

# MammoBLIP: End-to-End Mammography Report Generation with Vision–Language Models and Public Multi-Institutional Datasets

Adarsh Bhandary Panambur, Sebastian Wind, Siming Bayer, Andreas Maier

**Abstract**—Breast cancer is the most frequently diagnosed malignancy among women globally and presents a growing challenge for healthcare systems. While early detection and accurate lesion assessment are vital, traditional radiology workflows are labor-intensive and prone to variability. Existing AI solutions in breast imaging are task-specific and lack the ability to generate comprehensive radiology reports. Vision-language models (VLMs), such as CLIP and BLIP, offer a promising direction for unifying visual understanding with natural language generation. We propose MammoBLIP, an end-to-end framework for automated mammography report generation based on the MedBLIP architecture. We curated a dataset of 81,076 images from five public sources—VinDR-Mammo, RSNA, CMMD, InBreast, and KAU—with standardized clinical annotations. Images are processed through a frozen EVA-CLIP Vision Transformer, followed by a lightweight transformer and projection layers to produce vision embeddings. These are aligned with text embeddings using a contrastive loss. A GPT-2–based BioMedLM model generates reports conditioned on these visual features. Only the transformer and projection heads are trained, keeping backbone models frozen to ensure computational efficiency. MammoBLIP achieves strong results across datasets, with an overall BLEU score of 65.36, ROUGE-1 of 0.75, BERT-F1 of 0.88, and SBERT similarity of 0.91. The generated reports are clinically coherent, with an average Flesch–Kincaid grade of 7. These findings provide initial insights into automated report generation and lay the groundwork for developing robust vision–language models tailored to clinical imaging tasks. Our results highlight MammoBLIP’s potential to streamline workflows, enhance diagnostic consistency, and aid in radiologist training in future applications.

**Index Terms**—Mammography, Vision-Language Models, Large Language Models, Report Generation

## I. INTRODUCTION

**B**REAST cancer remains the most frequently diagnosed malignancy among women globally, imposing a significant and escalating burden on healthcare systems. Early detection and accurate characterization of lesions are vital for enhancing patient prognoses. Existing artificial intelligence (AI) approaches in breast imaging primarily focus on specific modalities and isolated tasks, such as lesion detection, risk stratification, or malignancy classification. While these AI

systems effectively identify regions of concern and recommend BI-RADS categories, they lack capabilities for generating comprehensive, structured radiology reports or synthesizing multiple clinical attributes—such as breast density, lesion morphology, and biopsy recommendations—into cohesive, integrated insights. Recent advancements in vision–language models through methods like CLIP, BLIP, and large vision language models (VLMs) have the potential to address these limitations through end-to-end image-to-text generation frameworks, integrating sophisticated visual feature extraction with advanced natural-language synthesis [1]. Such frameworks could significantly streamline radiology workflows by automating detailed report composition, thus enabling radiologists to allocate more time to patient care and critical diagnostic decisions. Moreover, automated report generation may extend breast cancer screening capabilities, particularly in regions with limited availability of radiology specialists, and can also support in training radiologists. Current approaches have so far not been extensively researched in Mammography to support the generation of complete, structured clinical reports.

In this study, we introduce MammoBLIP, an innovative, end-to-end framework for automated mammography report generation built upon the MedBLIP architecture [1]. We curated a comprehensive, multi-institutional dataset from five publicly available repositories [2]–[6], applying a unified preprocessing and label-to-report transformation pipeline to generate consistent, paired image-report data. Specifically, our contributions encompass: (1) The curation of a robust mammography dataset comprising 81,076 images with standardized, clinically relevant annotations; and (2) the introduction of MammoBLIP, a method that leverages vision foundation models and LLM capabilities to generate clinically consistent and comprehensive mammography reports directly from imaging inputs.

## II. METHODS

Figure 1 presents the MammoBLIP report generation pipeline. We curated 81,076 mammograms from VinDR-Mammo, RSNA, CMMD, InBreast, and KAU, applying unified preprocessing and annotation standards. Stratified sampling ensured consistent distributions of clinical attributes across train (59,089), validation (8,745), and test (13,242) sets [2]–[6]. Images were converted from DICOM to PNG, background black regions cropped, resized, and intensity-normalized for uniform resolution. Text data (radiology reports

A. B. Panambur, S. Wind, S. Bayer and A. Maier are with the Pattern Recognition Lab, FAU Erlangen-Nürnberg, Erlangen, Germany (e-mail: {adarsh.bhandary.panambur, sebastian.wind, siming.bayer, andreas.maier}@fau.de).

S. Wind is also with Erlangen National High Performance Computing Center, FAU Erlangen-Nürnberg, Erlangen, Germany.

S. Bayer is also with Siemens Healthineers, Erlangen, Germany

This paper is currently under review.

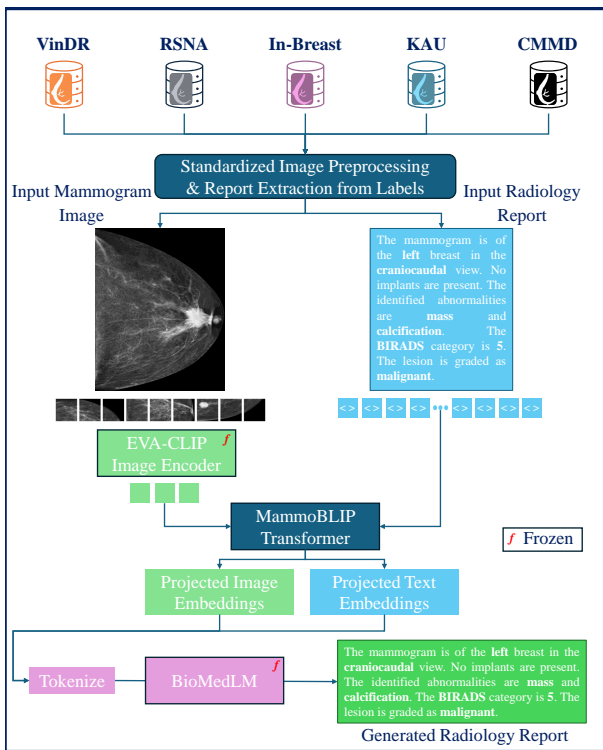


Fig. 1. Overview of the MammoBLIP Report Generation Pipeline.

and QA templates) were lowercased, cleaned, and standardized to 100 tokens. Each image was encoded via a frozen EVA-CLIP ViT to extract patch features, which were processed by a lightweight MedBLIP transformer into 256-D vision embeddings [1]. The same transformer encoded textual descriptions, projecting them into a shared 256-D space. A symmetric contrastive loss enforced alignment between paired image–text embeddings. For report generation, the task was framed as autoregressive language modeling with frozen BioMedLM (GPT-2) [7]. The vision embedding was injected between the prompt and target report tokens. Cross-entropy loss was computed only over the report segment. During training, only the MammoBLIP transformer and projection heads were updated. We used AdamW with cosine annealing, 10% warm-up, mixed precision, and gradient clipping (max norm 1.0). Training ran for 10 epochs on a single NVIDIA Quadro RTX 8000 GPU, with validation monitoring and qualitative logging every 500 steps.

### III. RESULTS AND DISCUSSION

Table I shows the main evaluation metrics across five mammography datasets. We utilize a combination of lexical (BLEU, ROUGE-1, ROUGE-2, METEOR), semantic (BERT-F1, BERT-R, SBERT cosine similarity), and readability (Flesch–Kincaid grade) metrics to capture both surface-level overlap, deeper conceptual alignment, and clinical clarity. Overall, the model achieved BLEU 65.36, ROUGE-1 0.75, ROUGE-2 0.70, METEOR 0.76, BERT-F1 0.88, BERT-R 0.90, SBERT-sim 0.91, and FK-grade 7.04 over 13, 242 report pairs. Performance peaked on the KAU dataset (BLEU 85.86, ROUGE-1 0.92, BERT-F1 0.96), reflecting its standardized

TABLE I  
VLM EVALUATION METRICS BY DATASET.

Metric	CMMD (560)	InBreast (63)	KAU (334)	RSNA (8285)	VinDR (4000)	Overall (13242)
BLEU	56.12	53.01	85.86	60.66	75.45	65.36
ROUGE-1	0.71	0.70	0.92	0.71	0.84	0.75
ROUGE-2	0.66	0.65	0.90	0.63	0.82	0.70
METEOR	0.76	0.79	0.95	0.69	0.90	0.76
BERT-F1	0.91	0.92	0.98	0.87	0.96	0.90
BERT-R	0.86	0.86	0.96	0.86	0.92	0.88
SBERT-sim	0.90	0.93	0.93	0.91	0.92	0.91
FK-grade	9.41	7.75	7.36	6.72	7.35	7.04

protocols and homogeneous lesion descriptions. In contrast, InBreast was most challenging (BLEU 53.01, ROUGE-1 0.70, BERT-F1 0.86); it is both a smaller cohort and was acquired earlier, leading to greater variability in scan parameters, annotation style, and terminology. CMMD and RSNA yielded moderate scores (BLEU 56–61, ROUGE-1 0.71), while VinDR’s larger, high-quality labeled dataset drove robust semantic alignment (SBERT-sim 0.92) and slightly higher readability requirements (FK 7.35). High ROUGE and METEOR indicate reliable n-gram coverage of key clinical terms; BERT metrics and SBERT similarity confirm conceptual fidelity, crucial for diagnosing subtle radiographic findings. A consistent Flesch–Kincaid grade of 7 supports accessible, clinician-friendly language.

### IV. CONCLUSION

In this study, we introduced MammoBLIP, a robust, end-to-end vision-language framework for automated mammography report generation, demonstrating strong performance across multiple clinical datasets. The method effectively synthesizes visual and textual data to produce comprehensive radiology reports. Evaluation metrics confirmed high lexical coherence, semantic alignment, and readability, underscoring MammoBLIP’s potential to significantly streamline clinical workflows and enhance diagnostic consistency. Future research will further refine this approach, optimizing its application to diverse mammography reporting scenarios and performing clinically relevant evaluations.

### REFERENCES

- Chen, Q., and Y. Hong, “MedBlip: Bootstrapping Language–Image Pre-Training from 3D Medical Images and Texts,” in *Proceedings of the Asian Conference on Computer Vision*, 2024. Available: <https://github.com/Qtybc/MedBLIP/tree/main>
- Nguyen, H. T., et al., “VinDr-Mammo: A Large-Scale Benchmark Dataset for Computer-Aided Diagnosis in Full-Field Digital Mammography,” *Scientific Data*, vol. 10, no. 1, p. 277, 2023.
- Moreira, I. C., et al., “INbreast: Toward a Full-Field Digital Mammographic Database,” *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.
- Alsolami, A. S., et al., “King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD),” *Data*, vol. 6, no. 11, p. 111, 2021.
- Cai, H., et al., “An Online Mammography Database with Biopsy-Confirmed Types,” *Scientific Data*, vol. 10, no. 1, p. 123, 2023.
- Carr, C., et al., “RSNA Screening Mammography Breast Cancer Detection,” Kaggle Competition, 2022. [Online]. Available: <https://kaggle.com/competitions/rsna-breast-cancer-detection>
- Bolton, Elliot, et al. “Biomedlm: A 2.7 b parameter language model trained on biomedical text.” arXiv preprint arXiv:2403.18421 (2024).