

DOI: 10.37943/HLVA7583

Beibit Abdikenov

PhD, Director, Artificial Intelligence Science and Innovation Center
beibit.abdikenov@astanait.edu.kz, orcid.org/0000-0002-0284-0949
Astana IT University, Kazakhstan

Aruzhan Imasheva

Master's Degree, Junior Researcher, Artificial Intelligence Science and Innovation Center

aruzhan.imasheva@astanait.edu.kz, orcid.org/0009-0004-3790-2339
Astana IT University, Kazakhstan

Tomiris Zhaksylyk

PhD Student, Researcher, Artificial Intelligence Science and Innovation Center
Zhaksylyk.tomiris@astanait.edu.kz, orcid.org/0009-0002-8749-1967
Astana IT University, Kazakhstan

A BILATERAL VISION-LANGUAGE FRAMEWORK WITH CROSS-LATERAL ATTENTION FOR MAMMOGRAPHY

Abstract: Screening mammography is read comparatively: radiologists hold one breast against the other, and several finding categories have no meaning outside that comparison. Most automated pipelines nonetheless encode a single breast at a time. We describe Bilateral MV-CLIP, a contrastive image-text framework in which all four projections of a patient enter one forward pass and the two lateralities exchange information through a Cross-Lateral Attention block, with a learned gate and a placeholder token preserving behaviour when a projection or an entire side is unavailable. Training combines image-text and image-image objectives, graded similarity targets, and a symmetry term tying embedding geometry to recorded asymmetry status. We train on VinDr-Mammo with reports synthesised from structured annotations and evaluate by linear probing of frozen features, cross-modal retrieval and report decoding, against an independently trained single-breast model matched in backbone, optimiser, schedule and epoch budget, with checkpoints selected on a held-out validation partition. On a pre-specified composite endpoint pooling the three comparison-defined categories, fusion did not improve on that baseline: area under the curve 0.781 against 0.792, a difference of -1.1 points with a 95% confidence interval from -9.6 to +7.5. Per-category effects follow the direction the design predicts, but the partition holds 49 positive comparison-defined breasts and no difference survived correction. Retrieval is limited by weak instance-level image-text alignment rather than by report redundancy, with a pool-size-independent ranking accuracy of 0.691. On an external cohort of 11,913 patients labelled by biopsy the representations transfer at reduced discrimination, reaching 0.65 for any abnormality. We therefore report explicit bilateral comparison as plausible but, on this dataset, unconfirmed, and give confidence intervals throughout.

Keywords: breast cancer; mammography; vision-language models; contrastive learning; cross-lateral attention; bilateral analysis; asymmetry detection; representation learning; medical image analysis; deep learning

Introduction

Breast malignancy is responsible for approximately 2.3 million fresh diagnoses each year making it amongst the most commonly identified cancers anywhere in the world [1]. The association between early identification and improved outcome is well established. It has been shown repeatedly that organized screening with mammographic imaging reduces deaths from the disease. This is why mammography remains the examination of choice. But the interpretation is not so easy. The diagnostic signs are often subtle, a lesion can be completely masked by

fibroglandular tissue, one reader may disagree with another [2], and a confident assessment usually requires considering several projections simultaneously.

The examination itself consists of four acquisitions, that is a cranio-caudal and medio-lateral oblique projection for each side. It routinely takes the other side for a control when the reader finds a sketchy area. It's not just convention. Three of the categories of findings in routine use, asymmetry, focal asymmetry and global asymmetry, are described in contrast to the corresponding territory of the other breast, so that no absolute description of a single image can express them [3]. Therefore, the underlying categories cannot be represented by automated systems consuming one breast per forward pass, by construction.

Pixel-to-language contrastive pre-training is an active area for learning from medical images as well. The method of aligning visual inputs with unrestricted text using a contrastive loss, pioneered in CLIP [4], results in representations that generalize broadly and was rapidly applied in radiology. In mammograms, for example, Mammo-CLIP took the approach to its next stage by adding projection consistency for each breast and severity grading of supervision [5]. Bilateral symmetry is another area that received little attention – few methods use actual interaction between the two sides in the space of visual representations.

This gap is what we address in this paper. Our approach is that a network allowed to think across the two sides during the encoding process should capture exactly those observations which have a comparative clinical definition. We thus attempt to create a bilateral framework that would consume all four views together but be operational using one-sided inspections, develop an attention system for feature level communication across the two sides, and provide supervision that would directly encode symmetry and asymmetry; we then evaluate the learned features on VinDr-Mammo using classification, retrieval, and generation tasks. We term this framework Bilateral MV-CLIP.

Four contributions then follow. We propose a bilateral contrastive image-text architecture for mammograms where a single visual backbone works for both sides, and there is a mechanism to fuse both sides together. We propose a Cross-Lateral Attention block where the gated fusion takes into account the interactivity between the two sides but can degrade gracefully when one of the two sides is missing. We define a symmetry loss based on the annotation of the asymmetry of the mammogram, which, to the best of our knowledge, is the first contrastive loss in the literature to supervise the bilateral relationship directly.

Related Work

Deep learning for mammographic interpretation. Convolutional approaches to mammography matured rapidly once end-to-end training displaced hand-designed feature extraction [6]. The clearest demonstration of clinical relevance came from a large international evaluation in which a system trained on tens of thousands of examinations from two countries reduced both false positives and false negatives relative to routine reading [2]. Parallel work established that networks trained on all four projections of an examination outperform those restricted to a single view, and that attention can weight projections adaptively rather than uniformly. What almost all of this work shares is a per-breast decision boundary: the contralateral side, which a human reader consults constantly, plays no role in the computation.

Modelling the two sides jointly. Systems that treat laterality as a modelled relationship rather than a batching detail are comparatively rare. The earliest examples embed each side separately and concatenate the resulting vectors immediately before classification. Concatenation does improve on single-breast baselines, but it delays interaction until after all spatial detail has been pooled away, so the network is never in a position to compute a difference between corresponding regions. Interaction at the granularity of patch tokens [7], which is what a comparison between anatomical territories actually requires, has not been explored within contrastive image-text frameworks for this modality. Our Cross-Lateral Attention block is aimed precisely at that gap.

Contrastive pre-training with paired text. The theoretical basis predates its application to imaging: contrastive predictive coding framed such objectives as maximisation of a lower bound on mutual information [8], and subsequent purely visual methods confirmed that contrastive training with strong augmentation rivals supervised initialisation [9-10]. CLIP carried the idea to paired image-caption data at web scale [4], and radiology followed. ConVIRT first applied paired image-report training to chest radiographs, showing that clinical prose serves as a viable supervisory signal [11]. BioViL added phrase grounding [12] so that specific report expressions align with image regions, a design that motivates the local alignment term we retain. GLoRIA aligned whole reports with whole images and individual words with sub-image regions simultaneously [13]. MedCLIP relaxed the pairing requirement altogether by deriving semantic similarity between reports from a knowledge graph, an idea complementary to the overlap-based soft targets used here [14].

Text encoders and generation for the clinical domain. General-purpose language models transfer poorly to clinical vocabulary [15], which prompted domain-adapted variants; the model we adopt for the text branch was specialised on hospital notes and is accordingly better matched to radiological phrasing than a general encoder [16]. On the generation side, early report systems paired a convolutional encoder with a recurrent decoder and cross-attention between regions and sentences, later replaced by transformer decoders [17] and, more recently, by conditioning frozen language models on visual prefixes [18]. Evaluation conventionally combines n-gram overlap measures with embedding-based semantic scores [19–21], and we follow that convention.

Graded rather than binary contrastive targets. Standard contrastive training regards every non-matching batch element as a negative of equal weight, which misfires in medicine, where two examinations sharing a pathology should not be forced apart. Supervised contrastive learning addressed this by promoting same-class elements to positives, though at the cost of imposing hard class boundaries where the underlying structure is continuous [22]. Soft-target formulations instead assign graded similarity, and Mammo-CLIP applied this to mammography using overlap between finding sets and distance between severity scores [5]. We inherit both terms and extend the principle to the bilateral relation, where the target is determined not by shared pathology but by whether a radiologist recorded an asymmetry.

Set against the contrastive models closest to it, the proposed framework occupies distinct ground [5, 11–14]. ConVIRT, GLoRIA, and BioViL each operate on a single image at a time, with only MedCLIP admitting anything resembling graded supervision. Mammo-CLIP adds processing across projections and graded targets but remains one-sided, and none of the five carries a generation branch. Bilateral MV-CLIP is the only framework among them to combine multi-projection input, explicit interaction across sides, graded supervision, and report decoding within a single architecture.

Methods and Materials

Imaging data and report synthesis

Our experiments draw throughout on VinDr-Mammo, a full-field digital collection assembled at a pair of Vietnamese hospitals comprising five thousand complete two-sided examinations [23]. Every examination supplies cranio-caudal and medio-lateral oblique projections for each side, so a patient contributes at most four images. Annotation was performed per breast and covers a BI-RADS category in the range 1 through 5 [3], a density grade on the A-to-D scale, and localised boxes for ten categories of finding: masses, suspicious calcification, the three asymmetry variants, architectural distortion, suspicious lymph nodes, and retraction of skin or nipple together with skin thickening. The partition supplied with the dataset assigns four fifths of the examinations to training and the remainder to testing.

Free-text reporting is absent from the release, so we derive it deterministically from the structured annotation. Each combination of examination and side receives a report organised into technique, breast composition, findings, and impression. Phrase selection is conditioned jointly on density grade, which findings appear, roughly where they sit, and the assigned BI-RADS category.

Position is recovered from the centre of each annotated box relative to the image frame and expressed in the quadrant vocabulary used clinically, with the medial-lateral axis reversed as appropriate for the side being described. The procedure produces [AD1.1]one report per breast, at most ten thousand overall, matching clinical mammographic register in both vocabulary and organisation.

Sampling at the level of the patient

Contrastive mammography models published previously assemble batches from individual breasts, so an examination contributes two independent elements. We depart from this: one element of a batch corresponds to one patient and carries every available image alongside two reports, two finding indicator vectors, and two severity scores. A pair of boolean flags records which sides are present. Where a side is missing its tensor slot is zeroed, and the model later substitutes a learned placeholder before any cross-lateral exchange occurs.

Comparative findings appear in under a fifth of examinations, and uniform sampling would therefore place asymmetric pairs in a batch too seldom for the symmetry term to receive meaningful gradient. We counter this by sampling patients with probability proportional to a product of three factors,

$$W_i = W_{base} \times W_{BI-RADS} \times W_{asym} \quad (1)$$

in which the first contributes a factor of three where any abnormal keyword occurs in the report and unity otherwise; the second is , so a category-5 patient contributes threefold; and the third contributes double whenever an asymmetry of any variant has been recorded, reflecting how central such cases are to the objective. A patient satisfying all three conditions is therefore eighteen times more likely to be drawn than a wholly unremarkable one.

Architecture

Three cooperating components make up Bilateral MV-CLIP: a visual encoder shared across every input image, a bilateral encoder that combines the four projections via Cross-Lateral Attention, and a branch that embeds reports. Both modalities are projected into one 256-dimensional space. During a forward pass every image is embedded on its own; token sequences belonging to the same side are averaged, then passed through cross-lateral fusion, which returns a global vector per side together with one joint patient-level vector. An absent side contributes the learned placeholder, whose influence the gate learns to remove. Figure 1 shows the pipeline as a whole and Figure 2 the fusion block.

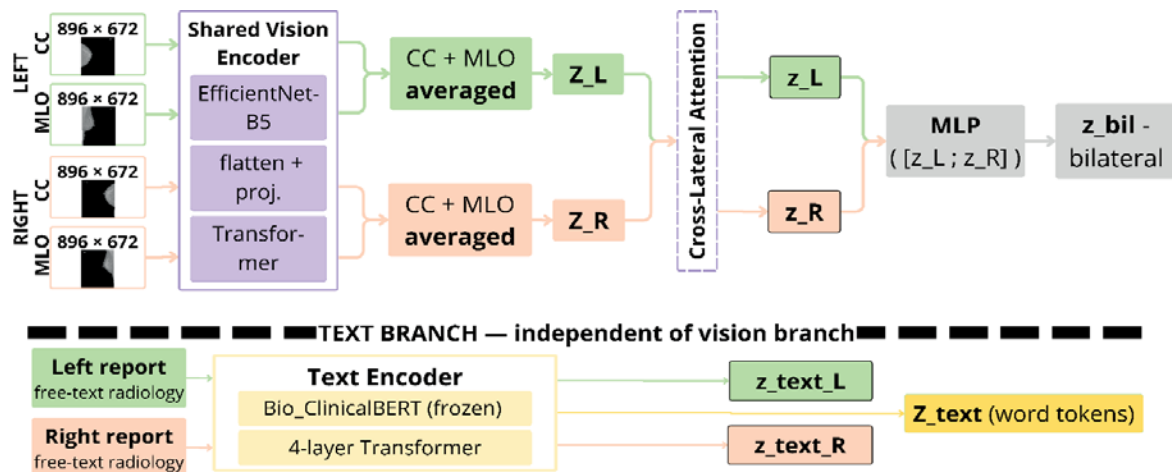


Figure 1. The complete Bilateral MV-CLIP pipeline. A single visual encoder embeds all four projections, token sequences belonging to one side are averaged, and Cross-Lateral Attention fuses them into a global vector per side together with one joint patient-level vector; the report branch runs independently.

Visual encoder. Every mammogram, whatever its side or projection, enters the same encoder, which wraps an EfficientNet-B5 backbone [24] initialised from ImageNet and operated purely as a feature extractor. Given input at 896 by 672 pixels, the terminal feature map is flattened into a token sequence and mapped linearly onto 256 channels. Four transformer layers with eight heads and feed-forward width 1024 [17] then contextualise those tokens, emitting L2-normalised local embeddings together with a global vector formed by max-pooling and then a linear map. Token sequences from the two projections of one side are then combined by elementwise averaging,

$$Z_{\text{breast}} = \frac{Z_{\text{CC}} + Z_{\text{MLO}}}{2} \quad (2)$$

which hands the fusion block one coherent description of each side in place of two signals competing for the same positions.

Cross-Lateral Attention. This block carries the architectural weight of the study. Supplied with the averaged token sequences of both sides, it returns sequences of matching shape enriched by information from the opposite side. Tokens attend across the midline through cross-attention layers whose parameters are shared between the two directions, after which a sigmoid gate determines the quantity of foreign context admitted at each position. Two such layers are stacked, and both sides are handled symmetrically within each:

$$Z_L^{(\ell)} = \text{FFN} \left(\text{LN} \left(Z_L^{(\ell-1)} + \text{MHA} \left(Z_L^{(\ell-1)}, Z_R^{(\ell-1)}, Z_R^{(\ell-1)} \right) \right) \right) \quad (3)$$

$$Z_R^{(\ell)} = \text{FFN} \left(\text{LN} \left(Z_R^{(\ell-1)} + \text{MHA} \left(Z_R^{(\ell-1)}, Z_L^{(\ell-1)}, Z_L^{(\ell-1)} \right) \right) \right) \quad (4)$$

Tying the parameters is a deliberate choice rather than an economy: comparison between sides is a geometrically symmetric operation, and the function mapping left onto right ought to coincide with its mirror image. Above, MHA denotes eight-head attention and LN layer normalisation.

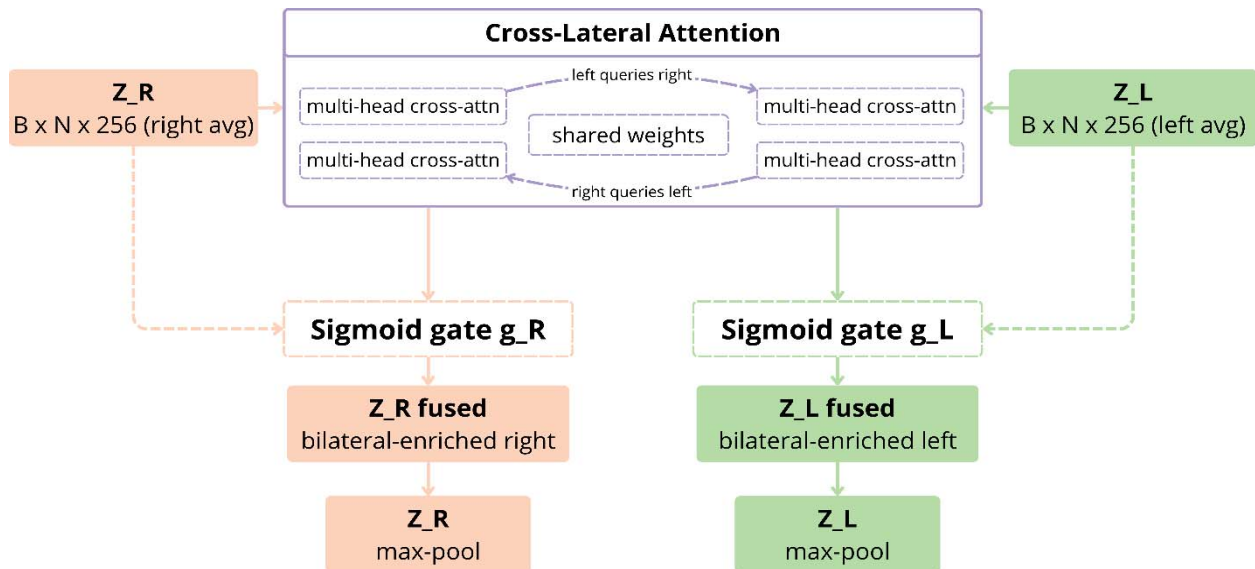


Figure 2. The Cross-Lateral Attention block. Two cross-attention layers with tied parameters let tokens of either side query the other, after which a sigmoid gate blends the enriched tokens back against the originals through a residual path and learns to discard the placeholder standing in for an absent side.

A gated residual then regulates retention of cross-lateral context position by position,

$$g_L = \sigma \left(W_g \left[Z_L^{(0)}; Z_L^{(L)} \right] + b_g \right) \quad (5)$$

$$\tilde{Z}_L = \text{Proj} \left(g_L \odot Z_L^{(L)} + (1 - g_L) \odot Z_L^{(0)} \right) \quad (6)$$

with the mirror expression applying on the opposite side. Initialising the gate bias at +0.5 leaves the gates half-open at the outset, so foreign context participates from the very first update. Should a side be absent, its slot carries the shared placeholder and the gate converges on values near zero, discarding the vacant context without any explicit masking.

Joint embedding. Global vectors per side follow from max-pooling and L2-normalising the fused sequences. Concatenating the two and passing them through a two-layer perceptron with layer normalisation produces a single vector describing the patient,

$$Z_{\text{bil}} = \frac{f_{\text{bil}}([Z_L; Z_R])}{\|f_{\text{bil}}([Z_L; Z_R])\|_2}, \quad f_{\text{bil}}: \mathbb{R}^{512} \rightarrow \mathbb{R}^{256} \quad (7)$$

A missing side leaves its global vector at zero, and the projection accommodates one-sided examinations without architectural modification.

Report encoder. Reports pass through Bio_ClinicalBERT [16] with its parameters held fixed for the duration; learning is confined to the layers stacked above it. Word-piece states drawn from the terminal hidden layer are mapped onto 256 channels and refined by four transformer layers configured as on the visual side. Max-pooling supplies a global text vector and the token outputs the local ones, both L2-normalised.

Objective

Nine weighted terms compose the loss,

$$L = L_{\text{IT}}^L + L_{\text{IT}}^R + L_{\text{bil}} + L_{\text{II}} + \lambda_{\text{loc}} L_{\text{loc}} + \lambda_{\text{soft}} L_{\text{soft}} + \lambda_{\text{br}} L_{\text{br}} + \lambda_{\text{sym}} L_{\text{sym}} + \lambda_{\text{cross}} L_{\text{cross}} \quad (8)$$

whose coefficients are given in Table 2. We characterise each briefly and write out those carrying the bilateral contribution.

Alignment of each side with its report. Wherever a side exists, its global vector is matched against the corresponding report under a symmetric InfoNCE objective [8] evaluated only over batch elements in which that side is present, with temperatures held as learnable scalars starting from 0.1. The two resulting terms drive each side's description toward agreement with the prose written about it.

Alignment between sides. When neither side of a patient carries an asymmetry, the two global vectors are drawn together,

$$L_{\text{II}} = \text{InfoNCE} (Z_L, Z_R; \tau_2), \quad S = \{i: \text{bilateral}_i = 1 \wedge a_i = 0\} \quad (9)$$

where a_i records whether any asymmetry was annotated for that patient. Withholding this term from asymmetric patients is deliberate. Applied without condition it would work against the symmetry objective introduced below, which separates the two sides in precisely those cases. Confining it to symmetric patients removes the antagonism and leaves a coherent geometry: unremarkable pairs converge while asymmetric ones diverge.

Graded targets. Two further terms, inherited from Mammo-CLIP [5], substitute continuous similarity for binary contrastive targets. One computes similarity from the overlap between ten-dimensional finding indicators, sparing clinically comparable images from being forced apart. The

other derives similarity from the gap between severity scores, drawing high-severity examinations together and away from unremarkable ones. A local term additionally couples word-level and patch-level tokens through cross-attention.

Symmetry. This term is, so far as we are aware, the first in contrastive mammography to supervise the bilateral relation directly. Its premise is uncomplicated: a pair without pathology should occupy nearby positions, whereas a pair containing an asymmetry should not. We express it through squared error applied to cosine similarity,

$$L_{\text{sym}} = \frac{1}{|B|} \sum_{i \in B} \left(Z_{L,i} \cdot Z_{R,i} - (1 - 2a_i) \right)^2 \quad (10)$$

taking B as those batch elements possessing both sides. Targets are consequently +1 for unremarkable pairs and −1 for asymmetric ones. Prior normalisation confines the inner product to the interval between −1 and 1, so every target lies within reach. A further bilateral image-text term anchors the joint vector to both reports, and an auxiliary reinforcement term was intended to open a second gradient route preventing the joint vector from drifting toward one side when the two reports disagree. We note for transparency that in the released implementation this auxiliary term is computed identically to the bilateral image-text term, so the objective applies that term at an effective weight of $1 + \text{lambda_cross}$ rather than combining two distinct routes.

Training setup

Implementation is in PyTorch, with training performed on one NVIDIA H200. Inputs are rescaled to 896 by 672 and standardised using ImageNet statistics. Augmentation covers horizontal flipping, affine perturbation with rotation and shear each bounded at twenty degrees, translation up to a tenth of the frame and scaling between 0.8 and 1.2, and exchange of the two projections belonging to one side, which discourages any dependence on projection ordering. Arithmetic is bfloat16 and the backbone is gradient-checkpointed, capping peak memory while leaving gradients intact.

AdamW drives optimisation [25]. Parameters are split in two: components initialised from scratch, namely the fusion block and the joint projection, take the higher of the two rates, while everything else takes the base rate. A single warm-up epoch precedes cosine decay across the following nineteen, and gradients are clipped at unit norm. Table 2 collects the configuration, and Table 1 the resulting model size and computational cost.

Table 1. Model size and computational cost, measured on one NVIDIA H200 in bfloat16 with gradient checkpointing enabled on the vision backbone.

Item	Value
Total parameters	144.5 M
Trainable parameters	36.2 M
Frozen text encoder	108.3 M
Vision backbone (EfficientNet-B5)	27.29 M
Cross-Lateral Attention	1.91 M (1.3% of total, 5.3% of trainable)
Joint projection	0.20 M
Checkpoint size (fp32 / bf16)	577.9 / 289.0 MB
Inference cost, vision side	126.4 GFLOPs per four-projection study
Inference latency, batch 1	47.2 ms per study (p95 51.3)
Inference latency, batch 8	15.1 ms per study, 66.4 studies/s
Inference peak memory	0.83 GB (batch 1), 1.71 GB (batch 8)
Training step, batch 64 studies	5.84 s, 11.0 studies/s
Training peak memory	70.2 GB
Training wall-clock, 20 epochs	approximately 1.8 h of GPU compute

Cross-Lateral Attention adds 1.91 M parameters, 1.3% of the model and 5.3% of its trainable parameters and does not materially change inference cost: a full four-projection study is encoded in 47 ms at batch 1 within 1.71 GB. Because the leave-one-out condition that disables cross-lateral exchange retains those parameters unused, the corresponding comparison isolates the mechanism rather than the additional capacity.

Model-development protocol. In the originally submitted work the validation loop was computed on the official test partition, and checkpoint selection used that loss. No per-category metric and no hyperparameter was tuned on it, but the selection itself was not independent, and figures obtained under that protocol may be optimistically biased. All results reported here come from models selected on a validation partition held out from the official training partition at patient level, stratified by maximum BI-RADS category, presence of any asymmetry annotation, and laterality completeness. The linear probe is fitted on the remaining training studies with its regularisation constant selected on that same validation partition, and the test partition is used once, for final evaluation. Because the probe now sees fewer training studies, absolute values differ from the submitted version even for an unchanged model.

Development surfaced a conflict between gradients. In its original form the between-sides objective applied to every two-sided patient, and once this was recognised the restricted formulation was substituted and the model fine-tuned for five additional epochs from the sixteenth-epoch checkpoint at a tenth of the base rate. Everything else was held as before, and results reported below all derive from that fine-tuned checkpoint.

Table 2. Configuration used for training Bilateral MV-CLIP, grouped by category.

Category	Setting	Value
Input	Resolution	896×672
Input	Precision	bfloat16
Schedule	Patient-level batch size	64
Schedule	Total epochs	20
Schedule	Warm-up epochs	1
Schedule	Decay	Cosine to 10^{-6}
Rates	Base	5×10^{-5}
Rates	Fusion block and joint projection	1.5×10^{-4}
Rates	Weight decay	10^{-4}
Loss weights	λ_{sym}	0.4
Loss weights	$\lambda_{\text{loc}}, \lambda_{\text{soft}}$	0.5
Loss weights	$\lambda_{\text{br}}, \lambda_{\text{cross}}$	0.3
Loss weights	τ_1, τ_2	Learnable, from 0.1
Sampling	Abnormal keyword present	$\times 3.0$
Sampling	Severity category 5	$\times 3.0$
Sampling	Any asymmetry recorded	$\times 2.0$
Sampling	Maximum combined	$\times 18.0$

Results

Protocol

Every measurement below uses the thousand two-sided examinations reserved by the official partition, none of which the model encountered during training. Three protocols apply. The first, and the one we treat as primary, fits a logistic-regression head to frozen features and predicts the ten finding categories, which isolates how discriminative the representation is in itself with the encoder untouched. The second measures zero-shot retrieval of reports from images across the test pool, indicating how well an image vector ranks its own report ahead of the remaining 999. The third decodes free text conditioned on the full set of projections and scores it against the reference.

Probing compares three feature sets under identical settings so that only the representation varies: a *unilateral* global vector taken from the left side before fusion, which was retained for continuity with the submitted version; this vector receives no image-text gradient, since all contrastive terms are computed on the post-fusion per-side vector, and its retrieval ranking accuracy of 0.551 against 0.500 for chance confirms that it is an unsupervised intermediate activation rather than a trained single-breast representation, so the independently trained unilateral model of Table 3 provides the reference against which all claims are made; a *fused* per-side vector taken after cross-lateral exchange, distinguished from the baseline by nothing except bilateral context; and the joint patient vector. All three share the same logistic-regression configuration fitted on the four thousand training examinations.

Baselines. The primary reference is an independently trained single-breast model: one breast per training example, no contralateral input, no bilateral loss term and no joint embedding, with backbone, input resolution, embedding width, transformer depth, text encoder, optimiser, learning-rate schedule, epoch budget and breasts-per-step held identical to the bilateral model. It differs from the full model by 2.11 M parameters, exactly the cross-lateral block and the joint projection. The pre-fusion vector used as the baseline in the submitted version is retained for continuity and reported alongside, but it is not a trained single-breast representation and no claim rests on it.

Finding classification from frozen features

Table 3 reports per-finding discrimination ($\text{AUC} \times 100$) on the held-out partition, ordered by how often each category occurs in the test set, with 95% bootstrap confidence intervals from 2,000 patient-clustered resamples evaluated on all present breasts. Categories defined by contralateral comparison are marked with a dagger (†). Categories with fewer than ten positive test breasts are marked with an asterisk (*) and are descriptive only: they carry no inference and are excluded from the multiple-testing family. The composite asymmetry endpoint pools the three comparison-defined categories into a single binary label and is the pre-specified primary comparison.

Against the independently trained single-breast model the fused per-side features are not uniformly better: macro-averaged AUC and micro-averaged average precision favour fusion while micro-averaged AUC and macro-averaged average precision favour the baseline. Restricted to the six categories with at least ten positive test breasts, macro AUC is 82.8 for the fused features against 83.2 for the baseline, a difference of -0.4 points with a 95% interval from -3.8 to $+2.9$. No head was fitted to the encoder in either case, so these figures compare representation quality directly.

Table 3. Per-finding discrimination on the held-out partition ($\text{AUC} \times 100$)

Finding	Cases	Unilateral (independent)	Fused z_L	Pre-fusion (submitted)
Mass	119	80.0 [75.5, 84.3]	81.1 [75.9, 85.9]	80.4 [74.8, 85.6]
Suspicious calcification	54	86.4 [80.2, 91.8]	83.8 [77.8, 89.3]	83.1 [76.7, 88.8]
Focal asymmetry †	26	82.9 [74.2, 90.6]	75.8 [66.3, 83.9]	78.1 [69.1, 86.2]
Asymmetry †	20	74.9 [64.3, 85.5]	84.7 [75.2, 92.9]	78.7 [68.3, 88.1]
Architectural distortion	13	79.1 [64.9, 90.4]	73.0 [56.1, 88.0]	80.6 [67.5, 92.0]
Suspicious lymph node	10	96.3 [90.9, 99.8]	98.8 [97.8, 99.6]	95.4 [89.7, 99.3]
Skin thickening *	7	94.5 [88.8, 99.4]	94.7 [87.6, 98.7]	96.8 [93.0, 99.5]
Nipple retraction *	4	98.7 [96.4, 99.9]	98.7 [97.0, 99.8]	98.8 [97.2, 99.7]

Finding	Cases	Unilateral (independent)	Fused z_L	Pre-fusion (submitted)
Global asymmetry † *	3	62.3 [10.5, 90.1]	99.2 [98.3, 99.9]	94.0 [86.5, 98.7]
Skin retraction *	2	99.0 [98.1, 99.7]	98.3 [97.0, 99.3]	99.1 [98.6, 99.5]
Macro AUC	---	85.4 [78.9, 89.5]	88.8 [85.7, 91.0]	88.5 [85.6, 90.6]
Macro AUC, n ≥ 10 only	---	83.2 [79.6, 86.6]	82.8 [79.2, 86.2]	82.7 [79.3, 86.1]
Micro AUC	---	78.3 [74.6, 81.7]	75.1 [70.5, 79.2]	76.9 [72.6, 80.8]
Macro AP	---	19.6 [14.7, 30.8]	17.9 [14.0, 30.6]	14.9 [12.7, 22.7]
Micro AP	---	22.5 [17.3, 28.6]	24.2 [18.5, 29.9]	19.7 [15.3, 24.9]
Composite asymmetry AUC	49	79.2 [72.0, 85.7]	78.1 [71.4, 84.5]	80.5 [74.9, 85.7]
Composite asymmetry AP	49	14.7	10.9	9.5

On the primary endpoint the fused representation does not improve on the independently trained baseline: the paired difference is -1.1 points of AUC (95% CI -9.6 to +7.5, $p = 0.84$) and -3.8 points of average precision (-11.7 to +5.8, $p = 0.50$). Among the six adequately powered categories no difference survives Benjamini-Hochberg correction; the largest nominal movement is Asymmetry at +9.8 points (-0.2 to +20.5, $p = 0.06$ uncorrected). Global Asymmetry moves by +36.9 points, but the baseline estimate on three positive cases is 62.3 with an interval spanning 10.5 to 90.1, so that figure reflects the instability of the comparison rather than a demonstrated benefit and is not used as evidence.

The distribution of gains matches what the design anticipated. Categories requiring comparison across the midline benefit most; several conspicuous one-sided categories are flat or marginally worse; the aggregate is clearly favourable. Isolating the three comparative categories, their pooled mean rises from 73.4 to 86.6, but the baseline value is depressed by a Global Asymmetry estimate built on three cases, and the pre-specified composite endpoint, which pools the same three categories into one adequately powered binary label, shows no advantage. We therefore do not treat the pooled mean of three separately estimated AUCs as evidence.

Outside the comparative group, architectural distortion records the second-largest movement, gaining 23.5 points to reach 89.1 under the joint vector. In the present evaluation architectural distortion moves in the opposite direction, from 79.1 to 73.0 with intervals spanning more than twenty points in each case, so the earlier interpretation is withdrawn. Skin thickening likewise improves markedly. Suspicious lymph nodes, skin retraction, and nipple retraction all slip somewhat under fused conditions, which sits comfortably with the expectation that conspicuous one-sided signs gain nothing from foreign context while the additional parameters slightly dilute the signal that already served them.

Clinically framed tasks

We next posed two questions closer to reporting practice, keeping the probing protocol fixed. The first separates unremarkable from abnormal examinations, mapping severity categories 1 and 2 onto the former and 3 through 5 onto the latter. The second predicts the severity category itself, scored primarily by quadratic weighted kappa because that measure scales its penalty with the size of the ordinal error.

Table 4 compares the three representations on both tasks on the held-out partition, where abnormal prevalence is 10.2%. For the binary task it gives AUC, PR-AUC, sensitivity, specificity, the confusion counts, accuracy, precision, F1, and sensitivity or specificity at fixed operating points of 90% and 95%; for the five-class ordinal task it gives quadratic weighted kappa, macro AUC and accuracy.

Table 4. Clinically framed probing tasks.

Task	Metric	Unilateral (independent)	Fused z_L	Pre-fusion (submitted)
Binary	AUC	0.815 [0.779, 0.849]	0.822 [0.786, 0.856]	0.817 [0.777, 0.852]
Binary	PR-AUC	0.504	0.509	0.527
Binary	Sensitivity	0.500	0.573	0.599
Binary	Specificity	0.928	0.888	0.886
Binary	Confusion (TN/FP/FN/TP)	1677 / 131 / 96 / 96	1605 / 203 / 82 / 110	1602 / 206 / 77 / 115
Binary	Accuracy	0.886	0.858	0.859
Binary	Precision (PPV)	0.423	0.351	0.358
Binary	F1	0.458	0.436	0.448
Binary	Specificity at 90% sens.	0.390	0.455	0.412
Binary	Sensitivity at 90% spec.	0.552	0.547	0.589
Binary	Sensitivity at 95% spec.	0.458	0.469	0.474
Ordinal	Quadratic weighted kappa	0.541 [0.485, 0.593]	0.464 [0.405, 0.517]	0.472 [0.418, 0.522]
Ordinal	Macro AUC	0.798	0.759	0.748
Ordinal	Accuracy	0.663	0.591	0.592

The three representations are close on the binary task and the differences are not significant: the fused vector exceeds the independently trained baseline by 0.7 points of AUC (95% CI -2.4 to +3.9, $p = 0.66$). They differ chiefly in operating characteristics rather than in overall discrimination: the baseline is the more specific at the 0.5 threshold, recovering 96 of 192 abnormal breasts at 92.8% specificity, while the fused vector recovers 110 at 88.8%. At a fixed 90% specificity the three are within four points of one another.

On the binary task the fused per-side vector recovers more abnormal examinations at the 0.5 threshold, at the cost of specificity, though the difference in overall discrimination is within noise. The joint vector instead maximises precision and average precision, surrendering recall for a lower false-positive rate. On the ordinal task the ranking reverses: the independently trained single-breast baseline leads throughout, reaching a quadratic weighted kappa of 0.541 against 0.464 for the fused vector, with macro AUC 0.798 against 0.759 and accuracy 0.663 against 0.591. The intervals on kappa overlap only slightly, so this is the clearest single indication that cross-lateral fusion does not improve severity ordering.

External validation

The evaluation above is confined to a single dataset. To test whether the learned representations transfer, we applied the frozen encoders to an independent screening cohort assembled from a different source, comprising 11,913 patients each with the four standard projections. A subsample of 1,000 patients, stratified on patient-level severity so that the scarce malignant cases are retained, yields 2,000 breasts of which 41 are malignant and a further 62 benign. Images were resized and standardised exactly as in training, and no encoder parameter was updated.

Two differences from the internal evaluation must be kept in view. The external label is derived from biopsy outcome, not from a radiologist's BI-RADS category, so it is a related but distinct target from the one the probe was trained to predict. The cohort also carries no annotation of individual finding categories, and in particular none of the comparison-defined categories on

which the bilateral claim rests. This experiment therefore tests robustness of the representation, and cannot adjudicate whether cross-lateral fusion improves detection of asymmetry.

We report two protocols. Under transfer, the probe fitted on the internal training partition is applied unchanged, which is external validation in the strict sense and carries the label mismatch just described. Under refit, a fresh probe is fitted on half the external patients and evaluated on the other half, patient-disjoint; this removes the mismatch but tests the encoder rather than the deployed pipeline.

Table 5 gives, for each protocol, representation and target, the AUC with its 95% patient-clustered bootstrap interval over 2,000 resamples, the average precision, and the sensitivity at 90% specificity. Under refit the positive counts refer to the evaluation half only, which is why they are roughly half those under transfer.

Table 5. External validation on an independent screening cohort.

Protocol	Representati on	Target	Positiv es	AUC [95% CI]	AP	Sens @ 90% spec.
Transfer	Unilateral (independent)	Malignant	41 / 2000	0.576 [0.479, 0.668]	0.034	0.171
Transfer	Fused z_L	Malignant	41 / 2000	0.623 [0.533, 0.709]	0.043	0.220
Transfer	Pre-fusion	Malignant	41 / 2000	0.584 [0.493, 0.671]	0.041	0.195
Transfer	Unilateral (independent)	Any abnormality	103 / 2000	0.648 [0.593, 0.703]	0.084	0.194
Transfer	Fused z_L	Any abnormality	103 / 2000	0.652 [0.596, 0.703]	0.089	0.204
Transfer	Pre-fusion	Any abnormality	103 / 2000	0.644 [0.591, 0.693]	0.094	0.165
Refit	Unilateral (independent)	Malignant	23 / 1000	0.610 [0.500, 0.723]	0.048	0.130
Refit	Fused z_L	Malignant	23 / 1000	0.551 [0.425, 0.684]	0.082	0.174
Refit	Pre-fusion	Malignant	23 / 1000	0.636 [0.515, 0.754]	0.095	0.217
Refit	Unilateral (independent)	Any abnormality	51 / 1000	0.653 [0.578, 0.729]	0.086	0.235
Refit	Fused z_L	Any abnormality	51 / 1000	0.657 [0.586, 0.725]	0.094	0.235
Refit	Pre-fusion	Any abnormality	51 / 1000	0.677 [0.605, 0.748]	0.098	0.196

Discrimination falls substantially relative to the internal partition, where the same representations reach an area under the curve of 0.82 on the binary task. Two causes are confounded in that drop and this design cannot separate them: the target has changed from radiologist suspicion to biopsy outcome, and the imaging population, equipment and acquisition protocol have all changed. The broader endpoint, any abnormality against normal, is the better powered of the two at 103 positives and places all three representations between 0.64 and 0.66 with intervals clear of chance, so the representations do carry transferable information about abnormality. The malignant endpoint, with 41 positives and an average precision of 0.03 to 0.04 against a base rate of 0.02, is underpowered and we do not draw conclusions from it.

The comparison between representations is inconclusive, as internally. Under transfer the fused vector leads the independently trained baseline on the malignant endpoint by 4.7 points of

AUC, but the paired interval runs from -5.2 to +15.4 ($p = 0.34$), and under refit the ordering reverses. Consistent with the internal result, external validation neither confirms nor excludes an advantage for cross-lateral fusion.

Retrieval and generation

Table 6 gives zero-shot retrieval of reports from images over the thousand-patient pool. Every retrieval condition clears chance by a wide margin, so image-text alignment was certainly learned, yet absolute performance stays low throughout. Corpus redundancy does not account for this. Median pairwise cosine similarity among distinct report embeddings in the test pool is 0.317, and an oracle ranking every duplicate group ahead of all other reports would reach 52.0% at R@1 and 99.8% at R@10, so the observed values fall far below both ceilings. The constraint is instance-level alignment. Matched image-report pairs reach mean cosine similarity 0.136 against 0.058 for mismatched pairs, a standardised difference of 0.88, giving a ranking accuracy of 0.691: the proportion of trials in which the correct report outranks a randomly chosen alternative. Prevailing in a thousand-way ranking requires outranking the maximum of 999 distractors rather than the average one, which places R@1 near one percent for alignment of this strength. R@1 at a fixed large pool size is therefore an insensitive summary, and we treat ranking accuracy as the primary retrieval measure, together with the pool-size curve of Table 7. That the same representation reaches a macro-averaged AUC of 88.8 under probing indicates that the visual side is discriminative and locates the constraint on the language side.

Corpus properties. Reports are generated deterministically from the structured annotations, and every distinct report string maps to exactly one finding vector in both partitions. The text supervision is therefore a lossless re-encoding of the labels the probe subsequently predicts rather than an independent signal. The training partition contains 1,580 distinct strings across 8,000 breast reports; the test partition contains 758 across 2,000, and 520 across the thousand-patient retrieval pool. Retrieval and generation scores must be read in that light.

In Table 6 each condition is ranked against the report pool of its own laterality. The ceiling row gives the best attainable value given that identical report strings are indistinguishable. Pre-fusion vectors receive no image-text gradient and are reported for continuity only.

Table 6. Zero-shot image-to-report retrieval.

Retrieval setting	Report pool	R@1	R@5	R@10	MedR	Ranking accuracy
Chance	-	0.10%	0.50%	1.00%	500	0.500
Left, pre-fusion	Left	0.30%	1.00%	1.90%	436	0.551
Right, pre-fusion	Right	0.20%	0.50%	1.70%	458	0.536
Left, fused	Left	0.60%	3.20%	5.10%	268	0.691
Right, fused	Right	0.60%	2.20%	4.20%	263	0.694
Joint z bil	Left	0.70%	1.90%	4.60%	273	0.689
Duplicate-aware ceiling	-	52.0%	95.4%	99.8%	-	-

Paired bootstrap on the same resampled queries gives a median-rank reduction of 169 (95% CI 133 to 216) for the left fused vector and 174 (136 to 220) for the right, both $p < 0.001$, while differences in R@1 are not individually significant ($p > 0.27$). Ranking accuracy is therefore reported as the primary retrieval measure. The joint vector at 0.689 is statistically indistinguishable from either per-side vector, so the submitted claim that it underperforms is corrected to parity without gain.

Table 7 gives R@1 and mean reciprocal rank of the fused left-side vector at pool sizes from 10 to 1,000.

Table 8 gives automatic scores for the bilateral report decoder on the held-out partition, measured with BLEU, ROUGE and BERTScore [19–21]; BERTScore is computed against roberta-large.

Table 7. Retrieval performance versus pool size.

Pool size	10	50	100	500	1000
R@1	23.7%	7.3%	5.0%	0.67%	0.60%
MRR	0.447	0.208	0.141	0.039	0.026

Table 8. Automatic scores for report generation.

Generation measure	Score	Generation measure	Score
ROUGE-1	66.10	BLEU-1	60.82
BLEU-2	53.05	ROUGE-L	50.82
ROUGE-2	48.65	BLEU-4	44.22
BERTScore-F1	91.69	—	—

Lexical agreement and semantic similarity are both high, and every decoded report adopts the two-sided organisation seeded by a laterality prefix, confirming that the joint visual prefix does condition the decoder on the complete examination. High n-gram scores are unsurprising given templated references; the semantic score of 91.69 carries more information, indicating coherence beyond surface agreement. One qualitative shortcoming is limited variety among opening technique sentences, traceable to uniformity in the training prose rather than to any failure of conditioning.

A worked example

To show what the automatic scores correspond to in practice, Figure 3 presents one examination from the held-out partition together with its reference and decoded reports. The case is unremarkable on both sides, graded BI-RADS 1 throughout, with heterogeneously dense parenchyma. Both networks are exactly those that produced every measurement above, with nothing tuned for this particular case.

Extended case set

A single unremarkable example does not show how the model behaves where it matters. We therefore report a case set selected from the same predictions from which the tables above are computed, so that the cases cannot diverge from the numbers. It comprises every positive instance of each comparison-defined category rather than a chosen example, one exemplar at each of BI-RADS 3, 4 and 5 taken at median rather than extreme model confidence, and the five highest-confidence false positives and five highest-confidence false negatives on the binary task. The set totals 75 breasts from 62 patients.

Table 9 describes the composition of this case set on the held-out partition. The mean $p(\text{abnormal})$ column gives the predicted probability of the fused per-side representation under the binary abnormality probe of Table 4. Exemplar and error groups list both breasts of the selected patients, so the BI-RADS range also spans the contralateral side.

Table 9. Composition of the extended qualitative case set.

Case group	Breasts	Patients	BI-RADS range	Mean $p(\text{abnormal})$
Global asymmetry, all positives	3	3	3 to 4	0.890
Asymmetry, all positives	20	20	3 to 5	0.644
Focal asymmetry, all positives	26	26	3 to 5	0.595

Case group	Breasts	Patients	BI-RADS range	Mean p(abnormal)
BI-RADS 3 exemplar	2	1	2 to 3	0.639
BI-RADS 4 exemplar	2	1	1 to 4	0.564
BI-RADS 5 exemplar	2	1	1 to 5	0.992
Highest-confidence false positives	10	5	1 to 2	0.985
Highest-confidence false negatives	10	5	1 to 4	0.072

All three global asymmetry breasts are scored as abnormal, with predicted probabilities of 0.98, 0.95 and 0.74. Because these probabilities come from the binary abnormality probe rather than the per-category probe of Table 3, they show that the cases are recognised as abnormal but do not by themselves resolve the instability of the Global Asymmetry AUC. For the twenty asymmetry and twenty-six focal asymmetry positives the mean predicted probability is 0.64 and 0.60 respectively, above the decision threshold of 0.5 but with limited margin.

The errors are more informative than the successes. All ten false-positive breasts belong to examinations graded BI-RADS 1 or 2 with no annotated finding of any kind, yet receive probabilities between 0.978 and 0.992: the model is confidently wrong rather than uncertain on these studies. The false negatives show the complementary failure. In four of the five patients one breast carries an annotated mass at BI-RADS 3 or 4 while the contralateral side is unremarkable, and the breast carrying the mass is scored below 0.09. Conspicuous unilateral disease alongside a normal opposite breast is thus the configuration the model handles worst, although it is precisely the configuration in which bilateral comparison should help most, and this should be read alongside the null result on the composite endpoint.

We note one limitation of this section. The generated reports for these cases have not been assessed by a radiologist, and n-gram and embedding similarity measures computed against template-derived references cannot establish clinical correctness. The case set and the model's predictions for it are reported here; a reader study over the same cases remains outstanding.

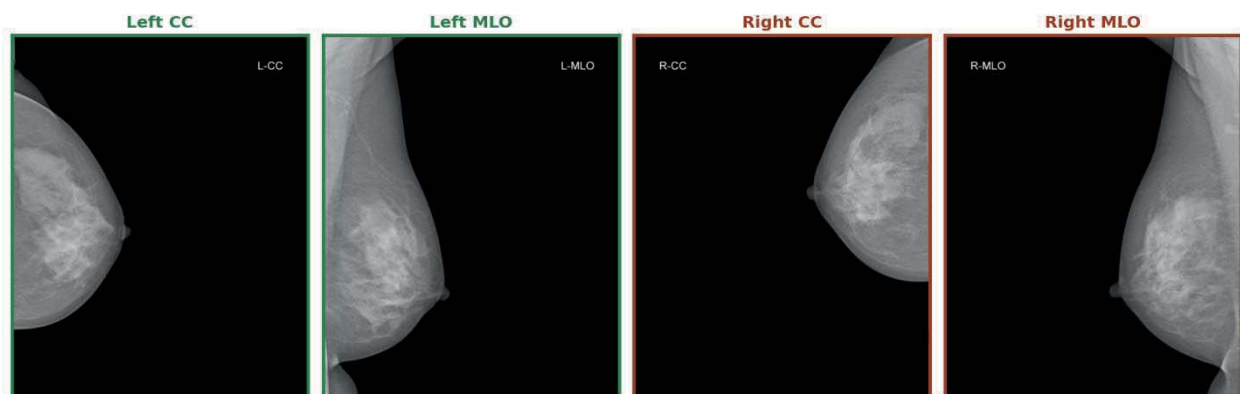


Figure 3. All four projections of an unremarkable examination from the held-out partition, arranged left cranio-caudal, left medio-lateral oblique, right cranio-caudal, right medio-lateral oblique.

Left breast: Standard bilateral two-view mammography was performed. No prior examinations are available for comparison. BREAST COMPOSITION: The breasts are heterogeneously dense, which may obscure small masses. FINDINGS: No dominant mass, suspicious calcification, or architectural distortion is seen. IMPRESSION: Negative mammogram. Annual screening mammography recommended. BI-RADS 1. Right breast: Digital mammography was performed with standard craniocaudal and mediolateral oblique views. No prior examinations

are available for comparison. BREAST COMPOSITION: The breasts are heterogeneously dense, which may obscure small masses. FINDINGS: The breast parenchyma appears symmetric with no focal abnormality identified. IMPRESSION: No mammographic evidence of malignancy. Routine screening in 1 year. BI-RADS 1.

Decoded report.

Left breast: Digital mammography was performed with standard craniocaudal and mediolateral oblique views. Prior studies were not available at the time of this report. BREAST COMPOSITION: The breasts are heterogeneously dense, which may obscure small masses. FINDINGS: No suspicious mass, calcification, or architectural distortion is identified. IMPRESSION: Negative mammogram. Annual screening mammography recommended. BI-RADS 1. Right breast: Digital mammography was performed with standard craniocaudal and mediolateral oblique views. Prior studies were not available at the time of this report. BREAST COMPOSITION: The breasts are heterogeneously dense, which may obscure small masses. FINDINGS: No suspicious mass, calcification, or architectural distortion is identified. IMPRESSION: Negative mammogram. Annual screening mammography recommended. BI-RADS 1.

Three things are worth drawing out. The decoder recovers the four-part organisation independently for each side and assigns the correct severity category to both, so the joint visual prefix is evidently carrying laterality-specific information rather than a single averaged summary. Second, and more informative for the bilateral question, nothing spurious appears on either side. The failure mode that concerns us in comparative cases is spillover, where a suspicious appearance on one side recruits abnormal vocabulary for the other; an examination that is unremarkable throughout is the cleanest test of whether gating leaves foreign context inert when there is nothing to transfer, and here it does.

The example also exposes the ceiling imposed by the synthetic corpus. Both sides of the decoded report open with the same technique sentence, whereas the reference varies its phrasing between them, and where the reference concludes the right side with an absence of malignant evidence and a one-year interval, the decoder substitutes a negative reading with annual screening. The two convey the same clinical content in different words. This is exactly the divergence that depresses n-gram agreement while leaving semantic similarity high, and it accounts for the gap between the BLEU figures and the BERTScore reported in Table 8.

Discussion

Probing does not establish the value of modelling both sides together. On the pre-specified composite endpoint the fused representation is 1.1 points of AUC below the independently trained single-breast model, with an interval from -9.6 to +7.5, and among the six adequately powered categories no difference survives correction. Averaged over the three comparative categories the fused mean is higher, 86.6 against 73.4, but that gap is produced almost entirely by a Global Asymmetry baseline estimated on three cases with an interval from 10.5 to 90.1, and the composite endpoint, which pools the same three categories into one adequately powered label, shows no advantage. Because these comparisons are made under frozen-feature probing, where the logistic head receives no category-specific information from the encoder, they speak to the representation rather than to any downstream classifier; what they do not show is a representational gain from cross-lateral fusion. Global asymmetry illustrates the hazard rather than the effect: three positive examples make any AUC estimate unstable in both directions, which is why the composite endpoint, not the per-category average, is the measure we pre-specified and report.

A methodological lesson surfaced along the way. Applied to every two-sided patient, the between-sides objective works against the symmetry term whenever an asymmetry is present, one drawing the two representations together while the other simultaneously separates them. Restricting the former to patients without asymmetry dissolves the antagonism and yields a consistent geometry. Stated generally, this suggests a principle for contrastive systems handling

inputs that are structurally paired yet clinically divergent: terms rewarding similarity and terms rewarding its opposite should not be permitted to act on the same elements simultaneously.

Direct numerical comparison against earlier mammographic CLIP variants is not available, since published work does not report frozen-feature AUC on this dataset over the same ten categories. That said, the macro-averaged AUC of 88.8 obtained here is consistent with strong contrastive representation learning on comparable medical benchmarks. The shape of the per-category differences is broadly what cross-lateral attention would predict, with Asymmetry and Global Asymmetry favouring the fused representation and the most conspicuous unilateral signs not, but the categories in which that pattern appears are precisely those with three to twenty positive cases. Whether the pattern reflects a targeted contribution or the noise of small samples cannot be settled on this test partition, and we do not claim the former.

Several constraints qualify these conclusions. Foremost is dependence on templated synthetic prose: faithful in vocabulary and organisation, it nonetheless lacks the variety of genuine clinical notes, which caps retrieval, limits decoder variety, and in all likelihood flatters the generation scores relative to practice. Assessment of the full labelled task is confined to one dataset. An external screening cohort is used to test transfer of the frozen representations, but it carries a biopsy-derived label rather than the BI-RADS category the probe was trained on and offers no annotation of individual finding categories, so it cannot validate the per-category or comparison-defined results. Validation across further institutions, equipment vendors and populations, with matched labelling, remains required before any claim to generality. Comparative findings are scarce in the test partition: three positive cases for Global Asymmetry, seven for Asymmetry and thirteen for Focal Asymmetry on the left side. Bootstrap intervals at these counts span twenty to forty AUC points, so per-category estimates in this group cannot support confirmatory claims and are reported as descriptive. A further constraint lies in the geometry of the learned spaces: the participation ratio of the singular values is 14.6 of 256 dimensions for the text embeddings and 33.9 of 256 for the image embeddings. A frozen report encoder with max-pooling over a templated corpus concentrates variance in a small number of directions, which suffices to separate ten finding categories but not to discriminate individual studies, and this accounts for the gap between probing performance and retrieval performance. Finally, no comparison against reader performance was undertaken, leaving the absolute figures unanchored to a clinical reference.

Table 10 reports a component ablation at a reduced and uniformly matched budget of six epochs, in which one component at a time is left out of the full model. Values are AUC, AP or $\kappa \times 100$ with 95% patient-clustered bootstrap intervals, and parentheses give the difference from the full model. Each condition was probed independently on its own frozen features, so the values are point estimates with marginal intervals and carry no between-condition significance test.

Table 10. Component ablation at a six-epoch budget.

Metric	Full model	- Cross-Lateral Attention	- symmetry loss
Composite asymmetry AUC	63.7 [56.6, 70.3]	68.6 [62.0, 74.9] (+4.9)	56.9 [50.2, 63.1] (-6.8)
Composite asymmetry AP	4.2 [2.8, 7.8]	4.7 [3.1, 7.9] (+0.5)	2.7 [2.0, 3.8] (-1.4)
Micro AUC	73.3 [70.4, 76.2]	73.1 [70.1, 76.1] (-0.2)	64.0 [60.0, 68.0] (-9.3)
Macro AUC, $n \geq 10$	67.9 [62.9, 72.6]	71.1 [66.4, 75.3] (+3.1)	62.2 [56.2, 67.3] (-5.8)
Comparative macro AUC	64.6 [57.0, 72.6]	72.3 [61.7, 80.1] (+7.7)	65.9 [56.6, 71.9] (+1.4)

Metric	Full model	- Cross-Lateral Attention	- symmetry loss
Binary AUC	61.4 [57.2, 65.4]	66.7 [62.2, 70.9] (+5.3)	58.8 [54.3, 63.5] (-2.6)
Binary PR-AUC	16.1 [12.9, 20.5]	23.2 [18.6, 29.3] (+7.1)	16.0 [12.8, 20.8] (-0.0)
Sensitivity at 90% spec.	21.4	31.8 (+10.4)	19.8 (-1.6)
BI-RADS quadratic kappa	5.0 [1.9, 7.9]	7.2 [4.3, 10.1] (+2.2)	2.6 [0.5, 4.7] (-2.4)
Validation loss at final epoch	18.40	19.26	18.42

Two cautions govern the reading of this table. First, at six epochs no condition has converged: every run selected its final epoch and validation loss was still falling. Absolute values are accordingly far below those of Table 4, where the full budget yields a binary AUC of 0.822 against 0.614 here, and the ordering observed at six epochs need not persist to convergence. Second, the comparison is confounded with optimisation speed: the cross-lateral block contributes 1.91 M randomly initialised parameters trained at three times the base learning rate, so a model containing it is further from convergence at a fixed small budget than one without it.

Within those limits the table records the following. Removing cross-lateral attention improves every downstream measure at this budget while worsening the training objective, whose final validation value rises from 18.40 to 19.26; the conditional image-image term likewise falls to 0.14 in the full model against 0.64 without fusion, confirming that the mechanism does align the two sides as intended. A lower contrastive loss therefore does not translate into better frozen-feature discrimination at six epochs. Removing the symmetry loss degrades every measure except the pooled comparative mean, which is the pattern the objective was designed to produce, though the intervals overlap throughout. Table 10 reports these observations as characterisation of the components under a reduced budget and not as evidence about the converged model.

Acknowledgement This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993145).

Conclusion

This study confronts a structural shortcoming in contrastive mammographic modelling: the habit of treating either breast as a self-contained object even though interpretation is inherently comparative. Bilateral MV-CLIP responds by encoding all four standard projections jointly and permitting the two sides to interact within a Cross-Lateral Attention block, so that comparison patterns are acquired inside a shared embedding space. Pairing a gate with a learned placeholder keeps the framework usable when a projection or an entire side is unavailable, removing any need for separate architectural variants.

The distinctive methodological element is the symmetry objective, which converts annotated asymmetry status into direct supervision: rather than presuming that a patient's two sides should always resemble one another, it promotes resemblance for unremarkable pairs and separation for asymmetric ones. Combined with per-side and cross-modal terms adapted from earlier contrastive mammography, this yields a multi-objective framework organising bilateral, per-side, and textual embedding spaces together. The observation that similarity-promoting and dissimilarity-promoting terms must be kept off the same elements transfers readily to related contrastive settings. Under a reduced-budget ablation the symmetry term is the component whose removal degrades performance most, although those runs are not trained to convergence and the finding is offered as characterisation rather than proof.

Empirically the case is weaker than we first reported. Measured against an independently trained single-breast model rather than against an unsupervised intermediate activation, explicit bilateral learning did not improve the pre-specified composite asymmetry endpoint on VinDr-Mammo, and no individual finding category showed a difference surviving correction. The

direction of movement across categories is consistent with the design, Asymmetry and Global Asymmetry favouring fusion while conspicuous unilateral signs do not, and the mechanism demonstrably operates: the conditional image-image objective falls sharply once the sides exchange information. What the test partition cannot do is resolve whether that mechanism yields a representational gain, since it contains 49 positive comparison-defined breasts and three cases of Global Asymmetry. The decoder produced structurally coherent reports spanning all four projections, while retrieval was limited by weak instance-level alignment and by embedding spaces occupying 15 of 256 available directions on the text side. On an external cohort the representations transfer at reduced discrimination, which is informative about robustness but cannot adjudicate the bilateral claim because that cohort lacks per-category annotation. Taken together, these results establish the architecture as workable and its benefit as unconfirmed at this sample size, which we regard as the honest reading of the evidence rather than a negative one. Beyond breast imaging, the same principles may prove useful wherever contrastive learning meets structurally paired anatomy

References

- [1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. <https://doi.org/10.3322/caac.21660>
- [2] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafi, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- [3] Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J., & Dogan, B. E. (2017). BI-RADS® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging*, 98(3), 179–190. <https://doi.org/10.1016/j.diii.2017.01.001>
- [4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, 139, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [5] Ghosh, S., Poynton, C. B., Visweswaran, S., & Batmanghelich, K. (2024). Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, & J. A. Schnabel (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2024* (Lecture Notes in Computer Science, Vol. 15012, pp. 632–642). Springer. https://doi.org/10.1007/978-3-031-72390-2_59
- [6] Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9(1), Article 12495. <https://doi.org/10.1038/s41598-019-48995-4>
- [7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, May 3–7). *An image is worth 16×16 words: Transformers for image recognition at scale* [Conference paper]. International Conference on Learning Representations, Virtual. <https://openreview.net/forum?id=YicbFdNTTy>
- [8] Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive multiview coding. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer vision – ECCV 2020* (Lecture Notes in

Computer Science, Vol. 12356, pp. 776–794). Springer. https://doi.org/10.1007/978-3-030-58621-8_45

[9] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of Machine Learning Research*, 119, 1597–1607. <https://proceedings.mlr.press/v119/chen20j.html>

[10] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729–9738). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00975>

[11] Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. *Proceedings of Machine Learning Research*, 182, 2–25. <https://proceedings.mlr.press/v182/zhang22a.html>

[12] Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., & Oktay, O. (2022). Making the most of text semantics to improve biomedical vision–language processing. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer vision – ECCV 2022* (Lecture Notes in Computer Science, Vol. 13696, pp. 1–21). Springer. https://doi.org/10.1007/978-3-031-20059-5_1

[13] Huang, S.-C., Shen, L., Lungren, M. P., & Yeung, S. (2021). GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3942–3951). IEEE. <https://doi.org/10.1109/ICCV48922.2021.00391>

[14] Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022). MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3876–3887). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.256>

[15] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>

[16] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 72–78). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-1909>

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

[18] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of Machine Learning Research*, 202, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>

[19] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>

[20] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Association for Computational Linguistics. <https://aclanthology.org/W04-1013/>

- [21] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020, April 26–30). *BERTScore: Evaluating text generation with BERT* [Conference paper]. International Conference on Learning Representations, Virtual. <https://openreview.net/forum?id=SkeHuCVFDr>
- [22] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 18661–18673. <https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html>
- [23] Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1), Article 277. <https://doi.org/10.1038/s41597-023-02100-7>
- [24] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of Machine Learning Research*, 97, 6105–6114. <https://proceedings.mlr.press/v97/tan19a.html>
- [25] Loshchilov, I., & Hutter, F. (2019, May 6–9). *Decoupled weight decay regularization* [Conference paper]. International Conference on Learning Representations, New Orleans, LA, United States. <https://openreview.net/forum?id=Bkg6RiCqY7>
- [26] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022, April 25–29). *LoRA: Low-rank adaptation of large language models* [Conference paper]. International Conference on Learning Representations, Virtual. <https://openreview.net/forum?id=nZeVKeeFYf9>