Machine Learning with Applications 23 (2026) 100826

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Vision-language zero-shot models for radiographic image classification: A systematic review

Ana Guerrero-Tamayo *, Ibon Oleagordia-Ruiz, Begonya Garcia-Zapirain

Faculty of Engineering, University of Deusto, Avda. Universidades, 24, Bilbao, 48007, Biscay, Spain

ARTICLE INFO

Keywords:
Survey
Systematic review
Vision-language models
Zero-shot
X-ray
Radiographic
Image classification

ABSTRACT

Zero-shot Vision-Language Models (VLMs) link visual and textual features, enabling generalization to unseen domains, making them promising for radiographic diagnosis, though clinical adoption is limited.

This systematic review examines zero-shot VLMs applied to radiographic image classification, following the PRISMA methodology. Articles were identified from IEEE, PubMed, Scopus, and Web of Science, with 16 selected after exhaustive screening. The analysis addressed five research questions (RQ1–RQ5) covering dataset characteristics, model attributes, natural language integration, reported limitations, and hyperparameter tuning.

Geographically, China (37%) and the United States (38%) contributed 75% of the reviewed studies, with no EU-led research identified, highlighting the need for increased European engagement in this field.

Architecturally (RQ2), high heterogeneity exists, with dual-encoder (43.75%) and attention-based fusion models most common. Most models (81.25%) employ a Joint Embedding Space for multimodal alignment.

Regarding datasets and natural language use (RQ1, RQ3), VLMs rely on few large but semantically narrow datasets, limiting generalizability and amplifying bias. Real clinical reports (direct supervision) and implicit pretrained textual embeddings each represent 37.5% of strategies, yet unstructured clinical text is underutilized. Limited vision-language integration negatively affects performance and explainability (RQ4). Hyperparameter tuning (RQ5) is rarely reported, with 9 of 16 studies not specifying methods, compromising reproducibility.

There is an urgent need for open, multilingual, multimodal datasets reflecting clinical and geographic diversity. Clinically useful zero-shot VLMs require transparent evaluation, including explainability metrics. Future models should adopt a multidisciplinary approach, combining technical innovation with usability, data representativeness, and methodological transparency to ensure diagnostic robustness.

1. Introduction

Vision-Language Models (VLMs) represent an advanced category of multimodal artificial intelligence systems designed to jointly integrate and process visual and textual information within a shared latent space (Bordes et al., 2024). This semantic alignment between modalities enables the execution of complex tasks such as classification, information retrieval, description generation, or question answering, without requiring task-specific architectures for each input type (Du et al., 2023). The core principle behind their functioning lies in the projection of both data types (images and text) into embedded vectors within a common semantic space, where their proximity reflects conceptual similarity (Zhang, Huang, Jin, & Lu, 2024a).

Within this category, zero-shot models stand out for their generalization capability: they can infer semantic relationships and make predictions on classes not seen during training, leveraging the richness of representations learned from large, heterogeneous, and unsupervised data corpora (Al Rahhal, Bazi, Elgibreen, & Zuair, 2023). This property is particularly valuable in biomedical tasks, where the availability of annotated data with diagnostic quality is limited, costly, and often restricted by ethical or legal considerations (Araf, Idri, & Chairi, 2024; Jin, Luo, Li, & Mathew, 2019; Price & Cohen, 2019). In this context, zero-shot VLMs enable the tackling of highly variable clinical problems, such as the diagnosis of rare diseases, emerging conditions not represented in training datasets, or complex clinical scenarios involving multiple comorbidities (Buckley, Diao, Rajpurkar, Rodman, & Manrai, 2024; Dimitri, 2024; Wang, Lin, et al., 2024; Zhang, Huang, Jin, & Lu, 2024b).

* Corresponding author.
E-mail addresses: ana.guerrero@deusto.es (A. Guerrero-Tamayo), ibruiz@deusto.es (I. Oleagordia-Ruiz), mbgarciazapi@deusto.es (B. Garcia-Zapirain).

https://doi.org/10.1016/j.mlwa.2025.100826
Received 25 September 2025; Received in revised form 1 December 2025; Accepted 23 December 2025
Available online 23 December 2025

However, the robustness of these models is undermined by multiple sources of bias inherent in the processes of pretraining, alignment, and data curation (Hamidieh, Zhang, Gerych, Hartvigsen, & Ghassemi, 2024; Ruggeri & Nozza, 2023). These include the overrepresentation of certain classes or domains (Wang, Yu, et al., 2024), spurious correlations between labels and visual features (Yang et al., 2025), and limitations in the semantic coverage of the textual data used for contrastive learning (Gavrikov et al., 2024; Howard, Bhiwandiwalla, Fraser, & Kiritchenko, 2024). Such deficiencies can lead to failures in out-of-distribution (OoD) generalization, directly affecting the fairness, reliability, and reproducibility of results in real-world settings (Parashar et al., 2024). In high-stakes domains such as medicine, this issue becomes particularly critical, as incorrect inference may result in harmful clinical consequences. Therefore, the deployment of zero-shot VLMs in healthcare environments requires a rigorous evaluation of their performance, taking into account metrics beyond accuracy, such as calibration, interpretability, and bias resilience (Jeong, Garg, Lipton, & Oberst, 2024).

This study specifically focuses on the application of Vision-Language Models (VLMs) to the classification of radiographic (X-ray) images in zero-shot settings, a task that encapsulates both the methodological and clinical challenges of the multimodal approach. The choice of X-rays as the subject of analysis is justified by multiple factors:

- They are a widely standardized and accessible medical imaging modality.

- They represent the first level of diagnostic screening across various specialties (such as pulmonology, traumatology, and cardiology).

- They offer a rich semantic interplay with textual clinical reports (Çallı, Sogancioglu, van Ginneken, van Leeuwen, & Murphy, 2021).

Moreover, this modality has historically served as a testbed for models such as CLIP and its derivatives, making it a fertile ground for evaluating the real-world performance of VLMs in biomedical applications (You et al., 2023; Zhang et al., 2024). Consequently, this work aims to provide a comprehensive, technically grounded, and clinically oriented systematic review that examines the state of the art, current limitations, and future opportunities for zero-shot VLMs in the automated interpretation of X-ray images.

The remainder of this article is structured as follows: Section 2 presents the research methodology, including the research questions, data sources, and eligibility criteria. Section 3 presents the results obtained following the application of the PRISMA framework. Section 4 provides the answers to the research questions, along with a discussion. Finally, Section 5 outlines the study's final conclusions.

2. Methods

The present systematic review was conducted in accordance with the methodological principles established by the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), which provides a standardized and widely recognized framework to ensure transparency, rigor, and reproducibility in the development and reporting of systematic reviews. The guidelines proposed by Moher, Liberati, Tetzlaff, and Altman (2010) were rigorously followed to ensure an appropriate structure for the processes of search, selection, analysis, and synthesis of the available scientific evidence.

2.1. Research questions

This subsection presents the core research questions that structure the analytical framework of this systematic review. The focus is placed on zero-shot vision-language models (VLMs) for X-ray classification, with emphasis on dataset composition, architectural and functional model attributes, integration of clinical language corpora, existing methodological bottlenecks, and strategies for hyperparameter optimization. These questions aim to facilitate a comprehensive and systematic examination of the current landscape in this emerging research area.

RQ1. What are the characteristics of the datasets used in zero-shot VLMs applied to X-ray classification?

RQ2. What are the defining features of zero-shot VLMs for X-ray classification?

RQ3. How many models actually leverage large unstructured clinical corpora, such as radiology reports or natural language descriptions? And how many incorporate detailed natural language descriptions as an integral part of the learning or inference process?

RQ4. What are the main limitations in the integration of natural language and medical imaging, and how do they impact model performance?

RQ5. What robust methodologies are described in the literature for hyperparameter selection and tuning in zero-shot VLMs for X-ray classification?

2.2. Data sources and search strategy

Electronic bibliographic searches were conducted in IEEE, PubMed, Scopus, and Web of Science up to November 17, 2025.

This systematic review focuses on zero-shot VLM systems applied to the classification of radiographic images. However, in medical terminology, the concept of "classification" can be expressed in multiple terms, such as diagnosis, detection, or disease identification. The term "classification" indeed belongs more to the field of computer science than to biomedicine. For this reason, and in order to avoid excluding potentially eligible articles, we deliberately chose not to bias the search strategy by using MeSH terms exclusively related to "classification". The selection of articles specifically focused on classification (diagnosis, detection, disease identification, and similar) was performed manually and in detail, evaluating each abstract and/or full text to determine whether the study indeed involved models applied to classification tasks. Therefore, MeSH terms and free-text keywords related to "vision-language model", "zero-shot", and "X-ray" were used.

2.3. Study selection process (PRISMA)

The study selection process was conducted rigorously following the PRISMA 2020 guidelines to ensure transparency, reproducibility, and traceability across all phases: identification, screening, eligibility assessment, and final inclusion.

Identification.

All articles resulting from the initial search in IEEE, PubMed, Scopus, and Web of Science (updated to November 17, 2025) were exported to a reference manager, where an automatic duplicate removal process was applied first. Manual checks were subsequently performed. The articles remaining after this initial deduplication proceeded to the screening phase.

Screening (title and abstract).

Two reviewers independently screened the titles and abstracts. In this phase, studies that clearly did not meet the basic inclusion criteria were excluded, including those that:

(a) did not use radiographic images (X-ray),

(b) did not address vision-language models or zero-shot scenarios,

(c) relied exclusively on supervised, fine-tuned, or transfer learning approaches,

(d) focused on non-classification tasks (e.g., segmentation or registration),

(e) belonged to grey literature (preprints, theses, posters, or technical reports).

Discrepancies between reviewers were resolved through discussion and consensus. After excluding the articles corresponding to these criteria, the remaining publications proceeded to the full-text evaluation phase (inclusion criteria).

Inclusion (full text).

The full texts of candidate articles were reviewed in detail to explicitly apply the inclusion and exclusion criteria. It was verified that each study:

(1) described or evaluated a vision-language model in a zero-shot setting;

(2) applied the model to radiographic images in classification tasks (including clinically equivalent terms such as diagnosis, detection, or disease identification);

(3) provided sufficient methodological detail regarding the data, the model, and the experimental protocol;

(4) was published in indexed, peer-reviewed journals or conferences.

During this phase, studies were excluded if, upon full-text analysis, they:

(a) did not involve zero-shot inference;

(b) used modalities other than X-ray (e.g., CT or MRI);

(c) did not integrate vision-language components;

(d) lacked sufficient methodological information.

The articles remaining after this screening met all criteria and were therefore included in the qualitative synthesis. Consequently, the next step, once the articles forming the systematic review were selected, was a systematic analysis according to the research questions (RQ1–RQ5).

3. Results

The application of the PRISMA methodology, as described in Section 2, allowed for the initial identification of candidate articles for analysis. These articles were first subjected to a screening process to remove duplicates. During the eligibility phase, based on both a preliminary review and a detailed examination of abstracts, articles that did not fall within the scope of the proposed review were excluded. Finally, in the inclusion phase, after a thorough full-text analysis, the final set of articles for the systematic review was selected.

The complete search results are presented in Fig. 1.

During the Identification phase, the MeSH terms "vision language model", "zero-shot", and "X-ray" were used across the four major scientific databases (IEEE, Scopus, Web of Science, and PubMed), yielding an initial total of 140 candidate articles. In the Screening phase, duplicates across the databases were identified, leaving 61 articles to be evaluated in the subsequent Eligibility and Inclusion phases.

The Eligibility phase consisted of two sub-phases. In the first sub-phase, articles that, after an initial review, did not align with the scope of the review (e.g., few-shot learning studies, unrelated domains, or systematic reviews) or met exclusion criteria (e.g., conference proceedings, posters, preprints, etc.) were filtered out, reducing the number of articles from 61 to 53. In the second sub-phase, following a detailed examination of abstracts, 11 additional articles were excluded, leaving 42 articles to advance to the final phase.

In the Inclusion phase, a thorough full-text analysis was conducted. Of the 42 articles, 26 were excluded for not meeting the defined scope, resulting in a final set of 16 articles selected for the comparative analysis of the presented models and techniques.

Subsequently, a complementary analysis was performed to examine the distribution of the selected studies according to publication year and the geographic affiliation of the first author. The aim was to identify temporal trends in scientific output and to map the geographic landscape of academic contributions at the intersection of vision-language models, zero-shot learning, and radiographic imaging.

This descriptive analysis complements the core objectives of the review by situating the selected studies within a broader bibliometric and geopolitical context, providing additional insights into the maturity and geographical distribution of research in the field under study.

It is noteworthy that no relevant articles were found prior to 2023, reflecting the very recent emergence of vision-language models (VLMs) applied to zero-shot classification in radiographic imaging. Among the 16 selected articles, 4 correspond to 2023, 9 to 2024, and only 3 to 2025 (as of November 17). The low number of publications in 2025 may suggest that zero-shot VLMs for radiographic image classification are encountering practical challenges or limitations, and their application may not yet be yielding the expected results.

Regarding the country distribution, Fig. 2 shows the results based on the country of affiliation of the first author of each article.

The geographical analysis reveals a clear concentration of scientific output around two leading research powers, China and the United States, which together contribute an equal share of 75% of the articles. This joint dominance underscores the strong leadership of both nations in the development of vision–language models applied to radiographic imaging within zero-shot learning frameworks. Notably, no member country of the European Union is represented among the analyzed studies.

Several factors may help explain this geographical distribution. First, funding priorities in both China and the United States have increasingly emphasized artificial intelligence for medical imaging, supported by large-scale national initiatives and substantial public–private investment. These programs have enabled rapid progress in multimodal architectures and clinical AI applications. Second, access to large-scale radiographic datasets — often supported by extensive hospital networks and centralized data infrastructures — provides researchers in these countries with the volume and diversity needed to train and evaluate advanced vision–language models. Third, the regulatory environments in both nations tend to be comparatively more flexible regarding the use of medical images for research purposes, especially when data are de-identified. This contrasts with stricter data protection frameworks, such as the GDPR in the European Union, which may impose additional barriers to large-scale data sharing and interinstitutional collaboration.

Demographically, in 2025 the population of China is estimated at approximately 1.416 billion inhabitants, that of the United States at around 347 million, and that of the countries comprising the European Union at roughly 450 million (Worldometer, 2025). Indeed, China's large population facilitates the generation and access to a high volume of clinical cases and medical data. However, the aggregated population of the European Union — greater than that of the United States — also holds substantial potential for the creation of valuable datasets for the development of artificial intelligence models in healthcare. Despite this, the United States has managed to produce more results in this field with a smaller population, likely due to a more flexible regulatory framework and a significantly more ambitious public–private strategy in the investment and promotion of AI technologies.

Taken together, these factors suggest that the observed geographical pattern does not solely reflect scientific interest, but is shaped by structural conditions (investment, data availability, and regulatory context) that currently favor the research leadership of China and the United States.

In this regard, the European Union currently faces the challenge of finding the appropriate balance between fostering technological and scientific innovation and developing and enforcing regulations that protect society from the misuse of, among other tools, AI technologies.

4. Discussion

This section presents the results of the in-depth analysis of the content of the 16 articles, providing answers to the research questions. Additionally, we offer a discussion of the extracted data.

4.1. RQ1. What are the characteristics of the datasets used in zero-shot VLMs applied to X-ray classification?

The importance of the datasets used to train zero-shot VLM models lies in their role as the foundation for learning visual–linguistic relationships across diverse data pairs without the need for task-specific retraining. In the biomedical domain, this importance is heightened due

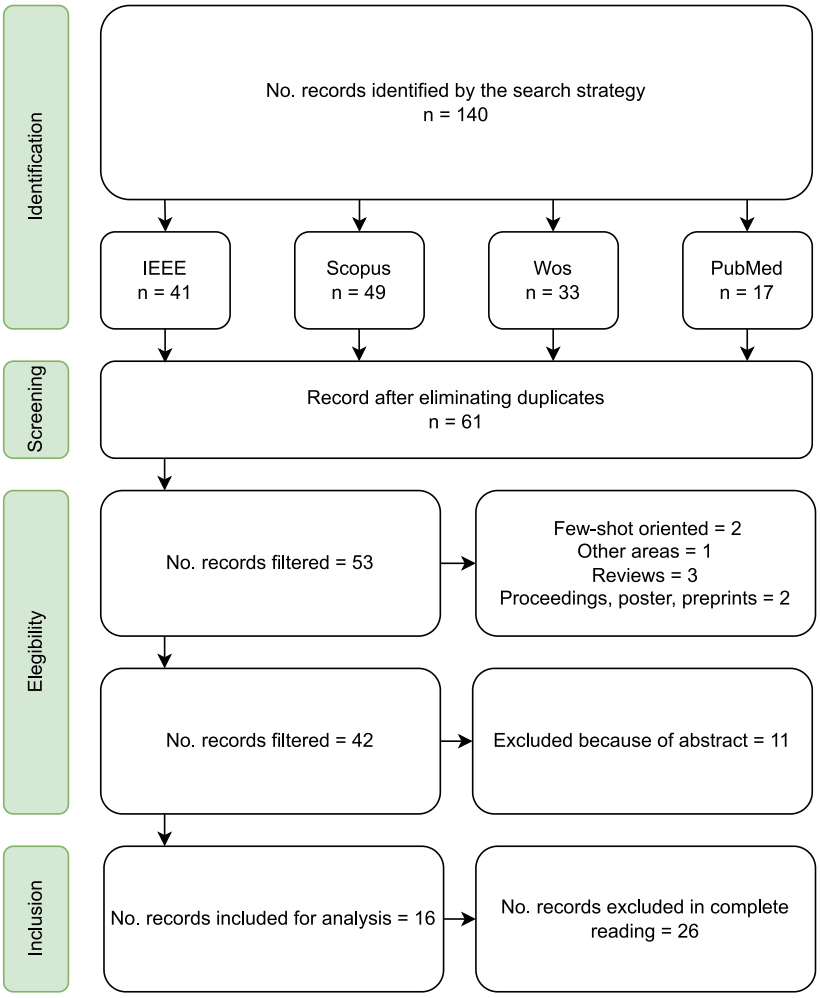

Fig. 1. Study selection flow according to PRISMA.

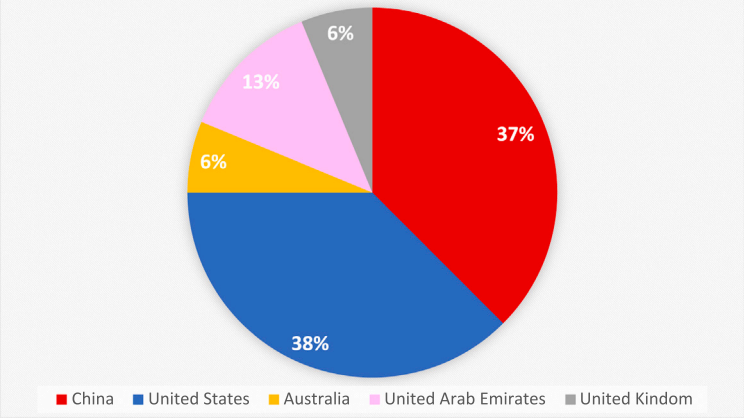

Fig. 2. Number of articles per country.

to the diagnostic support and pathology detection functions embedded within these models.

The 16 models resulting from the review were trained using a total of 35 datasets. The details of the data used in each model are presented in Table 1:

The first relevant observation from Table 1 is that all models were trained with hundreds of thousands of multimodal data points, with several even using over a million data items. This is understandable given the need to "acclimate" the model to diverse cases and the added complexity of processing medical images. A single model (Rahman et al., 2025) uses a substantially smaller total number of data points: 18,692 images and 24 labels, which also include X-ray datasets (Shenzhen-TB, Montgomery-TB, Guangzhou-Pneumonia) along with retinal (IDRiD) and dermatology (ISIC) data. According

Table 1
Datasets per article.

Paper	Datasets used	Total data
A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images (Jin, Lu, Li, & Wang, 2023)	ChestX-ray14	112,120 images, 14 labels (non-publicly available texts)
Weakly supervised zero-shot medical image segmentation using pretrained medical language models (Guo & Terzopoulos, 2024)	MIMIC-CXR, CheXlocalize, SIIM-ACR Pneumothorax Segmentation	719,000 images, 377,000 reports, 234,000 bounding boxes, 3,600 segmentations, 14 labels
Villa: Fine-grained vision-language representation learning from real-world data (Varma, Delbrouck, Hooper, Chaudhari, & Langlotz, 2023)	DocMNIST, DeepFashion, MIMIC-CXR, COCO, LVIS, CheXpert	1,994,482 images, 377,000 reports, 4,000,000+ bounding boxes, 3,500,000+ segmentations, 1,278+ labels
Knowledge-enhanced visual-language pre-training on chest radiology images (Zhang, Wu, Zhang, Xie, & Wang, 2023)	MIMIC-CXR, PadChest, CheXpert, ChestX-Det10	887,000 images, 537,000 reports, 30,000 bounding boxes, 198 labels
Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning (Huang, Li, et al., 2024)	MIMIC-CXR V2, ChestX-ray14, CheXpert, RSNA Pneumonia, SSIIM-ACR Pneumothorax	851,000 images, 377,000 reports, 30,000 bounding boxes, 2,600 segmentations, 16 labels
A latent diffusion approach to visual attribution in medical imaging (Siddiqui, Tirunagari, Zia, & Windridge, 2025)	COVIDx CXR-2, CheXpert	254,000 images and 17 labels
Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models (Cao et al., 2024)	ChestCT-16K, ChestCT-EXT, LIDC-IDRI, MIMIC-CXR	512,018 images, 377,000 reports, 7,000 segmentations, 14 labels
Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis (Wu, Zhang, Zhang, Wang, & Xie, 2023)	MIMIC-CXR V2, ChestX-ray14, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2, Edema Severity	661,839 images, 377,000 reports, 30,000 bounding boxes, 2,600 segmentations, 37 labels
Decomposing disease descriptions for enhanced pathology detection (Phan et al., 2024)	MIMIC-CXR V2, CheXpert, ChestX-ray14, PadChest, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2	1,041,000 images, 537,000 reports, 30,000 bounding boxes, 2,600 segmentations, 208 labels
Carzero: Cross-attention alignment for radiology zero-shot classification (Lai et al., 2024)	MIMIC-CXR, Open-I, PadChest, ChestX-ray14, CheXpert, ChestX-D t10	1,006,000 images, 544,000 reports, 30,000 bounding boxes, 251 labels
Enhancing biomedical multi-modal representation learning (Zhong, Batmanghelich, & Sun, 2024)	Open-I, MIMIC-III, MIMIC-CXR, RadNLI, MedNLI, CheXpert	704,000 images, 2,598,500+ reports, 34 labels
Adapting visual language models for generalizable anomaly detection (Huang, Jiang, et al., 2024)	Brain MRI, Liver CT, Retinal OCT, ChestX-ray14, Digital Histopathology (HIS)	172,000 images, 1,000 bounding boxes, 55,000 segmentations, 20 labels
Pairaug: What can augmented image–text pairs do for radiology? (Xie et al., 2024)	MIMIC-CXR, CheXpert, PadChest, RSNA Pneumonia	887,000 images, 537,000 reports, 30,000 bounding boxes, 222 labels
Improving medical vision language contrastive pretraining (Liu et al., 2023)	MIMIC-CXR, CheXpert, RSNA Pneumonia, SIIM-ACR Pneumothorax	739,000 images, 377,000 reports, 30,000 bounding boxes, 2,600 segmentations, 30 labels
Core-Periphery Multi-Modality Feature Alignment (Yu, Zhang, Wu, & Zhu, 2025)	MIMIC-CXR, ChestXray, SIIM-ACR, INbreast, CheXpert5_200, TMED	489,000 images, 377,000 reports, 30,000 bounding boxes, 2600 segmentations, 24 labels
Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification? (Rahman, Imam, Yaqub, Amor, & Mahapatra, 2025)	Shenzhen-TB, Montgomery-TB, Guangzhou-Pneumonia, IDRID, ISIC	18,692 images, 24 labels

to the authors, their strategy is language-guided unsupervised adaptation, which minimizes dependence on extensive pre-training with large paired datasets.

Apart from this model, all others require hundreds of thousands of data points. Among them, the model trained with the smallest dataset used 112,120 images and 14 labels (Jin et al., 2023). The model with the largest raw data volume (Varma et al., 2023) utilizes an estimated total of 9,871,482 elements, including images, reports, bounding boxes, and segmentations.

Another interesting fact is that, although the primary application of the models is X-ray classification, we found that they were trained with datasets containing other types of medical images. We also identified several datasets specialized in segmentation, including metadata related to bounding boxes or masks. This variety is motivated precisely by the zero-shot nature of the models—that is, equipping the model with tools that facilitate or guide its ability to handle unseen data and information that aids in identifying textures associated with different pathologies or events in medical imaging.

We proceeded with a detailed analysis of the datasets used, classifying them into five main categories:

- X-ray classification: Specifically designed for X-ray classification. Their data primarily consists of X-ray images as the main modality, along with text.

- X-ray segmentation: Specialized not only in classification and event detection in images but also in positional annotation of the detected events. Their data primarily consists of X-ray images, text, and bounding boxes.

- Non X-ray classification/segmentation: Their data mainly consists of non X-ray images (e.g., CT scans), text, and bounding boxes if segmentation is included. The use of these datasets aligns with the zero-shot or generalized zero-shot nature of the models.

- Semantic inference: Datasets designed for evaluating textual inference models in biomedical contexts, focused on automatically identifying logical relationships (entailment, contradiction, neutral) between pairs of sentences extracted from unstructured clinical narratives. The aim is to model deep semantic representations and contextual reasoning in medical natural language understanding tasks.

- Heterogeneous data sources: Contain data from very diverse categories, not limited to images and text. For example: EEG, among others.

Starting with the first category (X-ray classification), Table 2 compiles the datasets specific to X-ray classification along with their most relevant characteristics:

The Edema Severity dataset (Nafees, 2020), also included in this X-ray classification category, is an extract derived from MIMIC-CXR and is therefore also restricted-access.

Table 3 lists the datasets specific to X-ray segmentation.

Table 4 is a compilation of datasets specific to non X-ray classification/segmentation:

The Brain MRI, Liver CT, Retinal OCT, and Digital Histopathology (HIS) datasets (Huang, Jiang, et al., 2024) belong to this category; however, it is not possible to directly access the exact number of data samples they contain.

Table 5 includes the relevant characteristics of the datasets specific to semantic inference:

Table 6 presents the dataset containing multimodal data of highly heterogeneous nature.

Based on the results of the breakdown performed, there is a clear upward trend toward incorporating multiple types of data in the training of VLM models, particularly those with zero-shot ambitions. The model called Villa (Varma et al., 2023) was implemented using the largest amount of data, nearly 10 million elements including images, reports, segmentations, and bounding boxes. This multimodal and multiscenario approach, although not exclusively medical, suggests a transferable learning strategy from general data to improve performance in specific clinical tasks.

Categorizing datasets by type helps illustrate how the objectives and nature of the data impact the design and scope of the models:

- X-ray classification datasets are fundamental as a diagnostic basis but limited in semantic complexity.

- X-ray segmentation datasets add a spatial granularity level that allows evaluation not only of what is detected but also where.

- The use of data from other medical modalities (non-X-ray) aligns with cross-modal generalization, consistent with zero-shot learning principles.

- Datasets focused on semantic inference (such as RadNLI or MedNLI) contribute a layer of clinical reasoning, enabling models capable not only of describing images but also understanding clinical hypotheses.

- Heterogeneous data source datasets suggest a movement toward biomedical foundation models capable of operating with multiple input types and tasks simultaneously.

Overall, the analysis reveals that data volume is not the sole determining factor: diversity, semantic richness, and modal alignment are key to zero-shot performance in relevant clinical tasks.

The 10 most used datasets are shown in Fig. 3.

CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019, 2024) were the most frequently used datasets in the selected articles.

A strong dependence on a small number of datasets can be observed, with just five datasets accounting for the majority of the total data volume used. This concentration suggests a low diversity in the biomedical data employed, which could represent a bottleneck for both model evaluation and generalization capacity. The limited variety of sources constrains the spectrum of clinical scenarios represented, potentially affecting the robustness and external validity of zero-shot approaches in radiology.

In our view, PadChest (Bustos et al., 2020) stands out for its broad taxonomic coverage (in contrast to other commonly used datasets such as CheXpert or MIMIC-CXR). It includes over 190 labels encompassing not only radiological findings (e.g., consolidation, effusion, atelectasis), but also pathologies, procedures, medical devices, and even anatomical projections.

This characteristic may enhance the zero-shot capabilities of the models, as it enables them to infer conditions not seen during training. Furthermore, its annotation in Spanish makes it particularly relevant for multilingual research, although this may have limited its global adoption.

However, its relative underutilization compared to more widely used datasets reflects a paradox within the biomedical data ecosystem: despite the availability of more expressive resources, models tend to rely on smaller, more standardized datasets. This constrains the development of truly generalist and clinically meaningful models, suggesting that the bottleneck lies not only in data availability but also in the lack of effective integration of richer existing resources such as PadChest.

The fact that it is a Spanish-language dataset may have limited its global uptake, despite its clear advantages for the development of multilingual models.

The detailed annotations in PadChest enable VLMs to learn more fine-grained and differentiated representations, which is essential for zero-shot tasks where the model is expected to infer conditions not seen during training.

Overall, the relative diversity of biomedical datasets used represents a promising foundation for the development of robust models, particularly in zero-shot scenarios.

However, this heterogeneity remains insufficient in both scale and scope, especially when compared to multimodal datasets in the general domain. This scarcity poses a significant risk of propagating and amplifying the inherent biases of the datasets into the trained models.

Such limitations may undermine the external validity and generalization capacity of the models in real-world clinical applications. It is imperative to expand the representativeness and quality of available biomedical data.

Table 2
Datasets designed for X-ray classification.

Dataset	Number of Images/Cases	Main Content	Labels	Access
CheXpert (Irvin et al., 2019)	224,316 images from 65,240 patients	Chest X-rays with clinical reports	14 observations (positive, negative, uncertain) via NLP labeler	Open
MIMIC-CXR v2 (Johnson et al., 2019; Johnson, Pollard, Mark, Berkowitz, & Horng, 2024)	377,110 images from 227,835 studies	DICOM images (PA/AP/Lateral), detailed reports, demographic metadata	Text report labels; no structured label set	Restricted
SIIM–ACR Pneumothorax (Zawacki et al., 2019)	12,047 images	DICOM images, binary segmentation masks, metadata (sex, age, pneumothorax type)	1 class (pneumothorax: tension/no tension)	Open
PadChest (Bustos, Pertusa, Salinas, & de la Iglesia-Vayá, 2020)	160,861 images (multiple projections)	Images + 343 semantic labels, clinical + technical metadata	Diseases, radiological signs, devices, anatomical findings	Open
COVIDx CXR-2 (Zhao, 2020)	19,203 images from 16,656 patients	CXR labeled for COVID diagnosis, with demographic metadata	Binary: COVID-19 positive/negative	Open
Open-I (Demner-Fushman, Kohli, Rosenman, Shooshan, Rodriguez, Antani, Thoma, & McDonald, 2016)	7,470+ images and 3955 reports	X-rays and other modalities (CT, MRI, US), descriptive text, reports	Image + report + extensive metadata	Open
Shenzhen-TB (Jaeger, Candemir, Antani, Wáng, Lu, & Thoma, 2014)	662 images	Chest X-rays	Image-level labels: normal vs. tuberculosis; lung segmentation masks for some images	Open
Montgom. TB (Jaeger et al., 2014)	138 images	Chest X-rays	Image-level labels: normal vs. tuberculosis; lung segmentation masks for the lungs	Open
Guangzhou (Kermany, Zhang, & Goldbaum, 2018)	5863 images	Chest X-rays (pediatric anterior-posterior)	Image-level labels: normal vs pneumonia	Open

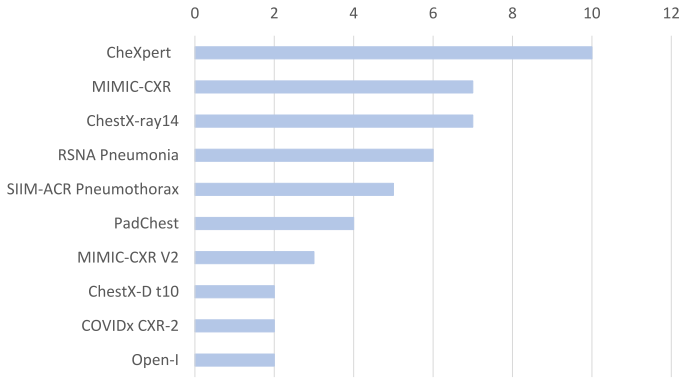

Fig. 3. Top 10 biomedical datasets.

Table 3
Datasets designed for X-ray segmentation.

Dataset	Number of images/cases	Main content	Labels	Access
RSNA Pneumonia	30,000 images	Frontal chest X-rays in DICOM format; pneumonia labels; bounding boxes	CSV annotations with coordinates; metadata (patient, dimensions, acquisition parameters)	Open
ChestX-ray14	112,120 PNG images (1024 × 1024)	Frontal chest X-rays with 14 disease labels extracted via NLP	Bounding boxes for ~1,000 images; metadata (age, sex, position, etc.)	Open
ChestX-D t10	3543 images	Subset of NIH ChestX-ray14 with thoracic abnormality annotations	Pathology labels and bounding boxes (x1, y1, x2, y2)	Open
CheXlocalize	234 images	X-rays from CheXpert with pixel-level segmentation masks for 10 pathologies	Segmentation masks and representative points; metadata (age, sex, labels)	Open
SIIM–ACR Pneumothorax Segmentation	12,047 images	DICOM X-rays; 2669 annotated for pneumothorax	Segmentation masks; limited metadata	Open
COVID-19 Radiography Database	21,165 images	COVID-19 and other pulmonary conditions; image folders organized by class	Segmentation masks; limited metadata	Open

Table 4
Datasets designed for non-x-ray classification and segmentation.

Dataset	Number of Images/Cases	Main Content	Labels	Access
ChestCT-16K and ChestCT-EXT	Not available for direct download	Computed tomography (CT) images	Not available	Not available
LIDC-IDRI	244,527 CT DICOM images	CT images (primary) and X-rays; extensive clinical and technical metadata	Radiologist annotations for nodule segmentation; nodule count and diagnosis, radiomic features, quality and acquisition protocol	Open
INbreast	410 mammography images (115 cases)	Mammography (breast X-ray, FFDM)	Lesion segmentations (masses, calcifications, asymmetries, distortions); BI-RADS density; classification labels (cancer vs. non-cancer)	Open
TMED	2,773 patients; 100 images/patient (24964–70000+ echocardiogram frames)	Transthoracic echocardiography (ultrasound)	View-labels (PLAX/PSAX/Other), Aortic-Stenosis diagnosis (no AS/mild-moderate AS/severe AS)	Restricted
IDRiD	516 retinal fundus images	Retinal fundus photography	Pixel-level lesion annotations (microaneurysms, hemorrhages, hard/soft exudates), optic disc and fovea centers; image-level diabetic retinopathy and macular edema grading	Open
ISIC	33,126 dermoscopic skin-lesion images	Dermoscopic skin lesion photography	Image-level diagnostic labels (benign vs malignant skin lesions, multiple lesion classes), (in some subsets) lesion segmentation/metadata	Open

Table 5
Datasets designed for semantic inference in the medical domain.

Dataset	Number of Pairs	Main Content	Labels	Access
RadNLI (Miura, Zhang, Tsai, Langlotz, & Jurafsky, 2021)	960 pairs (premise, hypothesis) annotated in both directions	Radiology report text extracted from MIMIC-CXR	Entailment, neutral, contradiction	Restricted
MedNLI (Shivade, 2019)	14,049 pairs (development), 480 pairs (test), annotated unidirectionally	Medical report text extracted from MIMIC-CXR	Entailment, neutral, contradiction	Restricted

Table 6
Heterogeneous data sources for clinical data.

Source	Number of Patients/Cases	Main Content	Data Types	Access
MIMIC-III (Johnson, Pollard, & Mark, 2016)	Over 60,000 patients (ICU admissions, 2001–2012)	Anonymized ICU clinical data	Demographics, clinical notes, prescriptions, lab results, vital signs, diagnoses and procedures, admissions and stays, microbiology, real-time events, billing, and resources	Restricted

4.2. RQ2. What are the defining features of zero-shot VLMs for X-ray classification?

Despite the fact that the 16 selected models differ significantly in terms of architecture and computational characteristics, we have categorized their architecture, type of mapping, classifier type, and domain adaptation strategy into distinct groups to facilitate comparative analysis. The results of this categorization are presented in Table 7.

The description of the architecture categories is as follows:

- **Dual-encoder (43.75% of the articles)**: Architecture with two independent encoders (one visual and one textual), which project the inputs into a shared latent space for subsequent comparison or fusion.
- **Fusion with attention (25% of the articles)**: A mechanism in which visual and textual representations are integrated through cross-attention modules to capture precise multimodal relationships.
- **Knowledge-enhanced (18.75% of the articles)**: Explicit incorporation of structured medical knowledge (such as triplets or clinical entities) to enrich the learned representations.
- **Contrastive (6,25% of the articles)**: Architecture employing contrastive learning with partial input masking (visual or textual), promoting robust semantic alignment.
- **Diffusion-based (6,25% of the articles)**: A generative model that learns representations through a diffusion and reconstruction process in a latent space, useful for visual explainability.

There is high variability among the different architecture categories, especially when compared to other technical aspects. However, certain approaches show a clear predominance. The **dual-encoder** category appears in seven studies (Huang, Jiang, et al., 2024; Jin et al., 2023; Liu et al., 2023; Rahman et al., 2025; Varma et al., 2023; Xie et al., 2024; Zhong et al., 2024). The widespread use of this approach can likely be attributed to its efficiency and relative simplicity in mapping visual and textual modalities into a shared latent space, enabling direct and effective comparison through similarity metrics such as cosine similarity. Nevertheless, this early modal separation may limit the model's ability to capture deep and contextual multimodal interactions during the initial processing stages, potentially constraining performance on tasks that require more integrated multimodal fusion.

The second most common architecture category is **fusion with attention**, used in four articles (Cao et al., 2024; Guo & Terzopoulos, 2024; Lai et al., 2024; Phan et al., 2024). Cross-attention mechanisms for integrating visual and textual representations enable finer and more contextualized alignment, enhancing the ability to model complex relationships between modalities. However, their greater complexity typically implies a significant increase in computational cost and data requirements for effective training. Additionally, early interaction between modalities may heighten the model's susceptibility to errors from one modality propagating and affecting the joint representation, potentially compromising robustness and stability in clinical settings with noisy or incomplete data.

The high adoption of joint embedding spaces is explained by their efficiency and simplicity: they project images and text into a shared space, allowing direct comparisons via similarity metrics, which promotes robustness, reproducibility, and scalability in zero-shot scenarios (Cherti et al., 2023). The dual-encoder architecture, dominant in the literature, prioritizes these qualities, but limits early interaction between modalities, which may hinder the capture of deep semantic relationships (Moeller, Tilli, Vu, & Padó, 2025). In contrast, fusion-with-attention approaches allow richer multimodal alignment and contextual reasoning, improving discrimination in fine-grained tasks, although with higher computational cost and greater sensitivity to data quality and consistency (Ding, Shih, Wen, Li, & Yang, 2025). Overall, the choice of architecture reflects a trade-off between efficiency, robustness, and discriminative capacity, explaining the predominance of dual-encoder + joint embeddings as the de facto standard for multimodal zero-shot tasks.

The third most frequently used architecture category is **knowledge-enhanced**, employed by three models (Wu et al., 2023; Yu et al., 2025; Zhang et al., 2023). The incorporation of structured medical knowledge offers the advantage of enriching multimodal representations with explicit clinical information, improving model interpretability and its ability to capture complex, domain-specific relationships. This approach can enhance diagnostic accuracy and facilitate decision explainability—both critical aspects in biomedical applications. However, integrating external knowledge presents challenges, such as dependence on comprehensive and up-to-date knowledge bases and difficulty generalizing to pathologies or contexts not covered by those sources. As with **Fusion with attention**, the added complexity can increase computational cost and hinder scalability to large, heterogeneous datasets.

The reviewed models use different visual and textual backbones, with a marked preference for architectures previously pretrained in biomedical domains. The main observations are highlighted below:

1. Visual Backbones:

Table 7
Technical features of the models.

Paper	Architecture	Mapping	Classifier	Domain adaptation
A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images (Jin et al., 2023)	Dual-encoder	Joint Embedding Space	Prototype similarity	Language model integration
Weakly supervised zero-shot medical image segmentation using pretrained medical language models (Guo & Terzopoulos, 2024)	Fusion with attention	Joint Embedding Space	Decoder-based	Language model integration
Villa: Fine-grained vision-language representation learning from real-world data (Varma et al., 2023)	Dual-encoder	Joint Embedding Space	Contrastive	Pretraining on domain data
Knowledge-enhanced visual-language pre-training on chest radiology images (Zhang et al., 2023)	Knowledge-enhanced	Semantic → Visual	Binary classification	Pretraining on domain data
Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning (Huang, Li, et al., 2024)	Contrastive (masked)	Joint Embedding Space	Contrastive	Pretraining on domain data
A latent diffusion approach to visual attribution in medical imaging (Siddiqui et al., 2025)	Diffusion-based	Joint Embedding Space	Generative	Language model integration
Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models (Cao et al., 2024)	Fusion with attention	Visual → Visual (semantic-guided)	Binary classification	Knowledge distillation
Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis (Wu et al., 2023)	Knowledge-enhanced	Joint Embedding Space	Binary classification	Expert curation
Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework (Phan et al., 2024)	Fusion with attention	Semantic → Visual	Supervised (contrastive + classification)	Expert curation
Carzero: Cross-attention alignment for radiology zero-shot classification (Lai et al., 2024)	Fusion with attention	Joint Embedding Space	Cross-attention	Prompt-based
Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination (Zhong et al., 2024)	Dual-encoder	Joint Embedding Space	Contrastive	Pretraining on domain data
Adapting visual language models for generalizable anomaly detection in medical images (Huang, Jiang, et al., 2024)	Dual-encoder	Joint Embedding Space	Few-shot	Prompt-based
Pairaug: What can augmented image–text pairs do for radiology? (Xie et al., 2024)	Dual-encoder	Joint Embedding Space	Linear probe	Synthetic data augmentation
Improving medical vision language contrastive pretraining with semantics-aware triage (Liu et al., 2023)	Dual-encoder	Joint Embedding Space	Contrastive + linear	Pretraining on domain data
Core-Periphery Multi-Modality Feature Alignment for Zero-Shot Medical Image Analysis (Yu et al., 2025)	Knowledge-enhanced	Joint Embedding Space	Contrastive	Pretraining on domain data
Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? (Rahman et al., 2025)	Dual-encoder	Joint Embedding Space	Unsupervised (entropy + prompt)	Language-guided unsupervised adaptation

- Vision Transformers (ViT): These appear frequently (ViT-B/16, ViT-B/32), used in models such as Cao et al. (2024), Guo and Terzopoulos (2024), Huang, Li, et al. (2024), Lai et al. (2024), Rahman et al. (2025), Xie et al. (2024). This indicates a growing preference for transformer architectures in vision, likely due to their ability to model complex spatial relationships without strong inductive biases like those in CNNs.

- ResNet-50: Although less frequent, it remains a relevant backbone (e.g., Liu et al. (2023), Wu et al. (2023), Yu et al. (2025), Zhang et al. (2023)), especially when robustness and computational efficiency are sought. It is often used in combination with contrastive pretraining strategies such as ConVIRT or GLoRIA.

- CNNs with medical pretraining: Some models, such as Varma et al. (2023), use CNNs pretrained specifically in medical domains (e.g., ConVIRT), which enhances the specialization of visual representations.

- Variational Autoencoder (VAE) + U-Net: In the particular case of the diffusion model (Siddiqui et al., 2025), a VAE is used as the visual

encoder and a U-Net as the decoder, aligning with tasks of visualization and interpretability.

- 3D adaptations: The model in Cao et al. (2024) employs a ViT adapted to 3D volumes, showing the need to adapt standard 2D architectures for modalities like volumetric CT.

2. Textual Backbones:

- Clinical BERT and variants: Most models use biomedical adaptations of BERT, such as BioBERT, ClinicalBERT, or BioClinicalBERT (Guo & Terzopoulos, 2024; Jin et al., 2023; Lai et al., 2024; Liu et al., 2023; Rahman et al., 2025; Wu et al., 2023; Yu et al., 2025), reflecting the need for language representations tailored to medical technical language.

- SBERT and CXR-BERT: These appear as specialized variants for tasks like report matching (Varma et al., 2023; Zhong et al., 2024), with a focus on the semantics of radiology reports.

- CLIP Text Encoder: Some models directly reuse the CLIP text encoder (e.g., Xie et al. (2024)), which is useful for generalization and transfer learning approaches, although it may require additional adaptation for biomedical domains.

3. Fusion and Additional Modules:

- Transformer-based fusion: Models such as Phan et al. (2024), Wu et al. (2023) incorporate multimodal fusion modules using Transformers, enhancing cross-modal alignment.

- Projection and alignment: Several models include linear layers, ReLU activations, or additional modules to project data into a shared latent space, often combined with contrastive techniques or cross-attention.

Analyzing the results of the review, several design trends can be identified:

1. Predominance of pretraining in the biomedical domain: In both vision and language, most models rely on backbones specifically pretrained on clinical data, which is crucial for improving performance and generalization.

2. Growing adoption of ViT and Transformers: There is a noticeable shift from CNNs like ResNet toward Transformer architectures (e.g., ViT) for both image and text processing, suggesting a convergence toward more flexible and scalable models.

3. Module hybridization: Many models combine visual and textual encoders with fusion, projection, or diffusion modules, indicating a trend toward modular architectures tailored to specific tasks (e.g., classification, segmentation, explainability).

Regarding the Mapping (Multimodal Alignment Mechanism) used in the 16 articles included in the review, the following categories were identified:

- **Joint Embedding Space Visual ↔ Semantic (81.25% of the articles)**: Both modalities (image/text) are projected into a shared space where they are directly compared based on semantic proximity.

- **Semantic → Visual (12.5% of the articles)**: The textual representation directly guides the generation or classification of visual representations.

- **Visual → Visual (semantic guided) (6.25% of the articles)**: A visual representation is learned guided by semantic information extracted from the text, without explicitly projecting into the textual space.

In the analyzed sample, there is a clear preference for the approach based on joint embedding spaces (**Joint Embedding Space**), present in 13 of the 16 reviewed models (Guo & Terzopoulos, 2024; Huang, Jiang, et al., 2024; Huang, Li, et al., 2024; Jin et al., 2023; Lai et al., 2024; Liu et al., 2023; Rahman et al., 2025; Siddiqui et al., 2025; Varma et al., 2023; Wu et al., 2023; Xie et al., 2024; Yu et al., 2025; Zhong et al., 2024). This predominance can be attributed to its architectural simplicity and its effectiveness in widely explored tasks such as classification, cross-modal retrieval, or contrastive pretraining. Furthermore, this approach scales well to zero-shot or few-shot scenarios. However, one of its well-known limitations is the limited early interaction between modalities, which can hinder deeper semantic alignment (Zhu, Wei, Liang, Zhang, & Zhao, 2023). The high percentage of articles indicates that the scientific community considers this approach as the de facto standard for multimodal zero-shot tasks in the domain of radiological image classification. Nevertheless, we should caution about the possibility of a publication bias, that is, methods with strong and reproducible results are more frequently reported than less common alternatives, which a priori may represent advantages over the inherent limitations of these models.

Models that employ a **Semantic → Visual** mapping constitute a minority strategy (only two works adopt it (Phan et al., 2024; Zhang et al., 2023)), indicating that directly projecting semantic information into the visual space is less common than architectures based on joint embeddings. This low adoption can be partly explained by the difficulty of generalizing the semantic–visual mapping across heterogeneous datasets and the need for a particularly robust visual feature generator, which increases system complexity and requires asymmetric information flows supported by curated textual knowledge. Even so, this approach offers specific value in clinical contexts where integrating structured knowledge can modulate visual representations, improving interpretability and diagnostic accuracy. Furthermore, its use is particularly relevant in scenarios requiring pseudo-supervised image generation or abstract concept retrieval, highlighting research niches exploring unidirectional mappings to enhance zero-shot capabilities, although still at experimental stages.

A single model in the review adopts a **Visual → Visual** approach guided by semantic information (Cao et al., 2024), a choice aimed at facilitating knowledge transfer between heterogeneous visual modalities (for example, translation between CT images and X-rays). This can be particularly interesting in distillation or domain adaptation scenarios. However, the limited adoption of this strategy suggests that refining visual representations via semantic guidance is not yet a widespread practice, likely due to its architectural complexity and the need for datasets with rich and consistent annotations. Despite these limitations, this type of mapping can be especially useful in fine-grained tasks, where small visual differences are critical, or in highly specialized domains such as medicine or security, where semantic modulation is indispensable. From a methodological perspective, this is an emerging field with the potential to offer competitive advantages in zero-shot scenarios, although it still lacks systematic studies and its low adoption may reflect both reproducibility challenges and the absence of standardized benchmarks to assess its effectiveness.

The following categorization corresponds to the type of **Classifier** (inference or decision mechanism). The identified categories are as follows:

- Contrastive (25% of the articles): Method that learns embeddings by aligning positive pairs and separating negative pairs using a contrastive loss. Enables zero-shot retrieval and robustness to OOD (out-of-distribution) data.

- Binary classification (18.75% of the articles): Treats each decision as a binary "yes/no" problem per class. Similarity outputs or logits transformed into probabilities can be used. Suitable for multi-label zero-shot tasks, but less scalable than contrastive methods.

- Prototype similarity (6.25% of the articles): Strategy that represents each class or concept with a vector prototype in the embedding space. Classification is performed by measuring the similarity between the sample embedding and the prototypes.

- Decoder-based (6.25% of the articles): Models that use a generative decoder, typically autoregressive, to map multimodal representations to outputs (text or labels).

- Generative (6.25% of the articles): Models that directly generate content (text or image) from the input, with predictions evaluated by comparing the generated output to the possible options.

- Supervised (contrastive + classification) (6.25% of the articles): Combines contrastive learning and supervised classification, typically training a model with known labels while aligning visual and textual modalities. Reduces the zero-shot gap through partial supervision.

- Cross-attention (6.25% of the articles): Uses cross-attention mechanisms between modalities to generate richer joint embeddings. Allows semantic information to directly influence the visual representation during inference.

- Few-shot (6.25% of the articles): Approach that adapts the model with very few labeled examples per class, using techniques such as in-context learning or prompt-tuning, to improve zero-shot performance without requiring large datasets.

- Linear probe (6.25% of the articles): A linear classifier (e.g., softmax over pretrained embeddings) is trained without modifying the backbone weights. Allows efficient evaluation of the semantic information contained in pretrained embeddings.

- Contrastive + linear (6.25% of the articles): Hybrid approach where contrastive learning is applied to align embeddings and then a linear probe is used for classification, combining zero-shot robustness with linear discriminative capacity.

- Unsupervised (entropy + prompt) (6.25% of the articles): Unsupervised strategy that uses uncertainty or entropy measures to evaluate confidence in predictions and textual prompts as guidance. Useful for zero-shot retrieval or classification in unlabeled domains.

The inference or classification component constitutes a fundamental axis in the design of multimodal models, as it determines how decision-making is operationalized from the joint representations. In the analyzed literature, there is a clear predominance of contrastive learning, present in five works (Huang, Li, et al., 2024; Liu et al., 2023; Phan et al., 2024; Varma et al., 2023; Zhong et al., 2024). Its use may indicate the community's prevailing view that explicit alignment between visual and textual embeddings is highly important for enabling zero-shot inference. It is a robust approach against heterogeneous distributions, and its high reproducibility and compatibility with large-scale pretraining favor generalization to unseen classes. However, relying on similarity measures without explicit supervision may limit performance in clinical tasks that require more structured decisions or sensitivity to contextual nuances, and it may require additional mechanisms, such as prototypes or cross-attention, to maintain effectiveness in multi-label or fine-grained scenarios.

The second most frequent trend corresponds to binary classification-based models, which appear in works that employ classifiers explicitly trained on latent representations (Cao et al., 2024; Wu et al., 2023; Zhang et al., 2023). This family of methods offers more direct control over the output space and usually achieves higher accuracy when labeled data are available, explaining their adoption in clinical environments where decision quality is a priority. However, it depends on the availability of annotations for each new task, which hinders immediate generalization and limits its utility in zero-shot scenarios. Moreover, unlike the contrastive approach, it does not model semantic relationships between classes, reducing its robustness to out-of-distribution concepts.

Beyond these two dominant approaches, there is a notable heterogeneity of minor methods, each representing 6.25% of the articles. This diversity reveals a field in the midst of methodological exploration, experimenting both with improving embedding quality through mechanisms such as cross-attention, and with generative models and decoders that offer alternatives to direct comparison of representations. In this line, some models incorporate cross-attention directly in the inference module (Lai et al., 2024), enabling deeper interaction between text and image and fostering multimodal contextualization that can be crucial for clinical reasoning tasks. However, these methods involve higher computational complexity and present additional interpretability challenges.

Generative approaches in the latent space also emerge, such as the diffusion-based method (Siddiqui et al., 2025), which not only produces inferences but also seeks to explain its decisions through internal generative processes. Although promising for attribution or interpretability objectives, its implementation remains costly and practical adoption is still limited. This is complemented by the presence of models oriented toward segmentation as the inference output, rather than traditional classification, as seen in Guo and Terzopoulos (2024). This approach is ideal for clinical tasks focused on anomaly localization, although it requires different metrics and evaluation frameworks.

Overall, the predominance of contrastive learning, followed by fine-tuned binary classifiers, shows that the community continues to favor approaches that are robust, reproducible, and compatible with zero-shot strategies. However, the presence of alternative methods based on cross-attention, generative models, unsupervised approaches, or segmentation indicates that the field is actively exploring hybrid and more sophisticated solutions. Altogether, this suggests that the next generation of multimodal models in clinical settings will more tightly integrate semantic alignment with contextual reasoning mechanisms and fine-grained capabilities, especially in scenarios with limited resources or high interpretability requirements.

Finally, regarding classification related to Domain Adaptation:

- Pretraining on domain data (37.5% of the articles): Pretraining the model on domain-specific datasets before zero-shot evaluation. This adapts visual and textual embeddings to the particular distribution of the domain, improving generalization in specialized tasks.

- Language model integration (18.75% of the articles): Explicit integration of pretrained language models (LLMs) to guide semantic understanding and generate prompts or enriched representations for visual tasks. Allows the VLM to incorporate contextual knowledge and complex relationships between concepts.

- Expert curation (12.5% of the articles): Use of manual selection of data, prompts, or image–text pairs by experts to guide training or evaluation. Increases accuracy and relevance, especially in sensitive domains such as biomedicine.

- Prompt-based (12.5% of the articles): Use of carefully designed textual prompts to guide the model's prediction. A central strategy in zero-shot and few-shot learning, where prompt formulation can drastically influence performance.

- Knowledge distillation (6.25% of the articles): Transfer of knowledge from a larger or more powerful "teacher" model to a smaller or more efficient "student" model, maintaining zero-shot capability while reducing computational requirements.

- Synthetic data augmentation (6.25% of the articles): Generation of synthetic data to expand coverage of the semantic space and improve zero-shot generalization. Reduces reliance on large annotated datasets. In sensitive domains with highly complex data, synthetic data generation can be unreliable (poorly generated data).

- Language-guided unsupervised adaptation (6.25% of the articles): Unsupervised adaptation guided by text, using prompts, entropy, or uncertainty measures to modify visual embeddings without labels. Allows the model to adjust to new domains or tasks without direct supervised training.

In the reviewed literature, pretraining on domain-specific data emerges as the most widely used domain adaptation strategy to improve zero-shot performance in multimodal models. This technique allows visual and textual embeddings to adapt to the particular distributions of clinical or specialized settings, capturing lexical, semantic, and visual patterns that generally pretrained models might overlook (Huang, Li, et al., 2024; Liu et al., 2023; Varma et al., 2023; Yu et al., 2025; Zhang et al., 2023; Zhong et al., 2024). Its strength lies in the direct improvement of semantic alignment between modalities, which is critical in scenarios where ambiguity or polysemy of medical concepts could compromise prediction accuracy. At the same time, this strategy facilitates handling out-of-distribution data, increasing generalization without requiring full retraining for each new task. Nevertheless, domain pretraining involves high computational cost and may induce overfitting if performed on limited datasets, which could partially reduce the model's zero-shot nature.

Complementarily, the integration of large language models (LLMs) is emerging as a strategy to enrich textual semantics and guide the interpretation of visual embeddings. This integration enables the capture

of complex relationships between concepts, the generation of contextual prompts, and the enhancement of visual–semantic reasoning and zero-shot retrieval tasks. Its application, however, requires additional computational resources and careful alignment between visual and linguistic embeddings to ensure semantic coherence.

Other approaches include knowledge distillation, which transfers capabilities from a larger or more specialized teacher model to a more efficient student model, preserving zero-shot capacity while reducing computational requirements (Cao et al., 2024). Similarly, strategies based on carefully designed prompts and expert curation allow the model's output to be conditioned according to diagnostic or semantic expectations, increasing accuracy and relevance in sensitive domains such as biomedicine (Lai et al., 2024). These techniques, although effective, rely on intensive manual intervention and present scalability limitations.

Likewise, some models incorporate explicit knowledge structures, such as medical graphs, to enrich inference (Phan et al., 2024; Wu et al., 2023). These predefined entities and relationships enable more precise alignment between language and vision, improving interpretability and diagnostic accuracy, although their scalability depends on the coverage and availability of specialized ontologies. In parallel, the use of synthetic data and language-guided unsupervised adaptation represent experimental strategies aimed at expanding the semantic space, adjusting embeddings without labels, and preserving zero-shot capability in settings with limited data.

Taken together, these strategies show that domain adaptation is not limited to a single method, but is instead shaped through the combination of complementary techniques that balance robustness, efficiency, and generalization capacity. Domain pretraining remains the most established and dominant strategy, while the integration of LLMs and emerging techniques — distillation, prompts, synthetic data, and unsupervised adaptation — point toward a hybrid and semantically enriched future. The evidence suggests that the next generation of zero-shot VLMs will combine these strategies to achieve models that are more adaptable, accurate, and capable of handling both the complexity of clinical domains and the heterogeneity of data.

The analysis of these four technical features makes it possible to identify both the maturity of the field and the emerging areas of innovation in zero-shot VLMs. In terms of maturity, the predominance of dual-encoder architectures combined with Joint Embedding Space and contrastive learning strategies shows that most existing models prioritize robustness, reproducibility, and scalability. This combination forms the backbone of the current literature, especially in retrieval and general-purpose classification tasks, defining a stable framework that enables consistent comparisons and broad applicability in clinical and biomedical domains. The strength of this set of methods reflects the consolidation of a de facto standard for implementing zero-shot VLMs, upon which most recent research is built.

However, beyond this predominant core, the review reveals considerable methodological experimentation. The minority strategies, which include fusion with attention, cross-attention, knowledge-enhanced approaches, prototype similarity, generative and decoder-based methods, LLM integration, prompts, unsupervised adaptation, and synthetic data, represent a fertile ground for innovation. These emerging approaches aim to explore more sophisticated reasoning capabilities, improve discrimination in fine-grained tasks, and enable greater adaptability to specific domains or limited-resource settings. The presence of these methods indicates that the scientific community is actively testing hybrid combinations and complementary solutions that could overcome the limitations of the dominant approaches.

In terms of emerging trends, clear patterns are observed that point to the next generation of zero-shot VLMs. The combination of contrastive learning with linear probes or cross-attention mechanisms is shaping up as a strategy to maintain generalization while increasing discriminative power. The integration of LLMs together with prompts and domain pretraining enhances semantic reasoning and the coverage of concepts specific to a clinical environment. Meanwhile, knowledge-enhanced and diffusion-based approaches allow the incorporation of external knowledge and the generation of synthetic data, offering alternatives to overcome the limitations of scarce or heterogeneous datasets.

These trends also highlight significant research opportunities. The development of hybrid models that combine multiple strategies and modalities may enable deeper semantic refinement and more efficient cross-attention. Emerging techniques for language-guided unsupervised adaptation and synthetic augmentation are particularly promising for low-resource settings, while the optimization of computational efficiency through distillation and linear probes opens the door to scalable models that do not sacrifice zero-shot capability.

In summary, the review reveals a dual landscape: on one hand, a consolidated foundation dominated by dual-encoder + Joint Embedding + contrastive learning, which ensures robustness and generalization; on the other, a diverse set of minority methods that reflect the frontier of innovation. This balance between maturity and experimentation suggests that the next generation of zero-shot VLMs will consist of hybrid, semantically enriched models, guided by LLMs and adaptable to specific domains, capable of tackling complex multimodal and fine-grained reasoning tasks. The evidence presented underscores the representativeness and comprehensiveness of the systematic review, offering a clear map of the state of the art and the most promising lines of research.

Once some of the main architectural features were analyzed, the next step was to perform a similar analysis regarding hyperparameter setting and fine-tuning, as well as the evaluation metrics used.

We conducted a categorization of these fields, the results of which are shown in Table 8.

Like in previous cases, we made a categorization of hyperparameters based on their type. The different established categories are:

- Optimization (32,26% of the articles): General optimization strategies for the models, including hyperparameter tuning, optimization algorithms (Adam, SGD, LAMB, etc.), and learning schedules. Their goal is to improve convergence, stability, and zero-shot performance without changing the architecture or the loss.

Loss-related (25.81% of the articles): Methods focused on the loss function, such as contrastive loss, triplet loss, cross-entropy, InfoNCE, or hybrid losses. Adjusting the loss enables better alignment between visual and textual embeddings and guides semantic discrimination in zero-shot settings.

Architecture-specific (22.58% of the articles): Optimization or modification directly linked to the model's architecture, such as tuning attention layers, fusion blocks, normalization layers, or dual-encoder design. This allows the model to capture more complex multimodal interactions.

Augmentation/control (6.45% of the articles): Strategies involving data control or the generation of input variants to improve generalization: visual data augmentation (rotations, crops, color jitter) or control of textual prompts. This helps reduce overfitting and improve robustness.

Prompt configuration (6.45% of the articles): Fine-tuning of textual prompts to guide the model's prediction. This includes question formulation, templates, and context chains, which is especially critical in zero-shot and few-shot settings.

Not specified (6.45% of the articles): Studies that mention performance improvements or adjustments but do not clearly detail the technique applied.

The authors note that the percentage does not add up to 100% linearly in terms of applied techniques because several articles use more than one strategy simultaneously.

In the reviewed works, the most common category corresponds to optimization hyperparameters, present in the majority of the articles (Cao et al., 2024; Huang, Jiang, et al., 2024; Jin et al., 2023; Rahman et al., 2025; Xie et al., 2024; Yu et al., 2025; Zhang et al., 2023; Zhong et al., 2024). These include learning rate, batch size,

Table 8
Hyperparameters and Metrics.

Paper	Hyperparameters	Metrics
A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images (Jin et al., 2023)	Optimization, Loss-related, Architecture-specific	Classification
Weakly supervised zero-shot medical image segmentation using pretrained medical language models (Guo & Terzopoulos, 2024)	Loss-related	Segmentation
Villa: Fine-grained vision-language representation learning from real-world data (Varma et al., 2023)	Architecture-specific, Loss-related	Classification, Retrieval/Ranking
Knowledge-enhanced visual-language pre-training on chest radiology images (Zhang et al., 2023)	Optimization	Classification
Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning (Huang, Li, et al., 2024)	Optimization, Loss-related	Classification, Segmentation, Explainability
A latent diffusion approach to visual attribution in medical imaging (Siddiqui et al., 2025)	Augmentation/control	Generative
Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models (Cao et al., 2024)	Optimization, Architecture-specific	Classification
Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis (Wu et al., 2023)	Not specified	Classification, Explainability
Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework (Phan et al., 2024)	Not specified	Classification, Segmentation
Carzero: Cross-attention alignment for radiology zero-shot classification (Lai et al., 2024)	Architecture-specific, Optimization, Prompt configuration	Classification, Explainability
Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination (Zhong et al., 2024)	Optimization, Loss-related, Architecture-specific	Classification
Adapting visual language models for generalizable anomaly detection in medical images (Huang, Jiang, et al., 2024)	Optimization, Loss-related, Architecture-specific	Classification
Pairaug: What can augmented image–text pairs do for radiology? (Xie et al., 2024)	Augmentation/control, Optimization	Classification
Improving medical vision language contrastive pretraining with semantics-aware triage (Liu et al., 2023)	Loss-related	Classification
Core-Periphery Multi-Modality Feature Alignment for Zero-Shot Medical Image Analysis (Yu et al., 2025)	Optimization, Architecture-specific	Classification, Explainability
Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? (Rahman et al., 2025)	Optimization, Loss-related, Prompt configuration	Classification

number of epochs, and optimizer parameters. Proper tuning of these hyperparameters is crucial to ensure stable and efficient convergence of the models, especially in clinical scenarios where data are limited or noisy. However, their importance is often underestimated in the articles, with tuning methods poorly described or unspecified, which affects the reproducibility of results and limits comparability across studies.

The loss-related hyperparameters constitute another relevant group, present in at least six models (Guo & Terzopoulos, 2024; Huang, Li, et al., 2024; Jin et al., 2023; Liu et al., 2023; Varma et al., 2023; Zhong et al., 2024). These include contrastive margins, softmax temperatures, and balancing weights between loss terms, all of which directly impact the nature of the learned representations and the stability of training, especially in contrastive or multitask architectures. Properly tuning these parameters requires a deep understanding of the behavior of the loss function—information that is not always detailed in the reviewed works, revealing an area where methodological reporting could be improved.

Architecture-dependent hyperparameters are also common (Cao et al., 2024; Huang, Jiang, et al., 2024; Lai et al., 2024; Varma et al., 2023; Yu et al., 2025; Zhong et al., 2024), defining structural aspects such as patch size, input resolution, number of layers, or embedding dimension. The selection of these parameters affects both the representational capacity of the model and its computational cost. However, in most cases, values from previous works, such as CLIP or ViT, are adopted without explicit empirical justification, limiting understanding of the trade-offs between performance and efficiency.

More advanced models incorporate augmentation or control hyperparameters (Siddiqui et al., 2025; Xie et al., 2024), especially in generative approaches or visual and textual augmentation strategies.

Parameters such as the number of denoising steps, guidance scale, or segmentation thresholds directly affect the quality of the generated output and require careful tuning to achieve clinically plausible and stable results. In zero-shot or few-shot contexts, some models like Carzero (Lai et al., 2024) integrate prompt or configuration hyperparameters, including textual templates, maximum input length, or syntactic formats. These parameters can have a disproportionate impact on the performance of foundation models adapted to the medical domain, although their tuning is generally based on heuristics rather than systematic exploration.

The global analysis shows that the literature primarily focuses on improving performance through internal optimization and loss tuning, reflecting that zero-shot models remain sensitive to experimental configuration. Architectural modifications and augmentation techniques act as strategic complements, especially useful for complex or fine-grained tasks, but they are not strictly necessary to achieve basic generalization in zero-shot scenarios. Meanwhile, prompt configuration, although less frequent, emerges as a key tool to fully exploit embeddings and zero-shot capabilities, particularly when large language models or domain-specific pretraining are integrated.

Overall, the review suggests that the success of zero-shot models does not depend solely on architecture or embeddings, but on the interaction of multiple factors, including optimization, loss functions, architectural design, and semantic guidance through prompts or augmentation. This multifactorial ecosystem explains the predominance of certain configurations over less common strategies, while also highlighting the need for greater methodological transparency and standardized documentation to facilitate comparison and reproducibility of results. In this regard, trade-offs between computational efficiency and model complexity, as well as between zero-shot robustness and fine-grained discrimination, are directly reflected in the choice of hyperparameters and architectural adjustments, showing that the fine-tuning of these elements remains a critical determinant of performance in clinical settings.

Regarding Hyperparameter tuning method, the resulting categorization is as follows:

- **Grid search**: Systematic search of combinations within a predefined grid.
- **Ablation**: Evaluation of the influence of each parameter through experiments where it is individually removed or varied.
- **Manual tuning/heuristics**: Empirical selection or based on previous studies, without explicit systematic search.
- **Validation-based selection**: Tuning based on performance over a validation set (e.g., minimum loss or maximum metric).
- **Not specified**: The procedure is not detailed.

Most of the articles do not specify the procedure used (Cao et al., 2024; Huang, Jiang, et al., 2024; Huang, Li, et al., 2024; Lai et al., 2024; Phan et al., 2024; Siddiqui et al., 2025; Varma et al., 2023; Wu et al., 2023; Xie et al., 2024; Zhang et al., 2023), which compromises reproducibility and makes rigorous comparison between approaches difficult. This lack of methodological transparency is especially problematic in biomedical contexts, where small variations in configuration can significantly affect clinical outcomes.

Among the works that do document their tuning strategy, the most common technique is ablation analysis, used in studies such as Guo and Terzopoulos (2024), Jin et al. (2023), Yu et al. (2025), which allows evaluating the model's sensitivity to the presence or absence of specific components, including hyperparameters. Although useful for identifying individual contributions, this approach is often limited to a subset of configurations and does not guarantee finding the global optimal values.

Only one work (Zhong et al., 2024) implements a systematic search via grid search, indicating a more exhaustive and controlled exploration of the hyperparameter space. However, this approach is computationally expensive and therefore uncommon in large models or when training on sensitive medical data.

Validation-based selection is explicitly documented in only one article (Liu et al., 2023; Rahman et al., 2025), which is surprising given its balance between efficiency and robustness. This method is particularly useful when a validation set representative of the clinical deployment environment is available, although it critically depends on the quality of that set.

In summary, hyperparameter tuning methods in the analyzed literature are poorly documented and, when applied, tend to be heuristic or partial. This situation highlights a significant methodological gap and suggests that future work should adopt and report more rigorously systematic and reproducible procedures for hyperparameter optimization, especially in contexts where diagnostic accuracy is critical.

Regarding the evaluation metrics used, we can classify them into the following categories:

Classification (60.87% of the articles): Classification metrics such as AUC, accuracy, precision, recall, F1, MCC.

Segmentation (13.04% of the articles): Segmentation metrics such as IoU, Dice score, pixel-wise accuracy.

Retrieval/Ranking (4.35% of the articles): Metrics such as Recallk, Precisionk, R-Precision.

Generative (4.35% of the articles): Image generation metrics such as Frechet Inception Distance (FID), SSIM, MS-SSIM.

Explainability (17.39% of the articles): Metrics such as Pointing Game, Phrase Grounding, Contrast-to-Noise Ratio, useful for attribution and visual explainability.

The metric values reported for each article are shown in Table 9.

The presented results show that, while zero-shot (and few-shot) models demonstrate promising performance, their generalization in specific medical tasks remains limited and highly dependent on the domain and dataset used. Specialized models such as MAVL, MedKLIP, MaCo, and PairAug-E stand out for their consistent performance on radiology datasets such as CheXpert, SIIM-ACR, RSNA, or NIH Chest X-ray, achieving high AUC and F1 scores in zero-shot scenarios, confirming the advantage of adapting architectures and pretraining to the clinical domain. However, variability across datasets is notable; for example, zero-shot classification models exhibit accuracies that vary significantly depending on the task, indicating that the heterogeneity of radiological modalities and the complexity of pathologies strongly affect robustness.

In clinical report generation tasks, although BIUD and BLIP achieve competitive metrics on internal datasets, performance drops on external tests reflect overfitting and highlight the difficulty of generalizing to unseen domains. Similarly, extremely high results in few-shot anomaly detection, such as those reported by MVFA on some metrics (AC/AS AUCs above 99%), should be interpreted with caution, as they may reflect easy datasets, potential data leakage, or measurements that do not capture true clinical difficulty. Generalist models, including ChatGPT-4 and OpenFlamingo, show uneven performance: they can achieve good results on specific datasets but do not offer consistency across different radiological domains.

Although many models report AUC as a performance metric, it has important limitations in clinical contexts: because it is aggregated over all possible thresholds, it does not reflect the model's capability in critical ranges of high sensitivity or low specificity, and it can be high even when the model fails to correctly identify extreme-risk cases. This bias is amplified in imbalanced medical datasets, where a high AUC can mask poor performance on the minority class, typically the most relevant for decision-making. Furthermore, discrepancies between validation and external testing, overfitting, or data leakage can inflate AUC without reflecting actual performance, while model calibration and robustness to atypical cases are also not captured (Hancock, Khoshgoftaar, & Johnson, 2023).

Therefore, clinical evaluation requires complementing AUC with threshold - specific metrics (sensitivity, specificity, F1, PPV/NPV, Precision–Recall curves) and analyzing relevant subgroups to ensure generalization beyond the training domain. In our view, any diagnostic

Table 9
Summary table of performance metric results.

Model - Article	Dataset	Key Metrics and Values
MedUnA (Rahman et al., 2025)	S-TB	Top-1 Accuracy: 67.67%
	M-TB	Top-1 Accuracy: 72.41%
	G-Pneumonia	Top-1 Accuracy: 76.12%
	IDRID	Top-1 Accuracy: 32.04%
	ISIC	Top-1 Accuracy: 29.70%
	Overall Average (MedCLIP-Swin)	Top-1 Accuracy: 55.59%
	Overall Average (GPT-3.5 descr.)	Accuracy: 55.79%
LaFTer (Rahman et al., 2025)	Overall Average (CLIP-ViT-B/32)	Top-1 Accuracy: 43.16%
	Overall Average (MedCLIP-Swin)	Top-1 Accuracy: 40.51%
TPT (Rahman et al., 2025)	Overall Average (CLIP-ViT-B/32)	Top-1 Accuracy: 41.07%
	Overall Average (MedCLIP-Swin)	Top-1 Accuracy: 43.44%
Zero-Shot (Baseline) (Rahman et al., 2025)	Overall Average (MedCLIP-Swin)	Top-1 Accuracy: 42.84%
MedAlpaca-13B (Baseline) (Rahman et al., 2025)	Overall Average	Accuracy: 45.34%
OpenBioLLM-8B (Baseline) (Rahman et al., 2025)	Overall Average	Accuracy: 38.29%
BIUD (Cao et al., 2024)	ChestCT-16K (Report Generation)	F1: 85.3%, P: 86.7%, R: 83.9%
	ChestCT-EXT (External)	F1: 70.4%, P: 65.4%, R: 76.3%
R2Gen (Cao et al., 2024)	ChestCT-EXT	F1: 57.2%
CMN (Cao et al., 2024)	ChestCT-EXT	F1: 26.5%
RadFM (Cao et al., 2024)	ChestCT-EXT	F1: 9.9%
BLIP (Cao et al., 2024)	ChestCT-EXT	F1: 80.3%
DCL (Cao et al., 2024)	ChestCT-EXT	F1: 78.5%
CP-CLIP (Yu et al., 2025)	SIIM-ACR, INbreast, TMED, ChestXray, CheXpert5 × 200 (Zero-Shot)	Balanced accuracy higher than MedCLIP (numerical value not provided)
VLM (Zhong et al., 2024)	CheXpert (Fine-tuned Multi-task)	Accuracy Promedio: 72.60%
	MedNLI (Fine-tuned)	Accuracy: 85.79%
	Open-I (Zero-Shot Semantic Structure)	Accuracy: 49.00%
ConVIRT (Zhong et al., 2024)	CheXpert (Fine-tuned Multi-task)	Accuracy Promedio: 36.20%
	MedNLI (Fine-tuned)	Accuracy: 86.80%
	Open-I (Zero-Shot Semantic Structure)	Accuracy: 43.10%
GLoRIA (Zhong et al., 2024)	CheXpert (Fine-tuned Multi-task)	Accuracy Promedio: 50.00%
	MedNLI (Fine-tuned)	Accuracy: 86.64%
	Open-I (Zero-Shot Semantic Structure)	Accuracy: 44.30%
MVFA (Huang, Jiang, et al., 2024)	Medical AD benchmark (Few-shot K=4, Ensemble)	AC AUC: 88.97%, AS AUC: 98.67%
	HIS (Few-shot K=4)	AC AUC: 82.71%
	LiverCT (Few-shot K=4)	AC AUC: 81.18%, AS AUC: 99.73%
APRIL-GAN (Huang, Jiang, et al., 2024)	Medical AD benchmark (Few-shot K=4)	HIS AC AUC: 76.11%, LiverCT AC AUC: 53.05%
WinCLIP (Huang, Jiang, et al., 2024)	Medical AD benchmark (Few-shot K=4)	HIS AC AUC: 67.49%, LiverCT AC AUC: 67.19%
CARZero (Lai et al., 2024)	ChestXray14 (Zero-Shot)	AUC: 0.810, MCC: 0.257, F1: 0.270, ACC: 0.867
KAD (Zhang et al., 2023)	CheXpert (Zero-Shot, 5 avg)	AUC: 0.905, MCC: 0.589, F1: 0.647, ACC: 0.875
	ChestXray14 (Zero-Shot, Macro Avg)	AUC: 0.789, MCC: 0.280, F1: 0.323, ACC: 0.816
CheXzero-E (Xie et al., 2024)	CheXpert (Zero-Shot, 5 diseases avg)	AUC: 88.92%, ACC: 85.75%, F1: 66.51%
PairAug-E (Xie et al., 2024)	CheXpert (Zero-Shot, 5 diseases avg)	AUC: 89.97%, ACC: 86.21%, F1: 67.78%
BiomedCLIP (Van, Verma, & Wu, 2024)	ALL-IDB2 (Zero-Shot)	Test accuracy: 76.92%
	BTD (Zero-Shot)	Test accuracy: 79.14%
	Overall Average	Test accuracy: 71.52%
ChatGPT-4 (Van et al., 2024)	ALL-IDB2 (Zero-Shot)	Test accuracy: 84.85%
	CX-ray (Zero-Shot)	Test accuracy: 61.82%
	Overall Average	Test accuracy: 66.09%
OpenFlamingo (Van et al., 2024)	CX-ray (Few-shot K=4)	Test accuracy: 71.28%
	Overall Average (Few-shot K=4)	Test accuracy: 57.68%

(*continued on next page*)

Table 9 (*continued*).

Model - Article	Dataset	Key Metrics and Values
MAVL (Phan et al., 2024)	SIIM-ACR (Zero-Shot)	AUC: 92.04%, F1: 77.95%, ACC: 87.14%
	CheXpert (Zero-Shot)	AUC: 90.13%, F1: 65.47%, ACC: 86.44%
MaCo (Huang, Li, et al., 2024)	SIIM (Zero-Shot Classification)	AUC: 90.4%
	RSNA (Zero-Shot Classification)	AUC: 88.6%
	NIH Chest X-ray (Zero-Shot)	AUC: 77.3%
	MS-CXR (Phrase Grounding)	CNR: 1.144, mIoU: 25.5%
MedKLIP (Wu et al., 2023)	RSNA Pneumonia (Zero-Shot)	AUC: 0.8694
	NIH Chest X-ray (Zero-Shot)	AUC: 76.8%
ViLLA (Varma et al., 2023)	MIMIC (Mapping Quality)	F1: 74.5
Zero-Shot Segmentation (Guo & Terzopoulos, 2024)	SIIM-ACR Pneumothorax (Zero-Shot)	Dice: 68.64%
Baseline (Image Generation) (Siddiqui et al., 2025)	COVID Dataset (Healthy Counterfactuals)	MS-SSIM: 0.830, SSIM: 0.798

support tool should additionally report accuracy by finding type and Top-1 accuracy in multi-label settings, ensuring that the model not only performs well globally but is also reliable in real clinical scenarios.

There is a notable prevalence of classification metrics, present in the vast majority of the reviewed works (Cao et al., 2024; Huang, Jiang, et al., 2024; Huang, Li, et al., 2024; Jin et al., 2023; Lai et al., 2024; Liu et al., 2023; Phan et al., 2024; Rahman et al., 2025; Varma et al., 2023; Wu et al., 2023; Xie et al., 2024; Yu et al., 2025; Zhang et al., 2023; Zhong et al., 2024). These metrics — such as AUC, accuracy, precision, recall, or F1 — constitute a widely adopted standard in diagnostic tasks, as they provide a direct evaluation of predictive performance over clinical categories.

In a subset of works, particularly those focused on localization or anatomical delineation tasks, segmentation metrics such as Dice score or Intersection over Union (IoU) are incorporated (Guo & Terzopoulos, 2024; Huang, Li, et al., 2024; Phan et al., 2024). These metrics are crucial in scenarios where spatial precision is relevant for clinical interpretation, for example, in identifying pathological regions in radiological images.

A smaller number of articles include explainability metrics (Huang, Li, et al., 2024; Lai et al., 2024; Wu et al., 2023; Yu et al., 2025), such as Pointing Game or Contrast-to-Noise Ratio, which allow indirect evaluation of the alignment between model reasoning and human or clinical attention. Although these metrics provide valuable information for validating model behavior, their use remains limited, likely due to the difficulty of defining consistent protocols and reliable reference datasets.

Retrieval or ranking-type metrics, such as Recallk, are used less frequently and appear only in specific cases of information retrieval or fine alignment tasks (Varma et al., 2023). These metrics are useful for evaluating the quality of multimodal representations in contexts involving text–image matching.

Finally, there is a limited application of generative metrics, such as FID or SSIM, in the context of diffusion-based models (Siddiqui et al., 2025). Since these models do not perform explicit predictions but rather conditioned visual reconstructions, generative metrics are necessary to quantify fidelity and realism.

It is worth noting that a considerable number of works include multiple metrics, reflecting a more comprehensive evaluation approach. However, the complete omission of metrics or vague specification in some cases (e.g., "not specified") limits the possibility of rigorous replication and comparison between approaches, highlighting the need for stronger standardization in evaluation protocols, especially in medical applications where robust validation is essential.

From the joint analysis of the hyperparameters used, tuning methods, and evaluation metrics, an overall heterogeneous and, in some respects, unsystematic picture of experimental design in biomedical vision-language models emerges. Although there is relative consensus on the importance of tuning optimization parameters, loss functions, and architecture, only a small fraction of works clearly specify the tuning strategies employed, with "not specified" annotations predominating. This suggests that despite the technical nature of these studies, a lack of transparency or standardization persists in documenting critical processes for reproducibility.

Likewise, the use of evaluation metrics shows some diversity depending on the task, but with a clear dominance of classification metrics. This reinforces the diagnostic orientation of many of these works, although it also limits the assessment of finer capabilities such as localization, explanation, or multimodal generalization. While some proposals have begun to integrate more specific metrics (generative, explainability, ranking), their adoption remains marginal.

Overall, these findings reveal a partial maturity in the experimental practices of the field. On one hand, there is evident technical expertise in the design of models and tasks; on the other, methodological rigor deficiencies are identified that could compromise the comparability, reproducibility, and clinical applicability of the results. Moving toward more standardized tuning and evaluation protocols, along with greater transparency in experimental documentation, will be key to consolidating the robustness and credibility of vision-language systems in the medical domain.

The predominance of optimization and loss function tuning demonstrates that, even in established architectures such as dual-encoders with joint embedding spaces, fine-tuning these parameters determines stability, discriminative capacity, and zero-shot generalization. Architectural modifications, augmentation strategies, and prompt configuration act as complementary factors that enhance performance in fine-grained or limited-resource tasks, but their impact directly depends on the proper adjustment of the main hyperparameters. Furthermore, the limited documentation of tuning methods and metrics highlights the need for greater transparency and standardization, especially in clinical contexts where small variations can significantly alter results. Therefore, conscious hyperparameter selection, combined with systematic tuning and evaluation practices, constitutes a key criterion to guide architectural decisions and optimize zero-shot VLM models for specific radiographic tasks, beyond simple descriptive percentages. The use of explainability provides the transparency in model decision-making that is essential for the end user of these diagnostic support applications (medical expert).

Table 10
Classification of 16 multimodal models according to the type of natural language usage.

Model	Category	Natural Language Usage Techniques	Key Comments
Weakly supervised zero-shot medical image segmentation using pretrained medical language models (Guo & Terzopoulos, 2024)	(A)	Use of real reports as weak supervision + ClinicalBERT	Image–text alignment during training; learning guided by clinical text
Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination (Zhong et al., 2024)	(A)	Clinical reports + contrastive loss + text perturbation	Semantic sensitivity and robustness to lexical variations
Improving medical vision language contrastive pretraining with semantics-aware triage (Liu et al., 2023)	(A)	Image–report pairs + BioClinicalBERT	Pretraining directly on unstructured clinical text
Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models (Cao et al., 2024)	(A)	Reports used in distillation + shared encoders	Knowledge transfer from X-ray to CT via language
Pairaug: What can augmented image–text pairs do for radiology? (Xie et al., 2024)	(A)	Image–text data augmentation using real clinical pairs + CLIP	Direct supervision combined with intelligent contrastive augmentation
Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning (Huang, Li, et al., 2024)	(A)	Vision Transformer + BERT + semantic alignment	Semantic granularity between image regions and clinical phrases
Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis (Wu et al., 2023)	(B)	Extraction of clinical triplets + ClinicalBERT	Explicit use of NER and medical knowledge structures
Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework (Phan et al., 2024)	(B)	Splitting clinical descriptions into visual aspects	Reformulation of VLP as multi-aspect supervised classification
Carzero: Cross-attention alignment for radiology zero-shot classification (Lai et al., 2024)	(B)	Prompts generated by clinical LLM (Spark1)	Smart insertion of standardized descriptions
Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? (Rahman et al., 2025)	(B)	LLM-generated disease descriptions. Cross-modal adapter + prompt tuning	Unpaired text–image learning via LLM descriptions
A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images (Jin et al., 2023)	(C)	Use of BioBERT as semantic embedding generator	Does not process full reports, only pretrained vectors
Knowledge-enhanced visual-language pre-training on chest radiology images (Zhang et al., 2023)	(C)	Textual knowledge encoder + embeddings	Does not use reports, but uses structured textual information
A latent diffusion approach to visual attribution in medical imaging (Siddiqui et al., 2025)	(C)	BERT-based text encoder + latent space	Use of language as a generative conditioning factor, not as direct supervision
Villa: Fine-grained vision-language representation learning from real-world data (Varma et al., 2023)	(C)	Clinical SBERT for embeddings + region-text alignment	Contrastive pretraining, but without full reports
Adapting visual language models for generalizable anomaly detection in medical images (Huang, Jiang, et al., 2024)	(C)	Textual prompts + CLIP adapted for medicine	Uses text as semantic guidance without training on reports
Core-Periphery Multi-Modality Feature Alignment for Zero-Shot Medical Image Analysis (Yu et al., 2025)	(C)	Textual prompts + BERT encoder for textual embedding	Uses text prompts, not full reports

4.3. RQ3. How many models actually leverage large unstructured clinical corpora, such as radiology reports or natural language descriptions? and how many incorporate detailed natural language descriptions as an integral part of the learning or inference process?

If we analyze individually each of the 16 selected articles (Table 10):

Legend of categories:

(A): Direct supervision with real clinical reports (37.5% of the articles). Allows aligning visual and textual embeddings with authentic domain data, maximizing accuracy and clinical relevance in zero-shot tasks.

(B): Semantic enrichment or description engineering (25% of the articles). Improvement of training texts or prompts through semantic enrichment, normalization of medical terminology, synonym expansion, or generation of more comprehensive descriptions. Its goal is

to increase the semantic coverage and generalization capacity of the model without changing the images.

(C): Implicit incorporation of natural language via embeddings (37.5% of the articles). Indirect integration of clinical knowledge through pretrained text embeddings or those derived from large LLMs, which encapsulate semantic domain information without requiring supervision.

12 out of the 16 extensively use large text corpora (Cao et al., 2024; Guo & Terzopoulos, 2024; Huang, Li, et al., 2024; Lai et al., 2024; Liu et al., 2023; Phan et al., 2024; Rahman et al., 2025; Siddiqui et al., 2025; Varma et al., 2023; Wu et al., 2023; Xie et al., 2024; Zhong et al., 2024). Delving into the actual use of natural language descriptions, we established a series of classifications regarding the use of natural language. Vision-language models applied to the medical field can be classified based on how and for what purpose they incorporate clinical natural language. This analysis distinguishes three main approaches: the direct use of real medical reports as supervision, semantic engineering through text transformation, and the implicit incorporation of textual representations via clinical embeddings.

(A) Direct supervision with real clinical reports. This group includes models that directly use medical reports, especially radiology reports, as the primary source of semantic supervision. These reports are employed either during pretraining as image–text pairs or as a central input for designing contrastive and multimodal architectures. Representative models such as "Weakly supervised zero-shot medical image segmentation using pretrained medical language models" (Guo & Terzopoulos, 2024), "Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination" (Zhong et al., 2024), "Improving medical vision language contrastive pretraining with semantics-aware triage" (Liu et al., 2023), and "Bootstrapping chest CT image understanding by distilling knowledge from X-ray expert models" (Cao et al., 2024) adopt this strategy. A common feature of these approaches is the use of specialized language encoders (such as BERT, BioBERT, or ClinicalBERT), often initialized with pretrained weights on clinical corpora. The alignment between clinical descriptors and visual features enables a more semantically informed encoding of medical content, enhancing the clinical reasoning capabilities of the models.

(B) Semantic enrichment or description engineering. Another line of work focuses on transforming clinical natural language into more explicit and manipulable semantic structures. This includes automatic extraction of entity-relation triples, decomposition of diseases into relevant clinical aspects, or generation of clinical prompts using language models. Notable examples of this approach include "Medklip: Medical knowledge enhanced language-image pre-training for X-ray diagnosis" (Wu et al., 2023) (which combines NER and ClinicalBERT to extract explicit knowledge), "Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework" (Phan et al., 2024), and "Carzero: Cross-attention alignment for radiology zero-shot classification" (Lai et al., 2024), which generates clinically informed prompts using LLMs. The MedUnA model (Medical Unsupervised Adaptation) leverages LLM-generated descriptions for each class to obtain rich textual embeddings that guide unsupervised adaptation to medical images without labels, overcoming the scarcity of paired data (Rahman et al., 2025). These strategies enable more robust generalization, better model interpretability, and mitigation of semantic imbalance, all without relying on rigid or closed labels. There is also a growing interest in equipping models with sensitivity to lexical perturbations and variability in clinical language.

(C) Implicit incorporation via textual embeddings. In this last approach, models do not directly process full clinical texts but integrate vector representations of them extracted from pretrained language models such as BioBERT or ClinicalBERT. These representations are used as additional inputs in multimodal architectures, for tasks such as matching, classification, or segmentation. Models like "A cross-modal deep metric learning model for medical visual representation" (Jin et al., 2023) and "Knowledge-enhanced visual-language pre-training on chest radiology images" (Zhang et al., 2023) adopt this approach. The CP-CLIP model (Yu et al., 2025) integrates vectorial representations of clinical text, obtained from a pre-trained BERT encoder, to map them to a unified latent space using an auxiliary Core-Periphery network that optimizes multimodal feature alignment. While they allow some degree of semantic integration, these solutions have limitations: by not incorporating the full reports, they may lose rich contextual information such as negations, temporality, or relationships between entities, and they rely heavily on the quality and coverage of the embeddings used.

The comparative analysis of these models allowed us to identify several trends. First, there is a universalization of natural language as an essential modality. All 16 reviewed models incorporate, either explicitly or implicitly, natural language descriptions, thus consolidating the image–text paradigm in the medical field, in line with the evolution observed in models like CLIP, ConVIRT, or GLoRIA.

A second trend is the predominance of real clinical reports over artificial annotations. More than 80% of the analyzed models use texts directly extracted from medical reports, which enables capturing fundamental linguistic nuances in clinical discourse, such as diagnostic uncertainty, negations, or the temporal relationship between findings.

Likewise, there is an observed increasing sophistication in text processing. The use of pretrained encoders is no longer sufficient: models incorporate advanced NLP techniques such as entity extraction via NER, prompt generation through LLMs, semantic decomposition of clinical descriptions (aspect queries), and the design of mechanisms robust to lexical variations. All of this contributes to a richer and more robust semantic representation.

Finally, models that integrate clinical natural language demonstrate significant improvements in zero-shot and transfer learning scenarios. These models show higher performance in classification without the need for specific labels, better generalization across domains (e.g., between datasets like CheXpert, MIMIC-CXR, and PadChest), and greater adaptability to new tasks through fine-tuning with minimal data.

These findings suggest a conceptual shift from a traditional label-based supervision paradigm toward a new paradigm of rich semantic supervision. In this new framework, medical reports are not seen as auxiliary information but as a deep and structured source of clinical knowledge that current models have yet to fully exploit.

To advance toward a genuine synergy between image and language, it will be necessary to develop models better aligned with clinical reasoning. This includes the explicit incorporation of structured medical knowledge (such as ontologies and clinical guidelines), the capacity for reasoning about complex semantic relationships, and the simultaneous adaptation of the model to both visual and textual contexts.

The analysis of natural language incorporation strategies in medical vision models shows a balanced predominance between direct supervision using real clinical reports and implicit incorporation via pretrained text embeddings, each present in 37.5% of the reviewed studies. Direct supervision allows precise alignment of textual semantics with medical images, enhancing accuracy and robustness in zero-shot scenarios, although its implementation requires access to annotated data, which can be costly or sensitive. On the other hand, the implicit integration of knowledge through clinical embeddings or semantically enriched language models offers advantages in generalization and scalability, reducing reliance on explicit annotations, though with certain limitations in capturing fine contextual details.

To a lesser extent, semantic enrichment or description engineering constitutes an intermediate strategy that expands coverage of complex medical concepts, improves understanding of synonyms, clinical jargon, and abbreviations, and increases zero-shot robustness in multi-center environments with heterogeneous vocabularies. Overall, these trends reflect a strategic balance between accuracy and generalization, suggesting that the most successful models combine partial

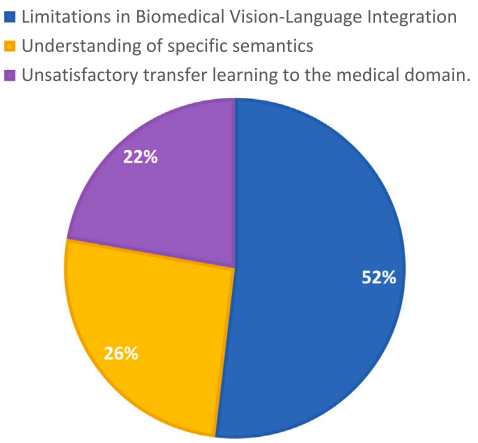

Fig. 4. Classification of biomedical zero-shot articles based on the type of the minimized issue.

direct supervision, semantically enriched clinical embeddings, and advanced textual processing techniques. This integration enables reliable zero-shot clinical reasoning that is adaptable to different hospitals and vocabularies, underscoring the need for representative datasets and transparent methodological procedures to maximize the clinical applicability of the models.

There is a pressing need to develop more representative and diverse multimodal datasets. Many current benchmarks lack the semantic complexity and linguistic variability inherent in real clinical practice. Future datasets should incorporate richer textual content, including multiple ways of expressing clinical findings, different languages, and diverse writing styles, along with high-granularity annotations. Evaluation metrics should go beyond numerical performance to assess clinical coherence, interpretative validity, and the model's ability to generalize across heterogeneous clinical scenarios.

4.4. RQ4. What are the main limitations in the integration of natural language and medical imaging, and how do they impact model performance?

Throughout the 16 analyzed models, several structural and semantic limitations are repeatedly identified that hinder the full development of vision–language architectures in clinical settings. See Fig. 4.

Note: Individual articles can target several minimized issues simultaneously.

The points raised by two of the analyzed works (specifically the most recently published ones) (Rahman et al., 2025; Yu et al., 2025) are particularly interesting, as they explicitly acknowledge limitations in integration and propose solutions to improve transfer. However, they do not detail granular semantic understanding or fine-grained alignment between specific clinical attributes and visual components. Both articles start from the premise that current vision-language models have significant limitations when applied to the medical domain, mainly because they rely on paired image–text data or suffer a sharp performance drop when used directly in clinical scenarios. One of them (Rahman et al., 2025) identifies the central problem as the scarcity of labeled data and proposes MedUnA, which leverages automatically generated descriptions from large language models alongside unpaired images. Thus, the work focuses on improving cross-modal alignment and facilitating adaptation when texts are abundant but annotated visual data are scarce, surpassing the typical transfer learning approach. However, its approach does not thoroughly explore specific medical semantics (such as the analysis of clinical attributes or triplets), instead working with global text–visual alignment.

Along the same lines, the article "Core-Periphery Multi-Modality Feature Alignment for Zero-Shot Medical Image Analysis" (Yu et al., 2025) focuses on the performance degradation of CLIP and similar VLMs due to the domain shift between natural and medical images. The paper proposes a brain-inspired (core–periphery) architecture aimed at improving global text–image alignment and thereby restoring zero-shot capability in clinical applications. Again, the emphasis is on technical integration and effective transfer of general models to the medical domain, rather than dissecting or mapping fine-grained semantic details of clinical concepts to specific image regions.

All of them warn of issues regarding the results obtained, proposing specific solutions from their design to minimize their impact.

Almost all the articles address the limitations of pre-trained foundation models, such as BERT, GPT, and CLIP, when applied to the medical domain. While these models have proven successful in natural language processing and computer vision, their utility in medicine is restricted by the detailed nature of medical tasks and the high demand for domain-specific knowledge.

In the biomedical field, the issues associated with zero-shot VLM developments differ from the problems related to biases in generalist categories, to the point that none of the selected developments explicitly mention any zero-shot VLM bias. The problem revolves around the specificity of the biomedical domain. Specifically:

- Limitations in Biomedical Vision-Language Integration: Difficulties in capturing the complex semantics present in radiological reports without resorting to costly manual annotations or prior coding of semantic relationships.

- Transfer learning from models pre-trained on large natural image datasets to the medical domain often proves unsatisfactory due to significant differences between the two types of images.

- Under the general heading of Limitations in Biomedical Vision-Language Integration, three fundamental challenges can be distinguished, which are interrelated but have different technical and methodological implications:

(a) Scarcity of high-quality clinical image–text pairs.

Unlike the general domain, where datasets such as MS-COCO or LAION-400M abound, the biomedical field suffers from a structural lack of datasets containing medical images paired with detailed and annotated clinical reports. This phenomenon has multiple causes:

Data protection and confidentiality: The sensitive nature of clinical data prevents its free circulation and complicates the construction of large and diverse corpora.

Institutional heterogeneity: Report formats, imaging modalities, and diagnostic standards vary widely across medical centers and regions.

Annotation cost: The manual creation of clinically relevant pairs requires medical experts, significantly increasing costs.

This scarcity limits the size and diversity of available datasets, which in turn restricts the models' ability to generalize to varied clinical contexts or complex tasks. As a partial response, some works (e.g., "Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination" (Zhong et al., 2024)) explore the use of augmentations, distillation, or knowledge transfer from generalist models (e.g., CLIP), although with still limited effectiveness.

(b) Generation of false negative pairs due to high semantic similarity

In contrastive training, one of the core principles is to distinguish between positive pairs (associated image and report) and negative pairs (non-associated pairs). However, in the clinical domain, different radiological studies may produce very similar descriptions (e.g., "no infiltrates observed" in multiple normal chest images), causing pairs that are actually semantically close to be mistakenly labeled as negatives. This results in:

Noise in learning: Models learn to penalize legitimate similarities, reducing their ability to recognize real clinical patterns.

Semantic misalignment: Precision decreases in tasks requiring sensitivity to subtle clinical nuances such as the presence or absence of subtle pathologies.

Some approaches attempt to mitigate this issue through intelligent sampling (e.g., PairAug), redefinition of loss functions, or the use of semantic triplets, but the problem remains critical, especially in small or less diverse datasets.

(c) Semantic misalignment between clinical text and image (Poor alignment between image and text).

Precise alignment between visual entities and their corresponding textual expressions remains a significant barrier. In many cases:

The report includes information not observable in the image (e.g., medical history, prior findings).

The image contains findings not mentioned in the report (reporting bias).

The text is ambiguous or uses indirect structures (“cannot rule out..”., “no conclusive evidence of..”.)

This mismatch hinders the convergence of multimodal losses and creates uncertainty in tasks such as localization (e.g., lesion grounding), causal inference, or model explanation.

Most current solutions rely on semantic encoders (e.g., ClinicalBERT, BioBERT) to reduce textual ambiguity, but few explicitly address the logical structure of clinical discourse or integrate meta-information such as negations, temporal relations, or patient-specific context.

These structural and semantic limitations in image–text integration not only directly affect the quantitative accuracy achieved by multimodal models but also significantly impact their applicability and robustness in real clinical settings. In particular, three critical performance areas are especially affected: cross-domain transferability, performance in few-shot or zero-shot scenarios, and the generation of interpretable outputs that can be reliably integrated into medical practice.

Firstly, the ability to effectively transfer across heterogeneous clinical domains (for example, from a dataset like CheXpert to others such as MIMIC-CXR or PadChest) is severely compromised when models are trained on limited, non-representative data or suffer from poor image–text alignment. This fragility is evident in at least four of the reviewed studies, which report a sharp drop in performance when models are evaluated on unseen domains. The root cause of this issue often lies in overfitting to domain-specific writing styles, pathology frequencies, or population characteristics in the source dataset. When a model learns non-generalizable associations between images and specific textual expressions, its performance outside that environment becomes highly unstable, severely limiting its usefulness in diverse clinical settings where usage conditions vary rapidly.

Secondly, these limitations hinder proper functioning in scenarios of unsupervised or minimally supervised learning, known as zero-shot or few-shot learning. In such contexts, models critically depend on the semantic richness captured during pretraining to infer unseen concepts, adapt to new descriptions, or recognize latent patterns in novel images. However, the scarcity of clinically paired data, the presence of semantically false negative pairs, and the misalignment between text and image significantly restrict this generalization capacity. As a result, models tend to behave erratically in unfamiliar conditions or show strong bias toward dominant concepts present in the training data. This rigidity undermines their utility in tasks such as clinical triage, detection of rare findings, or adaptation to less common imaging modalities.

Finally (and especially important from a clinical adoption standpoint) these limitations directly affect the model’s ability to generate interpretable outputs that align with clinical reasoning and are suitable for use in real-world healthcare environments. The lack of precise alignment between clinical reports and their corresponding images prevents the model from coherently justifying a prediction based on concordant visual and textual findings. This undermines the traceability of the model’s internal reasoning and hinders validation by medical professionals. Furthermore, the ambiguity or semantic poverty of textual embeddings may lead to vague or clinically imprecise responses, weakening trust in the system. Interpretability is not an optional feature in medicine; it is a fundamental requirement for validation, auditing, and eventual integration of these systems into clinical workflows.

Taken together, these shortcomings represent not only technical obstacles but also the threshold between the research utility of multimodal models and their practical viability in real-world medical settings. Overcoming these barriers requires more than fine-tuning architectures and loss functions, it demands a methodological reconfiguration that includes more representative datasets, higher-granularity semantic supervision, and a deeper understanding of the linguistic and visual structures inherent to clinical discourse.

We would also like to highlight the strong user-oriented focus evident in how the 16 reviewed papers describe these challenges. The terminology used in these works emphasizes how the identified shortcomings impact real-world tasks in medical interpretation, the integration with clinical language, and patient safety. This applied perspective is prioritized over purely mathematical or architectural deficiencies (such as domain shift (Gao et al., 2024; Jung, Jang, & Wang, 2024), hubness (Cheraghian, Rahman, Campbell, & Petersson, 2019; Dinu, Lazaridou, & Baroni, 2015), etc.).

Potential mitigation strategies include:

- Curation of datasets with richer and multi-granular annotations. This involves not only increasing the amount of available data but also enriching the labels and descriptions to better capture the semantic complexity inherent in clinical practice.

- Advanced techniques for clinical prompt engineering and semantic decomposition. These methodologies aim to strengthen the alignment between visual and textual information, enhancing the model’s ability to accurately interpret relationships and nuances present in medical reports.

- Explicit integration of structured medical knowledge, such as ontologies and clinical guidelines. Incorporating such formal information enables models to embed validated clinical rules and relationships, fostering a deeper and more coherent understanding of multimodal content.

- Novel loss functions designed to differentially penalize semantic false negatives. This innovation is key to improving learning quality by mitigating errors caused by high similarity between images or reports, which has traditionally hindered the correct association of image–text pairs.

From a purely humanistic point of view, the absence of mention of specific biases inherent to the models’ own architecture denotes a lack of multidisciplinary collaboration in practice. We consider it essential that, in the development of specialized biomedical zero-shot VLM models, development teams include experts from all relevant fields of knowledge from the very beginning of the design phase-covering architecture, learning techniques, performance metrics, result evaluation, etc. All involved agents should have sufficient knowledge of all the relevant areas. In fact, medical specialists should be proficient in all computing-related knowledge, the associated mathematical foundations, etc. Similarly, computing experts should have a proper understanding of the specific medical field in question.

From a more purely technical point of view, we consider it necessary to explicitly address the technological biases inherent in zero-shot VLM systems in biomedical applications. Ignoring them and maintaining an exclusive focus on biomedical problems from the user’s perspective could compromise the efficiency and quality of the system.

Phan et al. also point out an interesting issue directly related to the zero-shot capability: approaches that rely solely on the names of diseases may struggle to diagnose diseases that are truly unseen during training (Phan et al., 2024). This can be particularly sensitive in the case of rare diseases, infrequent syndromes, and even new pathologies.

Although the ideal scenario involves datasets with sufficient data on all existing diseases, the resources and effort required for this purpose should not limit the progress in the development of models that can overcome the limitations related to the quality of the dataset.

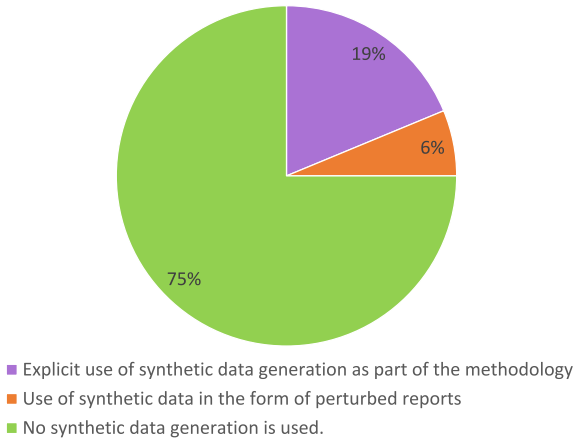

Fig. 5. Generative AI applied to bias mitigation.

Therefore, in our opinion, and in line with the zero-shot concept, a zero-shot VLM model should have the necessary generalization capability to handle these situations. However, this still does not seem to work in practice. It is possible that better management of the seen-unseen bias and domain-shift could help improve the generalization capacity for detecting rare or less frequent diseases.

Regarding the potential use of generative AI for bias mitigation, Fig. 5 graphically illustrates the use of synthetic data to mitigate biases related to domain adaptation in the 16 analyzed articles.

Within this set of articles, there is a predominant trend toward the use of real-world data, with relatively limited use of synthetic data for training and enhancing vision-language models in the medical domain. 12 of the 16 papers used no synthetic data generation. One of them used synthetic data in the form of perturbed reports. And only three of the 16 reviewed articles explicitly incorporated synthetic data generation as part of their methodology:

"Villa: Fine-grained vision-language representation learning from real-world data" (Varma et al., 2023). This work generates synthetic images by randomly assigning attributes to image regions, accompanied by automatically generated textual descriptions using predefined templates. This strategy aims to expand the variety of examples and strengthen semantic alignment, providing a richer foundation for learning fine-grained region-level representations.

"Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework" (Phan et al., 2024). In this case, synthetic generation is based on creating detailed textual descriptions of the visual aspects of diseases using a large language model (LLM), such as GPT-4. This approach focuses on enriching semantic textual information to improve pathology detection and classification, adding an extra layer of granularity and specificity that is often absent in original clinical reports.

"PairAug: What can augmented image–text pairs do for radiology?" (Xie et al., 2024). This article proposes a framework called PairAug, which generates synthetic data through both inter-patient and intra-patient augmentation techniques, in addition to applying data pruning methods to improve dataset quality. The synthetic data generation is designed to increase dataset diversity, enhance model robustness, and improve performance in radiology classification tasks.

"Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination" (Zhong et al., 2024) mentions the use of synthetic data in the form of perturbed reports. In this case, synthesis is carried out through the controlled perturbation of texts, generating modified versions of clinical reports to train the model to be more sensitive to subtle semantic variations, thereby enhancing its contrastive learning capabilities.

The remaining articles do not employ synthetic data generation. Instead, they rely on large volumes of curated and annotated real clinical data, combined with extensive pre-training and advanced learning techniques to improve multimodal representation. These approaches reflect the trust and priority given to the fidelity and accuracy of real-world data in the biomedical context, where clinical precision and consistency are critical.

The limited adoption of synthetic data generation may be attributed to several factors:

- Complexity and risk of clinical errors: Generating synthetic medical images and descriptions that maintain clinical coherence, diagnostic accuracy, and realistic variability is a highly complex task. Incorporating erroneous synthetic data could introduce biases or serious errors into sensitive models.

- Limited availability of mature tools: Unlike other areas of computer vision, synthetic data generation for medical imaging requires highly specialized models and rigorous clinical validation, which hinders broader implementation.

- Preference for real data quality and quantity: Many studies prioritize collecting and leveraging large-scale multimodal datasets with extensive linked clinical reports, which offer rich semantics and realistic diversity, reducing the immediate need for synthetic data.

However, the studies that do incorporate synthetic data highlight several important benefits:

- Increased diversity and effective dataset size without the need for costly manual annotation processes.

- Improved model generalization, particularly in handling rare or underrepresented cases within the original data.

- Enhanced training for fine-grained semantic alignment between image and text, by generating controlled examples with precise descriptions.

However, we believe that these benefits are highly dependent on the guarantees regarding the accuracy of the generated data and its absolute consistency with clinical reality.

Therefore, although synthetic data generation is a promising strategy to overcome some limitations of biomedical datasets, its current adoption remains marginal and is reserved for specific cases where added value can be ensured without compromising clinical integrity. This situation opens a clear line of future research aimed at developing more robust and clinically validated synthetic generation methods that can effectively complement existing large-scale real datasets.

4.5. RQ5. What robust methodologies are described in the literature for hyperparameter selection and tuning in zero-shot VLMs for X-ray classification?

After analyzing the hyperparameter tuning methods, the approach used in each of the 16 reviewed studies is presented in Table 11.

As in the previous research questions, we established a classification of methods based on the following categories:

- **Grid search**: Systematic search of combinations over a predefined grid.

- **Ablation**: Evaluation of the influence of each parameter through experiments where it is individually removed or varied.

- **Manual tuning/heuristics**: Empirical selection or based on prior studies, without explicit systematic search.

- **Validation-based selection**: Tuning based on performance on a validation set (e.g., minimum loss or maximum metric).

- **Not specified**: The procedure is not detailed.

Grouping the results obtained (Table 12):

Most papers do not specify the procedure used (Cao et al., 2024; Huang, Jiang, et al., 2024; Huang, Li, et al., 2024; Lai et al., 2024; Phan et al., 2024; Siddiqui et al., 2025; Varma et al., 2023; Wu et al., 2023; Xie et al., 2024; Zhang et al., 2023), which compromises reproducibility and makes rigorous comparison between approaches difficult. This lack

Table 11
Hyperparameter tuning method per model.

Paper	Hyperparameter tuning method
A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images (Jin et al., 2023)	Ablation
Weakly supervised zero-shot medical image segmentation using pretrained medical language models (Guo & Terzopoulos, 2024)	Ablation
Villa: Fine-grained vision-language representation learning from real-world data (Varma et al., 2023)	Not specified
Knowledge-enhanced visual-language pre-training on chest radiology images (Zhang et al., 2023)	Not specified
Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning (Huang, Li, et al., 2024)	Not specified
A latent diffusion approach to visual attribution in medical imaging (Siddiqui et al., 2025)	Not specified
Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models (Cao et al., 2024)	Not specified
Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis (Wu et al., 2023)	Not specified
Decomposing disease descriptions for enhanced pathology detection: A multiaspect vision-language pre-training framework (Phan et al., 2024)	Not specified
Carzero: Cross-attention alignment for radiology zero-shot classification (Lai et al., 2024)	Not specified
Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination (Zhong et al., 2024)	Grid search
Adapting visual language models for generalizable anomaly detection in medical images (Huang, Jiang, et al., 2024)	Not specified
Pairaug: What can augmented image–text pairs do for radiology? (Xie et al., 2024)	Not specified
Improving medical vision language contrastive pretraining with semantics-aware triage (Liu et al., 2023)	Validation-based selection
Core-Periphery Multi-Modality Feature Alignment for Zero-Shot Medical Image Analysis (Yu et al., 2025)	Grid search, Ablation
Can Language-Guided Unsupervised Adaptation Improve Medical Image Classification Using Unpaired Images and Texts? (Rahman et al., 2025)	Validation-based selection

Table 12
Papers by hyperparameter tuning method employed.

Type of tuning method	Number of papers	Additional comments
Ablation studies	3	Explore the effect of parameters and modules, more manual
Parameter discussion (manual experiments)	1	Systematic evaluation of the individual impact of hyperparameters
Targeted grid search	2	Automated and systematic search in a reduced space
Pre-training with validation	2	Tuning based on validation performance to select optimal parameters
Not specified/Manual/Implicit	9	Tuning methods not detailed

of methodological transparency is especially problematic in biomedical contexts, where small variations in configuration can significantly affect clinical outcomes.

Among the works that do document their tuning strategy, the most represented technique is ablation analysis, used in studies such as Guo and Terzopoulos (2024), Jin et al. (2023), Yu et al. (2025), which allows evaluation of the model's sensitivity to the presence or absence of specific components, including hyperparameters. Although useful for identifying individual contributions, this approach is usually limited to a subset of configurations and does not guarantee finding globally optimal values.

Only one work (Zhong et al., 2024) implements a systematic grid search, indicating a more exhaustive and controlled exploration of the hyperparameter space. However, this approach is computationally expensive and therefore uncommon in large models or when training on sensitive medical data.

Validation-based selection is developed in only two articles (Liu et al., 2023; Rahman et al., 2025), which is surprising given its balance between efficiency and robustness. This method is particularly useful when a validation set representative of the clinical deployment environment is available, although it critically depends on the quality of that set.

The hyperparameter tuning methods in the analyzed literature are sparsely documented and, when applied, tend to be heuristic or partial. This heterogeneity reflects both the diversity of problems addressed and the differences in computational resources, experimental objectives, and levels of methodological rigor among the studies. The predominance of unspecified approaches highlights a significant methodological gap and suggests a critical opportunity to promote greater transparency and systematic documentation of these processes. Adopting and reporting systematic and reproducible procedures for hyperparameter optimization is essential, particularly in domains where diagnostic accuracy and scientific reproducibility are priorities.

4.6. Strengths and weaknesses of the current state of zero-shot VLMs in radiographic classification

Strengths (potential and methodological maturity):

Zero-shot Vision-Language Models (VLMs) demonstrate transformative potential in radiographic classification due to the very nature of zero-shot learning, which endows them with unique characteristics highly valuable in the biomedical domain:

- Generalization to unseen findings: A core strength of these models is their ability to infer semantic relationships and make predictions on unseen classes without specific training. This is particularly relevant in healthcare, where high-quality annotated data is limited and costly.

- Data efficiency and scalability: The approach reduces the need for expensive annotations and can adapt to multiple datasets without full retraining. The dominant architecture, the dual-encoder, combined with a Joint Embedding Space and contrastive learning, prioritizes robustness, reproducibility, and scalability, setting a de facto standard in the field.

- Leveraging domain knowledge: The prevailing trend of pretraining on clinically specific data (present in 37.5% of the studies) is crucial. This ensures that visual and textual embeddings are adapted to the particular distribution of the clinical environment, capturing specialized patterns and improving generalization in specialized tasks.

Weaknesses and critical limitations (methodological and data challenges):

Despite their strengths, effective implementation in clinical practice is hindered by structural, methodological, and data-related limitations:

- Dependence on narrow datasets and bias: Current VLMs rely on a few large but semantically narrow datasets (RQ1). This concentration results in low diversity in the biomedical data used, limiting generalizability and posing a significant risk of propagating and amplifying inherent dataset biases into trained models.

- Insufficient vision-language integration: The main limitation, reported in 52% of studies, is difficulty in vision-language integration. Most models underutilize unstructured clinical text (RQ3), and semantic alignment is weak, negatively affecting both performance and explainability. This manifests as:

(a) Generation of semantic false negatives during contrastive training, where legitimately similar reports are incorrectly penalized.

(b) Misalignment between text and image (e.g., reports contain historical information not visible, or images contain findings not reported), complicating causal inference and model reasoning traceability.

- Fragility and lack of clinical robustness: Performance is sensitive to class imbalance, and robustness is limited against domain variations, artifacts, or rare findings. Out-of-distribution (OoD) generalization and effective transfer between heterogeneous clinical domains (e.g., PadChest to MIMIC-CXR) are compromised.

- Methodological deficiencies in reproducibility: Hyperparameter selection is rarely reported (RQ5), with 9 of 16 studies not detailing tuning methodology. This lack of methodological transparency hinders reproducibility and limits the ability to rigorously compare approaches (a critical issue in diagnostic settings).

- Limited clinical evaluation: Classification metrics predominate (60.87%), such as AUC, which have significant limitations in clinical contexts as they do not reflect performance at critical thresholds or for minority classes. Probabilistic calibration is suboptimal, and the use of explainability metrics (e.g., Pointing Game) remains marginal (17.39% of studies).

4.7. Positioning in the broader landscape of medical AI and foundational models

Research on zero-shot Vision-Language Models (VLMs) for radiographic classification occupies a critical point in the landscape of medical AI. The central principle of these models — the ability to infer semantic relationships and make predictions on unseen classes without specific training — is particularly valuable in the biomedical domain.

This utility stems from inherent characteristics of the healthcare sector:

- Scarcity of annotated data: The availability of medically annotated data with diagnostic quality is limited, costly, and constrained by ethical or legal considerations.

- Clinical variability: Zero-shot VLMs allow addressing highly variable clinical problems, such as diagnosing rare diseases or emerging conditions that are not represented in training datasets.

- Processing of clinical texts: Radiographic classification, as a first-line screening modality, offers rich semantic interaction with textual clinical reports.

The studies included in this review (N=16) focused strictly on radiographic image classification in zero-shot settings. Therefore, most generalist foundation models or those trained at large scale without specific adaptation to the medical domain did not meet the inclusion criteria and were excluded. The exclusion was based on the objective of maintaining a strict focus on models truly relevant to the X-ray domain, as the utility of generally pre-trained models (such as BERT, GPT, and CLIP) in medicine is limited by the detailed nature of medical tasks and the high demand for domain-specific knowledge.

Nevertheless, the relevance of foundation models in the future of medical AI is undeniable, as multimodal models can leverage text to make predictions on medical images. The recent emergence of health-specialized FMs, such as Med-PaLM M, GPT-4V, LLaVA-Med, and Med-CLIP, represents a line of research shaping the next generation of zero-shot VLMs.

Although most FMs were not the primary focus, some models in the review are based on adaptations of, or are directly compared to, them:

- Many of the 16 reviewed studies use visual backbones such as Vision Transformers (ViT) and textual backbones like ClinicalBERT and

its variants, which form the foundation of foundation model architectures.

- The dominant dual-encoder architecture combined with a Joint Embedding Space reflects a robust and scalable standard for zero-shot tasks.

- Some included papers compare their performance with generalist or adapted foundation models, such as ChatGPT-4 and OpenFlamingo, which showed uneven performance across different radiological domains.

Models such as MedKLIP (included in the review) use pretraining with explicit medical knowledge (knowledge-enhanced), demonstrating a pathway toward foundation models adapted to the domain.

Implications for the field:

Domain-specific zero-shot VLMs for radiography (like those analyzed) and foundation models (FMs) mutually influence each other. The reviewed studies show that for clinical application, pretraining on domain-specific data is crucial to adapt visual and textual embeddings to the specialized clinical environment.

The next generation of zero-shot VLMs will likely consist of hybrid models. These models will combine the robustness and scalability of dual-encoder architectures with the integration of large language models (LLMs) through prompt-based techniques and expert curation. The goal will be to leverage the vast contextual knowledge of FMs while maintaining the specificity and diagnostic accuracy required in clinical settings, enabling deeper semantic reasoning.

This hybrid approach will allow zero-shot VLMs to overcome current limitations, such as sensitivity to domain shift and insufficient semantic alignment, ensuring that diagnostic support tools are adaptable, accurate, and capable of addressing the complexity of clinical domains.

4.8. Most promising areas of research

Overcoming current limitations requires a shift toward the development of hybrid and semantically enriched models:

1. Representative and multilingual datasets:

∘ There is an urgent need for open, multilingual, and multimodal datasets that reflect clinical and geographic diversity.

∘ Priority should be given to creating datasets with broad taxonomic coverage (similar to PadChest) to capture fine-grained nuances (RQ1).

∘ Despite the risks, the generation of synthetic data is considered a line of research, especially for augmenting image–text pairs and training models more robust to semantic variations.

2. Deep integration of knowledge and clinical reasoning:

∘ Development of architectures that allow closer integration of semantic alignment with contextual reasoning mechanisms.

∘ Explicit integration of medical knowledge structures (ontologies and clinical guidelines) to enrich multimodal representations and improve interpretability.

∘ Exploration of integrating Large Language Models (LLMs) through prompt-based techniques to enhance semantic reasoning and coverage of specific concepts.

3. Rigorous methodology and evaluation:

∘ Research on designing methodologies for optimal hyperparameter selection (RQ5), which is crucial for zero-shot stability and generalization.

∘ Mandatory inclusion of explainability metrics (RQ4) to allow the end user (the medical expert) to verify the correct functioning of the model and the validity of the clinical patterns identified.

∘ Implementation of standardized evaluation protocols that include multi-label and cross-dataset benchmarks, and reporting metrics at clinically relevant thresholds (sensitivity, specificity, F1, PPV/NPV).

4. Development of hybrid and efficient models:

∘ Research on strategies that combine zero-shot with light fine-tuning (such as few-shot or prompt-based approaches) to maximize generalization and discriminative capacity.

∘ Use of knowledge distillation to transfer capabilities from large models to smaller, more efficient models while maintaining zero-shot capability.

∘ Exploration of unsupervised language-guided adaptation techniques (RQ2), which are promising for resource-limited environments.

The current state of research suggests that successful future zero-shot VLMs will combine the robustness of dual-encoder architectures with the semantic refinement of LLMs and explicit domain knowledge, while ensuring methodological transparency and rigorous clinical validation.

It is important to note that, since our review focused specifically on zero-shot models, the search strategy may have biased the selection toward studies explicitly using this term, potentially overlooking "few-shot", "fine-tuning", or "transfer learning" approaches with similar objectives. Therefore, future studies could consider a broader exploration of these adaptation strategies, allowing for a more comprehensive view of the VLM landscape in radiographic imaging.

5. Final conclusions

Zero-shot Vision-Language Models (VLM), with their ability to detect complex relationships between text corpora and images and to infer in unseen domains, offer great potential as diagnostic support tools in medical imaging. However, effective implementation in the field has not yet been achieved.

In preliminary outlooks, we found that in the biomedical domain, the area that has garnered the most interest is medical image classification, so we focused our systematic review specifically on this aspect.

As a result of applying the PRISMA methodology, we identified 16 articles. Subsequently, we thoroughly analyzed the datasets used, their architectures, hyperparameters, evaluation metrics, use of natural language, encountered issues, and hyperparameter tuning methodologies.

In some cases, it was difficult to identify the text corpus analysis component, turning the system into a supervised classifier with labels. This may be related to the scarcity of open-access medical reports available to train such models. (RQ3)

Regarding the zero-shot capability of the models, understood as their ability to generalize to domains not seen during training, if the models are trained with some data from another domain, we might be dealing with few-shot or fine-tuning models. (RQ4)

There is an urgent need for more open-access text–image paired datasets to better train models, working to avoid biases related to the relative heterogeneity caused by all models training on the same datasets. The very high potential of these architectures undoubtedly makes it worthwhile to redouble efforts to generate datasets with thousands of effectively integrated multimodal data points. Likewise, the data must include all scenarios that might appear in imaging in sufficient quantity. It is also necessary to have datasets in multiple languages to optimize the multilingual capability of the models. In this regard, PadChest is a comprehensive dataset, very varied in terms of imaging events and available in Spanish. (RQ1)

Among the high variability in architectures, dual-encoder is the most used, followed by architectures that utilize fusion with attention. The most frequent visual backbones are Vision Transformers (ViT) and, secondarily, ResNet-50. The most used textual backbones are Clinical BERT and its variants, followed by SBERT and CXR-BERT. Thirdly, we identify CLIP Text Encoder. A large number of architectures work with joint embedding spaces without a clear trend regarding inference mechanisms, although contrastive classification and the use of fine-tuned classifiers show some prevalence. (RQ2)

Regarding the metrics used to evaluate the models, there is a notable prevalence of classification metrics. Although the use of explainability metrics is not yet widespread, we consider the development and implementation of these metrics fundamental, especially in the biomedical field, to provide medical professionals with support tools

where they (the end users) can verify the correct functioning of the model. Transparency of models in this field is paramount. (RQ4)

It is interesting to observe that most models do not use synthetically generated data. This is logical given the difficulty mainly associated with generating images and the risks involved in training models with images (and texts) that may present undetected errors or defects. In our opinion, at this moment it may even be risky to introduce synthetic data in training diagnostic support tools. (RQ2)

We want to highlight that the problems detected by the authors of the 16 articles are deeply user-focused. Inherent biases of the architectures themselves, such as hubness or domain shift bias, are not mentioned. (RQ4)

It is surprising that methodologies for selecting optimal hyperparameters are unspecified. (RQ5)

After analyzing these 16 excellent articles, we reaffirm the urgent need to provide the scientific community with more and better datasets that include images, medical reports, and increasingly integrated multimodal data—in multiple languages—not only to optimize the multilingual capabilities of the models but also to better reflect scenarios related to different geographic regions. (RQ1, RQ3)

We consider that generating high-quality synthetic data at this time involves too many inherent risks and introduces much uncertainty to these models; therefore, in our opinion, efforts should undoubtedly be directed toward increasing the quantity and quality of training data. (RQ2)

An interesting research avenue will be designing methodologies for selecting optimal hyperparameters for the specific training of zero-shot VLM models, as a means to effectively seek their best performance and generalization capacity. (RQ5)

Furthermore, it is essential to address these issues from a multidisciplinary perspective, incorporating rigorous analysis of structural biases present within the model architectures themselves. (RQ4)

The dataset must be carefully balanced, ensuring an equivalent number of image–text pairs for each type of clinical finding. We propose, as a design hypothesis of our own, a threshold of approximately 5000 image–text pairs per finding, as this allows capturing both the visual and textual diversity of each category, facilitates balanced training without introducing class bias, and aligns with the dataset sizes used in reference clinical benchmarks. This strategy ensures representative coverage without requiring excessive computational resources.

Furthermore, for each additional language included, it is recommended to add an equivalent set of 5000 image–text pairs per finding, maintaining the consistency and multilingual representativeness of the dataset.

We recommend standardizing the following model evaluation protocols: first, report metrics at clinically relevant thresholds, including sensitivity, specificity, F1-score, PPV/NPV, and Precision–Recall curves, complementing AUC to capture performance in critical subgroups and minority classes. Second, include Top-1 accuracy and accuracy per finding type in multi-label tasks, ensuring that models not only discriminate globally but also correctly predict individual findings. Third, evaluate cross-dataset generalization by using external test sets with distributions different from training and validation, to detect overfitting and information leakage. Fourth, report calibration and robustness against atypical cases or outliers through confidence interval analyses and controlled perturbations of input data. Finally, we suggest accompanying all results with a standardized checklist for VLMs in radiology, including:

1. Clear and comprehensive description of datasets and splits, explicitly indicating the number of image–text pairs per finding type (measure of class imbalance).

2. Primary and subgroup-specific metrics, including test accuracies per finding type and top-1 test accuracy, since clinical classification is multi-label.

3. Preprocessing protocols: image normalization, text tokenization, terminology standardization, metadata handling, etc.

4. Zero-shot strategies: type of text–image mapping, prompts or clinical templates, embeddings, etc.

5. External generalization tests across multiple and diverse test sets.

6. Robustness analyses against outlier cases, domain shifts (imaging device parameters, noise, contrast, different cohorts, etc.), or controlled perturbations (small rotations, cropping, compression, etc.).

7. Implementation of explainability tools that allow verification that the model identifies patterns and textures associated with each finding type unequivocally.

8. Full availability of code and models to ensure reproducibility.

The correct implementation of zero-shot VLM systems in the biomedical field requires not only technical advances but also sustained commitment to creating representative data, transparency in inference processes, and active inclusion of professionals from diverse disciplines. The development of systems capable of effectively integrating into clinical environments will largely depend on a balanced combination of technical progress, access to representative data, and a multidisciplinary approach that considers both the mathematical nature of these models and their final usability in diagnostics.

6. Future work

Although this review focused on zero-shot models, it is important to acknowledge that the search strategy may have biased the selection toward studies explicitly using this term, potentially overlooking few-shot or transfer learning approaches with similar objectives.

Future research could:

- Explore few-shot, fine tuning and transfer learning approaches to evaluate their performance relative to zero-shot models.

- Develop comparative benchmarks and datasets that allow for uniform evaluation of these variants.

- Investigate hybrid integrations of these techniques with LLMs and explicit medical knowledge, aiming to enhance robustness, interpretability, and generalization.

Glossary

AUC: Area Under the Curve.

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

RQ: Research Question.

VLM: Vision-Language Model.

CRediT authorship contribution statement

Ana Guerrero-Tamayo: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Visualization, Writing – original draft. **Ibon Oleagordia-Ruiz:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing. **Begonya Garcia-Zapirain:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Funding sources

This work has been supported by the Basque Government through the Hazitek 2024 program, Spain, within the framework of the IRUD-IA project: “Medical Image Analysis Technologies with Artificial Intelligence for the Development of Medical Devices”, project code ZE-2024/00030.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ana Guerrero-Tamayo reports financial support was provided by Basque Government. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data availability

No data was used for the research described in the article.

References

Al Rahhal, M. M., Bazi, Y., Elgibreen, H., & Zuair, M. (2023). Vision-language models for zero-shot classification of remote sensing images. *Applied Sciences*, *13*(22), 12462.

Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, *57*(4), 80.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., et al. (2024). An introduction to vision-language modeling. arXiv preprint.

Buckley, T., Diao, J. A., Rajpurkar, P., Rodman, A., & Manrai, A. K. (2024). Multimodal foundation models exploit text to make medical image predictions. arXiv preprint.

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, *66*, Article 101797.

Çallı, Erdi, Sogancioglu, Ecem, van Ginneken, Bram, van Leeuwen, Kicky G., & Murphy, Keelin (2021). Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, *72*, Article 102125.

Cao, W., Zhang, J., Xia, Y., Mok, T. C. W., Li, Z., Ye, X., et al. (2024). Bootstrapping chest CT image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11238–11247). Conference paper.

Cheraghian, A., Rahman, S., Campbell, D., & Petersson, L. (2019). Mitigating the hubness problem for zero-shot learning of 3d objects. arXiv preprint arXiv:1907.06371.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., et al. (2023). Reproducible scaling laws for contrastive language-image learning. *vol. 281*, In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8–2829).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., et al. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304–310.

Dimitri, G. M. (2024). Mining comorbidities: a brief survey. arXiv preprint arXiv:2406.10696.

Ding, L., Shih, K., Wen, H., Li, X., & Yang, Q. (2025). Cross-attention transformer-based visual-language fusion for multimodal image analysis. *International Journal of Applied Science*, *8*(1), 27.

Dinu, G., Lazaridou, A., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. arXiv preprint arXiv:1412.6568.

Du, Y., Konyushkova, K., Denil, M., Raju, A., Landon, J., Hill, F., et al. (2023). Vision-language models as success detectors. In S. Chandar, R. Pascanu, H. Sedghi, D. Precup (Eds.), *vol. 232*, *Proceedings of the 2nd conference on lifelong learning agents* (pp. 120–136). PMLR, Conference paper.

Gao, H., Zhang, R., Yi, Q., Yao, H., Li, H., Guo, J., et al. (2024). Prompt-based visual alignment for zero-shot policy transfer. arXiv preprint arXiv:2406.03250.

Gavrikov, P., Lukasik, J., Jung, S., Geirhos, R., Lamm, B., Mirza, M. J., et al. (2024). Are vision language models texture or shape biased and can we steer them? arXiv preprint arXiv:2403.09193.

Guo, D., & Terzopoulos, D. (2024). Weakly supervised zero-shot medical image segmentation using pretrained medical language models. In *2024 IEEE international symposium on biomedical imaging* (pp. 1–5). Conference paper.

Hamidieh, K., Zhang, H., Gerych, W., Hartvigsen, T., & Ghassemi, M. (2024). Identifying implicit social biases in vision-language models. *vol. 7*, In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 547–561). Conference paper.

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced big data. *Journal of Big Data*, *10*(1), 42.

Howard, P., Bhiwandiwalla, A., Fraser, K. C., & Kiritchenko, S. (2024). Uncovering bias in large vision-language models with counterfactuals. arXiv preprint arXiv:2404.00166.

Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., & Wang, Y. (2024). Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11375–11385). Conference paper.

Huang, W., Li, C., Zhou, H.-Y., Yang, H., Liu, J., Liang, Y., et al. (2024). Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nature Communications*, *15*(1), 7620.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., et al. (2019). CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth AAAI symposium on educational advances in artificial intelligence* (pp. 73–80). Conference paper.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., & Thoma, G. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, *4*(6), 475.

Jeong, D. P., Garg, S., Lipton, Z. C., & Oberst, M. (2024). Medical adaptation of large language and vision-language models: Are we making progress? arXiv preprint arXiv:2411.04118.

Jin, Y., Lu, H., Li, Z., & Wang, Y. (2023). A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images. *Multimedia Tools and Applications*, *82*(21), 33421–33442.

Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. *IEEE Access*, *7*, 61656–61669.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317.

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III clinical database (version 1.4) [dataset]. PhysioNet. RRID:SCR_007345.

Johnson, A. E. W., Pollard, T. J., Mark, R. G., Berkowitz, S. J., & Horng, S. (2024). MIMIC-CXR database (version 2.0.0) [dataset]. PhysioNet..

Jung, H., Jang, T., & Wang, X. (2024). A unified debiasing approach for vision-language models across modalities and tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *vol. 37*, *Advances in neural information processing systems* (pp. 21034–21058). Curran Associates, Inc..

Kermany, D., Zhang, K., & Goldbaum, M. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data, V2.

Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., et al. (2024). CARZero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11137–11146). Conference paper.

Liu, B., Lu, D., Wei, D., Wu, X., Wang, Y., Zhang, Y., et al. (2023). Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, *42*(12), 3579–3589.

Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., & Jurafsky, D. (2021). RadNLI: A natural language inference dataset for the radiology domain (version 1.0.0) [dataset]. PhysioNet..

Moeller, L., Tilli, P., Vu, T., & Padó, S. (2025). Explaining vision-language similarities in dual encoders with feature-pair attributions.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *International Journal of Surgery*, *8*(5), 336–341.

Nafees, W. (2020). *Edema MIMIC reports dataset [dataset]*. Kaggle.

Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., et al. (2024). The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12988–12997). Conference paper.

Phan, V. M. H., Xie, Y., Qi, Y., Liu, L., Liu, L., Zhang, B., et al. (2024). Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11492–11501). Conference paper.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43.

Rahman, U., Imam, R., Yaqub, M., Amor, B. B., & Mahapatra, D. (2025). Can language-guided unsupervised adaptation improve medical image classification using unpaired images and texts? In *Proceedings of the 2025 IEEE 22nd international symposium on biomedical imaging* (pp. 1–5). Conference paper.

Ruggeri, G., & Nozza, D. (2023). A multi-dimensional study on bias in vision-language models. In J. Rogers, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL, 2023* (pp. 6445–6455). Association for Computational Linguistics.

Shivade, C. (2019). MedNLI - a natural language inference dataset for the clinical domain (version 1.0.0) [dataset]. PhysioNet.

Siddiqui, A. A., Tirunagari, S., Zia, T., & Windridge, D. (2025). A latent diffusion approach to visual attribution in medical imaging. *Scientific Reports*, *15*(1), 962.

Van, M.-H., Verma, P., & Wu, X. (2024). On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM conference on connected health: applications, systems and engineering technologies*. 172–176.

Varma, M., Delbrouck, J.-B., Hooper, S., Chaudhari, A., & Langlotz, C. (2023). ViLLA: Fine-grained vision-language representation learning from real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22225–22235). Conference paper.

Wang, M., Lin, T., Lin, A., Yu, K., Peng, Y., Wang, L., et al. (2024). Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. arXiv preprint.

Wang, Y., Yu, Z., Wang, J., Heng, Q., Chen, H., Ye, W., et al. (2024). Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, *132*(1), 224–237.

Worldometer (2025). Worldometer – real time world statistics. https://www.worldometers.info/. (Accessed 27 November 2025).

Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). MedKLIP: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 21372–21383). Conference paper.

Xie, Y., Chen, Q., Wang, S., To, M.-S., Lee, I., Khoo, E. W., et al. (2024). PairAug: What can augmented image-text pairs do for radiology? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11652–11661). Conference paper.

Yang, Y., Lee, C. P., Feng, S., Zhao, D., Wen, B., Liu, A. Z., et al. (2025). Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations. arXiv preprint.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E. K., et al. (2023). Cxr-clip: Toward large scale chest x-ray language-image pre-training. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, R. Taylor (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2023* (pp. 101–111). Cham. Springer Nature Switzerland.

Yu, X., Zhang, L., Wu, Z., & Zhu, D. (2025). Core–periphery multi-modality feature alignment for zero-shot medical image analysis. *IEEE Transactions on Medical Imaging*, *44*(10), 3973–3983.

Zawacki, A., Wu, C., Shih, G., Elliott, J., Fomitchev, M., Hussain, M., et al. (2019). SIIM-ACR pneumothorax segmentation 2019 [dataset].

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024a). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625–5644.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024b). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625–5644.

Zhang, X., Wu, C., Zhang, Y., Xie, W., & Wang, Y. (2023). Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, *14*(4542).

Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., et al. (2024). A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, *30*(11), 3129–3141.

Zhao, A. (2020). *COVIDx CXR-2 [dataset]*. Kaggle.

Zhong, X., Batmanghelich, K., & Sun, L. (2024). Enhancing biomedical multi-modal representation learning with multi-scale pre-training and perturbed report discrimination. In *2024 IEEE conference on artificial intelligence* (pp. 480–485). Conference paper.

Zhu, H., Wei, Y., Liang, X., Zhang, C., & Zhao, Y. (2023). CTP:towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22257–22267). Conference paper.