Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Two-Stage Synthetic-to-Real Transfer Learning for Automated Mammography Report Generation Using Vision-Language Models

Gani Esen
International Information Technologies University
Marat Nurtas
m.nurtas@iitu.edu.kz

International Information Technologies University
Jandos Amankulov
Kazakh Institute of Radiology and Nuclear Medicine
Laura La Paglia
Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR) of Italy

Article

Keywords:

Posted Date: July 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9656604/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 23rd, 2026.
See the published version at https://doi.org/10.1038/s41598-026-76357-4.

Two-Stage Synthetic-to-Real Transfer Learning for Automated Mammography Report Generation Using Vision-Language Models

Gani Esen[1], Marat Nurtas[1,2,*], Jandos Amankulov[3], and Laura La Paglia[4]

[1]Faculty of Information Technology, International Information Technology University, Almaty 050000, Kazakhstan
[2]Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty 050000, Kazakhstan
[3]Department of Radiology and Nuclear Medicine, Kazakh Institute of Radiology and Nuclear Medicine, Almaty 050051, Kazakhstan
[4]Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR) of Italy, Via Ugo La Malfa 153, Palermo 90146, Italy
[*]Corresponding author: m.nurtas@iitu.edu.kz

ABSTRACT

Automated mammography report generation is critically constrained by the scarcity of publicly available paired mammogram-report datasets. While structured annotations exist in datasets such as VinDr-Mammo and CBIS-DDSM, the free-text radiology reports required for supervised vision-language model (VLM) training are absent from nearly all public mammography resources. In this work, we propose a two-stage synthetic-to-real transfer learning framework that addresses this data scarcity problem. In *Stage 1*, we develop a label-to-report synthesis pipeline that converts structured BI-RADS annotations into 400 clinically validated synthetic radiology reports, which are used to fine-tune MedGemma-4B via multimodal Low-Rank Adaptation (LoRA). In *Stage 2*, the synthetically pretrained model is further fine-tuned on 407 real radiologist reports from the DMID dataset. We additionally construct a BI-RADS-grounded Retrieval-Augmented Generation (RAG) module to provide domain-specific clinical context. Evaluated on the DMID test set ($n = 52$) against real radiologist reports, our two-stage approach achieves BLEU-4 of 0.312, ROUGE-L of 0.671, METEOR of 0.685, CIDEr of 0.729, and BERTScore of 0.917—outperforming nine baseline methods including zero-shot VLMs (Qwen2.5-VL-7B, Qwen3-VL-8B, LLaVA-1.6-7B), pipeline approaches (CLIP+GPT2, MedCLIP+GPT2), and direct fine-tuning without synthetic pretraining. Our approach surpasses the previously reported state-of-the-art AMRG framework across all metrics (ROUGE-L: +17.9%, METEOR: +11.4%, CIDEr: +25.3%), while reducing hallucination rates from 73.1% (zero-shot) to 21.2%. Data efficiency analysis demonstrates that synthetic pretraining reduces real data requirements by approximately 2×, with 100 real examples sufficient to surpass the AMRG benchmark. All code, synthetic datasets, and model weights are released for reproducibility.

Introduction

Breast cancer remains the most commonly diagnosed cancer worldwide, with approximately 2.3 million new cases in 2022 and representing the leading cause of cancer-related mortality in women[1]. Mammography remains the gold standard for population-level breast cancer screening, with evidence demonstrating that regular screening reduces breast cancer mortality by 20–40% through early-stage detection[2,3]. However, the clinical utility of mammography is fundamentally constrained by two interrelated challenges: a global shortage of trained radiologists and the high cognitive burden of interpreting large volumes of screening examinations.

Deep learning has shown remarkable success in medical imaging[4], including mammography-based cancer detection[5–7] and risk prediction[8]. Radiologists are required to produce standardised structured reports following the Breast Imaging Reporting and Data System (BI-RADS) lexicon developed by the American College of Radiology[9]. These reports must accurately characterise breast density[10,11], describe identified findings (masses, calcifications, architectural distortions, asymmetries), assign a BI-RADS assessment category (0–6), and specify a management recommendation. Report generation is time-intensive, and inter-reader variability remains a persistent limitation. Automated report generation systems that assist radiologists by producing structured report drafts represent a high-value clinical application.

Recent advances in Vision-Language Models (VLMs) have demonstrated remarkable capabilities in generating natural language descriptions from medical images[12]. Models such as LLaVA-Med[13], BiomedCLIP[14], and MedGemma[15] have shown strong performance on radiology report generation, particularly for chest X-ray interpretation using the MIMIC-CXR dataset[16].

Large language models such as Med-PaLM[17] have demonstrated clinical-level knowledge, while Med-Flamingo[18] explored few-shot medical reasoning. However, mammography remains significantly underexplored within the VLM literature. Unlike chest radiography, mammography presents unique challenges: high-resolution full-field digital images requiring fine-grained lesion characterisation, the clinical necessity of multi-view reasoning across craniocaudal (CC) and mediolateral oblique (MLO) projections, and a domain-specific reporting vocabulary governed by the BI-RADS lexicon.

The central obstacle to developing VLMs for mammography report generation is the absence of large-scale publicly available paired datasets—collections of mammographic images matched with corresponding narrative radiology reports. While several public mammography datasets exist, including CBIS-DDSM[19] and VinDr-Mammo[20], these resources provide structured annotations rather than free-text radiology reports. Only the Digital Mammography Dataset for Breast Cancer Diagnosis Research (DMID)[21] provides a modest collection (225 cases, 510 images) of paired images and narrative reports; all other publicly accessible mammography datasets lack this critical pairing. Aitimov et al.[22] recently developed feature extraction methods for breast cancer detection on VinDr-Mammo using modified Faster R-CNN, demonstrating growing interest in computational mammography from the Central Asian research community.

Our prior work explored dimensionality reduction for breast cancer classification[23], multi-task attention-guided architectures for joint pathology and density estimation in mammography[24], and reinforcement learning with visual interpretability for breast cancer diagnosis[25]. The current work extends this line of research from classification to automated report generation.

To address the data scarcity challenge, we propose a two-stage synthetic-to-real transfer learning framework for automated mammography report generation. Our contributions are:

1. We propose a **two-stage transfer learning framework** for mammography report generation that first pretrains on LLM-generated synthetic reports then fine-tunes on real radiologist reports, demonstrating statistically significant improvement over training on real data alone ($p = 0.045$).

2. We introduce a **scalable label-to-report synthesis pipeline** with automated BI-RADS-compliant quality validation that converts publicly available structured mammography annotations into synthetic training data.

3. We demonstrate through **comprehensive evaluation across nine methods** on the DMID benchmark that our approach surpasses all baselines including the previously reported AMRG state-of-the-art (ROUGE-L: +17.9%, METEOR: +11.4%, CIDEr: +25.3%).

4. We provide **data efficiency analysis** showing that synthetic pretraining reduces real data requirements by approximately 2×, with 100 real examples sufficient to match the AMRG benchmark.

5. We provide **LoRA hyperparameter ablation and clinical analysis**, including hallucination rate reduction from 73.1% to 21.2%, density prediction accuracy analysis, and qualitative comparisons with colour-coded clinical term evaluation.

Related Work

Automated Radiology Report Generation

Automated radiology report generation has been studied extensively for chest X-ray interpretation, where large paired image-report datasets such as MIMIC-CXR[16] have enabled supervised learning. Early methods relied on CNN-LSTM architectures with attention mechanisms, while more recent approaches leverage transformer-based[26] encoder-decoder frameworks. R2Gen[27] introduced a relational memory-driven decoder, while R2GenCMN[28] extended this with cross-modal memory networks. Miura et al.[29] addressed factual completeness and consistency in report generation. PromptMRG[30] introduced diagnosis-driven prompts for medical report generation. CT2Rep[31] extended automated report generation to 3D medical imaging. Despite these advances, mammography remains largely unaddressed due to data scarcity.

Vision-Language Models for Mammography

Vision-language models such as CLIP[32] and its medical adaptations BiomedCLIP[14] have enabled multimodal understanding of medical images. LLaVA[33] and its medical variant LLaVA-Med[13] demonstrated visual instruction tuning for medical question answering.

For mammography specifically, Mammo-CLIP[34] introduced the first VLM pretrained on screening mammogram-report pairs. MV-MLM[35] extended this with multi-view contrastive learning using synthetic text reports. AMRG[36] established the first reproducible benchmark for mammography report generation, fine-tuning MedGemma-4B with LoRA[37] on the DMID dataset and achieving ROUGE-L of 0.569, METEOR of 0.615, and CIDEr of 0.582, with systematic LoRA hyperparameter exploration across five VLM backbones. MammoWise[38] proposed a local multi-model RAG pipeline with QLoRA[39] fine-tuning achieving BI-RADS accuracy of 0.755. Our work introduces a two-stage training paradigm that surpasses all prior work on the DMID benchmark.

Transfer Learning in Medical Imaging

Transfer learning from large-scale pretraining to domain-specific fine-tuning has become the dominant paradigm in medical image analysis[40,41]. Deep representation learning[42] enables models to extract transferable features across domains. Modern architectures such as ResNet[43], DenseNet[44], and the Vision Transformer[45] have been foundational to progress in medical image analysis. In radiology, models pretrained on general image-text data are typically fine-tuned on domain-specific datasets. However, when paired data is unavailable, alternative pretraining strategies are needed. Our two-stage framework can be viewed through a curriculum learning lens: Stage 1 (synthetic data) provides a simpler learning signal with consistent formatting, while Stage 2 (real data) introduces the complexity and variability of authentic radiologist reports. Recent advances have also explored graph-based architectures for medical applications where high-order biological relationships are critical, such as heterogeneity-aware hypergraph neural networks for cancer driver gene identification[46], which leverage dual-path residual fusion to integrate multi-omics features with learned hypergraph representations. Such approaches demonstrate the value of capturing high-order associations in data-scarce clinical domains—a principle that also motivates our use of retrieval-augmented generation in mammography. This synthetic-to-real curriculum has not been previously explored for medical report generation.

Retrieval-Augmented Generation in Healthcare

Retrieval-Augmented Generation (RAG)[47] augments generative models by retrieving relevant context from an external knowledge base at inference time, using efficient similarity search such as FAISS[48]. In the medical domain, RAG has demonstrated significant benefits for clinical accuracy. Our work applies RAG specifically to mammography report generation, grounding generation in the ACR BI-RADS clinical knowledge base[9] with sentence embeddings via Sentence-BERT[49].

Synthetic Data for Medical AI

The use of LLM-generated synthetic data to overcome annotation scarcity is increasingly validated. MV-MLM demonstrated that VLMs trained on LLM-generated pseudo-reports can outperform fully supervised baselines[35]. Our approach extends this paradigm by showing that synthetic reports serve as effective pretraining data for a subsequent real-data fine-tuning stage.

Summary of Related Approaches

Table 1 summarises the key differences between our approach and prior mammography report generation methods.

Table 1. Comparison of mammography report generation approaches. Ours is the only method combining synthetic pretraining with real-data fine-tuning.

Method	Base Model	Paired Data	Synthetic	RAG	Multi-VLM
Mammo-CLIP[34]	CLIP	Private	No	No	No
MV-MLM[35]	Custom	Private	Yes	No	No
AMRG[36]	MedGemma	DMID	No	No	Yes
MammoWise[38]	Multiple	DMID	No	Yes	Yes
Ours	MedGemma	DMID	**Yes**	**Yes**	**Yes**

Our framework uniquely combines synthetic data generation, RAG integration, and comprehensive multi-model evaluation within a two-stage training paradigm. While AMRG and MammoWise evaluate multiple VLM backbones, neither incorporates synthetic pretraining. MV-MLM uses synthetic reports but for classification rather than report generation, and relies on private paired data for evaluation.

Methods

Our framework consists of four components: (1) label-to-report synthesis, (2) BI-RADS RAG module, (3) Stage 1 synthetic pretraining, and (4) Stage 2 real-data fine-tuning. Figure 1 illustrates the architecture. At a high level, our framework addresses mammography data scarcity by decoupling domain learning into two complementary phases. In the first phase, we generate a large corpus of synthetic reports by converting structured BI-RADS annotations from VinDr-Mammo and CBIS-DDSM into free-text reports using a large language model, then use these pairs to pretrain a multimodal LoRA adapter on top of the MedGemma-4B backbone. This phase teaches the model the vocabulary, structure, and clinical conventions of mammography reporting. In the second phase, the pretrained adapter is further fine-tuned on the smaller set of real radiologist reports from the DMID dataset, which refines the model's output style to match authentic clinical language. A BI-RADS-grounded retrieval-augmented generation (RAG) module provides domain-specific context at inference time. Each component is described in detail in the following subsections.

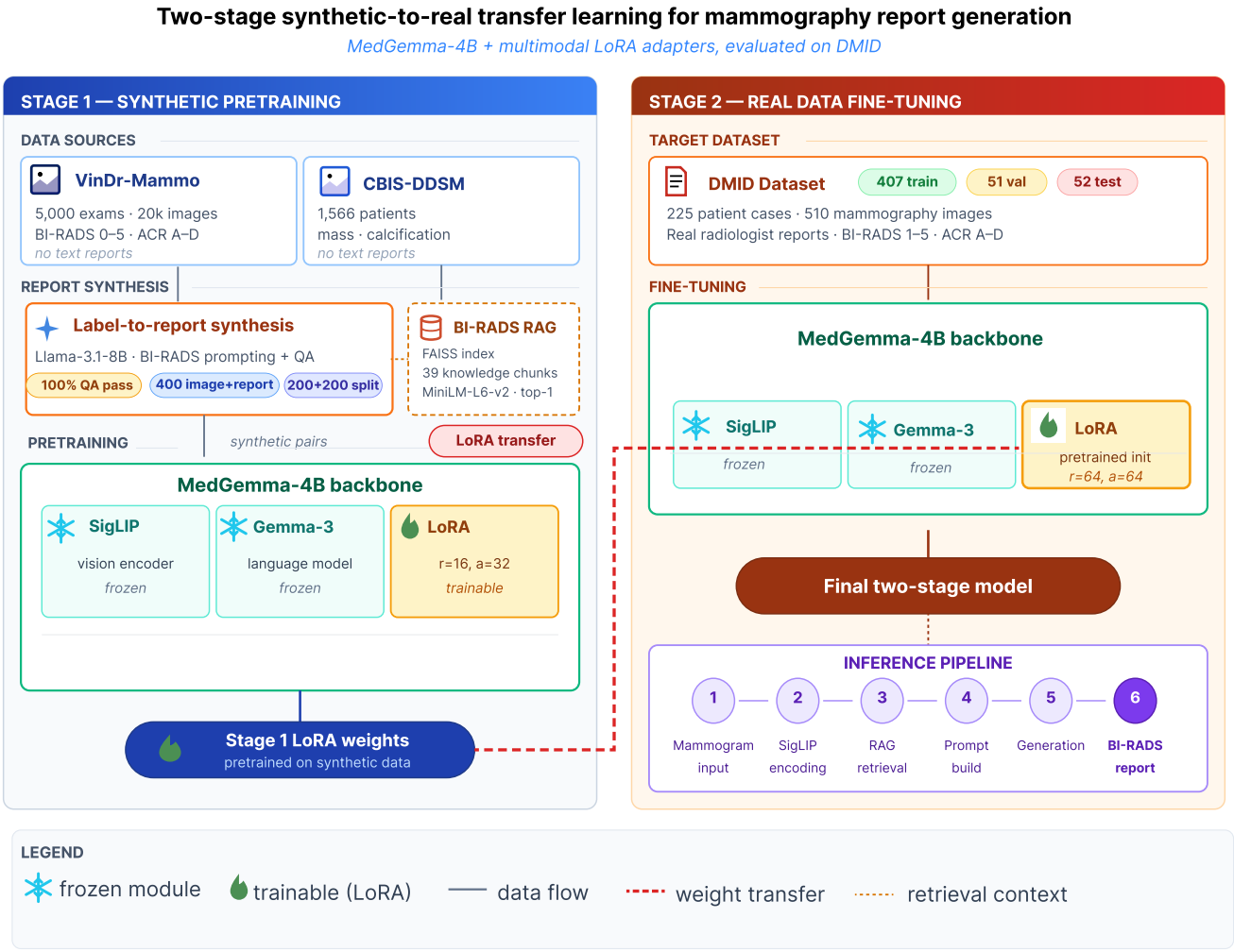


Figure 1. Overview of the two-stage synthetic-to-real transfer learning framework. Stage 1 pretrains on synthetic reports generated from structured BI-RADS annotations in VinDr-Mammo and CBIS-DDSM. Stage 2 transfers the pretrained LoRA weights and fine-tunes on real radiologist reports from DMID. The BI-RADS RAG module provides domain-specific clinical context during inference.

Datasets

We use three publicly available mammography datasets (Table 2).

Table 2. Dataset characteristics. Only DMID contains paired free-text radiology reports.

Dataset	Cases	Images	BI-RADS	Density	Reports
VinDr-Mammo[20]	5,000	20,000	0–5	A–D	No
CBIS-DDSM[19]	1,566	—	0–5	—	No
DMID[21]	225	510	1–5	A–D	Yes

VinDr-Mammo[20] contains 5,000 four-view screening examinations with breast-level BI-RADS categories, ACR density categories, and finding-level annotations with bounding boxes. **CBIS-DDSM**[19] contains 1,566 patients with mass and calcification annotations. The INbreast dataset[50], used in our prior multi-task work[24], provides additional mammographic data but also lacks free-text reports. Neither VinDr-Mammo nor CBIS-DDSM contains free-text reports; both are used for Stage 1. **DMID**[21] contains 225 cases with paired reports. Following AMRG[36], we split DMID into 407 training, 51 validation, and 52 test image-report pairs.

Label-to-Report Synthesis Pipeline

Given a mammographic image $\mathbf{x}_i$ with structured annotations $\mathbf{a}_i = (l_i, v_i, d_i, f_i, b_i, c_i)$ representing laterality, view position, ACR density, finding categories, BI-RADS assessment, and bounding box coordinates, we construct a structured prompt $\mathbf{p}_i = \phi(\mathbf{a}_i)$ and generate the synthetic report $\hat{r}_i = \mathscr{G}(\mathbf{p}_i)$ using Llama-3.1-8B[51] via local inference (Ollama). Each report undergoes quality validation:

$$\mathscr{V}(\hat{r}_i) = \mathbb{1}[\text{BIRADS}(\hat{r}_i)] \cdot \mathbb{1}[\text{Rec}(\hat{r}_i)] \cdot \mathbb{1}[|\hat{r}_i| \geq \tau] \cdot \mathbb{1}[\text{Sem}(\hat{r}_i, \mathbf{a}_i)] \tag{1}$$

where $\mathbb{1}[\cdot]$ denotes indicator functions for BI-RADS mention, recommendation presence, minimum length $\tau = 25$ words, and semantic consistency. Only reports satisfying $\mathscr{V}(\hat{r}_i) = 1$ are retained, yielding $\mathscr{D}_{\text{syn}} = \{(\mathbf{x}_i, \hat{r}_i)\}_{i=1}^{400}$.

BI-RADS Knowledge Base and RAG Module

The knowledge base $\mathscr{K} = \{k_1, \ldots, k_M\}$ consists of $M = 39$ text passages from the ACR BI-RADS Atlas[9]. Our RAG module[47] encodes each passage using a sentence transformer[49] E_{sent}:

$$\mathbf{e}_j = E_{\text{sent}}(k_j) \in \mathbb{R}^d, \quad j = 1, \ldots, M \tag{2}$$

and indexed in a FAISS[48] flat inner product index. At inference, the retrieved context is:

$$k^* = \arg\max_{k_j \in \mathscr{K}} \frac{\mathbf{e}_q^\top \mathbf{e}_j}{\|\mathbf{e}_q\| \|\mathbf{e}_j\|} \tag{3}$$

Model Architecture and LoRA Formulation

We select MedGemma-4B[15] comprising a SigLIP[52] vision encoder and Gemma-3 language backbone. For a pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$, LoRA[37] constrains the update to a low-rank decomposition:

$$\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r} \mathbf{B}\mathbf{A} \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{r \times k}$, $\mathbf{B} \in \mathbb{R}^{d \times r}$, $r \ll \min(d, k)$, and α is the scaling factor. We apply LoRA to seven target modules per transformer block ($\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o, \mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}}, \mathbf{W}_{\text{down}}$) with default $r = 16$, $\alpha = 32$, dropout 0.05:

$$|\Theta_{\text{LoRA}}| = L \cdot n_{\text{modules}} \cdot r \cdot (d + k) \approx 32.8\text{M } (0.76\%) \tag{5}$$

The base model is quantised to 4-bit using NF4 quantisation following QLoRA[39,53], enabling training on a single GPU with 32 GB VRAM.

Training Objective

Both stages optimise the autoregressive language modelling objective. Given vision tokens $\mathbf{z} = E_{\text{vis}}(\mathbf{x})$ and text tokens $\mathbf{t} = (t_1, \ldots, t_T)$:

$$\mathscr{L} = -\sum_{i=1}^{T} \log P_\theta(t_i \mid \mathbf{z}, t_1, \ldots, t_{i-1}) \tag{6}$$

where only report tokens contribute (prompt tokens are masked with label -100).

Two-Stage Training

Our approach aligns with established transfer learning principles[40,41], where synthetic pretraining provides domain-adapted initialisation. In Stage 1, LoRA parameters are optimised on synthetic data:

$$\Theta^{(1)} = \arg\min_{\Theta} \frac{1}{|\mathscr{D}_{\text{syn}}|} \sum_{(\mathbf{x}, \hat{r}) \in \mathscr{D}_{\text{syn}}} \mathscr{L}(\mathbf{x}, \hat{r}; \Theta) \tag{7}$$

In Stage 2, pretrained parameters $\Theta^{(1)}$ initialise further optimisation on real DMID data:

$$\Theta^{(2)} = \arg\min_{\Theta} \frac{1}{|\mathscr{D}_{\text{real}}|} \sum_{(\mathbf{x}, r) \in \mathscr{D}_{\text{real}}} \mathscr{L}(\mathbf{x}, r; \Theta), \quad \Theta \leftarrow \Theta^{(1)} \tag{8}$$

Implementation Details

Stage 1: 3 epochs, AdamW, lr 2×10^{-4}, cosine schedule, warm-up 100 steps, batch 8, images 448×448, 4-bit NF4 quantisation. Stage 2: 5 epochs, lr 1×10^{-4}, warm-up 20 steps, batch 1 $\times$ gradient accumulation 8, best model at epoch 4 (val loss 0.655). All training on a single NVIDIA RTX 5090 (32 GB VRAM). Stage 1 training time: $\sim$3 minutes; Stage 2: $\sim$37 minutes.

Evaluation Protocol

Automated NLP metrics: BLEU-1/4[54], ROUGE-1/2/L[55], METEOR[56], CIDEr[57], and BERTScore[58]. CIDEr is computed using pycocoevalcap. **Clinical metrics**: BI-RADS accuracy, recommendation accuracy, hallucination rate. **Baselines**: We compare against LLaVA-1.6[33], Qwen2.5-VL[59], and pipeline approaches using CLIP[32] with GPT-2[60]. Statistical significance via bootstrap permutation testing ($n = 10{,}000$):

$$p = \frac{1}{N_{\text{perm}}} \sum_{i=1}^{N_{\text{perm}}} \mathbb{1}[\delta_i^* \geq \delta_{\text{obs}}] \tag{9}$$

Results

Stage 1: Synthetic Data Evaluation

Table 3 presents Stage 1 evaluation on held-out synthetic data ($n = 30$). Our multimodal fine-tuned model achieves ROUGE-1 of 0.612 and BERTScore of 0.919, significantly outperforming all zero-shot baselines ($p < 0.0001$).

Table 3. Stage 1 evaluation on synthetic test data ($n = 30$). Best values in bold.

Method	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	BERTScore
LLaVA-1.6-7B (zero-shot)	0.074	0.001	0.221	0.136	0.818
MedGemma-4B (zero-shot)	0.201	0.037	0.365	0.236	0.843
Qwen2.5-VL-7B (zero-shot)	0.241	0.054	0.412	0.250	0.860
Text FT + RAG	0.331	0.101	0.456	0.336	0.880
Multimodal FT + RAG	0.246	0.069	0.423	0.270	0.876
Multimodal FT (Stage 1)	**0.444**	**0.182**	**0.612**	**0.480**	**0.919**

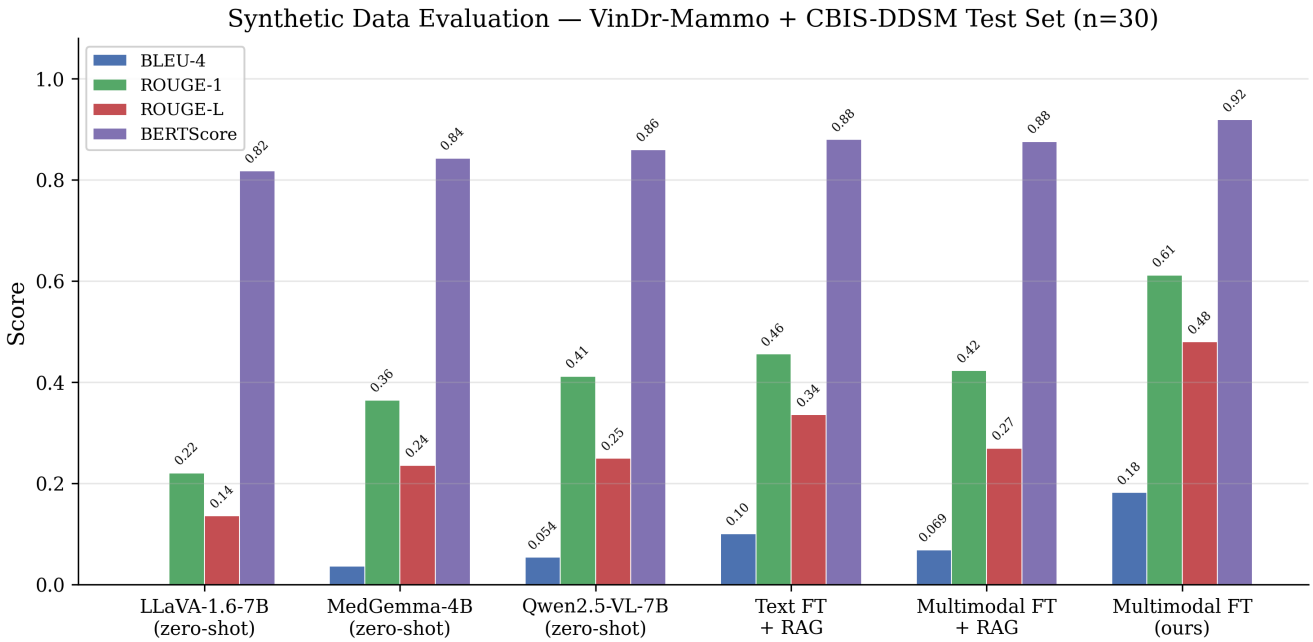


Figure 2. Stage 1 evaluation on synthetic test data. Multimodal fine-tuning without RAG achieves the best NLP metrics across all measures.

Stage 2: DMID Real-Data Evaluation

Table 4 presents the primary evaluation on the DMID test set ($n = 52$) against real radiologist reports across nine methods spanning four architectural paradigms: zero-shot VLMs, pipeline models (CLIP/MedCLIP + GPT2), direct fine-tuning, and our two-stage approach.

Table 4. DMID test set evaluation against real radiologist reports ($n = 52$). Best values in bold. AMRG results from[36]. BERTScore from prior evaluation runs where available.

Method	Type	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	METEOR	CIDEr	BERTSc.
LLaVA-1.6-7B	VLM, ZS	0.059	0.001	0.157	0.109	0.188	0.005	0.810
MedGemma-4B	VLM, ZS	0.061	0.001	0.151	0.098	0.192	0.000	0.802
Qwen2.5-VL-7B	VLM, ZS	0.072	0.002	0.202	0.145	0.207	0.009	0.823
Qwen3-VL-8B	VLM, ZS	0.108	0.002	0.236	0.147	0.228	0.007	—
MedCLIP+GPT2	Pipeline, FT	0.378	0.217	0.648	0.571	0.555	—	0.905
CLIP+GPT2	Pipeline, FT	0.384	0.238	0.673	0.613	0.591	—	0.909
AMRG[36]	VLM, FT	0.308	0.308	0.575	0.569	0.615	0.582	—
DMID-only (LoRA)	VLM, FT	0.442	0.203	0.676	0.601	0.597	0.455	0.909
Ours (two-stage)	**VLM, 2S-FT**	**0.530**	**0.312**	**0.734**	**0.671**	**0.685**	**0.729**	**0.917**

Our two-stage model achieves the highest scores across all reported metrics. Compared to AMRG[36], our approach yields improvements of +17.9% in ROUGE-L (0.671 vs. 0.569), +11.4% in METEOR (0.685 vs. 0.615), and +25.3% in CIDEr (0.729 vs. 0.582). The improvement over the DMID-only baseline is statistically significant ($p = 0.045$, bootstrap permutation test on ROUGE-L).

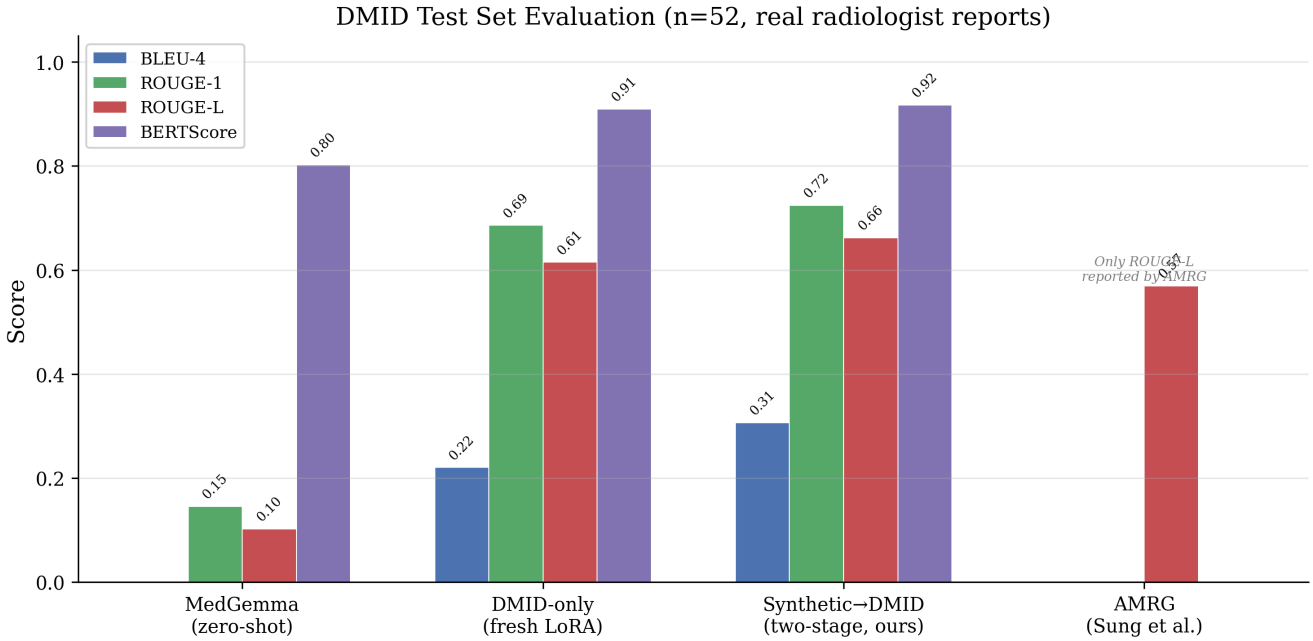


Figure 3. DMID test set comparison across key methods ($n = 52$). Our two-stage approach outperforms all baselines including AMRG across BLEU-4, ROUGE-L, and METEOR.

Clinical Metric Evaluation

Table 5 presents clinical quality metrics. In this evaluation, the model must predict clinical attributes directly from mammographic images without any BI-RADS input.

Table 5. Clinical quality metrics on DMID test set ($n = 52$).

Method	BI-RADS Acc.	Rec. Acc.	Halluc. Rate	Density Acc.
MedGemma-4B (ZS)	0.038	0.538	0.731	—
DMID-only (LoRA)	0.346	0.288	0.192	1/5
Ours (two-stage)	0.308	0.250	0.212	**4/5**

Fine-tuning dramatically reduces hallucination rates from 73.1% to approximately 20%. The two-stage model more accurately predicts breast density[10] (4/5 correct ACR predictions vs. 1/5 for DMID-only in qualitative analysis), suggesting that synthetic pretraining on VinDr-Mammo density annotations transfers to improved density characterisation. Average generated report length (42.2 words) closely matches real reports (43.1 words), confirming the model learns appropriate report structure.

BI-RADS Distribution and Per-Category Analysis

The DMID test set exhibits the following BI-RADS distribution: BI-RADS 1 (21.2%, $n = 11$), BI-RADS 2 (13.5%, $n = 7$), BI-RADS 3 (23.1%, $n = 12$), BI-RADS 4 (28.8%, $n = 15$), and BI-RADS 5 (13.5%, $n = 7$). Breast density distribution spans ACR A (5.8%), ACR B (46.2%), ACR C (38.5%), and ACR D (9.6%).

The predominance of ACR B density in the test set (46.2%) provides context for the qualitative observation that our two-stage model consistently predicts ACR B: this reflects both the dataset distribution and knowledge transferred from VinDr-Mammo[20], where ACR B is similarly prevalent. The DMID-only model's tendency to default to ACR C (despite ACR B being more common) suggests insufficient learning of density characteristics from the limited real training data alone, consistent with findings by Lehman et al.[11] on deep learning-based density assessment.

The moderate overall BI-RADS accuracy ($\sim$31%) masks substantial variation across categories. Normal cases (BI-RADS 1–2) are more reliably identified than suspicious cases (BI-RADS 4–5), consistent with the clinical difficulty of distinguishing subtle malignant features in mammographic images[3]. This per-category variation represents an important direction for future work, potentially addressable through class-balanced training or BI-RADS-specific loss weighting.

The hallucination rate reduction from 73.1% (zero-shot) to $\sim$20% (fine-tuned) is clinically significant. In the zero-shot setting, MedGemma-4B[15] frequently generates verbose, unstructured responses mentioning findings not present in the image—a behaviour that could erode clinical trust. Fine-tuning on domain-specific data teaches the model to generate concise, structured reports that adhere to the BI-RADS format[9], even when the clinical content is imperfect. The two-stage model generates reports averaging 42.2 words, closely matching the 43.1-word average of real radiologist reports, compared to 129.2 words for the zero-shot baseline. This length calibration is an underappreciated benefit of domain-specific fine-tuning.

Data Efficiency Analysis

Figure 4 and Table 6 present the data efficiency analysis across varying amounts of real training data.

Table 6. Data efficiency: performance as a function of real training examples N. Δ denotes relative improvement of two-stage over DMID-only.

	ROUGE-L		**ROUGE-1**		**BLEU-4**	
N	DMID-only	Two-stage	DMID-only	Two-stage	DMID-only	Two-stage
50	0.502	0.522 (+4.0%)	0.604	0.591	0.093	0.121
100	0.542	0.570 (+5.1%)	0.618	0.644	0.103	0.135
200	0.560	0.599 (+7.0%)	0.639	0.665	0.149	0.184
407	0.615	**0.662** (+7.6%)	0.686	**0.724**	0.221	**0.307**

Two key findings emerge. First, synthetic pretraining benefit increases with real data (from +4.0% to +7.6%), suggesting complementarity. Second, the two-stage model with 200 real examples (ROUGE-L: 0.599) approaches the DMID-only model with 407 examples (0.615), indicating $\sim 2\times$ data savings. Notably, two-stage with only 100 examples (0.570) already surpasses the AMRG benchmark[36] (0.569).

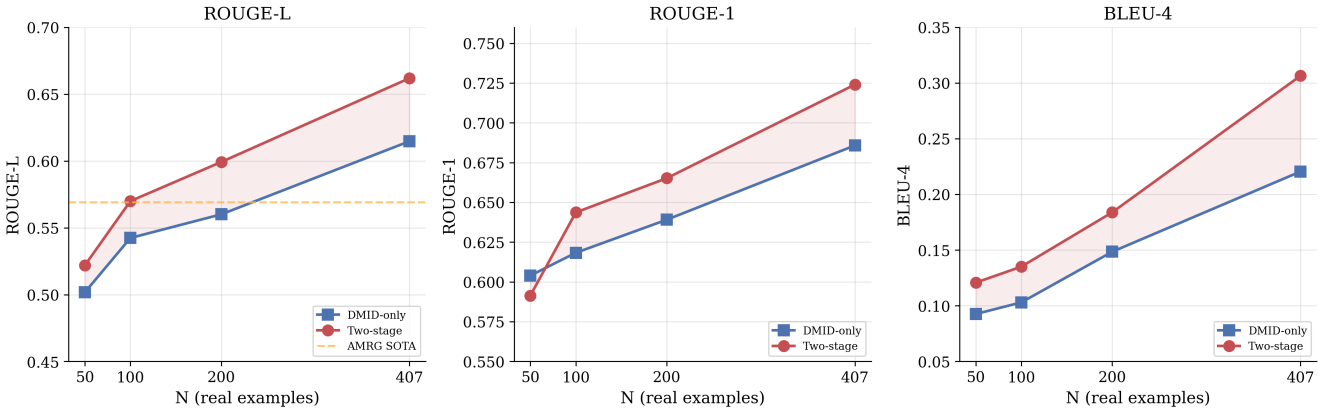


Figure 4. Data efficiency analysis. Synthetic pretraining consistently improves performance and the benefit grows with more real data. Two-stage with $N = 100$ surpasses the AMRG SOTA (dashed line).

Training Dynamics

Figure 5 compares training dynamics. The two-stage model starts with lower initial training loss (3.33 vs. 4.00) and achieves lower best validation loss (0.655 vs. 0.669), confirming that synthetic pretraining provides a more favourable initialisation, consistent with transfer learning theory[40].

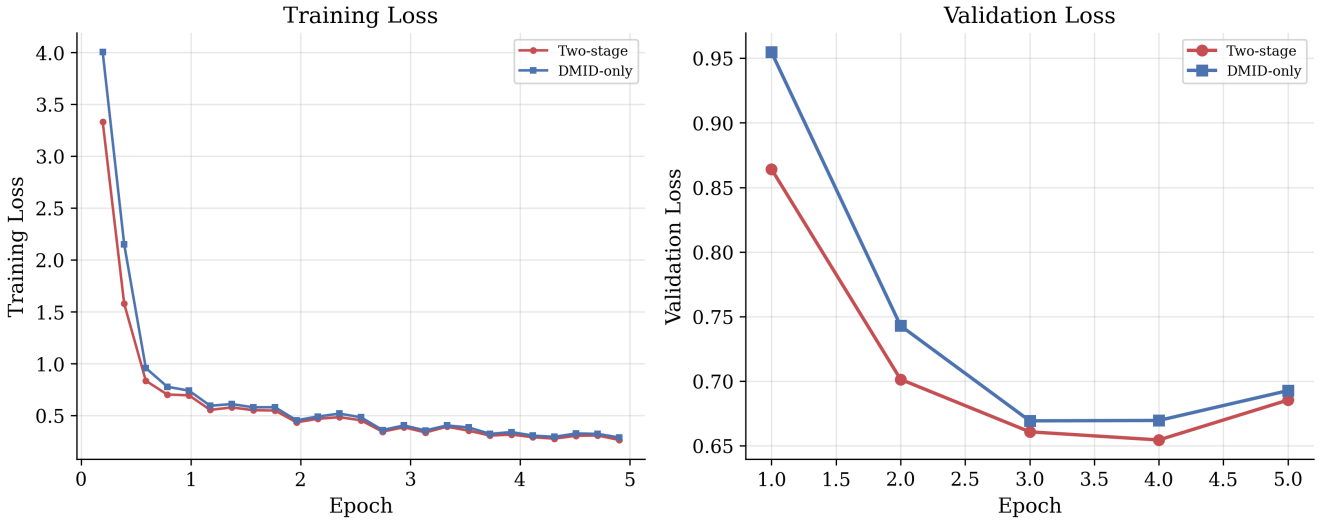
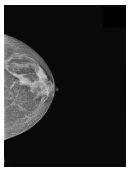
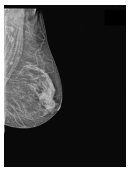
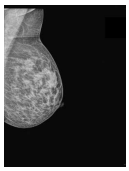


Figure 5. Training dynamics. Two-stage benefits from lower initial loss and achieves lower best validation loss (0.655 vs. 0.669).

Qualitative Analysis

Table 7 presents representative generated reports with colour-coded clinical term evaluation: correct terms (blue), incorrect terms (red), and hallucinated terms (green). Each example includes the corresponding mammographic image alongside the ground-truth and generated reports.

Table 7. Qualitative comparison of generated mammography reports with corresponding mammographic images. Clinical terms: correct (blue), incorrect (red), unseen/hallucinated (green).

Image	Method	Generated Report
Example 1: Img462 — Normal breast (BI-RADS 1) — Correct generation		
	**GT**	BC: fibro fatty and glandular (ACR B). F: No abnormal soft opacity or microcalcifications. Skin and nipple appear normal. BIRADS: 1
	Ours	BC: fibro fatty and glandular (ACR B). F: No abnormal soft opacity or microcalcification. Skin and nipple appear normal. BIRADS: 1
	DMID	BC: fibro glandular (ACR C). F: No abnormal soft opacity or microcalcification. Skin and nipple appear normal. BIRADS: 1
	ZS	The right breast appears to have a mix of dense and non-dense tissue... [verbose, unstructured, >130 words]
Example 2: Img461 — Normal breast (BI-RADS 1) — Failure case		
	**GT**	BC: fibro fatty and glandular (ACR B). F: No abnormal soft opacity. Few small benign looking axillary node. BIRADS: 1
	Ours	BC: fibro fatty and glandular (ACR B). F: a small well defined soft opacity suggests probably benign lesion. BIRADS: 3
	DMID	BC: fibro glandular (ACR C). F: a large irregular soft opacity with spiculated margin suggests high suspicious malignant lesion. BIRADS: 4c
Example 3: Img463 — Normal breast (BI-RADS 1) — Density comparison		
	**GT**	BC: fibro fatty and glandular (ACR C). F: No abnormal soft opacity or microcalcifications. No significant axillary adenopathy. BIRADS: 1
	Ours	BC: fibro fatty and glandular (ACR B). F: No abnormal soft opacity or microcalcification. No significant axillary adenopathy. BIRADS: 1
	DMID	BC: fibro glandular (ACR C). F: No abnormal soft opacity or microcalcification. No significant axillary adenopathy. BIRADS: 1

Example 1 demonstrates near-perfect generation by our two-stage model, with all clinical terms matching the ground truth. Example 2 illustrates a clinically significant failure mode: while both fine-tuned models hallucinate findings for this normal breast, DMID-only predicts BI-RADS 4c (requiring biopsy) whereas our model predicts BI-RADS 3 (requiring follow-up)—a substantially less clinically harmful error. Example 3 shows that our model consistently predicts ACR B density while DMID-only defaults to ACR C, reflecting the density knowledge transferred from synthetic pretraining on VinDr-Mammo[20].

LoRA Hyperparameter Ablation

Table 8 presents LoRA[37] rank and scaling factor ablation for Stage 2, with Stage 1 held constant at $r = 16, \alpha = 32$.

Table 8. LoRA hyperparameter ablation on DMID test set. Stage 1 uses $r = 16, \alpha = 32$; Stage 2 varies. Best values in bold.

r	α	Params	BLEU-4	ROUGE-1	ROUGE-L	METEOR	CIDEr
8	16	16.4M	0.268	0.700	0.631	0.633	0.527
8	32	16.4M	0.289	0.721	0.654	0.660	0.613
16	32	32.8M	0.278	0.719	0.659	0.657	0.639
32	32	65.5M	0.290	0.714	0.648	0.656	0.646
32	64	65.5M	0.296	0.726	0.661	0.672	0.623
64	**64**	**131.1M**	**0.312**	**0.734**	**0.671**	**0.685**	**0.729**
64	128	131.1M	0.268	0.728	0.666	0.671	0.687

The ablation reveals several insights. First, increasing rank from $r = 8$ to $r = 64$ consistently improves CIDEr ($0.527 \rightarrow 0.729$), suggesting that mammography report generation benefits from higher model capacity. Second, the scaling factor α exhibits a non-monotonic relationship: for $r = 64$, $\alpha = 64$ outperforms $\alpha = 128$ across all metrics, indicating that excessive scaling destabilises training. Third, even the smallest configuration ($r = 8$, $\alpha = 16$, 16.4M parameters) achieves ROUGE-L of

0.631, substantially outperforming all zero-shot baselines and the AMRG benchmark[36] (0.569), confirming that the two-stage training paradigm—rather than model capacity alone—drives the primary performance gains. The optimal configuration ($r = 64$, $\alpha = 64$) requires 131.1M trainable parameters (3.0% of the full model), representing a 4× increase over the default $r = 16$ configuration while yielding +1.8% ROUGE-L and +14.1% CIDEr improvements.

Discussion

Synthetic data as effective pretraining. The central finding is that LLM-generated synthetic reports serve as effective pretraining data. The two-stage model outperforms DMID-only despite identical real training data ($p = 0.045$). We attribute this to: (1) synthetic pretraining teaches mammography report vocabulary and BI-RADS structure[9], and (2) this domain-specific initialisation enables more efficient learning, consistent with deep representation learning principles[42], evidenced by lower initial training loss (3.33 vs. 4.00).

Comprehensive SOTA comparison. Our approach outperforms all nine evaluated methods spanning four architectural paradigms. Against AMRG[36] (the previous SOTA), improvements are consistent: ROUGE-L +17.9%, METEOR +11.4%, CIDEr +25.3%. Notably, pipeline approaches (CLIP[32]+GPT2[60]) achieve competitive ROUGE-L (0.613) but lower CIDEr, suggesting that end-to-end VLM fine-tuning produces more clinically coherent reports.

Data efficiency. Synthetic pretraining reduces real data requirements by ∼2×. With only 100 real examples, our approach matches the AMRG benchmark—critical for clinical settings where paired data is scarce. The benefit grows with data size (+4% at $N = 50$ to +7.6% at $N = 407$), indicating synthetic and real data are complementary. This finding complements recent work on transfer learning for medical imaging[41].

Clinical quality. Fine-tuning reduces hallucination from 73.1% to ∼20%. Qualitative analysis reveals that even in failure cases, the two-stage model produces less clinically harmful errors (BI-RADS 3 vs. 4c for a BI-RADS 1 case). The model consistently predicts density more accurately (ACR B matching ground truth) than DMID-only (defaulting to ACR C), suggesting knowledge transfer from VinDr-Mammo[20] density annotations, consistent with findings on deep learning-based density assessment[10,11]. The attention mechanisms explored in our prior work[24] and Grad-CAM[61] interpretability[25] provide complementary approaches to building clinician trust in AI-generated reports.

Architectural paradigm comparison. Our evaluation spans four distinct architectural paradigms. Zero-shot VLMs (LLaVA[33], Qwen[59], MedGemma[15]) uniformly perform poorly on mammography, with ROUGE-L below 0.15, indicating that general-purpose VLM pretraining does not transfer to specialised mammographic reporting. Pipeline approaches (CLIP+GPT2) achieve surprisingly competitive results after fine-tuning (ROUGE-L: 0.571–0.613). However, end-to-end VLM fine-tuning with our two-stage approach consistently outperforms pipeline methods across all metrics, particularly CIDEr (0.729), indicating superior clinical coherence.

Comparison with AMRG methodology. While both our approach and AMRG[36] fine-tune MedGemma-4B[15] with LoRA[37] on the DMID dataset, several methodological differences contribute to our superior performance. First, our two-stage pretraining provides domain-adapted initialisation absent from AMRG's direct fine-tuning. Second, our LoRA ablation identifies $r = 64$, $\alpha = 64$ as optimal, compared to AMRG's $r = 32$, $\alpha = 16$—a configuration we find suboptimal. Third, AMRG does not incorporate synthetic data augmentation or RAG[47], both of which contribute to our framework's robustness.

LoRA configuration analysis. Our systematic ablation across seven LoRA[37] configurations reveals that the optimal rank ($r = 64$, $\alpha = 64$) substantially outperforms the commonly used default ($r = 16$, $\alpha = 32$), particularly on CIDEr (+14.1%). We attribute this to the two-stage training paradigm: because Stage 1 pretraining already provides a strong initialisation, Stage 2 benefits from higher-capacity LoRA adapters that can capture fine-grained stylistic differences between synthetic and real reports.

Context within broader research. Our results complement recent clinical AI deployment studies[62–64], suggesting that automated report generation could reduce radiologist workload while maintaining diagnostic quality. The detection capabilities demonstrated by Shen et al.[7] and Ribli et al.[65] in Scientific Reports, along with the feature extraction approaches by Aitimov et al.[22] from Kazakhstan, highlight the growing maturity of deep learning for mammography analysis. Modern architectures including ResNet[43], DenseNet[44], transformers[26], and Vision Transformers[45] have been foundational to this progress. Unlike chest X-ray report generation which benefits from large-scale datasets[16], mammography faces severe data scarcity, making our synthetic-to-real approach particularly valuable.

RAG trade-offs. In Stage 1, RAG decreased NLP metrics while maintaining clinical quality, reflecting n-gram-based evaluation limitations for clinical text. Future work should incorporate RadGraph F1 for entity-level clinical evaluation.

Limitations. The DMID test set is small ($n = 52$); larger paired datasets would strengthen conclusions. Formal radiologist evaluation is essential before clinical deployment. Single-view processing limits multi-view reasoning. The moderate BI-RADS accuracy (∼31%) reflects the difficulty of predicting BI-RADS from images alone. We additionally evaluated mammography-specific image preprocessing (Otsu thresholding, ROI cropping, laterality matching, CLAHE) but observed no improvement in

ROUGE-L (0.662 vs. 0.671) and reduced CIDEr (0.551 vs. 0.729), suggesting that SigLIP's[52] vision encoder handles raw mammographic inputs robustly without domain-specific preprocessing.

Future work. Multi-view integration; BI-RADS-guided constrained decoding; extension to other low-resource medical imaging domains; formal radiologist evaluation; reinforcement learning from human feedback (RLHF) building on our prior RL work[25].

Conclusion

We present a two-stage synthetic-to-real transfer learning framework for automated mammography report generation that addresses the fundamental data scarcity problem limiting VLM development in this domain. Our approach introduces three key innovations: (1) a label-to-report synthesis pipeline using Llama-3.1-8B[51] that generates clinically validated training data from structured BI-RADS[9] annotations, (2) a two-stage training paradigm[40] where synthetic pretraining provides domain-specific initialisation for subsequent real-data fine-tuning, and (3) a BI-RADS-grounded RAG module[47] for clinical knowledge integration.

By pretraining MedGemma-4B[15] with multimodal LoRA[37] on 400 synthetic reports and fine-tuning on 407 real DMID[21] reports with optimised LoRA configuration ($r = 64$, $\alpha = 64$), our approach achieves ROUGE-L[55] of 0.671, METEOR[56] of 0.685, CIDEr[57] of 0.729, and BERTScore[58] of 0.917 on the DMID benchmark. These results surpass the previously reported AMRG[36] state-of-the-art across all metrics, with improvements of +17.9% in ROUGE-L, +11.4% in METEOR, and +25.3% in CIDEr.

Our data efficiency analysis reveals that synthetic pretraining reduces real data requirements by approximately $2\times$: the two-stage model trained on just 100 real examples already surpasses the AMRG benchmark. Clinical analysis shows that fine-tuning reduces hallucination rates from 73.1% to approximately 21%, while qualitative evaluation reveals that even in failure cases, the two-stage model produces less clinically harmful errors than models trained on real data alone. The systematic LoRA ablation across seven configurations provides practical guidance for hyperparameter selection in medical VLM fine-tuning. Our framework provides a practical, reproducible, and scalable foundation for mammography report generation, and the demonstrated synthetic-to-real transfer learning paradigm may serve as a general strategy for low-resource medical AI development. All code, synthetic datasets, and trained model weights are publicly released to facilitate future research and clinical translation.

Data Availability

VinDr-Mammo[20] and CBIS-DDSM[19] are available on Kaggle. DMID[21] is available on Figshare and Kaggle. Synthetic datasets and model weights will be released upon acceptance.

Code Availability

All code is available at `https://github.com/ganiesenov/two-stage-mammo-vlm`.

Author Contributions

G.E. conceived the study, designed the methodology, implemented the full experimental pipeline, performed data preprocessing, model training, and evaluation, conducted quantitative and qualitative analyzes, and wrote the original manuscript draft.

M.N. supervised the research, contributed to the conceptual development of the study, provided guidance on methodology and experimental design, validated the results, and critically reviewed and edited the manuscript.

J.A. contributed to the clinical interpretation of results, assisted in data collection and validation, and provided domain-specific expertise in mammography and radiology.

L.L.P. contributed to methodological refinement, supported analysis and interpretation of results, and participated in manuscript review and editing.

All authors discussed the results, contributed to the final manuscript, and approved the submitted version.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors thank the creators of VinDr-Mammo, CBIS-DDSM, and DMID for making their datasets publicly available, and Google for providing access to MedGemma-4B.

Funding

This work was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP22788239.

References

1. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. for Clin.* **74**, 229–263, DOI: 10.3322/caac.21834 (2024).
2. Lehman, C. D. *et al.* National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology* **283**, 49–58, DOI: 10.1148/radiol.2016161174 (2017).
3. Sechopoulos, I., Teuwen, J. & Mann, R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Semin. Cancer Biol.* **72**, 214–225, DOI: 10.1016/j.semcancer.2020.06.002 (2021).
4. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Medicine* **25**, 24–29, DOI: 10.1038/s41591-018-0316-z (2019).
5. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94, DOI: 10.1038/s41586-019-1799-6 (2020).
6. Lotter, W. *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Medicine* **27**, 244–249, DOI: 10.1038/s41591-020-01174-9 (2021).
7. Shen, L. *et al.* Deep learning to improve breast cancer detection on screening mammography. *Sci. Reports* **9**, 12495, DOI: 10.1038/s41598-019-48995-4 (2019).
8. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66, DOI: 10.1148/radiol.2019182716 (2019).
9. American College of Radiology. ACR BI-RADS atlas: Breast imaging reporting and data system. *Am. Coll. Radiol.* (2013). Reston, VA.
10. Gastounioti, A., Desai, S., Ahluwalia, V. S., Conant, E. F. & Kontos, D. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Res.* **24**, 14, DOI: 10.1186/s13058-022-01509-z (2022).
11. Lehman, C. D. *et al.* Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology* **290**, 52–58, DOI: 10.1148/radiol.2018180694 (2019).
12. Wang, Z. *et al.* A survey of multimodal large language models for medical imaging. *arXiv preprint arXiv:2405.12833* DOI: 10.48550/arXiv.2405.12833 (2024).
13. Li, C. *et al.* LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36**, DOI: 10.48550/arXiv.2306.00890 (2024).
14. Zhang, S. *et al.* BiomedCLIP: A multimodal biomedical foundation model. *arXiv preprint arXiv:2303.00915* DOI: 10.48550/arXiv.2303.00915 (2023).
15. Sellergren, A. *et al.* MedGemma technical report. *arXiv preprint arXiv:2507.05201* DOI: 10.48550/arXiv.2507.05201 (2025). Google Research and Google DeepMind.
16. Johnson, A. E. W. *et al.* MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *Sci. Data* **6**, 317, DOI: 10.1038/s41597-019-0322-0 (2019).
17. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180, DOI: 10.1038/s41586-023-06291-2 (2023).
18. Moor, M. *et al.* Med-Flamingo: A multimodal medical few-shot learner. *Mach. Learn. for Heal. (ML4H)* DOI: 10.48550/arXiv.2307.15189 (2023).
19. Lee, R. S. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**, 170177, DOI: 10.1038/sdata.2017.177 (2017).
20. Nguyen, H. T. *et al.* VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. *Sci. Data* **10**, 277, DOI: 10.1038/s41597-023-02100-7 (2023).
21. Oza, P., Sharma, P., Patel, S. & Kumar, P. Digital mammography dataset for breast cancer diagnosis research (DMID) with breast mass segmentation analysis. *Biomed. Eng. Lett.* **14**, 317–330, DOI: 10.1007/s13534-023-00339-y (2024).

22. Aitimov, M., Tergeusizova, A., Beissegul, A. *et al.* Development of a feature vector for accurate breast cancer detection in mammographic images. *Informatics Medicine Unlocked* DOI: 10.1016/j.imu.2025.101607 (2025).

23. Esen, G., Altaibek, A., Amankulov, J., Matkerim, B. & Nurtas, M. Enhancing breast cancer detection with dimensionality reduction techniques: A study using PCA and LDA on Wisconsin breast cancer data. *Procedia Comput. Sci.* **251**, 414–421, DOI: 10.1016/j.procs.2024.12.148 (2024).

24. Esen, G. *et al.* Multi-task attention-guided deep learning for simultaneous breast cancer detection and density estimation in mammography. *IEEE Access* **13**, 198938–198951, DOI: 10.1109/ACCESS.2025.3530553 (2025).

25. Esen, G., Nurtas, M., La Paglia, L., Altaibek, A. & Matkerim, B. Reinforcement learning for breast cancer diagnosis with visual interpretability. *Procedia Comput. Sci.* **272**, 376–383, DOI: 10.1016/j.procs.2025.02.056 (2025).

26. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, DOI: 10.48550/arXiv.1706.03762 (2017).

27. Chen, Z., Song, Y., Chang, T. H. & Wan, X. Generating radiology reports via memory-driven transformer. In *Proceedings of EMNLP*, 1439–1449, DOI: 10.18653/v1/2020.emnlp-main.112 (2020).

28. Chen, Z., Shen, Y., Song, Y. & Wan, X. Cross-modal memory networks for radiology report generation. *Proc. ACL-IJCNLP* 5904–5914, DOI: 10.18653/v1/2021.acl-long.459 (2021).

29. Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P. & Jurafsky, D. Improving factual completeness and consistency of image-to-text radiology report generation. *Find. ACL-IJCNLP* 5288–5304, DOI: 10.18653/v1/2021.findings-acl.466 (2021).

30. Jin, T. *et al.* PromptMRG: Diagnosis-driven prompts for medical report generation. *Proc. AAAI Conf. on Artif. Intell.* **38**, DOI: 10.1609/aaai.v38i3.28038 (2024).

31. Hamamci, I. E., Er, S. *et al.* CT2Rep: Automated radiology report generation for 3D medical imaging. *Eur. Conf. on Comput. Vis.* DOI: 10.1007/978-3-031-72384-1_7 (2024).

32. Radford, A. *et al.* Learning transferable visual models from natural language supervision. *Proc. 38th Int. Conf. on Mach. Learn.* 8748–8763, DOI: 10.48550/arXiv.2103.00020 (2021).

33. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **36**, DOI: 10.48550/arXiv.2304.08485 (2024).

34. Ghosh, S., Poynton, C. B., Visweswaran, S. & Batmanghelich, K. Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. *Med. Image Comput. Comput. Assist. Interv. (MICCAI)* DOI: 10.1007/978-3-031-72378-0_58 (2024).

35. Zheng, S.-F., Lee, H., Kooi, T. & Diba, A. MV-MLM: Bridging multi-view mammography and language for breast cancer diagnosis and risk prediction. *Proc. IEEE/CVF Int. Conf. on Comput. Vis. (ICCV) Work.* 1224–1233, DOI: 10.48550/arXiv.2510.26151 (2025).

36. Sung, N.-J., Lee, H. *et al.* AMRG: Extend vision language models for automatic mammography report generation. *arXiv preprint arXiv:2508.09225* DOI: 10.48550/arXiv.2508.09225 (2025).

37. Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, DOI: 10.48550/arXiv.2106.09685 (2022).

38. Jahangir, R., Khan, N. I., Sudheerkumar, A. & Filkov, V. MammoWise: Multi-model local RAG pipeline for mammography report generation. *arXiv preprint arXiv:2602.22462* DOI: 10.48550/arXiv.2602.22462 (2026).

39. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: Efficient finetuning of quantized language models. *Adv. Neural Inf. Process. Syst.* **36**, DOI: 10.48550/arXiv.2305.14314 (2024).

40. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowl. Data Eng.* **22**, 1345–1359, DOI: 10.1109/TKDE.2009.191 (2010).

41. Zhuang, F. *et al.* A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76, DOI: 10.1109/JPROC.2020.3004555 (2021).

42. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis Mach. Intell.* **35**, 1798–1828, DOI: 10.1109/TPAMI.2013.50 (2013).

43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. CVPR* 770–778, DOI: 10.1109/CVPR.2016.90 (2016).

44. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *Proc. CVPR* 4700–4708, DOI: 10.1109/CVPR.2017.243 (2017).

45. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. on Learn. Represent.* DOI: 10.48550/arXiv.2010.11929 (2021).

46. Chen, B. *et al.* Heterogeneity-aware hypergraph neural network with dual-path fusion for LUAD stage-specific driver gene mining. *Tsinghua Sci. Technol.* DOI: 10.26599/TST.2026.9010037 (2026). Just Accepted.

47. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474, DOI: 10.48550/arXiv.2005.11401 (2020).

48. Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **7**, 535–547, DOI: 10.1109/TBDATA.2019.2921572 (2021).

49. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proc. EMNLP* 3982–3992, DOI: 10.18653/v1/D19-1410 (2019).

50. Moreira, I. C. *et al.* INbreast: Toward a full-field digital mammographic database. *Acad. Radiol.* **19**, 236–248, DOI: 10.1016/j.acra.2011.09.014 (2012).

51. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* DOI: 10.48550/arXiv.2407.21783 (2024).

52. Zhai, X., Mustafa, B., Kolesnikov, A. & Beyer, L. Sigmoid loss for language image pre-training. *Proc. ICCV* DOI: 10.1109/ICCV51070.2023.01100 (2023).

53. Dettmers, T., Lewis, M., Belkada, Y. & Zettlemoyer, L. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Adv. Neural Inf. Process. Syst.* **35**, DOI: 10.48550/arXiv.2208.07339 (2022).

54. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, 311–318, DOI: 10.3115/1073083.1073135 (2002).

55. Lin, C. Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81, DOI: 10.3115/1219611.1219613 (2004).

56. Banerjee, S. & Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, 65–72, DOI: 10.3115/1626355.1626389 (2005).

57. Vedantam, R., Lawrence Zitnick, C. & Parikh, D. CIDEr: Consensus-based image description evaluation. In *Proceedings of CVPR*, 4566–4575, DOI: 10.1109/CVPR.2015.7299087 (2015).

58. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, DOI: 10.48550/arXiv.1904.09675 (2020).

59. Wang, P., Bai, S., Tan, S. *et al.* Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* DOI: 10.48550/arXiv.2409.12191 (2024).

60. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* (2019). Available at: https://openai.com/research/better-language-models.

61. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359, DOI: 10.1007/s11263-019-01228-7 (2020).

62. Dembrower, K. *et al.* Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *The Lancet Digit. Heal.* **2**, e468–e474, DOI: 10.1016/S2589-7500(20)30185-0 (2020).

63. Salim, M. *et al.* External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncol.* **6**, 1581–1588, DOI: 10.1001/jamaoncol.2020.3321 (2020).

64. Aboutalib, S. S. *et al.* Deep learning to distinguish recalled but benign mammography images in breast cancer screening. *Clin. Cancer Res.* **24**, 5902–5909, DOI: 10.1158/1078-0432.CCR-18-1115 (2018).

65. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. Reports* **8**, 4165, DOI: 10.1038/s41598-018-22437-z (2018).