OPEN ACCESS

Artificial Intelligence ARTICLE

The Innovation Informatics

Bridging medical imaging and reports: Learning radiologist's nuances via fine-grained multi-modal alignment

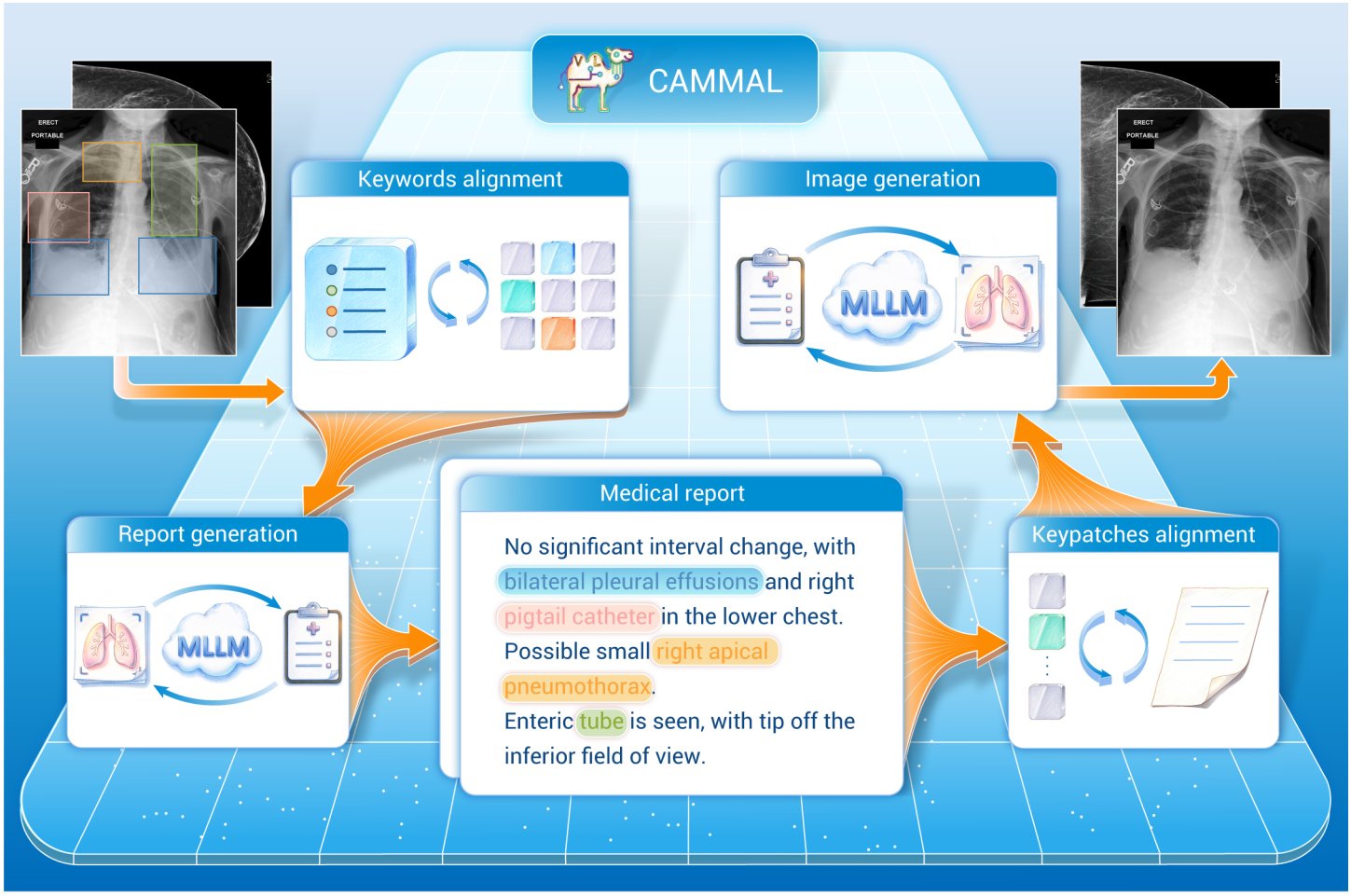
Wenting Chen,[1,2,9] Hui Ren,[2,9] Yujin Oh,[2] Yihan Cao,[2] Elshaimaa Sharaf,[2] Jiebo Luo,[3] Hong-Yu Zhou,[4] Wenxuan Wang,[8] Lichao Sun,[5] Tianming Liu,[6] Linlin Shen,[7] Quanzheng Li,[2] Xiang Li,[2,*] and Yixuan Yuan[8,*]

*Correspondence: xli60@mgh.harvard.edu (X.L.); yxyuan@ee.cuhk.edu.hk (Y.Y.)

Received: March 12, 2026; Accepted: October 31, 2026; Published Online: October 31, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100069

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- CAMMAL links specific areas in medical images to words in radiology reports.
- It generates reports from chest X-rays and creates synthetic medical images from text.
- It outperformed comparison models in describing mammograms and generating chest X-ray reports..
- Correct chest X-ray and report matches ranked in the top 10 in over 94% of searches.

ARTICLE Artificial Intelligence

OPEN ACCESS

Bridging medical imaging and reports: Learning radiologist's nuances via fine-grained multi-modal alignment

Wenting Chen,[1,2,9] Hui Ren,[2,9] Yujin Oh,[2] Yihan Cao,[2] Elshaimaa Sharaf,[2] Jiebo Luo,[3] Hong-Yu Zhou,[4] Wenxuan Wang,[8] Lichao Sun,[5] Tianming Liu,[6] Linlin Shen,[7] Quanzheng Li,[2] Xiang Li,[2,*] and Yixuan Yuan[8,*]

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR 999077, China
[2]Department of Radiology, Massachusetts General Hospital, Boston 02114 MA, USA
[3]Department of Computer Science, University of Rochester, Rochester 14627 NY, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston 02115 MA, USA
[5]School of Computer Science, Lehigh University, Bethlehem 18015 PA, USA
[6]School of Computing, University of Georgia, Athens 30602 GA, USA
[7]Computer Vision Institute, Shenzhen University, Shenzhen 518000, China
[8]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China
[9]These authors contributed equally
*Correspondence: xli60@mgh.harvard.edu (X.L.); yxyuan@ee.cuhk.edu.hk (Y.Y.)

Received: March 12, 2026; Accepted: October 31, 2026; Published Online: October 31, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100069

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Chen W., Ren H., Oh Y., et al. (2026). Bridging medical imaging and reports: Learning radiologist's nuances via fine-grained multi-modal alignment. *The Innovation Informatics* **2**:100069.

Precise and explainable alignment between data from different data modalities is crucial for advancing artificial general intelligence in medicine. In this work, we present CAMMAL (Cyclic Adaptive Medical Modality ALignment), a framework that can achieve fine-grained vision-language alignment through two key innovations, including an Adaptive Patch-Word Matching (AdaMatch) mechanism that dynamically correlates regions in medical images with specific words in radiology reports, and a bidirectional generative architecture that leverages the alignment between textual and visual codebooks to guide the translation between modalities within a single model. Evaluation of CAMMAL on the chest X-rays (combined MIMIC-CXR and OpenI datasets) and mammography (EMBED dataset) demonstrates its superior performance across multiple metrics over other methods. Human reader studies by radiologists validate the clinical effectiveness of the generated reports and synthetic images, particularly in capturing anatomical structures (73% rated good/excellent) and meaningful findings (56% rated very good/excellent). By enabling systematic capture of inter-modality medical data correspondence and fluid multi-way translation between modalities, CAMMAL could advance the development of interpretable and clinically reliable AI systems by bridging the gap between medical images and textual descriptions. Its ability to accurately retrieve, generate, and align reports with imaging data offers potential improvements in learning detailed radiologists' knowledge, which could be vital for AI-assisted diagnosis and medical education. The bidirectional generative approach further enhances model transparency and explainability, fostering trust for AI-driven healthcare applications. As multimodal medical foundation models continue to evolve, CAMMAL provides a foundation for integrating vision and language understanding, with the potential for broader clinical applications beyond radiology.

INTRODUCTION

Recent advances in artificial general intelligence have led to the emergence of "Foundation Models",[1–4] which demonstrate unprecedented capabilities in understanding, processing, and generating complex data patterns. These models, trained on vast datasets using self-supervised learning approaches with scaled-up parameter size, have redefined the boundaries of artificial intelligence by exhibiting emergent capabilities and cross-task generalization.[2,4] In healthcare, medical foundation models have shown remarkable potential in handling medical data with enhanced robustness, depth, and generalizability, potentially disrupting how medical data analysis will be performed.[5–14] These models have achieved significant breakthroughs in processing individual medical data modalities, from analyzing complex imaging data in pathology,[5,13] echocardiogram,[15] oncology,[14] to the understanding of clinical narratives,[16] often matching or exceeding human expert performance.

With the development of single-modality medical foundation models, there has been growing interest in integrated approaches that can link multiple medical data modalities, i.e., "multi-modal foundation models".[15,17–28] The most important and challenging feature of a multi-modal foundation model is the capability to establish meaningful and precise alignments between different modalities. In the medical domain, cross-modality (e.g., between imaging and language) alignment is the basis for various tasks, including image-report generation, data synthesis, visual question answering, etc. Another important feature derived from an effective imaging-language alignment is the potential capability of foundation models to trace and explain model decisions. It is vital to comprehend semantic relationships across different data types and perform the proper mapping, especially as such relationships can be both overlapping and complementary yet could contain unequal amount of information when describing the same semantics, as investigated in our previous work.[29–32] Human experts interpret multimodal data by relying on well-established, high-level mappings among different information sources, rather than merely merging them. As noted by Kline et al., [33] clinical decisions are based on the interplay of "hallmarks" derived from each information source. In contrast to humans, for an AI algorithm, without careful structuring of the alignment, multi-modal models have the risk of learning superficial "shortcuts" due to misinterpreting the relationship between training data pairs. For example, heart failure patients commonly use pacemakers and frequently develop conditions like cardiogenic pulmonary edema or chronic pulmonary congestion. Consequently, models that rely on co-occurrence statistics between the features in two modalities might incorrectly assume that patients with a pacemaker (image) would have edema or pulmonary congestion (text), simply using the presence of a pacemaker as a shortcut rather than inferring from the actual evidence of the disease. On the other hand, learning such alignment is non-trivial, especially for the medical domain, where subtle features in one modality—such as a small opacity in a chest radiograph—may represent critical diagnostic information in another modality, like a key finding in the radiology report and its impression. To effectively represent and capture such intricate relationships, we need to recognize that cross-modality alignment can occur at multiple scales and granularities. While traditional approaches have primarily focused on learning the sample-level alignment,[1,34–37] e.g., entire images are mapped to the whole reports, such a coarse-grained approach fails to capture the nuanced correspondences between specific image regions and detailed textual descriptions that are crucial for clinical applications.

Among methods that extend beyond sample-level alignment, local visual representations are commonly obtained from regions proposed by pre-trained object detectors[38–41] or from fixed-grid image patches.[42–45] The former depends on detector quality and predefined object categories, whereas the latter cannot flexibly adapt to medical abnormalities with irregular sizes, shapes, and locations. Furthermore, existing medical vision-language models generally interpret local alignment using attention or gradient-based heatmaps,[46–51] which often provide diffuse, post-hoc localization rather than

explicit region–word correspondences and may require ground-truth reports. These limitations motivate a detector-independent adaptive patching mechanism and explicit cross-modal representations that can support interpretable alignment during inference.

To overcome such challenges and achieve precise alignment between different medical modalities, we premise that the following two conceptual design problems need to be addressed: (1) Fine-grained and semantically meaningful tokenization of each modality will be needed for the alignment. While text data naturally breaks down into tokens through established linguistic structures, other data modalities require more sophisticated, learnable, and adaptive tokenization schemes. Moreover, such tokenization schemes should capture the potential fine-grained manifestations with different sizes, shapes, and positions. With appropriate alignment to textual data, the tokens can then be explicitly interpreted. (2) In addition to the commonly adopted contrastive learning architecture such as those adopted in CLIP [1], fine-grained alignment could be further learned through generative modeling approaches. By training models to "translate" between modalities and generate observable outputs, we can transform abstract alignment relationships into quantifiable and verifiable results. Specifically, in the medical vision-language tasks, tokens from the imaging modality shall be able to translate to the text modality and vice versa. Thus, we will be able to evaluate the alignment based on the corresponding generation results in an explainable and explicit way.

In this work, we explore the feasibility of fine-grained, generative modeling-driven alignment by implementing the design through a large medical vision-language model task, introducing the Cyclic Adaptive Medical Modality ALignment (CAMMAL) framework that learns the alignment between medical images and their corresponding textual descriptions. CAMMAL leverages an adaptive image tokenization scheme and bidirectional generative modeling to establish and verify precise correspondences between image regions and textual words, enabling more accurate and interpretable development of multimodal medical foundation models. As shown in Figure 1B, for image tokenization, we introduce an Adaptive patch-word Matching (AdaMatch) scheme to adaptively correlate image patches with textual words. AdaMatch aims to extract image patches that include potential lesions with varying sizes and locations and align the patches with the corresponding textual words in the reports. In addition, we develop a cyclic generation framework for the alignment with a bidirectional large language model (LLM), as shown in Figure 1A. This model employs AdaMatch to identify keywords for images and "keypatches" for reports, which guide the bidirectional generation process. Given that potential language semantics for images could be diverse, we establish a pre-defined text codebook containing the most frequent entities from medical reports as a knowledge base for the alignment and generation. AdaMatch aligns the text codebook (instead of the raw text) with the adaptive patches, which are encoded by VQ-GAN,[52] to facilitate the task of image-to-text generation by a specialized LLM. Conversely, we also establish a "visual codebook" with commonly observed patches as "keypatches" and utilize AdaMatch to obtain the matching keypatches from the given text reports for the text-to-image generation task by the same LLM with VQ-GAN to decode the generated tokens back to images. The proposed CAMMAL framework was quantitatively and qualitatively evaluated on two modalities: chest X-ray (CXR) and mammography, with human reader studies performed by radiologists from Massachusetts General Hospital.

MATERIALS AND METHODS

This retrospective study utilized publicly available de-identified medical imaging datasets with institutional review board approval and waiver of informed consent. The study period for data analysis extended from January 2023 to December 2024.

Dataset selection and preprocessing

Three publicly available datasets were utilized. The MIMIC-CXR dataset (version 2.1.0; PhysioNet) comprised 473,057 de-identified chest radiograph-report pairs from 63,478 patients collected between 2011 and 2016. The dataset was partitioned into training (368,960 images from 59,320 patients), validation (2,991 images from 500 patients), and testing (5,159 images from 658 patients) subsets with no patient overlap. The OpenI dataset (Indiana University Chest X-ray Collection; Kaggle) containing 3,684 radiograph-report pairs from 2000-2013 provided external validation. All chest radiographs were resized to 256×256 pixels.

The EMBED dataset (version 1.0; Registry of Open Data on AWS) included 54,488 de-identified mammography studies collected between 2013 and 2020. Cranio-caudal view mammograms underwent maximum square cropping and resizing to 256×256 pixels. The dataset was stratified by patient into training (41,500 images), validation (5,530 images), development (2,023 images), and testing (5,435 images) subsets. Unlike the free-text CXR reports, EMBED provides structured annotations. For mammogram-to-text evaluation, eight attributes covering mass characteristics, calcification features, BI-RADS assessment, and breast density were converted into multiple-choice question–answer pairs. For text-to-mammogram synthesis, these attributes were converted into standardized natural-language statements, generating 2,608,175 mammogram–text pairs. The EMBED tasks evaluate structured attribute prediction and attribute-conditioned synthesis rather than unrestricted report generation. More details can be found in the supplementary materials. For both the CXR and mammography experiments, all baseline methods were retrained using the same dataset splits and preprocessing procedures as CAMMAL. All baseline results were obtained from our experiments rather than taken directly from the original publications.

CAMMAL framework architecture

The CAMMAL framework is a unified multimodal architecture that enables bidirectional generation and retrieval between medical images and radiology reports. The framework consists of three core components working in sequence: (1) a Vector Quantized Generative Adversarial Network (VQ-GAN) that tokenizes images into discrete latent codes, (2) an Adaptive patch-word Matching module (AdaMatch) that aligns image patches with textual tokens through fine-grained contrastive learning, and (3) a LLM that processes both visual and textual tokens for generation tasks. The pipeline operates as follows: medical images are first encoded into discrete token sequences via VQ-GAN, creating a unified representation space. AdaMatch then establishes semantic correspondences between adaptive image patches and report text using cross-modal alignment. The LLM integrates these aligned visual and textual tokens to perform four key tasks: image-to-report generation, and report-to-image synthesis. For image generation, the model outputs visual token sequences that are decoded back into images through VQ-GAN. The framework employs textual and visual codebooks constructed from frequent medical entities and representative image patches to enhance cross-modal understanding. Source code is available at https://github.com/CUHK-AIM-Group/CAMMAL.

Evaluation metrics

Image-to-report generation was assessed using BLEU (B-1 through B-4), METEOR, ROUGE-L, and an LLM-as-a-Judge metric. The LLM-as-a-Judge metric uses an LLM (Qwen3-32B) to extract clinical claims from each ground-truth report and computes the percentage recovered in the generated report. Report-to-image synthesis quality was evaluated using Fréchet Inception Distance (FID) and Naturalness Image Quality Evaluator (NIQE), with lower scores indicating higher quality. Cross-modal retrieval accuracy was measured using Recall@K metrics (K=1, 5, 10).

Human evaluation

Three independent studies were conducted by board-certified radiologists. For chest radiograph reports, a radiologist with 2-3 years of experience assessed 50 randomly selected MIMIC-CXR test cases using 5-point Likert scales evaluating: (a) accuracy of positive findings, (b) accuracy of negative findings, and (c) clinical relevance. For synthetic chest radiographs, the same radiologist evaluated 150 synthetic images (three per report from 50 cases) assessing anatomical correctness, positive finding presence, and appropriate negative finding absence. For synthetic mammograms, a board-certified radiologist with over 10 years of experience independently evaluated 50 synthetic images assessing finding accuracy and overall image quality.

RESULTS

To demonstrate the capability of CAMMAL in learning the fine-grained

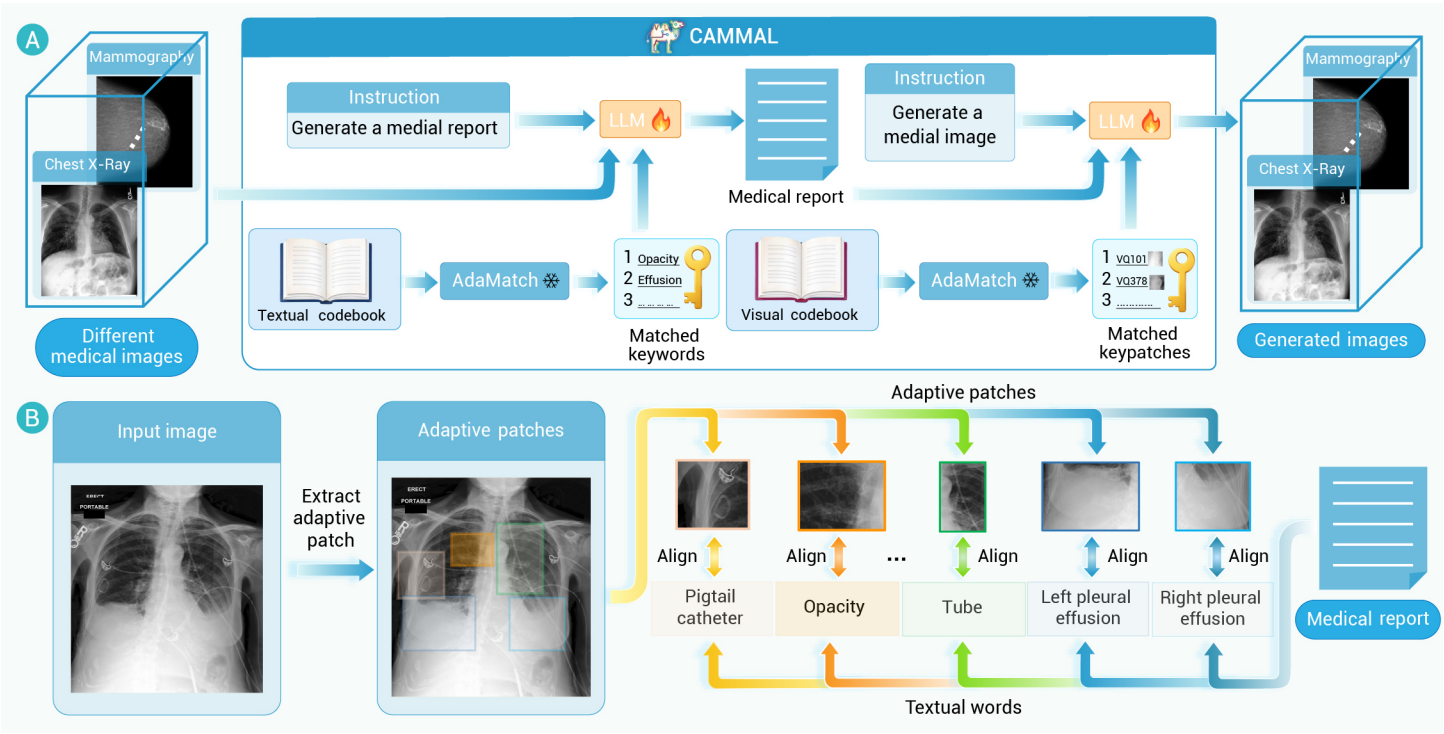


Figure 1. Illustrative diagram of the CAMMAL framework within a CXR/mammograph report and image generation task (A) AdaMatch-based bidirectional large language model (LLM) for report generation. (B) Adaptive patch-word Matching (AdaMatch) module for fine-grained alignment.

alignment between modalities, in this work, we train and validate CAMMAL on vision-language tasks involving radiology report generation and medical image synthesis. Two large-scale datasets of CXR and mammography modalities are used for the validation, which contain diverse lesions varying in shape, size, and location. For the CXR modality, we use the MIMIC-CXR dataset[53] (473,057 image-report pairs) for training and performed evaluation on the OpenI dataset[54] (3,684 pairs). We use chest x-ray with different types of view positions, including, Posterior-Anterior (PA), Anterior-Posterior (AP), and lateral. For the mammography modality, we utilize the EMBED dataset[55] comprising 54,488 mammograms with structured annotations including mass characteristics (shape, margin, density), calcification details (findings, distribution, count), BI-RADS assessment, and breast density. We convert this structured information into natural language descriptions using predefined templates, generating 2,608,175 mammogram-text pairs. With these paired data, we can train CAMMAL for both image-to-report and report-to-image generation.

Evaluation on the Chest X-Ray (CXR) datasets

We evaluate CAMMAL on CXR datasets for bidirectional generation between radiology reports and CXR images. The evaluation encompasses both report-to-CXR and CXR-to-report generation tasks. To assess CAMMAL's capability in fine-grained alignment, we analyze the quality of generated reports from CXR-to-report generation task, and the synthesized CXR images from report-to-CXR generation task. Additionally, we validate AdaMatch's fine-grained alignment performance through cross-modal retrieval tasks between CXR images and their corresponding reports. For internal validation, we test the model on the MIMIC-CXR testing set. To assess external validity, we perform inference on the OpenI dataset.

CXR-to-Report generation. For automatic evaluation, we use three commonly used natural language understanding (NLU) metrics, including BLEU, METEOR, and ROUGE-L.[56] As illustrated in Figure 2, CAMMAL framework demonstrates superior performance across both datasets, indicating robust internal and external validity, compared with current CXR-to-report generation methods such as R2Gen,[56] R2GenCMN,[57] Joint-TriNet,[58] XProNet,[59] ITHN,[60] and M2KT.[61] On the MIMIC-CXR dataset, CAMMAL achieves the highest scores in most metrics, including BLEU-1 (0.3793), BLEU-4 (0.1060), METEOR (0.1625), and ROUGE-L (0.2859), with improvements of 0.7%, 0.07%, 0.97%, and 0.68% respectively over the second-best model. Its performance on the OpenI dataset is even more pronounced, consistently outperforming all other methods across all metrics. Notably, it shows substantial improvements in BLEU-3 (0.2073) and BLEU-4 (0.1446), representing increases of 1.9% and 3.7%, respectively, over the next best model. This strong performance, particularly on the external OpenI dataset, suggests that CAMMAL has excellent generalization capabilities, a crucial factor for real-world clinical applications. Other methods, such as R2GenCMN and Joint-TriNet, show competitive performance on MIMIC-CXR but demonstrate less robust generalization to the OpenI dataset. Some models (e.g., ITHN and M2KT) exhibit a significant performance drop on OpenI, indicating potential overfitting or domain adaptation challenges. Beyond lexical overlap, we further evaluate CAMMAL using an LLM-as-a-Judge metric, which measures the percentage of clinical claims accurately recovered from the ground-truth report. This metric confirms CAMMAL's advantage, achieving the highest scores on both MIMIC-CXR (0.3299) and OpenI (0.6355), outperforming the respective second-best models by 1.7% and 2.4%. Notably, this metric also reveals that models such as ITHN and M2KT, despite their weak BLEU/ROUGE-L scores on OpenI, recover a comparatively high proportion of clinical content. This suggests that lexical metrics alone may underestimate the clinical adequacy of generated reports.

To comprehensively assess the clinical quality of the generated reports, we further conduct the human evaluation study. We randomly select 50 CXR images from the testing set of the MIMIC-CXR dataset and generate corresponding reports using CAMMAL. A radiologist trained in diagnostic imaging with 2-3 years of experience, is recruited to evaluate the generated reports. The radiologist is provided with each original CXR image and its generated report. The evaluation focused on three key dimensions: (a) accuracy of positive findings, assessing whether the positive findings apparent in the CXR images are accurately described in the generated reports; (b) accuracy of negative findings, determining whether the absence of specific pathologies is correctly reflected in the reports; and (c) clinical relevance, evaluating the overall clinical significance of the reports, particularly the impression sections, by comparing them with the ground-truth reports to assess their utility in guiding differential diagnosis and patient management. For dimensions (a) and (b), if uncertain about the presence of a finding in the image, the radiologist consults the ground-truth report for clarification. Each dimension is rated using a 5-point Likert scale ranging from 1 (Poor) to 5 (Excellent), with explicit definitions provided for each score to ensure consistency. Figure 5A

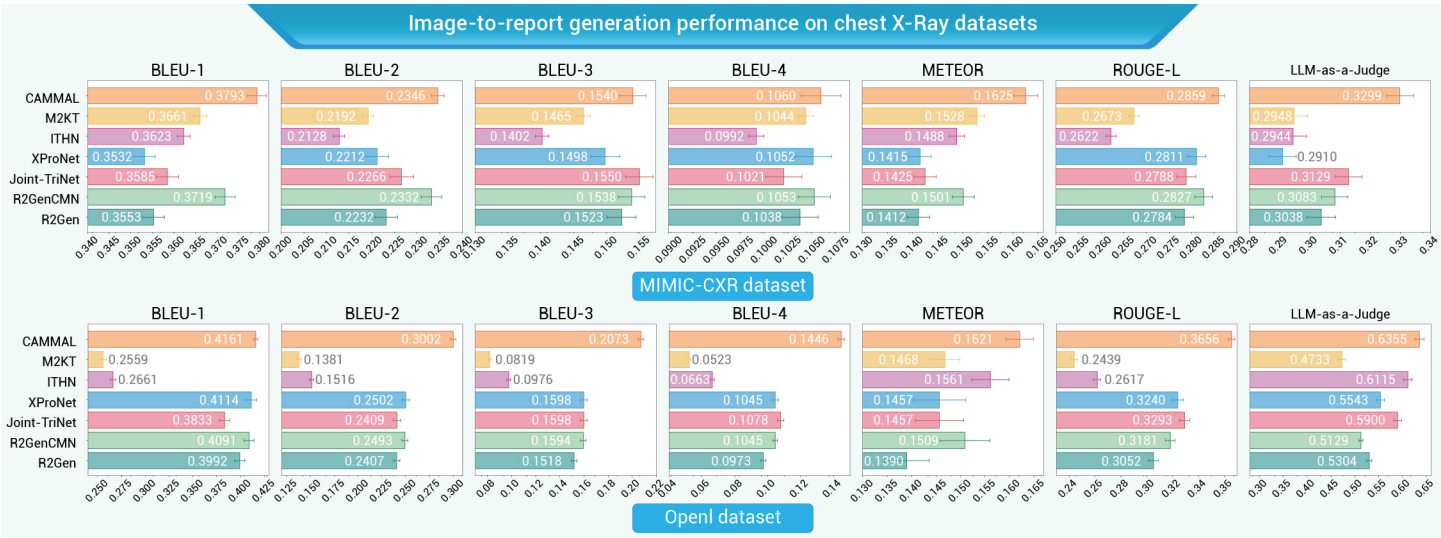


Figure 2. Image-to-report generation performance in comparison to existing report generation methods on the MIMIC-CXR and OpenI datasets.

shows the human evaluation results of the reports generated by CAMMAL. Our method demonstrates strong capability in identifying negative findings, with 56% of reports rated as very good or excellent. For positive findings, CAMMAL achieves moderate performance with 63% of reports receiving fair to excellent ratings. The clinical relevance scores show a polarized distribution, with 56% scoring below average and 44% above average, indicating the model's potential in generating clinically valuable reports. These results suggest that CAMMAL shows promise in certain areas, particularly in ruling out conditions. Quantitatively, as listed in Figure 5D, CAMMAL achieves the highest mean score in negative findings (3.0 ± 2.34), followed by clinical relevance (2.62 ± 1.83) and positive findings (2.31 ± 1.72). These results suggest that CAMMAL shows stronger capability in ruling out pathologies, while maintaining adequate performance in identifying positive findings and generating clinically meaningful reports.

In addition, we visualize the real reports and reports generated by CAMMAL, as shown in Figure 5G. In comparing the generated and real reports, CAMMAL successfully captures several key findings, including the presence of mild pulmonary edema, left pleural effusion, and atherosclerotic calcification in the aortic arch. The generated report correctly identifies the left lung base opacity and appropriately acknowledges the differential diagnosis between atelectasis and infection, similar to the ground truth's interpretation. The high degree of similarity between the generated and real reports underscores the potential of CAMMAL to serve as a valuable tool in clinical settings, subject to further validation.

Report-to-CXR generation. To automatically evaluate report-to-CXR generation performance, Figure 3 demonstrates the superior performance of CAMMAL compared to existing text-to-image generation approaches[62–66] across two widely used datasets, MIMIC-CXR and OpenI. We evaluate the performance using two commonly used metrics: Fréchet Inception Distance (FID) and Naturalness Image Quality Evaluator (NIQE). FID measures the similarity between the distribution of generated images and that of real images, while NIQE assesses the quality of images without reference to ground truth. For both metrics, lower values indicate better performance. CAMMAL consistently achieves the lowest FID and NIQE scores, indicating higher image quality and fidelity to the input reports. On the MIMIC-CXR dataset, CAMMAL achieves an FID of 1.09 and NIQE of 3.39, substantially outperforming the second-best method, LLM-CXR, which scores 2.18 and 3.60 respectively. Similarly, on the OpenI dataset, CAMMAL maintains its superiority with the lowest FID (1.59) and NIQE (3.31) scores. This consistent performance across datasets suggests the robustness and generalizability of our approach. Notably, CAMMAL achieves significant improvements over recent advanced methods like UniXGen and RoentGen, highlighting its effectiveness in bridging the gap between textual reports and visual CXR representations.

To rigorously evaluate the clinical accuracy of the generated CXR images, we conduct a human evaluation study involving a radiologist trained in diagnostic imaging with 2-3 years of experience. We randomly select 50 medical reports from the testing set of the MIMIC-CXR dataset and generate three synthetic CXR images for each report using CAMMAL, resulting in a total of 150 images. The radiologist is provided with the original reports and asked to assess each image based on four critical dimensions: (a) anatomical structure correctness and realism, (b) presence and accuracy of positive findings described in the report, and (c) absence of findings explicitly stated as negative in the report. Each dimension is rated on a 5-point Likert scale ranging from 1 (Poor) to 5 (Excellent). Figure 5B shows the human evaluation results on the CXR images generated by CAMMAL. Our model demonstrates notable strength in anatomical structure representation, with 73% of synthetic images receiving good to excellent ratings, indicating high fidelity in capturing fundamental anatomical features. In representing negative findings, CAMMAL shows reasonable competence with 38% of images achieving very good to excellent ratings. For positive findings, CAMMAL achieves moderate performance with 57% of images rated fair or above. Quantitatively, in Figure 5D, the synthetic images score highest in anatomical structure accuracy (3.69 ± 2.39), followed by negative findings (3.00 ± 2.34), and positive findings (2.18 ± 1.58). These results indicate that CAMMAL excels particularly in reproducing anatomical structures while maintaining adequate performance in representing both negative and positive pathological findings. Moreover, we visualize the real CXR and the CXR generated by CAMMAL, as shown in Figure 5F. Our CAMMAL demonstrates its effectiveness by accurately reflecting key findings from the input report. Notably, the generated images exhibit a clear representation of the pleural effusion (red arrow) and the opacity adjacent to the left heart border (blue arrow). These features are crucial diagnostic elements and their accurate depiction in the generated images highlights the model's ability to translate textual descriptions into visual representations effectively. This alignment between the input report and the visual output underscores the potential of our method to support clinical interpretations by providing reliable, report-driven imaging, thus enhancing diagnostic accuracy and consistency.

Report-CXR retrieval. To evaluate the AdaMatch's capability in fine-grained alignment, Figure 3B presents the performance of AdaMatch in both CXR-to-report and report-to-CXR retrieval tasks on the MIMIC-CXR dataset, in comparison to existing methods.[44,45,47,67–71] The retrieval performance is assessed using the Recall at K (R@K) metric, where K is set to 1, 5, and 10. R@K measures the percentage of queries for which the exact matching item is found within the top K retrieved results. In the CXR-to-report task, AdaMatch achieves a remarkable R@1 of 51.47%, surpassing the second-best method (LIMITR[70]) by 11.77%. Similarly, for the report-to-CXR task, AdaMatch attains an R@1 of 51.18%, exceeding LIMITR[70] by 13.48%. The performance gap widens further for R@5 and R@10 metrics, with AdaMatch reaching 86.19% and 94.77% for CXR-to-report, and 86.46% and 94.60% for

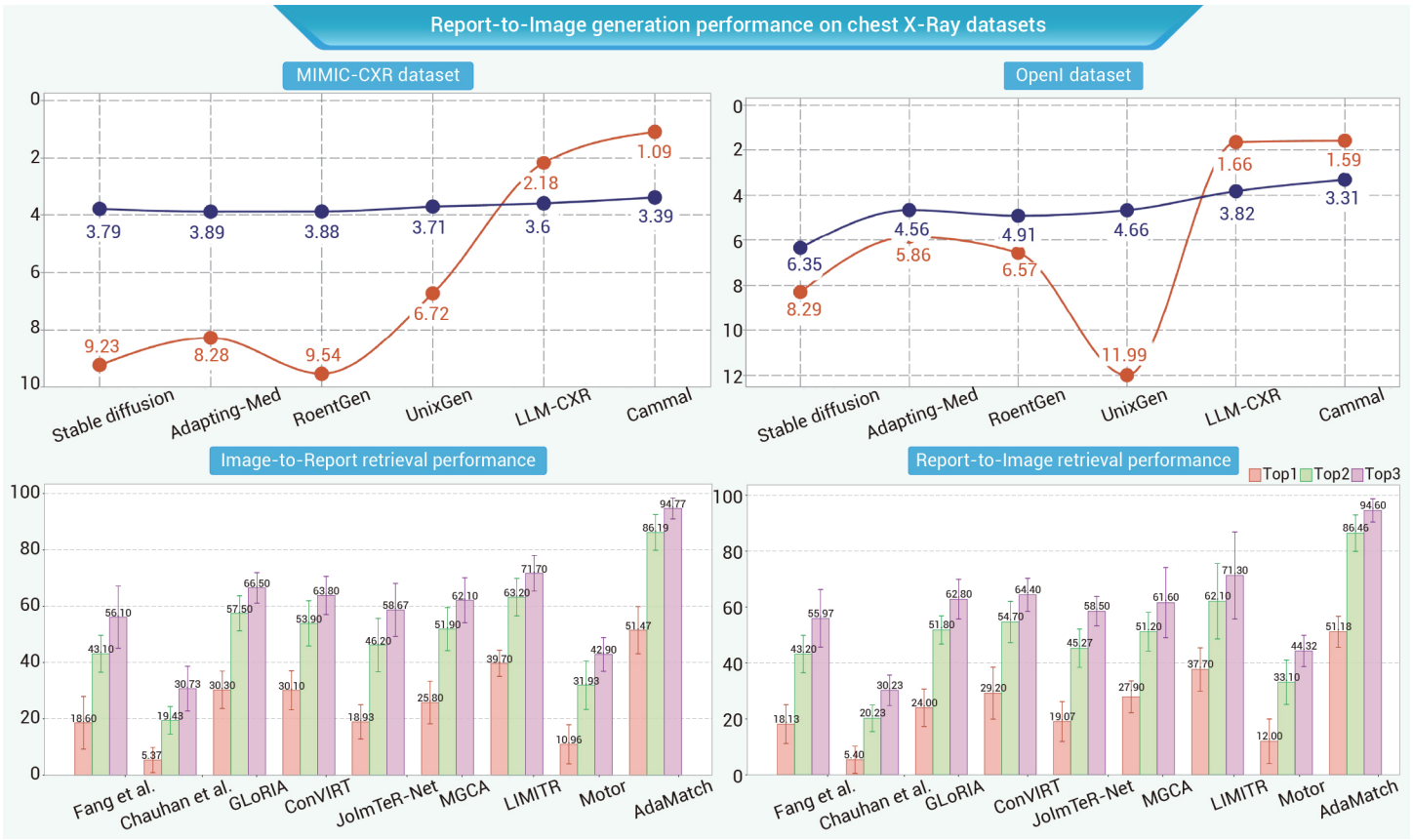


Figure 3. Quantitative evaluation on the chest x-ray (CXR) datasets (A) Report-to-image generation performance compared with current image synthesis methods on the MIMIC-CXR and OpenI datasets. (B) Image-report retrieval performance comparison on the MIMIC-CXR dataset.

report-to-CXR, respectively. These results indicate that AdaMatch can accurately retrieve the correct report or image in the top 10 matches for over 94% of the queries, substantially improving upon the state-of-the-art. The consistent performance across both retrieval directions underscores the robustness and versatility of AdaMatch in capturing the complex relationships between CXR images and their corresponding reports.

Evaluation on the mammography dataset

To validate the effectiveness of CAMMAL on diverse medical imaging applications, we extend our experiments to mammography using the EMBED dataset, focusing on bidirectional mammogram-text conversion tasks. As illustrated in Figure 4C, CAMMAL can handle both mammogram-to-text and text-to-mammogram generation. In the mammogram-to-text task, we predefine the question-answer pairs for each mammogram and provide options for the LLM to select. With the selected option and its question, we convert the answer into a statement with descriptive findings. Conversely, in the text-to-mammogram task, our method synthesizes mammographic images based on the statements of specific attributes. The analysis encompasses eight key mammographic characteristics: mass shape, mass margin, mass density, calcification findings, calcification distribution, calcification count, BI-RADS assessment, and breast density.

Mammogram-to-Text generation. For automatic evaluation, we compare the accuracy across multiple attributes for existing methods, including LLM-CXR[64] and LLAVA-Med[72]. In Figure 4A, CAMMAL achieves the highest average accuracy of 0.5673, surpassing LLM-CXR (0.5485) and LLAVA-Med (0.4568). Notably, CAMMAL outperforms both methods in crucial attributes such as mass shape (0.3014), mass margin (0.4578), and calcification finding (0.2035). It also shows marked improvement in breast density (0.6868) compared to LLAVA-Med (0.3564) and slightly outperforms LLM-CXR (0.6538) in this aspect. While all methods perform similarly well in identifying the number of calcifications, CAMMAL maintains a high accuracy (0.9972) comparable to its counterparts. These results highlight the robustness and versatility of CAMMAL across various mammography findings, suggesting its potential to enhance diagnostic accuracy and support clinical decision-making in breast cancer screening. In Figure 4D, we visualize the results of mammogram-to-text generation. It can be observed that CAMMAL can accurately answer different kinds of questions, such as breast density and BI-RADS assessment, when existing methods all generate wrong answers. Moreover, our method also generalizes well in the mass shape and calcification distribution questions, indicating our superiority over other methods.

Text-to-Mammogram Generation. To automatically evaluate text-to-mammogram generation performance, we compare FID and NIQE of existing text-to-image generation methods[62–64,66] and CAMMAL. As listed in Figure 4B, CAMMAL achieves the lowest FID of 0.0927, indicating exceptional fidelity and realism in generated mammograms. This represents a significant improvement over the second-best method, LLM-CXR (FID 0.0965), and a substantial leap from traditional approaches like Stable Diffusion (FID 4.4654). Additionally, CAMMAL attains the best NIQE score of 6.3896, surpassing all other methods and suggesting a higher perceptual quality of the generated images. These results collectively highlight the effectiveness of our approach in producing high-quality, realistic mammograms from textual descriptions.

To evaluate the clinical accuracy of our text-to-mammogram generation model, we conduct a rigorous human evaluation study. We randomly select 50 statements from the testing set of the EMBED dataset and generate corresponding mammogram images using CAMMAL. A board-certified radiologist with over 10 years of experience in mammography evaluates these generated mammograms. The radiologist is blinded to the image generation process and is provided only with the original text descriptions. They assessed each image on two key dimensions using a 5-point Likert scale (1=Poor, 5=Excellent): (1) presence and accuracy of described findings in the statement, and (2) overall image quality including realism, contrast, and tissue texture. Figures 5C-D demonstrate the evaluation results of the generated mammograms. CAMMAL shows consistent general image quality (mean score: 2.58 ± 0.37), with 52% of images rated as "Good" and 42% as "Fair". In representing pathological findings, our method achieves a mean

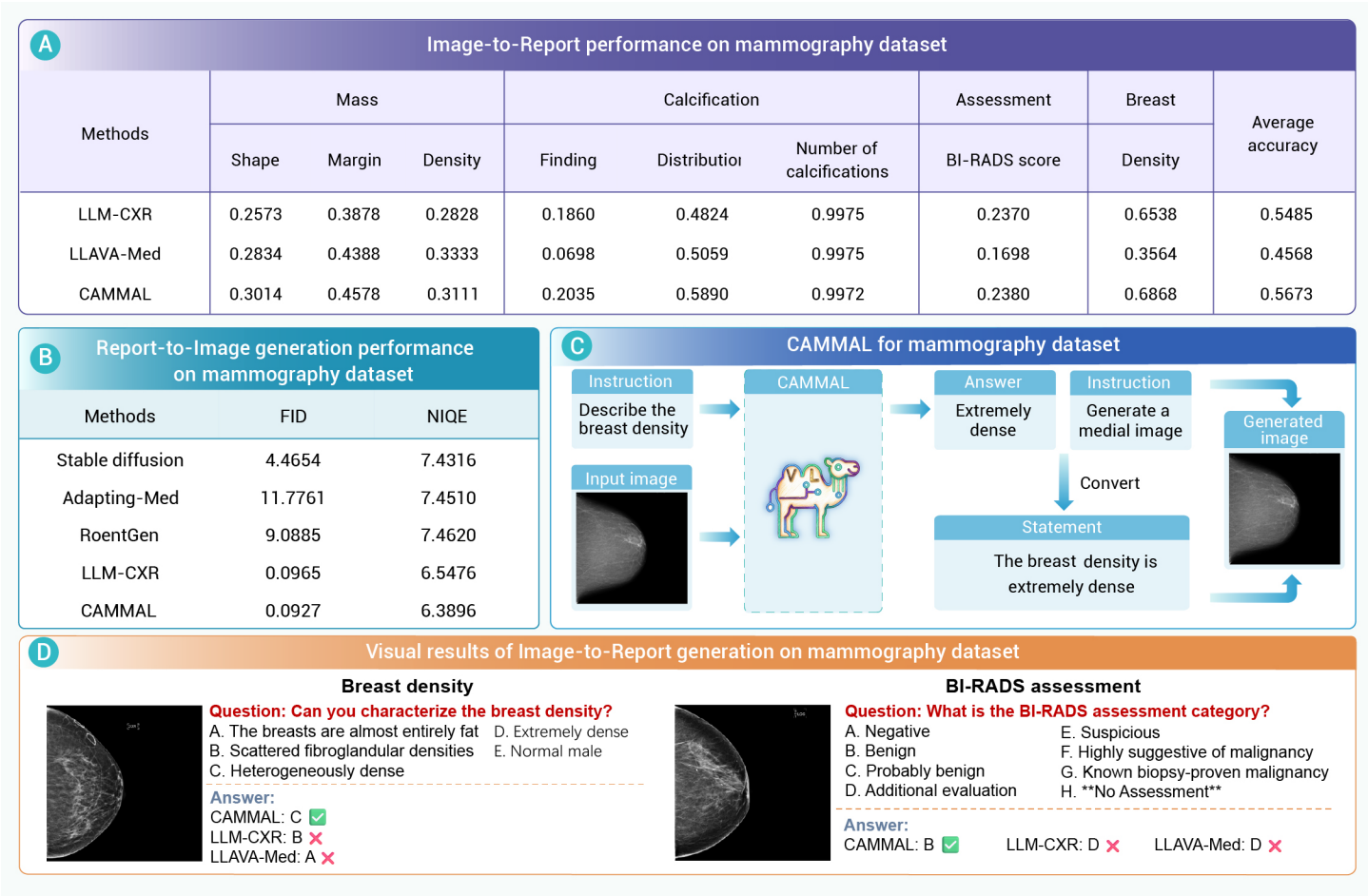


Figure 4. Quantitative evaluation and visual results on a mammography dataset (A) Image-to-Report generation performance on the mammography dataset. (B) Report-to-Image generation performance on the mammography dataset. (C) Illustration of CAMMAL processing the mammography dataset. (D) Visual results of Image-to-Report generation on the mammography dataset.

score of 1.82 ± 1.38, with varying performance across different clinical features. This performance profile indicates that while CAMMAL successfully captures general mammographic characteristics, future refinements could enhance the representation of specific clinical findings.

As visualized in Figure 5E, we present a comparison of statement texts alongside their corresponding real and synthetic mammography images. The leftmost example showcases a synthetic mammogram that accurately depicts a mass with a focal asymmetry shape, closely resembling the real image. In the rightmost case, the synthetic image successfully replicates the heterogeneously dense breast tissue observed in the actual mammogram. These examples demonstrate that our model can generate synthetic mammography images with realistic features that accurately reflect the descriptive statements provided.

DISCUSSION

In this work, we propose the CAMMAL framework to learn the fine-grained alignment between medical visual (CXR/mammography) and language modalities. CAMMAL features an adaptive tokenization model for image patches to achieve patch-level alignment with textual words in the reports. To improve the alignment reliability and robustness, we also introduce a cyclic generation scheme to perform bi-directional image-to-report-to-image generation based on the learned alignment. CAMMAL not only bridges the modality gap during multi-modal analysis but also enhances the interpretability of the learned alignment through generative modeling. By training on large-scale paired image-text datasets (over 473,000 pairs for CXR task and 54,000 pairs for the mammogram task), CAMMAL learns the capability to achieve accurate image-text alignment in these two tasks in individual experiments, showing a significant advancement over existing approaches. While current methods typically rely on fixed patch-word matching[1,34–37] or global image-report alignment[38–43], our adaptive patch extraction mechanism enables more precise and flexible alignment between image regions and textual descriptions. In Figure 3B, this superiority could be quantitatively demonstrated through CAMMAL's leading performance in image retrieval tasks, achieving R@1 scores of 51.47% for CXR-to-report retrieval and 51.18% for report-to-CXR retrieval, surpassing the previous state-of-the-art by 11.77% and 13.48% respectively. For broader retrieval metrics, CAMMAL achieves R@5 scores of 86.19% and 86.46%, and R@10 scores of 94.77% and 94.60% for CXR-to-report and report-to-CXR tasks, consistently outperforming existing methods across all metrics. In Figure 6C, the visualization of adaptive patches reveals clinically meaningful correlations with given words, providing interpretable evidence of the alignment process, particularly in identifying pathological features such as pleural effusions, consolidations, and cardiomegaly. To further analyze the alignment process, we select the top six keywords in the textual codebook and show the corresponding adaptive patches with the highest similarity, visualized in Figure 6D. It can be observed that both pathology (e.g. 'scattered opacities') and anatomy (e.g. bilateral upper lobe) image features can be captured through the alignment, as demonstrated by the corresponding adaptive patches.

Compared to existing generative methods for medical imaging, which primarily aims at translating data from one modality to another,[56–66] our design of CAMMAL focuses on leveraging image and text generation as a means to enhance multi-modal alignment learning. Rather than treating generation as an isolated task, CAMMAL enforces alignment-driven synthesis, where generation is explicitly guided by an alignment matrix that captures the correspondence between keywords and keypatches. Furthermore, the bidirectional nature of CAMMAL ensures modality-invariant alignment,

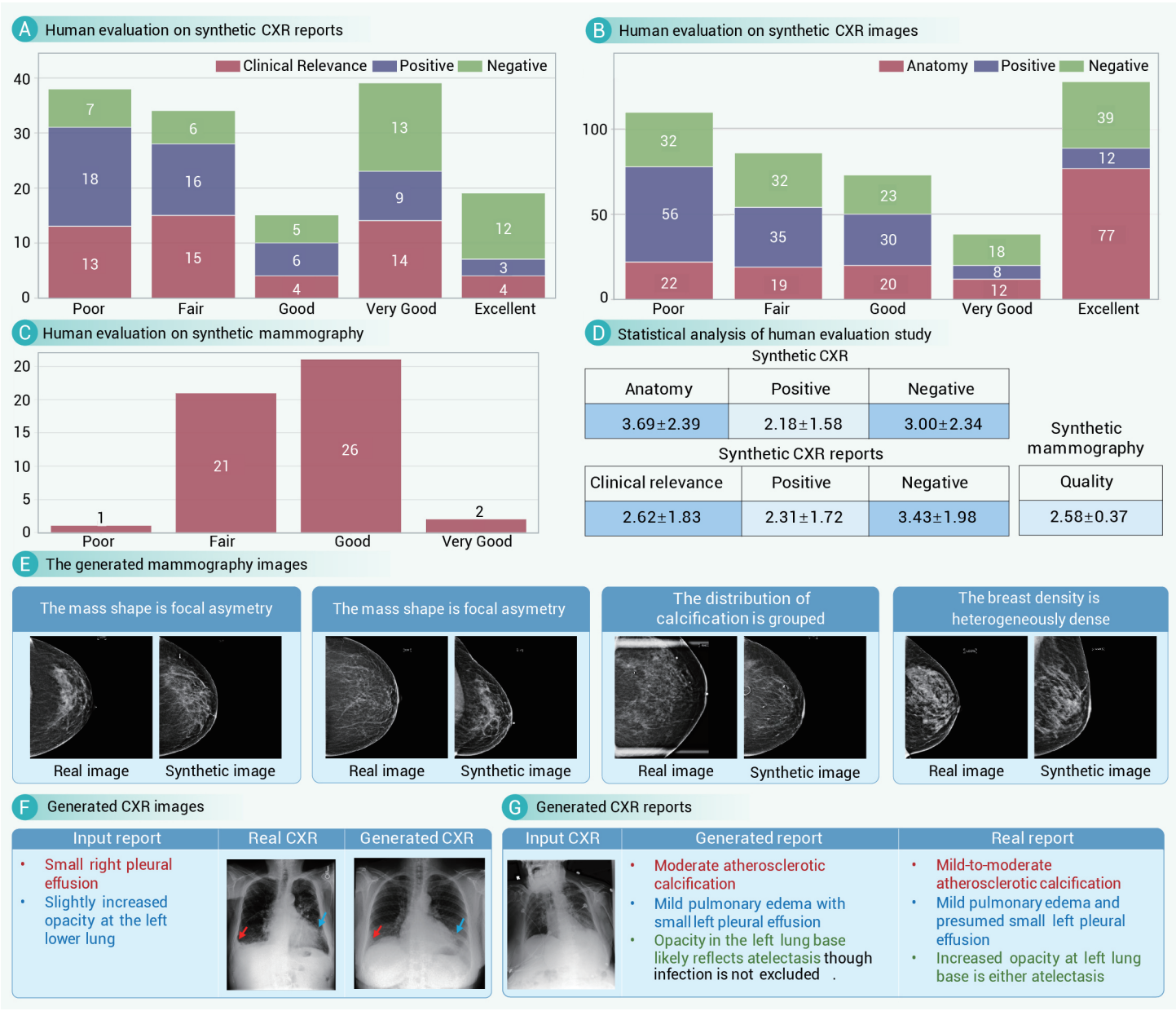


Figure 5. Human evaluation study and case visualization (A) Human evaluation study on synthetic CXR reports. (B) Human evaluation study on synthetic CXR images and comparison of anatomical structure with or without ribs. (C) Human evaluation study on synthetic mammography. (D) Statistical analysis of human evaluation study on synthetic CXR, CXR reports, and mammography. (E) Visualization of the mammography images generated by CAMMAL. (F) Visualization of the CXR images generated by CAMMAL. (G) Visualization of the reports generated by CAMMAL.

requiring both image-to-text and text-to-image generation to reinforce consistency and improve the interpretability of cross-modal representations. In Figure 3A, the effectiveness of this approach is evident in report-to-CXR generation, where CAMMAL achieves FID (1.09) and NIQE (3.39) scores on MIMIC-CXR dataset, substantially outperforming the second-best method, LLM-CXR (FID: 2.18, NIQE: 3.60). The human evaluation by radiologists reveals that CAMMAL excels in understanding the global structure of the images, with 38% of generated images achieving very good to excellent ratings in anatomical structure (mean score: 3.69 ± 2.39), followed by the rating in negative findings (mean score: 3.00 ± 2.34) and positive findings (mean score: 2.18 ± 1.58). The generated CXR images demonstrate an accurate representation of pathological features, such as right pleural effusion and opacity, as confirmed by radiologist evaluation in Figure 5F. In addition, Figure 6B visualizes the distributions of synthetic images through t-SNE analysis. The distribution of samples reveals superior clustering patterns compared to existing methods, with CAMMAL-generated images showing the highest degree of overlap with the real CXR images compared to Stable Diffusion, Adapting-Med, and UniXGen. These comprehensive results validate the effectiveness of CAMMAL in generating high-quality, clinically relevant chest X-rays.

To demonstrate the versatility and generalizability of our approach across different medical imaging modalities, we also train and validate CAMMAL on the mammography (EMBED) dataset. In this more challenging task, due to the much larger image size and more subtle lesions, CAMMAL maintains its superior performance while adapting to the unique characteristics of breast imaging. In the mammogram-to-text generation task, our method achieves higher accuracy across various clinically relevant attributes, with particularly strong performance in breast density assessment (0.6868) and calcification detection (0.9972 for number identification), as shown in Figure 4A. CAMMAL's ability to accurately characterize mass characteristics, including shape (0.3014) and margin (0.4578), demonstrates its capability to capture subtle radiological features crucial for breast cancer diagnosis. For text-to-mammogram generation, as listed in Figure 4B, CAMMAL achieves an FID score of 0.0927, outperforming existing methods by a large margin, indicating high fidelity in its image generation. In Figure 5C, the human evaluation by experienced breast radiologists further validates our approach, with 52% of

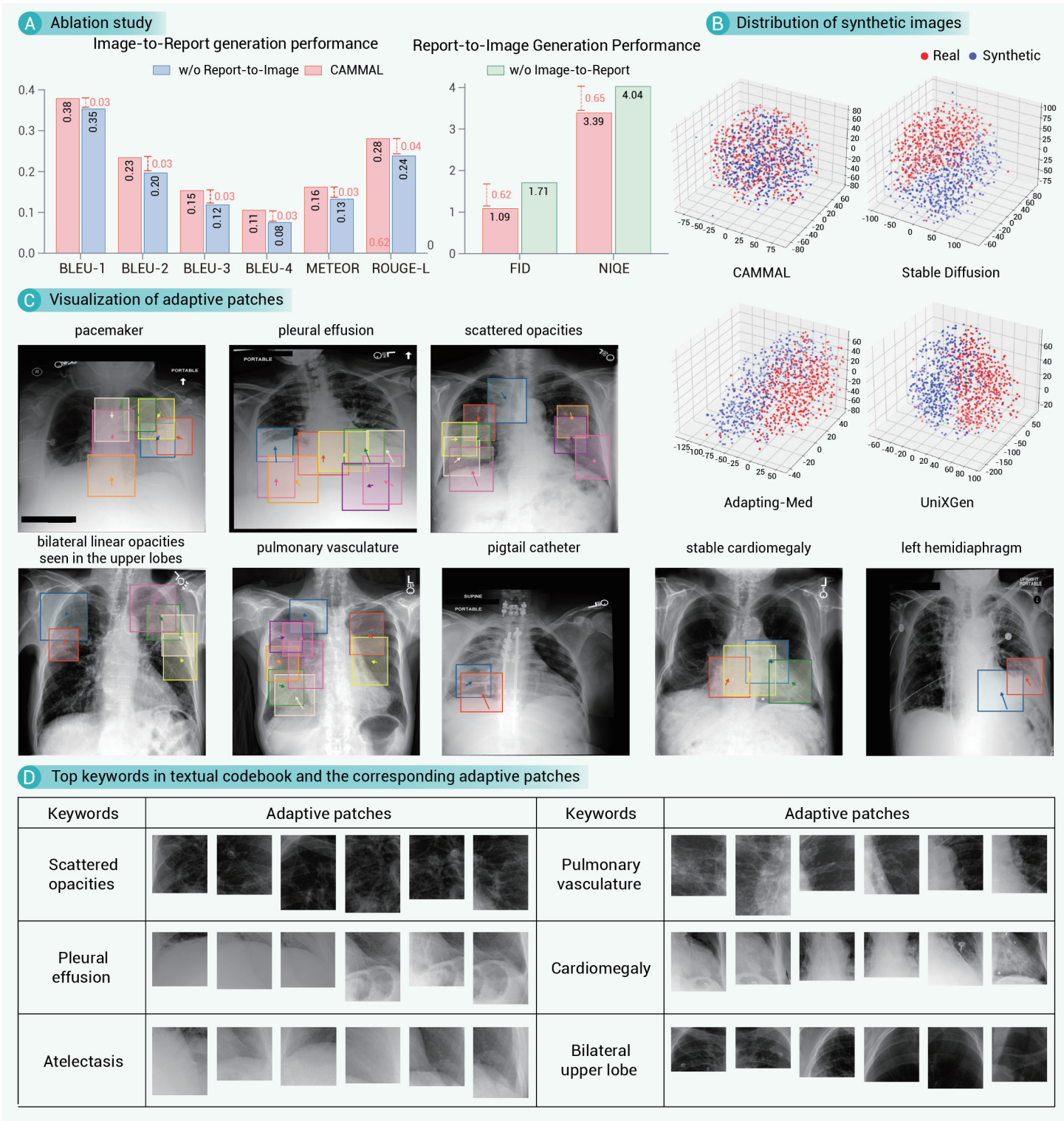


Figure 6. Ablation study and visualization of adaptive patches (A) Ablation study for cyclic generation. (B) Data distribution of synthetic images. (C) Visualization of adaptive patches. (D) Top keywords in the textual codebook and the corresponding adaptive patches.

generated mammograms receiving "Good" ratings for overall image quality, with a mean score of 2.58 ± 0.37. Notably, CAMMAL successfully replicates complex mammographic features such as focal asymmetries and heterogeneous breast tissue patterns, as shown in our qualitative analysis in Figure 5E. The consistent performance across two different imaging modalities suggests that CAMMAL is quite robust and generalizable to different medical imaging types and qualities.

The experimental results of CAMMAL demonstrate its promising potential in explainable and trustworthy medical AI. The framework presents an integrated approach to learning and utilizing knowledge embedded in multi-modal clinical data—specifically, radiology reports—by aligning information across "codebooks" as shown in Figure 6D. As recognized by recent works in clinical AI developments,[73] it is important to perform knowledge-driven learning from experts rather than via mere pattern recognition.[20] Unlike conventional vision-language models that primarily characterize the data correlations, CAMMAL focuses on learning multi-modal knowledge and mimics clinicians' diagnostic reasoning process by capturing the underlying clinical reasoning present in both images and reports. Through its fine-grained alignment and generation approach, CAMMAL aligns medical images with textual descriptions at the conceptual level and systematically captures semantic structures

in report writing. Such a feature not only enhances vision-language alignment but also fosters improved interpretability and explainability in its learned knowledge, addressing a key limitation of prior vision-language foundation models.[21]

This capacity for CAMMAL's structured knowledge alignment also presents a novel opportunity to enhance the current clinician training and education. By processing a large amount of image-report pairs, CAMMAL can explicitly link imaging features to their corresponding textual descriptions in a systematic and interpretable manner. The model output can help trainees develop a deeper and more comprehensive understanding of radiological interpretation and skills in structured report writing, which is critical in modern diagnostic practice.[74] Additionally, its bidirectional generation capability enables interactive case-based learning, where trainees (like radiology fellows) can interactively observe pathology visualizations in various patient settings, strengthening their ability to correlate findings with imaging manifestations.[75]

Despite its promising performance and interpretability, CAMMAL has several limitations. First, fine-grained image–text alignment depends on the concordance between paired images and reports. Some report content, such as clinical history, comparisons with prior examinations, diagnostic uncertainty, and management recommendations, may not have direct visual counterparts, while visible findings may be omitted from the report. Therefore, the learned image–text associations should be interpreted as statistical correspondences rather than definitive causal or anatomical explanations. Second, CAMMAL relies on routinely collected radiology reports, which may contain omissions, reporting-style variation, templated language, hedging, negation, and references to prior examinations. Such noise may propagate into the textual codebook, learned alignments, and generated outputs, while the frequency-based construction of the codebook may underrepresent rare but clinically important findings. Finally, CAMMAL remains limited in the clinical accuracy and resolution of its generated images. Human evaluation reveals CAMMAL has challenges in accurately identifying positive findings, particularly for certain pathologies in CXR images. The radiologist's assessment shows only 35% of images received 'Good' to 'Excellent' ratings for positive finding representation, as illustrated in Figure 5B. Regarding the quality of image generation, radiologists note that the low resolution of the generated images, especially in the case of mammography, would affect the reading of subtle lesions due to GPU memory limitations. It was also noted that the relatively bright contrast in mammograms poses challenges for radiologists in interpreting breast architecture, including vessels and calcifications. Future work will investigate higher-resolution and multi-scale representations, improved report normalization and uncertainty modeling, rare-finding-aware codebooks, advanced latent diffusion-based generation, and larger multi-reader, multi-center validation. At its current stage, CAMMAL should be considered a research framework, and its outputs should not be used independently for clinical diagnosis or patient management.

Beyond these technical limitations, the generative nature of CAMMAL raises ethical considerations that merit explicit discussion. Our reader studies quantify the central concern that synthetic outputs may look realistic while being clinically wrong: synthetic radiographs scored highest for anatomical realism (3.69 ± 2.39) but lowest for positive-finding accuracy (2.18 ± 1.58), generated reports were likewise weakest at positive findings (2.31 ± 1.72). The dominant failure mode is thus not implausible imagery but fluent, convincing output containing fabricated or, more often, omitted findings, which may induce automation bias in readers. This realism gap also creates misuse potential, including passing synthetic studies off as authentic patient data, contaminating public archives and training corpora with unlabelled synthetic images, and unsupervised deployment for triage. Although we used only de-identified public datasets under IRB approval with no attempt at re-identification, generative models retain a residual memorization risk, and we therefore do not present CAMMAL outputs as privacy-preserving surrogates for real data. CAMMAL is a research framework whose outputs are unverified drafts requiring radiologist review; it should not be used independently for diagnosis, screening, triage, or patient management.

CONCLUSION

In conclusion, CAMMAL represents a significant step forward in learning multi-modal knowledge and addressing the modality gap of the medical vision-language tasks. Quantitative evaluation and human reader study results on two vision-language datasets demonstrate its potential to learn the knowledge in alignment. While current limitations in interpreting and generating images present needs for future improvements, our framework provides a promising solution for future multi-modal medical foundation models. As CAMMAL shows the feasibility in both CXR/mammography data, we envision that it can potentially be utilized for other modalities or even non-vision-language tasks, e.g., alignment between two imaging modalities, warrant exploration in broader clinical domains.

REFERENCES

1. Radford A., Kim J.W., Hallacy C., et al. (2021). Learning transferable visual models from natural language supervision. *ICML*. **139**: 8748–63. https://proceedings.mlr.press/v139/radford21a.html
2. Oquab M., Darcet T., Moutakanni T., et al. (2024). DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=a68SUt6zFt
3. Kirillov A., Mintun E., Ravi N., et al. (2023). Segment anything. *ICCV*. **2023**: 3992–4003. DOI:10.1109/iccv51070.2023.00371
4. He K., Fan H., Wu Y., et al. (2020). Momentum contrast for unsupervised visual representation learning. *CVPR*. **2020**: 9726–9735. DOI:10.1109/cvpr42600.2020.00975
5. Chen R.J., Ding T., Lu M.Y., et al. (2024). Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**:850–862. DOI:10.1038/s41591-024-02857-3
6. Wu L., Zhuang J. and Chen H. (2024). VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. *CVPR*. **2024**: 22873–22882. DOI:10.1109/cvpr52733.2024.02158
7. Ma J., He Y., Li F., et al. (2024). Segment anything in medical images. *Nat. Commun.* **15**:654. DOI:10.1038/s41467-024-44824-z
8. Zhou Y., Chia M.A., Wagner S.K., et al. (2023). A foundation model for generalizable disease detection from retinal images. *Nature* **622**:156–63. DOI:10.1038/s41586-023-06555-x
9. He Y., Yang G., Ge R., et al. (2023). Geometric visual similarity learning in 3D medical image self-supervised pre-training. *CVPR*. **2023**: 9538–47. DOI:10.1109/cvpr52729.2023.00920
10. Wang J., Wang K., Yu Y., et al. (2025). Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat. Med.* **31**:609–17. DOI:10.1038/s41591-024-03359-y
11. Kondepudi A., Pekmezci M., Hou X., et al. (2025). Foundation models for fast, label-free detection of glioma infiltration. *Nature* **637**:439–45. DOI:10.1038/s41586-024-08169-3
12. Huang W., Li C., Zhou H.Y., et al. (2024). Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nat. Commun.* **15**:7620. DOI:10.1038/s41467-024-51749-0
13. Wang X., Zhao J., Marostica E., et al. (2024). A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**:970–8. DOI:10.1038/s41586-024-07894-z
14. Vorontsov E., Bozkurt A., Casson A., et al. (2024). A foundation model for clinicalgrade computational pathology and rare cancers detection. *Nat. Med.* **30**:2924–35. DOI:10.1038/s41591-024-03141-0
15. Christensen M., Vukadinovic M., Yuan N., et al. (2024). Vision-language foundation model for echocardiogram interpretation. *Nat. Med.* **30**:1481–8. DOI:10.1038/s41591-024-02959-y
16. Jin Q., Wang Z., Floudas C.S., et al. (2024). Matching patients to clinical trials with large language models. *Nat. Commun.* **15**:9074. DOI:10.1038/s41467-024-53081-z
17. Zhang K., Zhou R., Adhikarla E., et al. (2024). A generalist vision-language foundation model for diverse biomedical tasks. *Nat. Med.* **30**:3129–41. DOI:10.1038/s41591-024-03185-2
18. Bluethgen C., Chambon P., Delbrouck J.B., et al. (2025). A vision-language foundation model for the generation of realistic chest x-ray images. *Nat. Biomed. Eng.* **9**:494–506. DOI:10.1038/s41551-024-01246-y
19. Zhao T., Gu Y., Yang J., et al. (2025). A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**:166–76. DOI:10.1038/s41592-024-02499-w
20. Tanno R., Barrett D.G.T., Sellergren A., et al. (2025). Collaboration between clinicians and vision-language models in radiology report generation. *Nat. Med.* **31**:599–608. DOI:10.1038/s41591-024-03302-1
21. Lu M.Y., Chen B., Williamson D.F.K., et al. (2024). A visual-language foundation model for computational pathology. *Nat. Med.* **30**:863–74. DOI:10.1038/s41591-024-02856-4
22. Huang Z., Bianchi F., Yuksekgonul M., et al. (2023). A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**:2307–16. DOI:10.1038/s41591-023-02504-3
23. Li J., Guan Z., Wang J., et al. (2024). Integrated image-based deep learning and

language models for primary diabetes care. *Nat. Med.* **30**:2886–96. DOI:10.1038/s41591-024-03139-8
24. Kim C., Gadgil S.U., DeGrave A.J., et al. (2024). Transparent medical image AI via an image-text foundation model grounded in medical literature. *Nat. Med.* **30**:1154–65. DOI:10.1038/s41591-024-02887-x
25. Li X., Zhao L., Zhang L., et al. (2025). Artificial general intelligence for medical imaging analysis. *IEEE Rev. Biomed. Eng.* **18**:113–29. DOI:10.1109/rbme.2024.3493775
26. Bradshaw T.J., Tie X., Warner J., et al. (2025). Large language models and large multimodal models in medical imaging: A primer for physicians. *J. Nucl. Med.* **66**:173–82. DOI:10.2967/jnumed.124.268072
27. Shi Y., Li Q., Sun J., et al. (2025). Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *ICLR* **2025**: 85780–803. https://proceedings.iclr.cc/paper_files/paper/2025/hash/d51cd79a85833b022841f7a2383b32d3-Abstract-Conference.html
28. Zhong T., Zhao W., Zhang Y., et al. (2026). Chatradio-valuer: A chat large language model for generalizable radiology impression generation on multi-institution and multi-system data. *IEEE Trans. Biomed. Eng.* **73**:1050–61. DOI:10.1109/tbme.2025.3597325
29. Guo Z., Li X., Huang H., et al. (2019). Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans. Radiat. Plasma Med. Sci.* **3**:162–9. DOI:10.1109/trpms.2018.2890359
30. Peng Y., Qi J. and Yuan Y. (2018). Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Trans. Image Process.* **27**:5585–99. DOI:10.1109/tip.2018.2852503
31. Cui H., Mao L., Liang X., et al. (2024). Biomedical visual instruction tuning with clinician preference alignment. *NeurIPS* **37**:96449–67. DOI:10.52202/079017-3057
32. Ma C., Jiang H., Chen W., et al. (2024). Eye-gaze guided multi-modal alignment for medical representation learning. *NeurIPS*. **37**:6126–53. DOI:10.52202/079017-0198
33. Kline A., Wang H., Li Y., et al. (2022). Multimodal machine learning in precision health: A scoping review. *npj Digit. Med.* **5**:171. DOI:10.1038/s41746-022-00712-8
34. Jia C., Yang Y., Xia Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*. **139**: 4904–16.https://proceedings.mlr.press/v139/jia21b.html
35. Jang J., Kong C., Jeon D., et al. (2023). Unifying vision-language representation space with single-tower transformer. *AAAI*. **37**:980–8. DOI:10.1609/aaai.v37i1.25178
36. Wang W., Bao H., Dong L., et al. (2023). Image as a foreign language: BEiT pretraining for vision and vision-language tasks. *CVPR*. **2023**:19175–86. DOI:10.1109/cvpr52729.2023.01838
37. Yang J., Duan J., Tran S., et al. (2022). Vision-language pre-training with triple contrastive learning. *CVPR*. **2022**:15650–9. DOI:10.1109/cvpr52688.2022.01522
38. Chen Y.C., Li L., Yu L., et al. (2020). UNITER: Universal image-text representation learning. *ECCV*. **2020**: 104–20. DOI: 10.1007/978-3-030-58577-8_7
39. Li X., Yin X., Li C., et al. (2020). OSCAR: Object-semantics aligned pre-training for vision-language tasks. *ECCV*. **2020**:121–37. DOI: 10.1007/978-3-030-58577-8_8
40. Li W., Gao C., Niu G., et al. (2021). UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. *ACL-IJCNLP*. **2021**: 2592–607. DOI: 10.18653/v1/2021.acl-long.202
41. Zhan X., Wu Y., Dong X., et al. (2021). Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. *ICCV*. **2021**:11762–71. DOI: 10.1109/iccv48922.2021.01157
42. Kim W., Son B. and Kim I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. *ICML*. **139**: 5583–94. https://proceedings.mlr.press/v139/kim21k.html
43. Yao L., Huang R., Hou L., et al. (2022). FILIP: Fine-grained interactive language-image pre-training. *ICLR*. URL: https://openreview.net/forum?id=cpDhcsEDC2
44. Ji Z., Shaikh M.A., Moukheiber D., et al. (2021). Improving joint learning of chest x-ray and radiology report by word region alignment. *MLMI*. **2021**:110–9. DOI:10.1007/978-3-030-87589-3_12
45. Wang F., Zhou Y., Wang S., et al. (2022). Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*. **35**:33536–49. DOI:10.52202/068431-2430
46. Moon J.H., Lee H., Shin W., et al. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.* **26**:6070–80. DOI:10.1109/jbhi.2022.3207502
47. Huang S.C., Shen L., Lungren M.P., et al. (2021). GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. *ICCV*. **2021**: 3922–31. DOI: 10.1109/iccv48922.2021.00391
48. Yan B. and Pei M. (2022). Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. *AAAI*. **36**: 2982–90. DOI: 10.1609/aaai.v36i3.20204
49. Wan Z., Liu C., Zhang M., et al. (2023). Med-UniC: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *NeurIPS*. **36**: 56186–97. DOI:10.52202/075280-2450
50. Chen X., He Y., Xue C., et al. (2023). Knowledge boosting: Rethinking medical contrastive vision-language pre-training. *MICCAI*. **2023:** 405–15. DOI:10.1007/978-3-031-43907-0_39
51. Liu C., Cheng S., Shi M., et al. (2025). IMITATE: Clinical prior guided hierarchical vision-language pre-training. *IEEE Trans. Med. Imaging* **44**: 519–29. DOI: 10.1109/tmi.2024.3449690
52. Esser P., Rombach R. and Ommer B. (2021). Taming transformers for high-resolution image synthesis. *CVPR*. **2021**: 12868–78. DOI: 10.1109/cvpr46437.2021.01268
53. Johnson A.E.W., Pollard T.J., Berkowitz S.J., et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**:317. DOI:10.1038/s41597-019-0322-0
54. Demner-Fushman D., Kohli M.D., Rosenman M.B., et al. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**:304–10. DOI:10.1093/jamia/ocv080
55. Jeong J.J., Vey B.L., Bhimireddy A., et al. (2023). The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammo-graphic images. *Radiol. Artif. Intell.* **5**:e220047. DOI:10.1148/ryai.220047
56. Chen Z., Song Y., Chang T.H., et al. (2020). Generating radiology reports via memory-driven transformer. *EMNLP*. 1439–49. DOI:10.18653/v1/2020.emnlp-main.112
57. Chen Z., Shen Y., Song Y., et al. (2021). Cross-modal memory networks for radiology report generation. *ACL-IJCNLP*. **2021**: 5904–14. DOI: 10.18653/v1/2021.acl-long.459 DOI:10.18653/v1/2021.acl-long.459
58. Yang Y., Yu J., Zhang J., et al. (2023). Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Trans. Multimedia* **25**:167–78. DOI:10.1109/tmm.2021.3122542
59. Wang J., Bhalerao A. and He Y. (2022). Cross-modal prototype driven network for radiology report generation. *ECCV*. **2022**: 563–79. DOI:10.1007/978-3-031-19833-5_33
60. Voutharoja B.P., Wang L. and Zhou L. (2023). Automatic radiology report generation by learning with increasingly hard negatives. *ECAI*. **2023**:2427–34. DOI:10.3233/faia230545
61. Yang S., Wu X., Ge S., et al. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. *Med. Image Anal.* **86**:102798. DOI:10.1016/j.me-dia.2023.102798
62. Rombach R., Blattmann A., Lorenz D., et al. (2022). High-resolution image synthesis with latent diffusion models. *CVPR*. **2022**: 10674–85. DOI:10.1109/cvpr52688.2022.01042
63. Chambon P., Bluethgen C., Langlotz C.P., et al. (2022). Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv* preprint. DOI: 10.48550/arXiv.2210.04133
64. Bluethgen C., Chambon P., Delbrouck J.B., et al. (2025). A vision-language foundation model for the generation of realistic chest x-ray images. *Nat. Biomed. Eng.* **9**:494–506. DOI:10.1038/s41551-024-01246-y
65. Lee H., Kim W., Kim J.H., et al. (2023). Unified chest X-ray and radiology report generation model with multi-view chest X-rays. *arXiv* preprint. DOI: 10.48550/arXiv.2302.12172
66. Lee S., Kim W.J. and Ye J.C. (2023). LLM itself can read and generate CXR images. *arXiv* preprint. DOI: 10.48550/arXiv.2305.11490
67. Fang H., Gupta S., Iandola F., et al. (2015). From captions to visual concepts and back. *CVPR*. **2025**: 1473–82. DOI:10.1109/cvpr.2015.7298754
68. Chauhan G., Liao R., Wells W., et al. (2020). Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. *MICCAI*. **2020**: 529–39. DOI:10.1007/978-3030-59713-9_51
69. Zhang Y., Jiang H., Miura Y., et al. (2022). Contrastive learning of medical visual representations from paired images and text. *MLHC*. **182**: 2–25. https://proceedings.mlr.press/v182/zhang22a.html
70. Dawidowicz G., Hirsch E. and Tal A. (2023). LIMITR: Leveraging local information for medical image-text representation. *ICCV*. **2023**: 21108–16. DOI: 10.1109/iccv51070.2023.01935
71. Lin B., Chen Z., Li M., et al. (2023). Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. *arXiv* preprint. DOI: 10.48550/arXiv.2304.14204
72. Li C., Wong C., Zhang S., et al. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*. **36**:28541–64. DOI:10.52202/075280-1240
73. Mao D., Liu C., Wang L., et al. (2024). AI-MARRVEL—A knowledge-driven AI system for diagnosing mendelian disorders. *NEJM AI*. **1**:AIoa2300009. DOI:10.1056/aioa2300009
74. Nobel J.M., van Geel K. and Robben S.G.F. (2022). Structured reporting in radiology: A systematic review to explore its potential. *Eur. Radiol.* **32**:2837–54. DOI:10.1007/s00330-021-08327-5
75. Sugi M.D., Kennedy T.A., Shah V., et al. (2021). Bridging the gap: Interactive, case-based learning in radiology education. *Abdom. Radiol.* **46**:5503–8. DOI:10.1007/s00261-021-03147-z

FUNDING AND ACKNOWLEDGMENTS

This work was supported by Hong Kong Innovation and Technology Commission Innovation and Technology Fund PRP/082/24FX. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

W.C. designed the study, extended the code, conducted all experiments, analyzed data, and contributed to manuscript preparation. X.L. designed the study, provided supervision throughout the project, analyzed data, and contributed to manuscript preparation. H.R., Y.C., and E.S. were responsible for the human evaluation of the synthetic data. L.S. provided GPU resources for the training model. Y.Y. provided supervision throughout the project and contributed to manuscript preparation. T.L, Y.O, J.L., H.Z., Q.L., and L.Sun provided supervision throughout the project, from conception to discussion, and assisted in preparing the manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.

DATA AND CODE AVAILABILITY

OpenI dataset in this study is publicly available and can be accessed from: https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university. The MIMIC-CXR dataset can be requested via https://physionet.org/content/mimic-cxr-jpg/2.1.0/. The EMBED dataset can be requested via https://registry.opendata.aws/emory-breast-imaging-dataset-embed/. The Pytorch codes for the proposed CAMMAL used in this study is available at the following GitHub repository at https://github.com/CUHK-AIM-Group/CAMMAL.