

IRUD-IA: Vision-language models in the medical domain

Karen López-Linares^{1,2,*}, Montse Cuadros¹, Alexander Platas^{1,3}, Marcos Merino Prado¹, Elena Zotova¹, Paola Martínez-Arias¹, Miriam Gutiérrez¹, Eduardo Nieto¹, María Jesús García-González^{1,2} and Andoni Beristain^{1,2}

¹Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain

²eHealth group, Biogipuzkoa Health Research Institute, Spain

³Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU), Spain

Abstract

Vision-language models (VLMs) have recently emerged as a powerful paradigm for jointly processing visual and textual information. However, their direct application to the medical domain remains challenging due to domain-specific language, limited annotated datasets, and the need for clinically reliable validation. This paper presents IRUD-IA, an industrial research project focused on adapting VLMs to medical imaging and clinical text. The project studies multimodal representation learning and efficient adaptation strategies, with particular attention to Spanish clinical language. Several application scenarios are explored, including zero-shot chest X-ray classification, text-based image retrieval, radiology report generation, and text-guided medical image synthesis. Experiments across multiple public datasets highlight both the potential and key limitations of VLMs in clinical contexts, including sensitivity to domain shifts, prompt design, and evaluation challenges. These findings emphasize the need for domain-aware validation strategies when deploying multimodal models in medical environments.

Keywords

Vision-language models, Medical imaging, multimodal learning, radiology report generation

1. Introduction

Recent advances in large-scale multimodal models have significantly reshaped Artificial Intelligence (AI), enabling systems capable of jointly processing visual and textual information. Vision-Language Models (VLMs), initially developed for natural images and general-domain English corpora, have demonstrated strong performance in tasks such as image-text retrieval, zero-shot classification, visual question answering, and automatic report generation. However, their adaptation to the medical domain remains limited due to domain-specific language, data scarcity, and modality differences, which introduce significant scientific and technological challenges.

Transferring VLMs to the medical domain requires addressing several domain-specific constraints. Clinical language differs substantially from general-domain corpora and often involves specialized terminology, implicit reasoning patterns, and multilingual settings such as Spanish clinical documentation. At the same time, medical images present modality-specific properties that diverge significantly from the natural images typically used for pretraining. Bridging these two modalities in a clinically meaningful way demands careful adaptation strategies and domain-aware representation learning.

Within this context, IRUD-IA investigates VLM approaches tailored to medical imaging, focusing on multimodal representation learning, cross-modal alignment, and efficient adaptation of pretrained

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

✉ klopez@vicomtech.org (K. López-Linares); mcuadros@vicomtech.org (M. Cuadros); aplatas@vicomtech.org (A. Platas); mmerino@vicomtech.org (M. M. Prado); ezotova@vicomtech.org (E. Zotova); pnmartinez@vicomtech.org (P. Martínez-Arias); mgutierrezf@vicomtech.org (M. Gutiérrez); enietoY@vicomtech.org (E. Nieto); mjgarcia@vicomtech.org (M. J. García-González); aberistain@vicomtech.org (A. Beristain)

0000-0002-4800-6052 (K. López-Linares); 0000-0002-3620-1053 (M. Cuadros); 0009-0002-5501-017X (A. Platas); 0009-0004-9610-322X (M. M. Prado); 0000-0002-8350-1331 (E. Zotova); 0000-0002-5952-1578 (P. Martínez-Arias); 0000-0002-6692-9934 (M. Gutiérrez); 0000-0001-6357-1327 (M. J. García-González); 0000-0002-5452-2141 (A. Beristain)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

models to clinical data. The project explores methods for jointly encoding medical images and associated clinical text, leveraging both large-scale unlabeled resources and domain-specific datasets. Particular emphasis is placed on Spanish and multilingual clinical language modeling, few-shot and zero-shot classification, multimodal retrieval, and automatic report generation.

The remainder of this paper is organised as follows. Section 2 presents the background and scientific context of the project. Section 3 describes the consortium and funding agency. Section 4 outlines the main goals and expected results. Section 5 introduces the application scenarios addressed and Section 6 discusses the main scientific and technological challenges. Finally, Section 7 summarises the main conclusions and future work.

2. Background

VLMs aim to learn joint representations of visual and textual data, enabling cross-modal reasoning and transfer across tasks. The field was significantly advanced by large-scale contrastive pretraining approaches such as CLIP [1], which demonstrated that aligning image and text embeddings using web-scale image-caption pairs yields highly transferable representations and enables zero-shot classification. Transformer-based architectures further expanded this paradigm through dual-encoder designs, where visual and textual encoders are trained to map inputs into a shared embedding space, and fusion-based models, where cross-attention mechanisms allow deeper multimodal interaction [2, 3].

In the biomedical and clinical setting, adapting these paradigms presents unique challenges related to domain specificity, specialized terminology, dataset scarcity, and clinical validity. Several domain-adapted VLMs have been proposed in recent years. Early approaches such as ConVIRT [4], CheXZero [5], GLORIA [6], and MedCLIP [7] explored contrastive alignment between radiological images and reports, demonstrating the effectiveness of cross-modal pretraining in medical tasks. More recent biomedical foundation models such as BioMedCLIP extend CLIP-style training to large-scale biomedical image-text corpora, improving generalization across modalities and datasets. Recent reviews of VLMs for medical imaging highlight the rapid evolution of this area, moving from modality-specific multimodal systems toward scalable domain-adapted foundation models, while identifying persistent challenges in data heterogeneity, evaluation standardization, and multimodal generalization [8].

Recent work has extended VLMs toward large multimodal foundation models tailored to healthcare. MedGemma [9] is a family of medical VLMs based on the Gemma architecture, pretrained and fine-tuned on diverse clinical image and text corpora to support classification, multimodal reasoning, and report generation. Similarly, Med-Gemini [10] extends multimodal large language models to clinical contexts, incorporating medical reasoning capabilities across radiology, pathology, and genomics. Other recent systems such as LLaVA-Med [11], MiniGPT-Med [12], and RadFM [13] adapt general-purpose multimodal Large Language Models (LLM) frameworks to medical imaging through instruction tuning, parameter-efficient fine-tuning, and domain-specific alignment strategies. These models mark a shift from contrastive dual-encoder approaches toward unified multimodal reasoning systems that integrate visual encoders with large language models, enabling more complex clinical interaction and report generation capabilities.

In parallel, text-guided medical image generation has emerged as a complementary research direction. Diffusion-based models such as Stable Diffusion [14] have been adapted to the medical domain to generate synthetic images conditioned on clinical text prompts, enabling data augmentation, rare pathology simulation, and privacy-preserving dataset construction. These approaches leverage multimodal embeddings learned by VLMs or related architectures to align textual descriptions with anatomically plausible image synthesis, further reinforcing the role of cross-modal representation learning in medical AI.

Despite these advances, empirical studies show that medical VLMs remain vulnerable to clinically realistic perturbations, such as acquisition artefacts and image quality degradation, raising concerns about robustness and safe deployment in real-world environments [15]. Comprehensive surveys further emphasize the need for systematic evaluation, domain-aware pretraining strategies, and rigorous validation protocols to ensure reliability and clinical applicability [8].

3. Consortium and funding agency

IRUD-IA is a strategic industrial research project led by Ulma Medical Technologies and developed by a consortium of companies and research institutions within the Basque Science, Technology and Innovation Network. The consortium combines industrial partners and research institutions specialized in AI-based medical technologies with research organizations including Vicomtech, University of Deusto and Biogipuzkoa Health Research Institute.

The project is funded under the HAZITEK 2024 programme of the Basque Government (grant number ZE-2024/00030), which supports strategic industrial R&D initiatives to strengthen the regional ecosystem in AI-based medical technologies.

4. Goals and expected results

The main objective of IRUD-IA is to advance the development of domain-adapted VLMs for medical imaging. The project investigates multimodal models capable of jointly encoding radiological images and associated clinical text, enabling cross-modal reasoning in healthcare environments.

Particular attention is given to multilingual clinical language, especially Spanish, and to efficient adaptation techniques such as parameter-efficient fine-tuning and few-shot learning. The project explores multiple clinical tasks and evaluates model performance across public medical datasets. Figure 1 presents an overview of the project framework and the main application scenarios addressed.

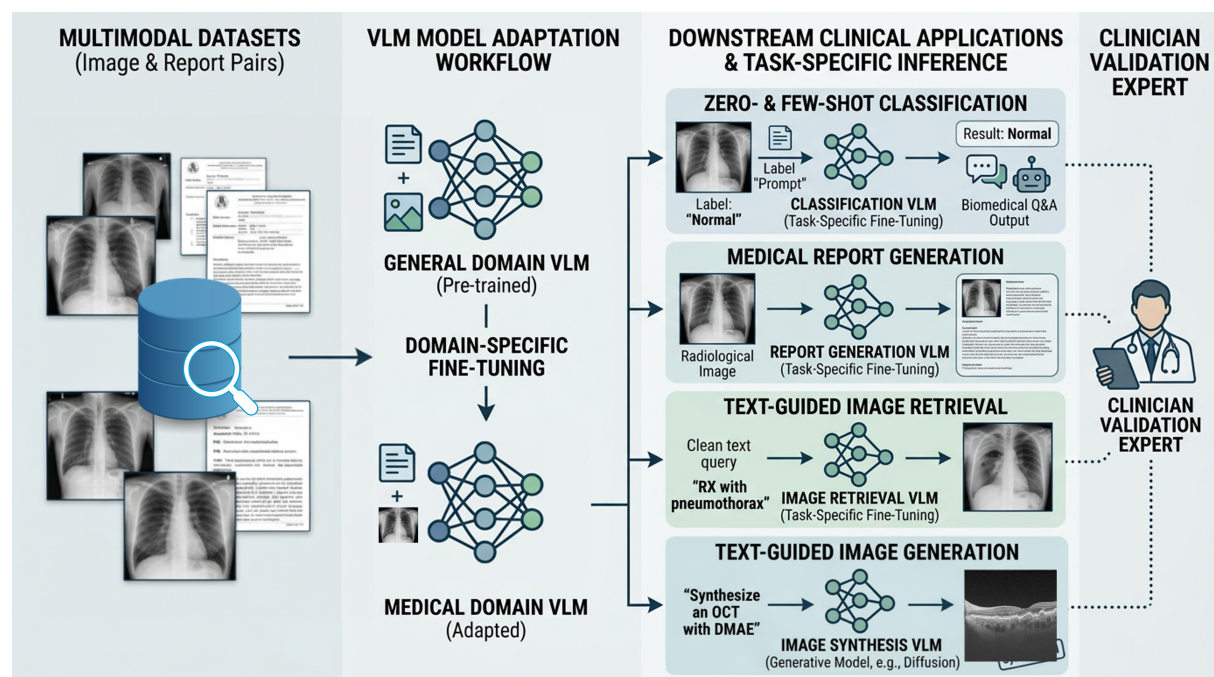


Figure 1: Overview of medical vision-language model (VLM) adaptation and clinical applications. Multimodal datasets (image-report pairs) are used to fine-tune a general-domain VLM into a medical-domain model, which is then applied to downstream clinical tasks including zero-/few-shot classification, medical report generation, text-guided image retrieval, and text-guided image generation, with clinician validation ensuring reliability.

5. Applications

5.1. Zero-shot classification

A first application scenario addressed in IRUD-IA is the use of VLMs for zero-shot classification of chest X-ray images, enabling the identification of radiological findings without task-specific supervised training. This paradigm is particularly relevant in medical imaging, where the availability of labeled data is limited and expert annotation is costly.

The project builds upon CheXzero, a widely used VLM for chest X-ray interpretation. CheXzero follows a CLIP-style architecture and is trained using a large collection of chest radiographs paired with radiology reports. This training strategy allows the model to perform classification of new unseen data by measuring the similarity between an input image and textual prompts describing the candidate pathologies.

To better understand its real-world behaviour, we conducted an empirical evaluation across several public datasets, including NIH-Chest X-ray14 [16], Padchest [17], CheXpert [18] and IU-XRay [19]. The analysis showed that, although CheXzero is often presented as a general zero-shot classifier, its performance is uneven across pathologies. The model performed reasonably well on findings such as atelectasis, cardiomegaly, consolidation, edema and pleural effusion, as it was originally adjusted to optimize their detection, typically achieving AUC values around 0.7. However, performance decreased when attempting to detect other pathologies, highlighting limitations of zero-shot generalization. In addition, we observed a drop in performance when evaluating on external datasets such as Padchest and X-ray 14, suggesting sensitivity to differences in data acquisition conditions.

Another aspect explored in the project was the impact of prompt design on zero-shot inference performance. Since CheXzero relies on textual descriptions to represent diagnostic hypotheses (e.g., “pneumonia: yes”), the wording and level of detail of the prompts can influence model predictions. We therefore systematically evaluated prompts with varying levels of complexity, ranging from simple pathology labels to more descriptive clinical statements. Our analysis showed that prompts containing richer clinical descriptions, closer to the language used in radiology reports, tend to produce more reliable predictions, improving performance by up to 10% compared to simpler prompts based only on pathology labels, demonstrating the importance of prompt design.

Based on these findings, the project also explored strategies to adapt the model to new individual datasets. Parameter-efficient approaches such as Low-Rank Adaptation (LoRA) [20] and continuous pre-training were investigated. While these adaptations can provide some benefits for certain classes, they do not substantially overcome the limitations observed in the zero-shot setting. Moreover, they sometimes lead to reduced performance in pathologies where the model was already well optimized and increased sensitivity to prompt variations, which may further affect performance.

As a final step, the project plans to perform continuous pre-training using all the datasets and extend the approach toward multilingual clinical settings, focusing on the adaptation of the model to Spanish clinical text.

5.2. Text-based image retrieval

In this task, we built upon the same VLM used in the previous application, CheXzero, illustrating how a single pretrained model can support multiple downstream applications within a unified framework. Since CheXzero follows a dual-encoder architecture, with independent encoders for images and text, it naturally supports cross-modal retrieval. Both modalities are mapped into a shared embedding space where semantic similarity can be computed using vector similarity measures.

This design enables an efficient retrieval pipeline. Image embeddings can be precomputed and stored in advance, allowing the system to perform retrieval by comparing a query embedding against a pre-indexed vector database rather than recomputing representations for the entire dataset at query time. This significantly reduces computational cost and latency during inference.

To support scalable retrieval, we deployed a vector database based on Qdrant, which enables efficient storage and Approximate Nearest Neighbor search over high-dimensional embeddings. Qdrant relies on

the Hierarchical Navigable Small World algorithm, providing a good balance between search accuracy and computational efficiency. In addition, it supports features such as structured filtering and dynamic data updates, which are relevant for real-world clinical environments.

Within this framework, embeddings were computed for the chest X-ray images from the datasets used in the previous application scenario and indexed in Qdrant. At query time, a textual description is encoded by the VLM to generate a query embedding, and the system retrieves the most similar images according to cosine similarity in the shared multimodal space.

Overall, this approach demonstrates how multimodal foundation models can support multiple clinical applications using the same pretrained representations, while the retrieval infrastructure provides a basis for future extensions such as multilingual queries or integration with hospital imaging repositories.

5.3. Radiology report generation

Another task explored in IRUD-IA is the automatic generation of radiology reports from medical images. This functionality aims to assist radiologists in producing textual descriptions of imaging findings, improving reporting efficiency and consistency. The task is challenging because models must accurately detect visual patterns and express them using precise clinical language. The difficulty is even greater in Spanish due to the limited availability of annotated datasets, pretrained models, and evaluation resources.

Two datasets were used for the experiments: (1) PadChest-GR [21], which provides chest X-ray images paired with reports written in Spanish; and (2) MIMIC-CXR-JPG [22], originally released with reports in English. For this study, 11.4% of the dataset was automatically translated into Spanish using OPUS-MT [23], yielding 26,000 image-report pairs.

Two main system architectures were evaluated:

Encoder-decoder architectures with visual encoders such as pretrained Vision Transformers (ViT-base) [24] and language decoders such as Medical-mT5 [25]. The models were trained end-to-end using teacher forcing and standard cross-entropy.

LLMs with (i) In-Context Learning (ICL), including zero-shot and few-shot prompting, and (ii) Supervised Fine-Tuning (SFT). We evaluated recent models such as MedGemma 4B/27B [26] and Claude 4.5 Sonnet [27]. Two generation strategies were compared: generating the report directly from the image, or complementing the image with labels summarizing the main radiological findings to guide generation. During fine-tuning, the generation format was modified so that the model first produces a header with detected findings before generating the report, improving alignment between visual input and generated text. LoRA-based fine-tuning was used to efficiently adapt the models to the medical domain.

The generated reports were evaluated using a combination of automatic and model-based methods (see Section 5.5). Results show significant differences depending on the evaluation metric. While encoder-decoder architectures remain competitive in lexical overlap metrics, their clinical quality is lower according to LLM-based evaluation. In contrast, multimodal LLMs show stronger semantic alignment and clinical plausibility.

The best overall performance is achieved by the LoRA fine-tuned MedGemma 4B model, which surpasses large-scale models like Claude 3.5 Sonnet and MedGemma 27B (in ICL settings) by approximately 10 points in the LLM-as-a-judge evaluation. These findings suggest that, for radiology report generation in Spanish under limited data conditions, parameter-efficient fine-tuning of multimodal LLMs is significantly more effective than relying on the few-shot capabilities of much larger models.

5.4. Text-guided image generation

Another application scenario explored in IRUD-IA is the generation of synthetic medical images from textual descriptions. This task highlights the potential of multimodal models to connect natural language and visual medical data, enabling the creation of images conditioned on clinical descriptions. Such

functionality can help mitigate the scarcity of annotated data for certain imaging modalities or rare pathologies.

To address this problem, we employed Stable Diffusion, a generative model capable of producing images from natural language prompts. Although Stable Diffusion is not strictly a VLM, it relies on a text encoder to condition the image generation process, allowing clinical descriptions to directly guide the synthesis of visual patterns.

Since the model was originally trained on large collections of natural images, its direct application to the medical domain is limited. To specialize the model for medical imaging modalities such as chest X-rays and optical coherence tomography (OCT), we adapted it using LoRA and image-text pairs from medical datasets, including both healthy cases and images containing different pathologies. Through this process, the model learns the relationship between clinical descriptions and the corresponding visual patterns. Figure 2 shows examples of the generation process.

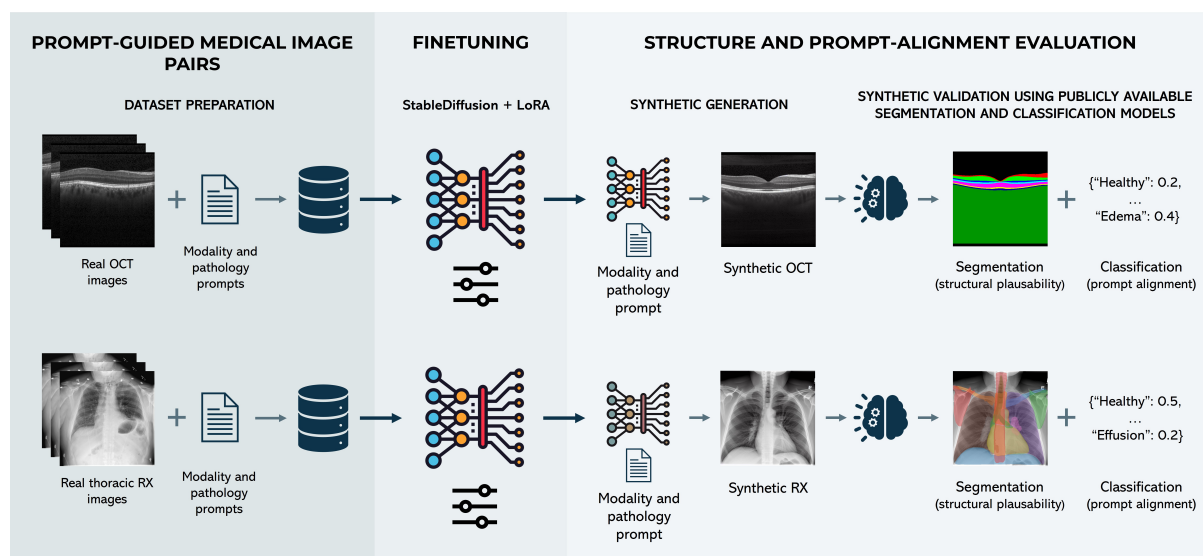


Figure 2: Overview of text-guided image generation.

The validity of the generated images was assessed using standard evaluation metrics commonly employed in generative image modeling, such as Fréchet Inception Distance (FID) or Structural Similarity Index Measure (SSIM). However, these metrics originate from the natural image domain and do not always provide reliable assessments for medical images. This limitation highlights the need for alternative validation strategies tailored to the medical domain, which are discussed in more detail in Section 5.5.

5.5. Validation methodologies

Evaluating generative models in the medical domain presents important challenges, as most widely used evaluation metrics originate from other domains such as natural images or general text generation. Our experiments on both medical image synthesis and radiology report generation show that standard metrics often fail to reflect the clinically relevant behavior expected from these systems.

For image generation, commonly used metrics such as FID or SSIM were initially considered. However, these metrics do not evaluate whether generated images correctly reflect the anatomical structures or pathological findings described in the conditioning text. To address this limitation, additional validation strategies incorporating clinical context were explored. These include auxiliary classifiers to verify whether generated images correspond to the pathologies specified in the prompts, segmentation models to validate anatomical plausibility, and modality-specific checks to ensure that image intensity distributions remain consistent with the expected acquisition modality.

For radiology report generation, the evaluation combined lexical and semantic metrics together with

model-based assessment methods. Specifically, we reported BLEU, ROUGE-L, METEOR, BERTScore, SemScore and LLM-as-a-Judge.

Lexical metrics measure N-gram overlap between generated and reference reports but are limited by the large terminological variability of clinical language. As shown in Example 1, a generated report may be clinically accurate yet receive a lexical score of zero due to the use of synonyms.

Example 1. Terminological variation

Reference: The lung parenchyma shows adequate aeration, with no areas of alveolar opacification.

Prediction: No pulmonary consolidations are identified.

Scores: Lexical: 0% | Semantic: 76% | Judge: 80% (Expected: High)

Embedding-based metrics partially alleviate this issue but may still fail to capture clinical correctness, as statements and their negations can share similar semantic representations. Example 2 illustrates this failure to distinguish a statement from its negation.

Example 2. Diagnostic contradiction

Reference: A pulmonary nodule is observed in the right lobe.

Prediction: No pulmonary nodules are observed.

Scores: Lexical: 31% | Semantic: 85% | Judge: 0% (Expected: Low)

LLM-based evaluation provides a qualitative assessment of coherence and clinical plausibility, although it is susceptible to verbosity and stylistic biases of the evaluator model. As illustrated in Example 3, an LLM may assign a high score to a report that "sounds" professional and detailed, even when the model has hallucinated clinical findings not present in the reference.

Example 3. Hallucinated details

Reference: Simple radial fracture.

Prediction: Transverse fracture of the distal radius with slight dorsal angulation and no articular involvement.

Scores: Lexical: 22% | Semantic: 81% | Judge: 80% (Expected: Mid)

These observations highlight the need for domain-aware validation strategies when evaluating generative models in medical imaging.

6. Conclusions

This paper presented IRUD-IA, a research project focused on adapting VLMs to medical imaging scenarios. The project explored the application of pretrained multimodal models to several clinical tasks, including zero-shot chest X-ray classification, text-based image retrieval, radiology report generation, and text-guided medical image synthesis.

The experimental results highlight several key observations. First, although VLMs enable flexible zero-shot inference, their performance is uneven across pathologies and highly sensitive to dataset characteristics and prompt design. Second, in generative tasks, multimodal LLMs show strong semantic capabilities, but ensuring clinical correctness remains challenging, particularly in the presence of hallucinations or contradictory statements. Third, across both image and text generation tasks, standard evaluation metrics fail to fully capture clinically relevant aspects, revealing important limitations when transferring methodologies from natural image and general NLP domains.

Overall, these findings confirm that adapting VLMs to the medical domain requires not only model fine-tuning but also domain-aware evaluation and validation strategies. Future work within the project will focus on improving robustness to data variability and extending multimodal models to multilingual clinical settings, particularly Spanish.

Acknowledgements

This work is funded by the Basque Government under the HAZITEK 2024 programme (grant number ZE- 2024/00030). We acknowledge the work done by Ulma Medical technologies, Deusto Seidor, Serikat Consultoria e Informatica, Cyber Surgery, Global Data Quantum, Data Value Management, A3z Advanced Analytical Consulting Services, University of Deusto and Biogipuzkoa Health Research Institute.

Declaration on Generative AI

Some images in this paper were generated using generative AI tools. These tools were used only for image creation and did not affect the research findings, analysis, or conclusions.

References

- [1] A. Radford, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021.
- [2] J. Lu, et al., Vilbert: Pretraining task-agnostic visiolinguistic representations, in: NeurIPS, 2019.
- [3] L. H. Li, et al., Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).
- [4] Y. Zhang, et al., Contrastive learning of medical visual representations from paired images and text, in: MLHC, 2022.
- [5] E. Tiu, et al., Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning, Nature Biomedical Engineering (2022).
- [6] S.-C. Huang, et al., Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: ICCV, 2021.
- [7] Z. Wang, et al., Medclip: Contrastive learning from unpaired medical images and text, arXiv preprint arXiv:2210.10163 (2022).
- [8] J. S. Ryu, et al., Vision-language foundation models for medical imaging: a review of current practices and innovations, Biomedical Engineering Letters (2025). doi:10.1007/s13534-025-00484-6.
- [9] G. Research, Medgemma: A family of medical vision-language foundation models, arXiv preprint (2024).
- [10] K. Saab, et al., Med-gemini: A multimodal foundation model for medicine, arXiv preprint (2024).
- [11] Li, et al., Llava-med: Training a large language-and-vision assistant for biomedicine, arXiv preprint (2023).
- [12] Wang, et al., Minigpt-med: Large language models for medical vision-language tasks, arXiv preprint (2023).
- [13] Wu, et al., Radfm: A large vision-language foundation model for radiology, arXiv preprint (2023).
- [14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
- [15] Z. Cheng, et al., Understanding the robustness of vision-language models to medical image artefacts, npj Digital Medicine (2025). doi:10.1038/s41746-025-02108-w.
- [16] X. Wang, et al., Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.
- [17] A. Bustos, et al., Padchest: A large chest x-ray image dataset with multi-label annotated reports, Medical image analysis 66 (2020) 101797.
- [18] P. Chambon, et al., Chexpert plus: Augmenting a large chest x-ray dataset with text radiology

reports, patient demographics and additional image formats, arXiv preprint arXiv:2405.19538 (2024).

- [19] D. Demner-Fushman, et al., Preparing a collection of radiology examinations for distribution and retrieval, *Journal of the American Medical Informatics Association* 23 (2016) 304–310.
- [20] E. J. Hu, et al., Lora: Low-rank adaptation of large language models, *ICLR* (2022).
- [21] D. C. de Castro, et al., Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation, *NEJM AI* (2025).
- [22] A. Johnson, et al., MIMIC-CXR-JPG - chest radiographs with structured labels, *PhysioNet* (2024). doi:10.13026/jsn5-t979.
- [23] J. Tiedemann, et al., Opus-mt – building open translation services for the world, in: *EAMT*, 2020.
- [24] B. Wu, et al., Visual transformers: Token-based image representation and processing for computer vision, 2020. arXiv:2006.03677.
- [25] I. García-Ferrero, et al., Medmt5: An open-source multilingual text-to-text llm for the medical domain, in: *LREC-COLING*, 2024.
- [26] A. Sellergren, et al., Medgemma technical report, 2025. arXiv:2507.05201.
- [27] Anthropic, Claude 4.5 Sonnet System Card, Technical Report, 2025.