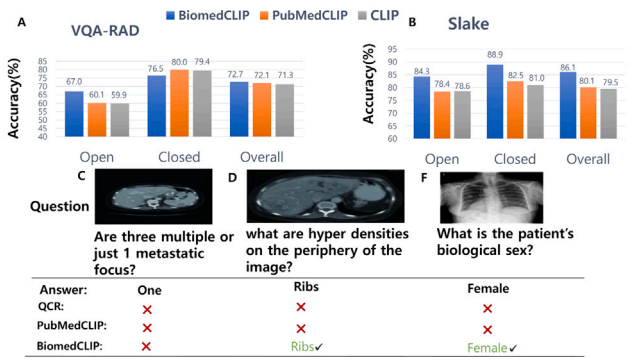

Fig. 17. Performance evaluation of medical image-based questioning and answering: (A) Evaluation of accuracy in three distinct places using the Slake and VQA-RAD datasets: Closed (multiple options, often yes/no) and Open-ended (free-form replies). (B) Three challenging VQA-RAD examples that were taken from the PubMedCLIP paper [241], and in which earlier VQA models such as PubMedCLIP (the former state-of-the-art), failed to provide correct answers. (C) and (D) showcase BiomedCLIP correctly answering the questions. While its response to (B) is not technically accurate, it accurately recognizes that the center of metastatic disease is the liver (visible on the CT Scan's right side).

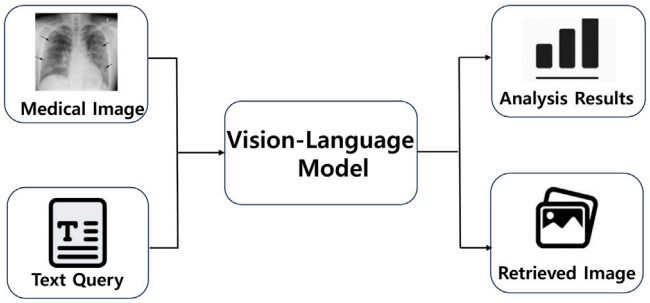

Fig. 18. Pipeline of Vision-Language Models (VLMs) in Medical Imaging. The architecture illustrates how medical images and clinical text are encoded via visual and textual encoders, followed by fusion and reasoning modules to perform tasks like VQA, captioning, or diagnosis.

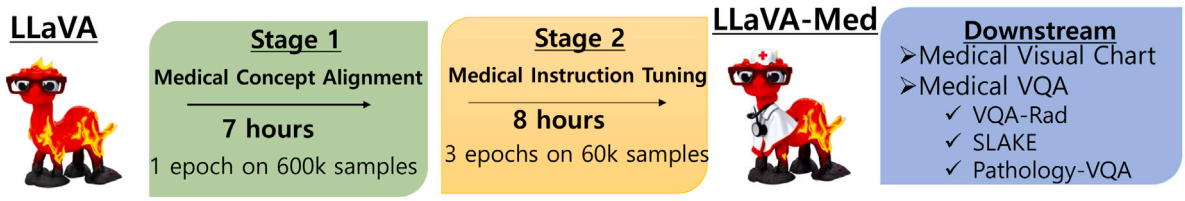

Fig. 19. LLaVA-Med was built by initializing with the general-domain LLaVA and further training it using a curriculum learning approach. This process involved two stages: first, aligning biomedical concepts, followed by comprehensive instruction tuning. The LLaVA-Med model was assessed on benchmark tasks, such as visual conversation and question answering.

a linear projection head and then converted into elements. The architecture merges two key written inquiries to guide its functionality: (1) assistant query contextualizes the model's role with the prompt. "You are a reliable virtual assistant committed to enhancing medical decision-making" (2). The query of the doctor directs the model to provide information specific to chest X-ray analysis. The visual tokens are combined with segmented prompts and provided into a medical LM, summarizing the chest X-ray findings. The underlying LM is Vicuna-7B [217], improved on a wide dataset comprising 100,000 actual patient-doctor interactions and 20,000 medical images focused on communication obtained from ShareGPT.com. While the parameters of the LLM and visual encoders remain constant throughout training, the parameters in the linear mapping layer are updated. Initially, the pre-processed MIMIC-CXR data collection [129], consisting of 213,514 picture annotation sets, was employed to train the model. Furthermore, 3000 visual labeled sets sourced from the Open-I [181] data collection were utilized. The MIMICCXR dataset yields ROUGE-L = 0.1997, ROUGE-1 = 0.3213, and ROUGE-2 = 0.0912 for XrayGPT, displaying its capability to generate precise and meaningful chest X-ray summaries.

In contrast to encoder-only models (PMC, BiomedCLIP), XrayGPT integrates Vicuna-7B for multimodal reasoning and radiology-style report generation. It enables interactive clinical dialogue and summarization, providing interpretability beyond retrieval. However, its training is focused primarily on CXR data, restricting scalability to multi-organ imaging domains unless fine-tuned. While more conversationally capable, it lacks the broad domain coverage of general VLMs like LLaVA-Med.

6.7. Large language and vision assistant for BioMedicine (LLaVa-Med)

LLaVa-Med is a VLM LLaVa [93] modified version that was particularly created for the medical field by applying instruction-following datasets [52] for training. It generates visual Specifications using the pre-learned [6] visual encoder ViT-L/14 [8], which can also be substituted with BiomedCLIP [29], for enhanced biomedical domain alignment. These visual characteristics are converted into tokens via a linear projection layer, and tokenized directions are then added. The resulting tokens are then processed by the LLaMa-7B [214] language model, which can also be replaced with Vicuna [217]. As illustrated in Fig. 19, after starting with the general-purpose LLaVA, the model is refined through a process called curriculum learning. Initially, the focus is on helping the model establish connections between visual features in biomedical images and corresponding textual descriptions [128] stored in the language knowledge base model. The 15 million PMC dataset, which was first used to train BiomedCLIP, contains 600,000 image-caption pairs, which are employed to accomplish this. When these pairs are converted into a task-specific dataset, the model is guided to provide a brief or comprehensive label for the image. The purpose is to make the model into the original caption when presented with the image and language instructions.

In this phase, only the linear projection layer is modified; the language model and visual encoder stay identical. The next phase of training emphasizes the model alignment to handle a wide range of instructions effectively. During the training phase, 60,000 photos are used to train the model, each paired with its corresponding label and multiple rounds of QA. During this phase, the visual weights encoder is kept static to retain previously learned visual features, while the LM weights and projection layer are actively updated. This type of strategy makes the model more capable of processing various instructions and producing detailed, contextually accurate multi-round conversational outputs. Finally, many visual question-answering standards are employed to improve and assess the model, including VQA-RAD [161], SLAKE [139], and PathVQA [143].

LLaVA-Med extends the general LLaVA architecture using curriculum-based medical instruction tuning. It supports multi-round medical conversation and multi-task reasoning, providing broader utility than PMC or BiomedCLIP for real-world clinical assistant settings. However, hallucination control is still a challenge, and trustworthy grounding is weaker compared with contrastive pipelines like BioViL. It excels in clinical pre-diagnostic discussion scenarios rather than strict retrieval-based tasks.

6.8. BioViL

BioViL is a biomedical VL model created for tasks that combine medical images and text. The model uses a vision encoder, usually based on CLIP or ViT, to derive visual characteristics from biomedical images. These visual features are associated with text features made by a language encoder, including BioBERT or PubMedBERT, through a shared latent space employing CL techniques [6,155]. BioViL is pre-trained on large-scale biomedical datasets, including PMC-OA (1.3 million image-text pairs) and MIMIC-CXR, ensuring its ability to comprehend medical terminology and visual patterns [85,132]. The training procedure contains multimodal alignment to carry connected image-text pairs nearer in representation space while separating unconnected pairs, enhancing multimodal representation learning [107, 243]. BioViL is mostly effective in tasks that include medical image captioning, VQA, and text-based image retrieval. For instance, it signifies robust performance on datasets like VQA-RAD [161] and PathVQA [143]. By leveraging its domain-specific pretraining, BioViL helps multimodal reasoning, allowing it to tackle intricate medical circumstances with incomplete labeled data, making it a valuable tool for biomedical AI applications [57,155].

BioViL improves cross-modal alignment more effectively than MedViLL using contrastive pretraining on a large biomedical corpus. It demonstrates superior retrieval and VQA accuracy with better semantic grounding. However, like most encoder-only methods, BioViL lacks native long-form report generation ability, which models like XrayGPT or Med-Flamingo address. BioViL is ideal for representation and grounding tasks but requires integration with an LLM for conversational or narrative generation [244].

6.9. Masked VL Modeling with Contrastive Losses

Masked VL Modeling with Contrastive Losses employs the primary six layers of BERT [78] for text encoding, the ViT-B/12 [8] for image encoding, and the final six layers of BERT for cross-models encoding [13]. The cross-model encoder arranges textual and visual

Table 15
Computational and performance comparison of Medical VL Models.

Model	Parameters (B)	FLOPs (T)	Pretraining dataset size	Inference speed	Performance highlights
MedViLL [204]	0.35	50	MIMIC-CXR (377k images)	25 ms/image	Accuracy: 84.1%, BLEU-4: 0.066
PubMedCLIP [29]	0.43	60	ROCO (81k images)	30 ms/image	Improved QCR Accuracy
BiomedCLIP [29]	0.46	65	PMC-15M (15 M image-text pairs)	28 ms/image	Optimized Contrastive Loss
RepsNet [213]	0.55	70	VQA-RAD (3.5k images)	22 ms/image	BLEU-4: 0.27, BLEU-2: 0.44
Visual Med-Alpaca [116]	7.0	–	ROCO, PubMedQA	–	Ongoing Evaluation
LLaVA-Med [52]	15.0	200	PMC, VQA-RAD	40 ms/image	ROUGE-L: 0.1997
XrayGPT [205]	13.0	180	MIMIC-CXR (377k images)	35 ms/image	ROUGE-1: 0.3213
Med-Flamingo [218]	14.0	210	MTB, PMC-OA	45 ms/image	PathVQA: 0.303
BioViL [219]	0.50	80	PMC-15M (15 M pairs), MIMIC-CXR	30 ms/image	Top-1 accuracy (VQA): 65% (approx.)
Masked VL Modeling w/ Contrastive Losses [220]	0.50	80	PMC-15M (15 M pairs), MIMIC-CXR	30 ms/image	Multi-label classification F1: 0.58 (approx.)

characteristics by using cross-attention layers. The model applies CL, MLM, and ITM objectives in incorporation for pre-training. Moreover, the model uses a recently presented masked image approach, which expands data by arbitrarily masking 25% of image areas. The model is disclosed to a wider range of visual contexts with the help of this method, which improves its capacity to learn robust representations that can process partially blocked inputs. The pre-training process utilizes data from the ROCO [6], MedICaT [51], and ImageCLEF caption [245] datasets. For downstream VQA tasks, the cross-modal encoder is supplemented with an answering decoder to make answer text words. After pretraining, the encoder weights are established, and the model is modified and measured using PathVQA [143], SLAKE [139], and [161]. These evaluations underscore the model's effectiveness in addressing medical VQA tasks.

This approach merges MLM and ITM with contrastive objectives, improving robustness under missing or partially masked visual cues, an area where MedViLL struggled. It performs competitively in multimodal alignment tasks, though training scale remains mid-range relative to BiomedCLIP or LLaVA-Med, leading to moderate performance in heavy generalization or cross-domain transfer scenarios. Suitable for balanced visual-text learning but less scalable than ultra-large paired corpora approaches.

6.10. Med-Flamingo

Med-Flamingo is a cross-modal small-data trainee approach made for the medical field based on the Flamingo [246] framework. The MTB [218] dataset is employed to pre-train the model. It encompasses 4721 s tions with text and images from diverse medical textbooks, with at least one and up to ten images per segment. Moreover, the PMC-OA [85] data collection comprises 1.3 million image-description pairs applied to train Med-Flamingo. Its few-shot abilities allow it to simplify various multimodal tasks with a minimal instance. Using perceiver resamples, trained from scratch, is utilized for the conversion of the graphical attributes into a predetermined number of elements. The model mentioned uses a constant CLIP ViT-L/14 vision encoder for the extraction of visual characteristics. A pre-trained LLaMA-7B [85] LLM is improved with fenced cross-attention layers and taught to handle these tokens and tokenized text prompts from the beginning. This setup enhances training stability and the model's ability to learn new associations. The performance of Med-Flamingo is calculated using several recognized standards, including PathVQA [143] and VQA-RAD [161]. It shows a few-shot performance, with perfect match scores of 0.3030 on PathVQA and 0.2000 on VQA-RAD. Nevertheless, its zero-shot effectiveness outcomes are an ideal match of 0.12 on PathVQA and 0.00 on VQA-RAD.

Med-Flamingo is among the strongest few-shot biomedical models, often outperforming fully supervised systems when only limited data is available. It excels in reasoning-driven tasks and interactive clinical explanation. However, its zero-shot capability is weaker, making it prompt-dependent. Compared with PubMedCLIP, Med-Flamingo is more reasoning-oriented, whereas its visual grounding is still less robust than contrastive models such as BioViL.

7. Model explainability and visualization techniques

7.1. Interpreting and visualizing VLM decisions

In medical applications, model performance alone is not sufficient; explainability is crucial to build trust among clinicians, ensure accountability, and support human-in-the-loop decision-making. VLMs, despite their remarkable performance, are often perceived as black-box systems due to their complex internal mechanisms that are difficult to interpret. This presents a significant barrier to clinical adoption, especially in high-stakes scenarios such as disease diagnosis or treatment recommendation, where transparency is essential to avoid misdiagnosis, gain regulatory approval (e.g., FDA, EU AI Act), and maintain clinician trust [247]. To address this, several explainability techniques have been adapted from general computer vision to the biomedical domain. One such method is **Gradient-weighted Class Activation Mapping (Grad-CAM)** [248], which highlights the regions in an image that most strongly influence the model's prediction. In VLMs, Grad-CAM can be extended to link image regions with text tokens, offering insights into how multimodal components interact. Similarly, **attention maps** derived from transformer architectures reveal the attention flow between image patches and textual queries [249], helping to decode the cross-modal reasoning process.

Additionally, **saliency visualization** techniques are used to highlight which parts of an image contribute most to a particular output prediction. These techniques, such as integrated gradients or occlusion sensitivity, are being adapted to multimodal settings [250]. **Token-to-region attribution** methods used in models like CLIP and BioViL attribute the influence of specific words or phrases in the prompt to image regions, enabling fine-grained interpretability for tasks such as lesion localization and report generation [57,251]. Other visual explanation tools include relevance heatmaps and saliency maps [252], which identify image areas that significantly impact the model's output in tasks like VQA [155]. Beyond visual cues, some VLMs have been extended to produce textual justifications alongside predictions [253], allowing clinicians to evaluate the reasoning behind automated decisions. Moreover, model-agnostic interpretation tools such as LIME [254] and SHAP [255], though originally developed for tabular and textual data, are being explored for multimodal adaptation to provide interpretable outputs in medical decision support systems.

Despite these advancements, challenges remain. The inherently complex interactions between vision and language components, the limited availability of benchmark datasets with human-annotated rationales, and the absence of standard metrics for evaluating explanation quality all hinder progress. Future research should focus on designing robust explainability frameworks tailored to clinical workflows, developing datasets with ground-truth rationales, and building interactive tools that allow end-users, clinicians in particular, to probe and validate model decisions in real time.

Table 16
Overview of vision-language models: architectures, pretraining strategies, and evaluation tasks.

Model	Visual encoder	Text encoder	Dataset(s)	Pretraining/Fine-tuning	Fusion	Eval tasks
LLaVA-Med	ViT-L/14	LLaMA-7B	PMC, MIMIC-CXR, VQA-RAD, SLAKE	Visual instruction tuning	Early fusion	VQA
Med-Flamingo	CLIP ViT-G/14	LLaMA-7B (Frozen)	VQA-RAD, MedMCQA, MTB, PMC-OA, PathVQA	Few-shot tuning	Late fusion	Few-shot VQA
PubMedCLIP	ResNet, ViT-B/32	CLIP Transformer	PubMed, PMC-OA, ROCO, SLAKE, VQA-RAD	Contrastive learning	CLIP-style	VQA
Medical Vision Language Learner (MedViLL)	ResNet-50	BERT	MIMIC-CXR, Open-I	Joint V+L pretraining	Early fusion	VQA, Report Gen (RG)
XrayGPT	MedCLIP	Vicuna-7B	MIMIC-CXR, Open-I	GPT fine-tuning on reports	Prompt fusion	Chest X-ray VQA
RepsNet	ResNeXt-101, ViT	BERT, Transformer	MIMIC, VQA-RAD, IU-Xray	Report-based learning	Hybrid fusion	VQA, RG
Visual Med-Alpaca	DePlot, ViT	Alpaca (LLaMA)	ROCO, PubMedQA	Visual instruction + report tuning	Early fusion	VQA
BiomedCLIP	ResNet-50x, ViT-B/16	PubMedBERT	PubMed, PMC-OA, MIMIC, VQA-RAD, SLAKE	Contrastive pretraining	CLIP-style	VQA
BioViL (Masked V+L Modeling with Contrastive Losses)	ViT	Transformer	MIMIC-CXR, ROCO	Masked modeling + Contrastive Loss	Early Fusion	VQA, Classification
Masked VL Modeling with Contrastive Losses	ViT	Transformer	MIMIC-CXR, ROCO	Masked modeling + Contrastive Loss	Early Fusion	VQA, Classification

8. Ethical and legal considerations in medical VLMs

The deployment of VLMs in medical settings necessitates rigorous attention to ethical and legal responsibilities, especially concerning patient privacy, data ownership, and algorithmic fairness. Medical data [256] is highly sensitive, and improper use or storage may violate regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the General Data Protection Regulation (GDPR) in Europe. Ensuring data anonymization, secure storage, and controlled access are foundational to ethical compliance [257]. In addition, approaches including FL and differential privacy are being explored to enable AI model training without directly exposing sensitive patient data [258].

Moreover, questions of data ownership, whether patients, hospitals, or third-party AI developers control the data, remain a gray area that requires clearer policies and informed consent practices [259,260]. The complexity increases when data is shared across institutions or countries, necessitating harmonized legal frameworks to prevent misuse and protect patient rights [261,262]. Algorithmic bias is another critical concern, as VLMs trained on non-representative or imbalanced datasets may yield unequal outcomes across patient demographics, potentially exacerbating health disparities [263]. Ethical AI development requires continuous bias auditing and mitigation strategies, including diverse data collection, fairness-aware model design, and transparency regarding model limitations [264].

Regulatory compliance must also be maintained throughout the model lifecycle, including requirements for model auditability, validation, and post-deployment monitoring to detect performance degradation or unintended consequences [265]. Ethical frameworks such as the EU AI Act and guidelines from the World Health Organization emphasize human oversight, transparency, and explainability as key principles in AI-driven healthcare [266,267]. Furthermore, the psychological impact on patients and clinicians should be considered, as overreliance on AI systems or opaque decision-making processes can undermine clinical judgment and patient autonomy [268]. The integration of VLMs should support collaborative human-AI interaction, ensuring that final clinical decisions remain with qualified healthcare professionals.

Future research should not only focus on improving model accuracy but also on embedding ethical safeguards to ensure patient trust, legal accountability, and responsible innovation. This includes developing standardized protocols for ethical AI development, promoting interdisciplinary collaboration between technologists, clinicians, ethicists, and legal experts, and fostering public engagement to shape socially acceptable AI healthcare applications [269–271].

8.1. Barriers and opportunities for clinical adoption of medical VLMs

Despite the remarkable advancements of VLMs in medical imaging and diagnostics, their adoption in actual hospital settings remains limited. Translating these models from research labs into real-world clinical practice involves a complex interplay of technical, regulatory, and human-centered challenges that are often underexplored in academic studies [247].

Workflow Integration. Clinical environments operate on tightly regulated workflows involving Picture Archiving and Communication Systems (PACS), Radiology Information Systems (RIS), and Electronic Medical Records (EMRs). Integrating VLMs into these systems requires robust interoperability and real-time processing to avoid disrupting existing practices. For example, models like BioViL have demonstrated effectiveness in generating radiology report impressions, but these results are often based on retrospective datasets [272]. To be clinically useful, VLMs must function reliably within hospital IT infrastructure and support clinicians without increasing their workload [273].

Limited Real-World Trials. Very few VLMs have progressed beyond the research phase to clinical pilot studies. Notably, BioViL has seen exploratory use in teaching hospitals, and LLaVA-Med has been applied in interactive diagnostic tools for medical education [272]. However, these deployments typically occur in sandboxed or controlled environments rather than in routine care settings. Comprehensive evaluations involving prospective trials, patient data variability, and clinician feedback are essential to validate their effectiveness under real-world conditions [265].

Regulatory and Legal Challenges. Gaining regulatory approval for clinical deployment remains a major hurdle. Agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) require rigorous validation, transparency, and explainability before approving AI systems for clinical use [267]. Although some image classification and segmentation models have received

clearance [228], VLMs, due to their multimodal and generative nature, lack standard approval pathways. As of mid-2025, no major medical VLM has received FDA or EMA certification for diagnostic deployment [266].

Trust and Human-in-the-Loop Design. Clinician acceptance is another key factor. Studies show that clinicians are hesitant to rely on black-box AI systems, especially in high-stakes scenarios like cancer diagnosis or surgical planning [268]. Tools that provide interpretable outputs, via visual methods such as Grad-CAM or attention heatmaps, or through textual justifications, are more likely to be adopted [248,253]. Ensuring that VLMs support, rather than replace, clinical judgment is crucial to building trust and facilitating responsible human-AI collaboration [269].

Infrastructure and Security Constraints. Many hospitals lack the computational infrastructure to support real-time inference from large VLMs, especially in low-resource or rural settings [257]. These models typically require GPU acceleration or high-performance computing clusters [274]. Moreover, strict data privacy regulations (such as HIPAA and GDPR) often prevent the use of cloud-based services for medical inference, demanding secure, on-premise solutions [261,275]. This creates a significant barrier to scalable deployment across diverse healthcare systems.

Path Forward. Bridging the gap between research and clinical adoption requires collaborative development between AI researchers, clinicians, hospital administrators, and regulatory experts [269,271]. Future efforts should prioritize lightweight, explainable, and regulation-ready VLMs [276]. Furthermore, investing in prospective clinical trials, benchmarking tools that assess model impact on diagnostic accuracy and workflow efficiency, and designing real-time, user-friendly interfaces will accelerate their integration into healthcare practice [265,267].

9. Current challenges and future trends

The integration of language models (LMs) into healthcare has the potential to revolutionize patient care, diagnostics, and treatment planning. However, several significant challenges [277] must be addressed to fully realize this potential and ensure safe, effective deployment in clinical settings.

1. High-quality and large-scale medical datasets are not readily available. One major hurdle in developing robust medical VLMs is the limited availability of large, high-quality medical and clinical datasets. This scarcity impedes thorough training, reducing the models' effectiveness, especially in handling complex or rare clinical scenarios [218]. A promising solution to this limitation is the use of large context windows and retrieval-augmented generation (RAG). By expanding the model's context through relevant, retrieved information, RAG can enhance performance in real-world applications [278]. Investigating pre-training within the RAG architecture provides an exciting research direction, offering increased resilience and flexibility for VLMs, particularly in novel medical cases.

2. Patient data privacy concerns during model training. Maintaining patient confidentiality while utilizing large-scale datasets is a critical concern. FL addresses this by enabling decentralized training across multiple data holders, ensuring sensitive data remains local and secure [279,280]. In FL, models are trained across institutions, sharing only model weights rather than raw data, addressing privacy concerns while supporting training on diverse datasets.

3. Inadequacy of traditional evaluation metrics for clinical tasks. Traditional evaluation metrics often do not capture the complexities of clinical language, making it difficult to reliably assess VLM performance [196]. This is particularly problematic when evaluating the precision of medical results or answering open-ended medical questions. To overcome this, developing and adopting specialized metrics tailored to medical VQA and RG is essential. These metrics would assess not only efficiency but also characteristics like effectiveness, stability, and adaptability. Establishing these evaluation standards is crucial for precise assessments and driving continuous improvements in medical VLMs.

4. Risk of hallucinations and misalignment in generated outputs. Artifacts where VLMs produce outputs unrelated to the input data or deviate from the intended information pose a significant risk in medical applications. These artifacts can lead to incorrect diagnoses or inappropriate treatment suggestions, which could have serious consequences. Misalignment between visual and textual data is a key factor contributing to hallucinations [281]. Addressing this requires better alignment of these modalities. For example, models such as LLaVA-RLHF [281] have made strides by using reinforcement learning from human feedback (RLHF) to improve modality alignment. Further research in this area is crucial to minimize hallucination risks and ensure medical VLMs generate outputs based on factual medical knowledge.

5. Catastrophic forgetting during continual learning. Catastrophic forgetting remains a critical challenge in VLM development [282,283]. This occurs when a model, while learning new data, unintentionally erases or distorts previously acquired knowledge, reducing overall performance. Balancing fine-tuning is key, too much fine-tuning can exacerbate forgetting, while moderate fine-tuning allows adaptation without loss of previous knowledge [278]. Continual learning approaches offer a promising solution by enabling VLMs to adapt to new data without forgetting prior knowledge. Integrating adapters within this framework can provide task-specific adaptations while preserving foundational knowledge, helping mitigate catastrophic forgetting [283–286].

6. Need for interdisciplinary collaboration and ethical deployment. The clinical validation and adoption of VLMs requires strong coordination between AI/ML researchers and healthcare professionals. Ethical implementation is critical to building trust [287], aligning models with clinical needs, and ensuring effective integration into healthcare systems. Fostering collaborative partnerships between technologists and medical experts will be key to responsibly deploying VLMs and ensuring their positive impact on healthcare outcomes.

Future research directions. Addressing these challenges makes it possible to create more reliable, effective, and safe medical VLMs. Future advancements should focus on the following strategies:

1. **Develop and adopt specialized metrics** for evaluating medical VLM performance.
2. **Investigate pre-training within the RAG architecture** to improve model resilience and flexibility.
3. **Explore federated learning** to secure patient privacy while enabling collaborative training across decentralized data sources.
4. **Align visual and textual data** effectively to reduce hallucinations and ensure medically accurate outputs.
5. **Leverage continual learning** approaches to prevent catastrophic forgetting while allowing models to adapt to new tasks.

CRediT authorship contribution statement

Gul E Arzu: Writing – original draft, Methodology, Conceptualization. **Muhammad Umar:** Visualization, Data curation. **Muhammad Fayaz:** Writing – review & editing, Visualization, Investigation. **Usman Ali:** Writing – review & editing. **L. Minh Dang:** Writing – review & editing, Formal analysis. **Hyeonjoon Moon:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1 A03038540) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries(IPET) through Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) (RS- 2025-02218444) and by National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (RS-2026-25473871).

Data availability

Data will be made available on request.

References

[1] S. Biswal, P. Zhuang, A. Pyrros, N. Siddiqui, S. Koyejo, J. Sun, EMIXER: End-to-end multimodal X-ray generation via self-supervision, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 297–324.
[2] G. Liu, T.-M.H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically accurate chest x-ray report generation, in: Machine Learning for Healthcare Conference, PMLR, 2019, pp. 249–269.
[3] T. Schlegl, S.M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, G. Langs, Predicting semantic descriptions from medical images with convolutional neural networks, in: International Conference on Information Processing in Medical Imaging, Springer, 2015, pp. 437–448.
[4] X. Wang, Y. Peng, L. Lu, Z. Lu, R.M. Summers, Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9049–9058.
[5] J. Kalpathy-Cramer, W. Hersh, Multimodal medical image retrieval: image categorization to improve search precision, in: Proceedings of the International Conference on Multimedia Information Retrieval, 2010, pp. 165–174.
[6] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[7] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.
[8] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[10] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 46 (8) (2024) 5625–5644.
[11] F. Bordes, R.Y. Pang, A. Ajay, A.C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, et al., An introduction to vision-language modeling, 2024, arXiv preprint arXiv:2405.17247.
[12] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NaacL-HLT, Vol. 1, Minneapolis, Minnesota, 2019, p. 2.
[13] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, PMLR, 2023, pp. 19730–19742.
[14] I. Cheong, K. Xia, K.K. Feng, Q.Z. Chen, A.X. Zhang, (A) I am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 2454–2469.
[15] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, Chatlaw: Open-source legal large language model with integrated external knowledge bases, 2023, arXiv preprint arXiv:2306.16092.
[16] A. Izzidien, H. Sargeant, F. Steffek, LLM vs. Lawyers: Identifying a subset of summary judgments in a large UK case law dataset, 2024, arXiv preprint arXiv:2403.04791.
[17] M. Irfan, L. Murray, Micro-credential: A guide to prompt writing and engineering in higher education: A tool for artificial intelligence in LLM, 2023, (Accessed 19 May 2025). URL https://researchrepository.ul.ie/articles/preprint/Micro-_Credential_A_guide_to_prompt_writing_and_engineering_in_higher_education_A_tool_for_Artificial_Intelligence_in_LLM/22868414.
[18] S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny, et al., Empowering education with llms-the next-gen interface and content generation, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 32–37.
[19] P. Shu, H. Zhao, H. Jiang, Y. Li, S. Xu, Y. Pan, Z. Wu, Z. Liu, G. Lu, L. Guan, et al., LLMs for coding and robotics education, 2024, arXiv preprint arXiv:2402.06116.
[20] Z. Liu, T. Zhong, Y. Li, Y. Zhang, Y. Pan, Z. Zhao, P. Dong, C. Cao, Y. Liu, P. Shu, et al., Evaluating large language models for radiology natural language processing, 2023, arXiv preprint arXiv:2307.13693.
[21] C. Ma, Z. Wu, J. Wang, S. Xu, Y. Wei, Z. Liu, F. Zeng, X. Jiang, L. Guo, X. Cai, et al., An iterative optimizing framework for radiology report summarization with ChatGPT, IEEE Trans. Artif. Intell. (2024).
[22] L. Zhang, Z. Liu, L. Zhang, Z. Wu, X. Yu, J. Holmes, H. Feng, H. Dai, X. Li, Q. Li, et al., Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology, Med. Phys. 51 (3) (2024) 2187–2199.
[23] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Trans. Knowl. Discov. Data 18 (6) (2024) 1–32.
[24] H. Zhao, Q. Ling, Y. Pan, T. Zhong, J.-Y. Hu, J. Yao, F. Xiao, Z. Xiao, Y. Zhang, S.-H. Xu, et al., Ophtha-llama2: A large language model for ophthalmology, 2023, arXiv preprint arXiv:2312.04906.
[25] T. Zhong, W. Zhao, Y. Zhang, Y. Pan, P. Dong, Z. Jiang, X. Kui, Y. Shang, L. Yang, Y. Wei, et al., Chatradio-valuer: a chat large language model for generalizable radiology report generation based on multi-institution and multi-system data, 2023, arXiv preprint arXiv:2310.05242.
[26] Z. Huang, F. Bianchi, M. Yuksekgonul, T.J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature Med. 29 (9) (2023) 2307–2316.
[27] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P.K. Anand, R. Krishna, L. Shapiro, Quilt-1m: One million image-text pairs for histopathology, Adv. Neural Inf. Process. Syst. 36 (2023) 37995–38017.
[28] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.
[29] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2023, arXiv preprint arXiv:2303.00915.
[30] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, P. Fung, Enabling multimodal generation on clip via vision-language knowledge distillation, 2022, arXiv preprint arXiv:2203.06386.
[31] Z. Qi, Y. Fang, M. Zhang, Z. Sun, T. Wu, Z. Liu, D. Lin, J. Wang, H. Zhao, Gemini vs GPT-4V: A preliminary comparison and combination of vision-language models through qualitative cases, 2023, arXiv preprint arXiv:2312.15011.
[32] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.
[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[34] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, J. Gao, Unified vision-language pre-training for image captioning and vqa, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 13041–13049.
[35] J. Roos, R. Martin, R. Kaczmarczyk, et al., Evaluating bard Gemini pro and GPT-4 vision against student performance in medical visual question answering: Comparative case study, JMIR Form. Res. 8 (1) (2024) e57592.
[36] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, Y.J. Lee, C. Li, Learning customized visual models with retrieval-augmented knowledge, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15148–15158.
[37] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).
[38] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, 2019, arXiv preprint arXiv:1908.07490.
[39] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv preprint arXiv:1908.03557.
[40] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, Vl-bert: Pre-training of generic visual-linguistic representations, 2019, arXiv preprint arXiv:1908.08530.
[41] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European Conference on Computer Vision, Springer, 2020, pp. 104–120.
[42] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer, 2020, pp. 121–137.

[43] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[44] T. Do, B.X. Nguyen, E. Tjiputra, M. Tran, Q.D. Tran, A. Nguyen, Multiple meta-model quantifying for medical visual question answering, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 64–74.

[45] H. Gong, G. Chen, M. Mao, Z. Li, G. Li, Vqamix: Conditional triplet mixup for medical visual question answering, IEEE Trans. Med. Imaging 41 (11) (2022) 3332–3343.

[46] L. Yin, L. Wang, S. Lu, R. Wang, Y. Yang, B. Yang, S. Liu, A. AlSanad, S.A. AlQahtani, Z. Yin, et al., Convolution-transformer for image feature extraction, CMES Comput. Model. Eng. Sci. 141 (1) (2024).

[47] G. Cao, Y. Wang, H. Zhu, Z. Lou, L. Yu, Transferable adversarial attack on image tampering localization, J. Vis. Commun. Image Represent. 102 (2024) 104210.

[48] J. Chen, Y. Jiang, D. Yang, M. Li, J. Wei, Z. Qian, L. Zhang, Can llms' tuning methods work in medical multimodal domain? in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 112–122.

[49] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, T.-H. Chang, Multi-modal masked autoencoders for medical vision-and-language pre-training, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 679–689.

[50] J. Chen, D. Yang, Y. Jiang, Y. Lei, L. Zhang, MISS: A generative pre-training and fine-tuning approach for med-VQA, in: International Conference on Artificial Neural Networks, Springer, 2024, pp. 299–313.

[51] S. Subramanian, L.L. Wang, B. Bogin, S. Mehta, M. Van Zuylen, S. Parasa, S. Singh, M. Gardner, H. Hajishirzi, Medicat: A dataset of medical images, captions, and textual references, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2112–2120.

[52] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, Adv. Neural Inf. Process. Syst. 36 (2023) 28541–28564.

[53] X. Li, L. Li, Y. Jiang, H. Wang, X. Qiao, T. Feng, H. Luo, Y. Zhao, Vision-language models in medical image analysis: From simple fusion to general large models, Inf. Fusion 118 (2025) 102995, http://dx.doi.org/10.1016/j.inffus.2025.102995.

[54] W. Yang, M. Chen, H. Liu, A survey of multimodal large language models for clinical decision support, J. Biomed. Inform. 147 (2024) 104506.

[55] R. Xu, T. Wang, L. Zhang, A survey on vision-language pretraining for medical applications: Datasets, methods, and challenges, IEEE Trans. Med. Imaging 42 (12) (2023) 3312–3325.

[56] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp. 2–25.

[57] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3876–3887.

[58] Z. Wan, C. Liu, M. Zhang, J. Fu, B. Wang, S. Cheng, L. Ma, C. Quilodrán-Casas, R. Arcucci, Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias, Adv. Neural Inf. Process. Syst. 36 (2024).

[59] C. Wu, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21372–21383.

[60] H. Luo, Z. Zhou, C. Royer, A. Sekuboyina, B. Menze, DeViDe: Faceted medical knowledge for improved medical vision-language pre-training, 2024, arXiv preprint arXiv:2404.03618.

[61] C. Shen, C. Wang, M. Huang, N. Xu, S. van der Zwaag, W. Xu, A generic high-throughput microstructure classification and quantification method for regular SEM images of complex steel microstructures combining EBSD labeling and deep learning, J. Mater. Sci. Technol. 93 (2021) 191–204.

[62] G.E. Arzu, M. Umar, A. Khan, U. Ali, L.M. Dang, H. Moon, Adaptive multimodal emotion detection for mental health monitoring using deep learning, Inform. Sci. (2026) 123385.

[63] R. Perera, D. Guzzetti, V. Agrawal, Optimized and autonomous machine learning framework for characterizing pores, particles, grains and grain boundaries in microstructural images, Comput. Mater. Sci. 196 (2021) 110524.

[64] M. Ilett, J. Wills, P. Rees, S. Sharma, S. Micklethwaite, A. Brown, R. Brydson, N. Hondow, Application of automated electron microscopy imaging and machine learning to characterise and quantify nanoparticle dispersion in aqueous media, J. Microsc. 279 (3) (2020) 177–184.

[65] N. Khatavkar, S. Swetlana, A.K. Singh, Accelerated prediction of Vickers hardness of Co-and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning, Acta Mater. 196 (2020) 295–303.

[66] S. Akers, E. Kautz, A. Trevino-Gavito, M. Olszta, B.E. Matthews, L. Wang, Y. Du, S.R. Spurgeon, Rapid and flexible segmentation of electron microscopy data using few-shot machine learning, npj Comput. Mater. 7 (1) (2021) 187.

[67] Z. Chen, X. Liu, J. Yang, E. Little, Y. Zhou, Deep learning-based method for SEM image segmentation in mineral characterization, an example from Duvernay Shale samples in Western Canada Sedimentary Basin, Comput. Geosci. 138 (2020) 104450.

[68] J. Rausch, D. Jaramillo-Vogel, S. Perseguers, N. Schnidrig, B. Grobéty, P. Yajan, Automated identification and quantification of tire wear particles (TWP) in airborne dust: SEM/EDX single particle analysis coupled to a machine learning classifier, Sci. Total Environ. 803 (2022) 149832.

[69] Z.A. Nazi, W. Peng, Large language models in healthcare and medical domain: A review, Informatics 11 (3) (2024) 57.

[70] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[71] H. Paulino, A.A. Matos, J. Cederquist, M. Giunti, J. Matos, A. Ravara, Sound atomicity inference for data-centric synchronization, 2023, arXiv preprint arXiv:2309.05483.

[72] A. Zhakatayev, L. Tlebaldiyeva, Long-range longitudinal electric wave in vacuum radiated by electric dipole: Part I, Radio Sci. 55 (5) (2020) 1–18.

[73] J. Dhar, N. Zaidi, M. Haghighat, S. Roy, P. Goyal, A. Alavi, V. Kumar, Multimodal fusion learning with dual attention for medical imaging, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, 2025, pp. 4362–4371.

[74] J. Kuang, Y. Shen, J. Xie, H. Luo, Z. Xu, R. Li, Y. Li, X. Cheng, X. Lin, Y. Han, Natural language understanding and inference with mllm in visual question answering: A survey, ACM Comput. Surv. 57 (8) (2025) 1–36.

[75] H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, H. Yang, Prompt tuning for generative multimodal pretrained models, 2022, arXiv preprint arXiv:2208.02532.

[76] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.

[77] S. Shi, W. Liu, B2-ViT Net: Broad vision transformer network with broad attention for seizure prediction, IEEE Trans. Neural Syst. Rehabil. Eng. 32 (2023) 178–188.

[78] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[79] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinform. 23 (6) (2022) bbac409.

[80] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, D. Shen, Doctorglm: Fine-tuning your chinese doctor is not a herculean task, 2023, arXiv preprint arXiv:2304.01097.

[81] P. Shi, J. Qiu, S.M.D. Abaxi, H. Wei, F.P.-W. Lo, W. Yuan, Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation, Diagnostics 13 (11) (2023) 1947.

[82] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array 3 (2019) 100004.

[83] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al., Capabilities of gemini models in medicine, 2024, arXiv preprint arXiv:2404.18416.

[84] W. Xiong, G. Zhang, D. Yan, L. Cao, X. Huang, D. Li, Multichannel feature fusion network-based technique for heart sound signal classification and recognition, Expert Syst. Appl. 273 (2025) 126839.

[85] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, W. Xie, Pmc-clip: Contrastive language-image pre-training using biomedical documents, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 525–536.

[86] D. Anand, V. Singhal, D.D. Shanbhag, S. KS, U. Patil, C. Bhushan, K. Manickam, D. Gui, R. Mullick, A. Gopal, et al., One-shot localization and segmentation of medical images with foundation models, 2023, arXiv preprint arXiv:2310.18642.

[87] G.E. Arzu, M. Fayaz, U. Ali, L.M. Dang, H. Moon, Efficient transformer-based semantic segmentation of colonic polyps using SegFormer, Neurocomputing (2025) 132339.

[88] K. Poudel, M. Dhakal, P. Bhandari, R. Adhikari, S. Thapaliya, B. Khanal, Exploring transfer learning in medical image segmentation using vision-language models, 2023, arXiv preprint arXiv:2308.07706.

[89] J. Jang, D. Kyung, S.H. Kim, H. Lee, K. Bae, E. Choi, Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders, Sci. Rep. 14 (1) (2024) 23199.

[90] Z. Zhao, S. Wang, J. Gu, Y. Zhu, L. Mei, Z. Zhuang, Z. Cui, Q. Wang, D. Shen, Chatcad+: Towards a universal and reliable interactive cad using llms, IEEE Trans. Med. Imaging (2024).

[91] M. Keicher, K. Zaripova, T. Czempiel, K. Mach, A. Khakzar, N. Navab, FlexR: few-shot classification with language embeddings for structured reporting of chest x-rays, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1493–1508.

[92] T. van Sonsbeek, M. Worring, X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation, in: International Conference on Information Processing in Medical Imaging, Springer, 2023, pp. 471–482.

[93] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2023) 34892–34916.

[94] Y. Ye, Y. Lu, H. Su, Y. Tian, S. Jin, G. Li, Y. Yang, L. Jiang, Z. Zhou, X. Wei, et al., A hybrid bioelectronic retina-probe interface for object recognition, Biosens. Bioelectron. 279 (2025) 117408.

[95] Y. Zhang, Y. Pan, T. Zhong, P. Dong, K. Xie, Y. Liu, H. Jiang, Z. Wu, Z. Liu, W. Zhao, et al., Potential of multimodal large language models for data mining of medical images and free-text reports, Meta-Radiology 2 (4) (2024) 100103.

[96] W. Zhao, J. Chen, Z. Meng, D. Mao, R. Song, W. Zhang, Vlmpc: Vision-language model predictive control for robotic manipulation, 2024, arXiv preprint arXiv:2407.09829.

[97] Z. Xu, E.-H. Li, J. Liu, Y.-J. Zhang, R. Xiao, X.-Z. Chen, Z.-H. Zhong, X.-J. Tang, L.-J. Fu, H. Zhang, et al., Postpartum hemorrhage emerges as a key outcome of maternal SARS-CoV-2 omicron variant infection surge across pregnancy trimesters, J. Infect. Public Health 18 (6) (2025) 102733.

[98] X. Li, W. Chen, X. Duan, X. Gu, C. Li, Collaborative surgical instrument segmentation for monocular depth estimation in minimally invasive surgery, Med. Image Anal. (2025) 103765.

[99] Y. Zheng, H.Y. Koh, M. Yang, L. Li, L.T. May, G.I. Webb, S. Pan, G. Church, Large language models in drug discovery and development: From disease mechanisms to clinical trials, 2024, arXiv preprint arXiv:2409.04481.

[100] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac, et al., Transformers in healthcare: A survey, 2023, arXiv preprint arXiv:2307.00067.

[101] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, 2021, arXiv preprint arXiv:2108.07258.

[102] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, ACM Trans. Intell. Syst. Technol. (TIST) 10 (1) (2020) 1–42.

[103] P. Soviany, R.T. Ionescu, P. Rota, N. Sebe, Curriculum learning: A survey, Int. J. Comput. Vis. 130 (6) (2022) 1526–1565.

[104] V. Rani, S.T. Nabi, M. Kumar, A. Mittal, K. Kumar, Self-supervised learning: A succinct review, Arch. Comput. Methods Eng. 30 (4) (2023) 2761–2775.

[105] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 9694–9705.

[106] X. Xue, H.-M. Hu, Z. He, H. Zheng, Towards multi-source illumination color constancy through physics-based rendering and spectral power distribution embedding, IEEE Trans. Comput. Imaging (2025).

[107] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[108] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.

[109] N. Lambert, Reinforcement learning from human feedback, 2025, arXiv preprint arXiv:2504.12501.

[110] A. Coronato, M. Naeem, G. De Pietro, G. Paragliola, Reinforcement learning for intelligent healthcare applications: A survey, Artif. Intell. Med. 109 (2020) 101964.

[111] M. Ren, B. Cao, H. Lin, C. Liu, X. Han, K. Zeng, W. Guanglu, X. Cai, L. Sun, Learning or self-aligning? rethinking instruction fine-tuning, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 6090–6105.

[112] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, M. Keicher, Radialog: A large vision-language model for radiology report generation and conversational assistance, 2023, arXiv preprint arXiv:2311.18681.

[113] I. Hartsock, G. Rasool, Vision-language models for medical report generation and visual question answering: A review, Front. Artif. Intell. 7 (2024) 1430984.

[114] M. Fayaz, L.M. Dang, H. Moon, Enhancing land cover classification via deep ensemble network, Knowl.-Based Syst. 305 (2024) 112611.

[115] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, Iclr 1 (2) (2022) 3.

[116] C. Shu, B. Chen, F. Liu, Z. Fu, E. Shareghi, N. Collier, Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities, Second. Vis. Med-Alpaca: Parameter-Effic. Biomed. LLM Vis. Capab. 2023 (2023).

[117] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059.

[118] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A frontier large vision-language model with versatile abilities, 2023, p. 3, arXiv preprint arXiv:2308.12966. 1 (2).

[119] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597.

[120] J. Cho, J. Lei, H. Tan, M. Bansal, Unifying vision-and-language tasks via text generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 1931–1942.

[121] T. Koleilat, H. Asgariandehkordi, H. Rivaz, Y. Xiao, Biomedcoop: Learning to prompt for biomedical vision-language models, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14766–14776.

[122] Q. Cao, Y. Chen, L. Lu, H. Sun, Z. Zeng, X. Yang, D. Zhang, Generalized domain prompt learning for accessible scientific vision-language models, Nexus 2 (2) (2025).

[123] N. Hussein, F. Shamshad, M. Naseer, K. Nandakumar, Promptsmooth: Certifying robustness of medical vision-language models via prompt learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 698–708.

[124] Y. Zhao, E. Zhong, C. Yuan, Y. Li, M. Zhao, C. Li, J. Hu, W. Liu, C. Liu, Med-VLM: Enhancing medical image segmentation accuracy through vision-language model, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 7283–7293.

[125] Q. Cao, Z. Xu, Y. Chen, C. Ma, X. Yang, Domain prompt learning with quaternion networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26637–26646.

[126] Z. Qin, H. Yi, Q. Lao, K. Li, Medical image understanding with pretrained vision language models: A comprehensive study, 2022, arXiv preprint arXiv:2209.15517.

[127] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C.M. Friedrich, Radiology objects in context (roco): a multimodal image dataset, in: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.

[128] A.G. Barreto, J.M. de Oliveira, F.N.B. Gois, P.C. Cortez, V.H.C. de Albuquerque, A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks, Bioengineering 10 (9) (2023) 1098.

[129] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Trans. Big Data 7 (3) (2019) 535–547.

[130] O. Pelka, S. Koitka, C.M. Friedrich, ROCO: An image dataset for medical image understanding, in: Medical Imaging 2018: Image Processing, Vol. 10574, SPIE, 2018, p. 105741M.

[131] D. Demner-Fushman, W. Rogers, A.R. Aronson, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2016) 304–310.

[132] A.E. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019, arXiv preprint arXiv:1901.07042.

[133] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[134] M. Saeed, M. Villarroel, A.T. Reisner, G.D. Clifford, L.-W.H. Lehman, G. Moody, T. Heldt, T.T. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (5) (2011) 952.

[135] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, R. Mark, Mimic-iv, PhysioNet (2020) Available online at: https://physionet.org/content/mimiciv/1.0/. (Accessed 23 August 2021).

[136] A. Bustos, A. Pertusa, J.-M. Salinas, M. De La Iglesia-Vaya, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797.

[137] M.T. Amorim, Enhancing Medical Image Captioning Through Multimodal Alignment: A Lightweight Approach (Master's thesis), Universidade do Porto (Portugal), 2025.

[138] J. Park, S. Kim, B. Yoon, J. Hyun, K. Choi, M4CXR: exploring multitask potentials of multimodal large language models for chest X-ray interpretation, IEEE Trans. Neural Netw. Learn. Syst. (2025).

[139] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, X.-M. Wu, Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, IEEE, 2021, pp. 1650–1654.

[140] S.S. Noor Mohamed, K. Srinivasan, A comprehensive interpretation for medical VQA: Datasets, techniques, and challenges, J. Intell. Fuzzy Systems 44 (4) (2023) 5803–5819.

[141] Q. Jin, B. Dhingra, Z. Liu, W.W. Cohen, X. Lu, Pubmedqa: A dataset for biomedical research question answering, 2019, arXiv preprint arXiv:1909.06146.

[142] A. Ben Abacha, S.A. Hasan, V.V. Datla, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes, 2019, p. 11, 9-12 September 2019.

[143] Z. He, X. Yang, Y. Li, F. Xing, L. Yang, PathVQA: Visual question answering for pathology images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 770–779.

[144] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, Z. Lu, NegBio: a high-performance tool for negation and uncertainty detection in radiology reports, AMIA Summits Transl. Sci. Proc. 2018 (2018) 188.

[145] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 590–597.

[146] M. Kayser, C. Emde, O.-M. Camburu, G. Parsons, B. Papiez, T. Lukasiewicz, Explaining chest x-ray pathologies in natural language, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 701–713.

[147] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc.: JAMIA 24 (1) (2016) 198.

[148] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, Sci. Data 6 (1) (2019) 96.

[149] S. Mohammed, J. Matos, M. Doutreligne, L.A. Celi, T. Struja, Focus: Big data: Racial disparities in invasive ICU treatments among septic patients: High resolution electronic health records analysis from MIMIC-IV, Yale J. Biol. Med. 96 (3) (2023) 293.

[150] A.E. Johnson, D.J. Stone, L.A. Celi, T.J. Pollard, The MIMIC Code Repository: enabling reproducibility in critical care research, J. Am. Med. Inform. Assoc. 25 (1) (2018) 32–39.

[151] V. Ramesh, N.A. Chi, P. Rajpurkar, Improving radiology report generation systems by removing hallucinated references to non-existent priors, in: Machine Learning for Health, PMLR, 2022, pp. 456–473.

[152] L. Tang, Y. Chen, Q. Xiang, J. Xiang, Y. Tang, J. Li, The association between IL18, FOXP3 and IL13 genes polymorphisms and risk of allergic rhinitis: a meta-analysis, Inflamm. Res. 69 (9) (2020) 911–923.

[153] K. Lo, L.L. Wang, M. Neumann, R. Kinney, D.S. Weld, S2ORC: The semantic scholar open research corpus, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4969–4983.

[154] Y. Ohkura, J. Shindoh, S. Haruta, D. Kaji, Y. Ota, T. Fujii, M. Hashimoto, G. Watanabe, M. Matsuda, Primary adrenal lymphoma possibly associated with Epstein–Barr virus reactivation due to immunosuppression under methotrexate therapy, Medicine 94 (31) (2015) e1270.

[155] B. Boecking, N. Usuyama, S. Bannur, D.C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, et al., Making the most of text semantics to improve biomedical vision–language processing, in: European Conference on Computer Vision, Springer, 2022, pp. 1–21.

[156] R. Bigolin Lanfredi, M. Zhang, W.F. Auffermann, J. Chan, P.-A.T. Duong, V. Srikumar, T. Drew, J.D. Schroeder, T. Tasdizen, REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays, Sci. Data 9 (1) (2022) 350.

[157] S. Luan, X. Yu, S. Lei, C. Ma, X. Wang, X. Xue, Y. Ding, T. Ma, B. Zhu, Deep learning for fast super-resolution ultrasound microvessel imaging, Phys. Med. Biol. 68 (24) (2023) 245023.

[158] R. Jiang, X. Yin, P. Yang, L. Cheng, J. Hu, J. Yang, Y. Wang, X. Fu, L. Shang, L. Li, et al., A transformer-based weakly supervised computational pathology method for clinical-grade diagnosis and molecular marker discovery of gliomas, Nat. Mach. Intell. 6 (8) (2024) 876–891.

[159] S. Wang, K. Zhang, A. Liu, Flat-Lattice-CNN: A model for Chinese medical-named-entity recognition, PLoS One 20 (9) (2025) e0331464.

[160] Y. Lin, Z. Zhu, Y. Guo, Z. Ou, S. Yao, M. Song, Exploring inter-domain wasserstein metric for adaptive object detection, IEEE Trans. Multimed. (2026).

[161] J.J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Sci. Data 5 (1) (2018) 1–10.

[162] F. Chen, Y. Yu, J. Yi, T. Zhang, J. Zhao, W. Jia, J. Yu, MCLL-diff: multiconditional low-light image enhancement based on diffusion probabilistic models, IEEE Sens. J. (2025).

[163] S. Wang, J. Bian, X. Huang, H. Zhou, S. Zhu, PubLabeler: enhancing automatic classification of publications in UniProtKB using protein textual description and PubMedBERT, IEEE J. Biomed. Health Inform. 29 (5) (2024) 3782–3791.

[164] E. Hellberg, Exploring GPT models as biomedical knowledge bases: By evaluating prompt methods for extracting information from language models pre-trained on scientific articles, 2023.

[165] F. Gao, C. Gao, X. Deng, C. Zhang, J. Huang, M. Xu, AIRPNet: Adaptive image restoration with privacy protection in steganographic domain, IEEE Trans. Pattern Anal. Mach. Intell. (2026).

[166] X. Yu, S. Luan, S. Lei, J. Huang, Z. Liu, X. Xue, T. Ma, Y. Ding, B. Zhu, Deep learning for fast denoising filtering in ultrasound localization microscopy, Phys. Med. Biol. 68 (20) (2023) 205002.

[167] A. Webb, Introduction to Biomedical Imaging, John Wiley & Sons, 2022.

[168] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S.A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, in: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-Working Notes, 2021, pp. 1081–1088, 21-24 September 2021.

[169] B. Ionescu, H. Müller, R. Péteri, A.B. Abacha, M. Sarrouti, D. Demner-Fushman, S.A. Hasan, S. Kozlovski, V. Liauchuk, Y.D. Cid, et al., Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 345–370.

[170] J. Geng, S. Zhao, W. Weizhe, Q. Li, Adaptive graph partitioning for clustering datasets with heterogeneous density, ACM Trans. Knowl. Discov. Data 20 (2) (2026) 1–27.

[171] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[172] J. Zhu, Y. Li, Y. Hu, S.K. Zhou, Embedding task knowledge into 3D neural networks via self-supervised learning, 2020, arXiv preprint arXiv:2006.05798.

[173] A.E. Johnson, T.J. Pollard, L. Shen, L.-w.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.

[174] S. Huang, B. Lv, R. Wang, K. Huang, Scheduling for mobile edge computing with random user arrivals—An approximate MDP and reinforcement learning approach, IEEE Trans. Veh. Technol. 69 (7) (2020) 7735–7750.

[175] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[176] K. Ganesan, Rouge 2.0: Updated and improved measures for evaluation of summarization tasks, 2018, arXiv preprint arXiv:1803.01937.

[177] R.R. Fernandes, Enhancing Contrastive Learning with Soft Prompts for Medical Image Captioning (Master's thesis), Universidade do Porto (Portugal), 2024.

[178] L. Fang, Y. Wang, Z. Liu, C. Zhang, S. Jegelka, J. Gao, B. Ding, Y. Wang, What is wrong with perplexity for long-context language modeling? 2024, arXiv preprint arXiv:2410.23771.

[179] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, Appl. Sci. 11 (14) (2021) 6421.

[180] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[181] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2016) 304–310.

[182] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2019, arXiv preprint arXiv:1904.09675.

[183] F. Liu, X. Chen, H. Xu, X. Zhou, F. Wang, J. Gao, T.-Y. Liu, Clinically accurate chest X-Ray report generation, 2021, arXiv preprint arXiv:2109.01180.

[184] A. Romanov, C. Shivade, Lessons from natural language inference in the clinical domain, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1586–1596.

[185] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, 2005, pp. 65–72.

[186] C. Meister, R. Cotterell, Language model evaluation beyond perplexity, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5328–5339.

[187] S. Jain, S. Biswal, P. Patel, T. Vaidya, W. Liao, M. Pomerantz, C.P. Langlotz, M.P. Lungren, RadGraph: Extracting clinical entities and relations from radiology reports, 2021, arXiv preprint arXiv:2106.14463.

[188] S.H. Park, K. Han, H.Y. Jang, J.E. Park, J.-G. Lee, D.W. Kim, J. Choi, Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis, Radiology 306 (1) (2023) 20–31.

[189] M. Gao, J.Y. Kim, H. Lee, S. Park, RadCliQ: A clinical and linguistic quality evaluation metric for radiology reports, Med. Image Anal. (2023).

[190] C.-W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, J. Pineau, How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2122–2132.

[191] Y. Zhang, M. Zhong, Q. Chen, W.Y. Wang, Optimizing the factual correctness of a summary: A study of summarizing radiology reports, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5108–5120.

[192] T. Sellam, D. Das, A.P. Parikh, BLEURT: Learning robust metrics for text generation, 2020, arXiv preprint arXiv:2004.04696.

[193] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[194] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A.Y. Ng, M.P. Lungren, CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, 2020, arXiv preprint arXiv:2004.09167.

[195] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E.P. Reis, E. Fonseca, H. Lee, Z. Shakeri, A. Ng, et al., Radiology report expert evaluation (rexval) dataset, 2023.

[196] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E.P. Reis, E.K.U.N. Fonseca, H.M.H. Lee, Z.S.H. Abad, A.Y. Ng, et al., Evaluating progress in automatic chest x-ray radiology report generation, Patterns 4 (9) (2023).

[197] X. Gan, GraphService: Topology-aware constructor for large-scale graph applications, ACM Trans. Archit. Code Optim. 22 (1) (2025) 1–24.

[198] X. Gan, T. Li, C. Gong, D. Li, D. Dong, J. Liu, K. Lu, GraphCSR: A degree-equalized CSR format for large-scale graph processing, Proc. VLDB Endow. 18 (11) (2025) 4255–4268.

[199] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, P. Rajpurkar, Multimodal image-text matching improves retrieval-based chest x-ray report generation, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 978–990.

[200] C. Tran, B. Pal, T. Stirrat, V. Shi, M. Umair, Recent advances in AI for radiology report generation: A brief review, BJR| Artif. Intell. (2026) ubag003.

[201] Z. Zhou, M. Shi, M. Wei, O. Alabi, Z. Yue, T. Vercauteren, Large model driven radiology report generation with clinical quality reinforcement learning, 2024, arXiv preprint arXiv:2403.06728.

[202] R. Stock, S. Denner, Y. Kirchhoff, C. Ulrich, M.R. Rokuss, S. Roy, N. Disch, K. Maier-Hein, From generalist to specialist: Incorporating domain-knowledge into flamingo for chest x-ray report generation, in: Medical Imaging with Deep Learning, 2024.

[203] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, 2023, p. 6, arXiv preprint arXiv:2303.00915. 2 (3).

[204] J.H. Moon, H. Lee, W. Shin, Y.-H. Kim, E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, IEEE J. Biomed. Health Inform. 26 (12) (2022) 6070–6080.

[205] O. Thawkar, A. Shaker, S.S. Mullappilly, H. Cholakkal, R.M. Anwer, S. Khan, J. Laaksonen, F.S. Khan, Xraygpt: Chest radiographs summarization using medical vision-language models, 2023, arXiv preprint arXiv:2306.07971.

[206] X. Zhao, Z. Liu, F. Liu, G. Li, Y. Dou, S. Peng, Report-concept textual-prompt learning for enhancing x-ray diagnosis, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2184–2193.

[207] J. Lau, S. Gayen, A. Abedin, et al., Vqa-rad: Visual question answering dataset for radiology, Med. Imaging (2018) 105790V.

[208] A. Pal, L.K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260.

[209] B. Li, W. Huang, Z. Gao, Y. Wang, Y. Shen, J. Lin, L. You, Y. Shen, S. Lin, W. Ouyang, et al., LLaVA-RadZ: Can multimodal large language models effectively tackle zero-shot radiology recognition? 2025, arXiv preprint arXiv:2503.07487.

[210] O.C. Thawakar, A.M. Shaker, S.S. Mullappilly, H. Cholakkal, R.M. Anwer, S. Khan, J. Laaksonen, F. Khan, Xraygpt: Chest radiographs summarization using large medical vision-language models, in: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, 2024, pp. 440–448.

[211] W. Gao, Z. Deng, Z. Niu, F. Rong, C. Chen, Z. Gong, W. Zhang, D. Xiao, F. Li, Z. Cao, et al., Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue, 2023, arXiv preprint arXiv:2306.12174.

[212] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. (HEALTH) 3 (1) (2021) 1–23.

[213] A.K. Tanwani, J. Barral, D. Freedman, Repsnet: Combining vision with language for automated medical reports, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 714–724.

[214] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.

[215] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, et al., MedDialog: Large-scale medical dialogue datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 9241–9250.

[216] A.B. Abacha, C. Shivade, D. Demner-Fushman, Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 370–379.

[217] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J.E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023, p. 6, See https://vicuna.lmsys.org (Accessed 14 April 2023). 2 (3).

[218] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E.P. Reis, P. Rajpurkar, Med-flamingo: a multimodal medical few-shot learner, in: Machine Learning for Health, ML4H, PMLR, 2023, pp. 353–367.

[219] M. Gao, J.Y. Kim, H. Lee, S. Park, BioViL: Masked vision–language modeling with contrastive losses for biomedical image–text representation, Med. Image Anal. 85 (2023) 102718, http://dx.doi.org/10.1016/j.media.2023.102718.

[220] L. Zhou, Y. Chen, X. Wu, X. Liu, J. Tang, Masked vision–language modeling with contrastive losses for clinical report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2023, pp. 11234–11244, http://dx.doi.org/10.1109/CVPR.2023.01123.

[221] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[222] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[223] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[224] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, 2016, arXiv preprint arXiv:1609.08144.

[225] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, Adv. Neural Inf. Process. Syst. 31 (2018).

[226] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[227] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain? in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.

[228] W. Song, X. Wang, Y. Guo, S. Li, B. Xia, A. Hao, Centerformer: a novel cluster center enhanced transformer for unconstrained dental plaque segmentation, IEEE Trans. Multimed. 26 (2024) 10965–10978.

[229] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in: Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21, Springer, 2011, pp. 52–59.

[230] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[231] G. Van Houdt, C. Mosquera, G. Nápoles, A review on the long short-term memory model, Artif. Intell. Rev. 53 (8) (2020) 5929–5955.

[232] C. Tang, Z. Wang, Y. Xie, Y. Fei, J. Luo, C. Wang, Y. Ying, P. He, R. Yan, Y. Chen, et al., Classification of distinct tendinopathy subtypes for precision therapeutics, Nat. Commun. 15 (1) (2024) 9460.

[233] L.-M. Zhan, B. Liu, L. Fan, J. Chen, X.-M. Wu, Medical visual question answering via conditional reasoning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2345–2354.

[234] B. Liu, L.-M. Zhan, L. Xu, X.-M. Wu, Medical visual question answering via conditional reasoning and contrastive learning, IEEE Trans. Med. Imaging 42 (5) (2022) 1532–1545.

[235] J. Liu, G. Jiang, C. Chu, Y. Li, Z. Wang, S. Hu, A formal model for multiagent Q-learning on graphs, Sci. China Inf. Sci. 68 (9) (2025) 192206.

[236] F. Liu, J.M. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, Y. Altun, Deplot: One-shot visual language reasoning by plot-to-table translation, 2022, arXiv preprint arXiv:2212.10505.

[237] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, Git: A generative image-to-text transformer for vision and language, 2022, arXiv preprint arXiv:2205.14100.

[238] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N.A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13484–13508.

[239] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible scaling laws for contrastive language-image learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2818–2829.

[240] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, et al., An empirical study of training end-to-end vision-and-language transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18166–18176.

[241] S. Eslami, G. de Melo, C. Meinel, Does clip benefit visual question answering in the medical domain as much as it does in the general domain? 2021, arXiv preprint arXiv:2112.13906.
[242] M. McCloskey, N.J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of Learning and Motivation, vol. 24, Elsevier, 1989, pp. 109–165.
[243] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[244] W. Zheng, G. Xu, S. Lu, J. Lyu, F. Bao, L. Yin, GNN: Core branches, integration strategies and applications, Comput. Model. Eng. Sci. 146 (1) (2026).
[245] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C.M. Friedrich, Overview of ImageCLEFmedical 2022–caption prediction and concept detection, in: CEUR Workshop Proceedings, Vol. 3180, CEUR Workshop Proceedings, 2022, pp. 1294–1307.
[246] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Adv. Neural Inf. Process. Syst. 35 (2022) 23716–23736.
[247] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Med. 25 (1) (2019) 44–56.
[248] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[249] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 782–791.
[250] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013, arXiv preprint arXiv:1312.6034.
[251] S.G. Picha, D. Al Chanti, A. Caplier, Trust but verify: Image-aware evaluation of radiology report generators, Mach. Learn. Appl. (2026) 100851.
[252] B. Liu, H. Du, J. Zhang, J. Jiang, X. Zhang, F. He, B. Niu, Developing a new sepsis screening tool based on lymphocyte count, international normalized ratio and procalcitonin (LIP score), Sci. Rep. 12 (1) (2022) 20002.
[253] M. Imran, Y. Lee, Multimodal vision–language models in medical imaging: A survey of retrieval, interpretability, and trust, IEEE Access (2026).
[254] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[255] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017) 4765–4774.
[256] G. Xu, X. Fan, S. Xu, Y. Cao, X.-B. Chen, T. Shang, S. Yu, Anonymity-enhanced sequential multi-signer ring signature for secure medical data sharing in IoMT, IEEE Trans. Inf. Forensics Secur. (2025).
[257] G. Kaissis, M.R. Makowski, D. Rückert, R.F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nat. Mach. Intell. 2 (6) (2020) 305–311.
[258] T. Li, A.K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Process. Mag. 37 (3) (2020) 50–60.
[259] E. Vayena, L. Madoff, Navigating the ethics of big data in public health, in: The Oxford Handbook of Public Health Ethics, 2019, pp. 354–367.
[260] T. Pham, Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use, R. Soc. Open Sci. 12 (5) (2025).
[261] D. Protection, General Data Protection Regulation, Intersoft Consulting, 2018, Accessed in October 24 (1).
[262] A. Renda, Artificial Intelligence. Ethics, Governance and Policy Challenges, CEPS Centre for European Policy Studies, 2019.
[263] D.M. Bowser, K. Mauricio, B.A. Ruscitti, W.H. Crown, American clusters: using machine learning to understand health and health care disparities in the United States, Health Aff. Sch. 2 (3) (2024) qxae017.
[264] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.
[265] W. Guidance, Ethics and Governance of Artificial Intelligence for Health, World Health Organization, 2021, pp. 1–165.
[266] W.H. Organization, Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance, World Health Organization, 2024.
[267] E.P. Vardas, M. Marketou, P.E. Vardas, Medicine, healthcare and the AI act: gaps, challenges and future implications, Eur. Heart J.-Digit. Health 6 (4) (2025) 833–839.
[268] E. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Hachette UK, 2019.
[269] B. Beets, T.P. Newman, E.L. Howell, L. Bao, S. Yang, Surveying public perceptions of artificial intelligence in health care in the United States: systematic review, J. Med. Internet Res. 25 (2023) e40337.
[270] F. HU, H. YANG, X.L. ZHOU, S. Zhao, L. Qiu, S. Wei, W. Xiaoping, J. Hu, Y. Chen, H. Hu, et al., Rehabilitation assistive technology transfer in the AI era: Network dynamics and public health impacts in the Yangtze River Delta, China, Front. Public Health 14 (2026) 1752396.
[271] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, Minds Mach. 28 (4) (2018) 689–707.
[272] B. Liu, Z. Lu, Y. Wang, Towards medical vision-language contrastive pre-training via study-oriented semantic exploration, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4861–4870.
[273] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (4) (2017).
[274] J. Xu, Z. Zhang, Z. Lin, Y. Chen, Z. Liu, W. Ding, How to characterize imprecision in multi-view clustering? IEEE Trans. Emerg. Top. Comput. Intell. (2025).
[275] X. Du, J. Zhu, J. Zhou, C.-m. Pun, Z. Lin, C. Wu, Z. Chen, J. Luo, Dp-trae: A dual-phase merging transferable reversible adversarial example for image privacy protection, IEEE Trans. Dependable Secur. Comput. (2025).
[276] Y. Gong, C. Wu, W. Zheng, S. Lu, G. Xu, L. Zhang, L. Yin, DDNet: A novel dynamic lightweight super-resolution algorithm for arbitrary scales, Comput. Model. Eng. Sci. 145 (2) (2025) 2223.
[277] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nature Med. 28 (9) (2022) 1773–1784.
[278] W. Zhang, Y. Huang, T. Zhang, Q. Zou, W.-S. Zheng, R. Wang, Adapter learning in pretrained feature extractor for continual learning of diseases, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 68–78.
[279] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowl.-Based Syst. 216 (2021) 106775.
[280] Z. Zhao, X. Li, B. Luan, W. Jiang, W. Gao, S. Neelakandan, Secure internet of things (IoT) using a novel brooks Iyengar quantum byzantine agreement-centered blockchain networking (BIQBA-BCN) model in smart healthcare, Inform. Sci. 629 (2023) 440–455.
[281] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, et al., Aligning large multimodal models with factually augmented rlhf, 2023, arXiv preprint arXiv:2309.14525.
[282] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y.J. Lee, Y. Ma, Investigating the catastrophic forgetting in multimodal large language models, 2023, arXiv preprint arXiv:2309.10313.
[283] H. Khan, N.C. Bouaynaya, G. Rasool, The importance of robust features in mitigating catastrophic forgetting, in: 2023 IEEE Symposium on Computers and Communications, ISCC, IEEE, 2023, pp. 752–757.
[284] L. Wang, X. Zhang, H. Su, J. Zhu, A comprehensive survey of continual learning: Theory, method and application, IEEE Trans. Pattern Anal. Mach. Intell. 46 (8) (2024) 5362–5383.
[285] D.-W. Zhou, Y. Zhang, J. Ning, H.-J. Ye, D.-C. Zhan, Z. Liu, Learning without forgetting for vision-language models, 2023, arXiv preprint arXiv:2305.19270.
[286] Y. Cai, M. Rostami, Dynamic transformer architecture for continual learning of multimodal tasks, 2024, arXiv preprint arXiv:2401.15275.
[287] F. Yu, X. Wang, R. Guo, Z. Ying, S. Cai, J. Jin, Dynamical analysis, hardware implementation, and image encryption application of new 4D discrete hyper-chaotic maps based on parallel and cascade memristors, Integration (2025) 102475.