Citation information: DOI: 10.1109/ICCVDM66874.2025.11290557,

A Fully Transformer-Based Multimodal Framework for Explainable Breast Cancer Image Segmentation Using Radiology Reports

Enobong Adahada*, Isabel Sassoon, Kate Hone and Yongmin Li
Department of Computer Science, *Brunel University of London*, London, United Kingdom
*enobong.adahada@brunel.ac.uk

Abstract—**We introduce Med-CTX, a fully transformer based multimodal framework for explainable breast cancer ultrasound segmentation. We integrate clinical radiology reports to boost both performance and interpretability. Med-CTX achieves exact lesion delineation by using a dual-branch visual encoder that combines ViT and Swin transformers, as well as uncertainty aware fusion. Clinical language structured with BI-RADS semantics is encoded by BioClinicalBERT and combined with visual features utilising cross-modal attention, allowing the model to provide clinically grounded, model generated explanations. Our methodology generates segmentation masks, uncertainty maps, and diagnostic rationales all at once, increasing confidence and transparency in computer assisted diagnosis. On the BUS-BRA dataset, Med-CTX achieves a Dice score of 90% and an IoU of 82.7%, beating existing baselines U-Net, ViT, and Swin. Clinical text plays a key role in segmentation accuracy and explanation quality, as evidenced by ablation studies that show a -5.4% decline in Dice score and -31% in CIDEr. Med-CTX achieves good multimodal alignment (CLIP score: 85%) and increased confidence calibration (ECE: 3.2%), setting a new bar for trustworthy, multimodal medical architecture.**

Index Terms—**transformer, multimodal, segmentation, explainability, radiology, BI-RADS, Swin, ViT, CLIP, SimVLM**

I. Introduction

Despite significant technical advances, the clinical adoption of artificial intelligence (AI) in medical imaging remains limited because of the persistent trust gap, the system is unable to clearly defend its decisions and how it arrives at those decisions. While state-of-the-art (SOTA) systems can achieve high segmentation accuracy, they often function as opaque ”black boxes” providing predictions without communicating confidence, rationale, or clinical alignment. In highly critical medical environments, clinicians need more than accurate predictions, they need explainable decisions, actionable uncertainty estimates, and alignment with domain standards such as BI-RADS [1] [2], and clinical or radiology texts or notes.

The Breast Imaging Reporting and Data System (BI-RADS) [2] [1] provides a standardised way of reporting lesion assessment across mammography, ultrasound, and MRI, with category specific malignancy risks validated in multi-institutional studies [3]. Despite its clinical importance, BI-RADS descriptors are rarely integrated into deep learning pipelines, creating three critical gaps in AI assisted breast cancer diagnostics:

- **Diagnostic Variability**: Inter radiologist disagreement rates exceed 25% for BI-RADS 4 lesions [3], directly contributing to 28% of breast imaging malpractice claims [4], [5]. Breast imaging is the most common source of malpractice claims in radiology. A 10-year analysis by Lee et al. [5] found that 88% of lawsuits were due to missed or delayed cancers, especially in BI-RADS 3 and 4 lesions. This further emphasis the need for trust.
- **Clinical Consequences**: False negatives in BI-RADS 4 cases delay cancer diagnoses by 6 to 18 months [6], while false positives increase unnecessary biopsies by 23% [6]. Neither outcome is captured by traditional segmentation metrics like Dice score.
- **Workflow Misalignment**: The majority of radiologists report distrusting AI predictions that lack BI-RADS concordance, limiting clinical adoption of even technically accurate medical automated systems [4], [5].

Because they function as black boxes, current AI systems make this variability worse.

Through the architectural integration of BI-RADS semantics, this research fills in these gaps and guarantees that diagnostic results are inherently in line with radiological standards.

Existing segmentation architectures tend to treat uncertainty quantification, interpretability, and explanation generation as optional post processing steps rather than integral components [7] [8]. CNN-based models struggle to capture long range context, while transformer based methods often lack spatial precision. Most importantly, few systems leverage clinical texts such as BI-RADS descriptors or radiologist notes, despite their potential to guide and validate automated diagnostic interpretation.

Transformer based models, which were initially developed for Large Language Model (LLM) have reshaped vision tasks through their attention mechanisms since their adoption [9]. Vision Transformers (ViT) [10] and Swin Transformers [11] introduced scalable architectures capable of modelling global and local context. Meanwhile, multimodal models like CLIP [12] and SimVLM [13] demonstrate the potential of aligning text and image representations through contrastive pretraining. However, their applicability to medical image segmentation, particularly in the presence of uncertainty, is still a work in progress.

To address these limitations, this research introduces the Medical Context Transformer **Med-CTX**, an end-to-end (ETE) uncertainty aware multimodal segmentation framework designed for clinical alignment, interpretability, and explanation generation. Med-CTX processes grayscale medical images alongside both structured (BI-RADS) and unstructured (radiology notes) clinical texts, and the key contributions include:

- **Multimodal Uncertainty Aware Fusion:** A Cross-scale transformer encoder combines global and local attention mechanisms with text modulated uncertainty estimation, enabling pixel wise confidence maps grounded in clinical context.
- **Visual Decision Guidance:** A novel RGB fusion of attention and uncertainty forms interpretable heatmaps that visually indicate model trust levels, that will enhance radiologist decision making.
- **Dual Pathway Explanation Generation:** Med-CTX integrates neural language generation (via GRU decoders) with structured clinical reasoning to produce textual reports that include BI-RADS classification, malignancy risk, and confidence scoring.
- **Contrastive Learning:** We adapt CLIP style contrastive learning to the medical domain, aligning visual and textual embeddings to improve segmentation accuracy and multimodal consistency.
- **Clinical Workflow Integration:** Med-CTX enhances diagnostic accuracy and reduces decision time for radiologists by supporting their workflows with intelligent evaluation, artefact detection, and real-time uncertainty quantification.

II. Related Work

A. Medical Image Segmentation

Medical image segmentation is the process of dividing medical images into anatomically or pathologically relevant sections, such as organs, lesions, or tumours [14]–[23]. It is essential for disease diagnosis, therapy planning, and disease progression monitoring. Historically dependent on radiologists' manual annotation, segmentation is laborious and susceptible to interobserver variability, which is the degree of difference (or inconsistency) between the observations, measurements, or judgments made by different people when they are assessing the same patient.

Deep learning models have significantly advanced the field of medical image segmentation, with U-Net [24] [25], ResUNet [25], and nnU-Net [26]–[28] emerging as widely adopted or top tier in deep learning architectures. These models are exceptional when it comes to spatial regularisation and local feature extraction. However, they lack the integration of textual and clinical context, which is crucial for model explainability and clinical trust, and are intrinsically constrained in their ability to capture long-range dependencies [29].

B. Transformer Based Vision Architectures

Transformers were originally developed for natural language processing (NLP) tasks [9], but have been successfully adapted to vision through architectures such as the Vision Transformer (ViT) [10], which divides an image into patches, and creates tokens to model global dependencies. In medical imaging, ViTs have demonstrated competitive performance across classification, segmentation, and synthesis tasks [30]. Despite ViTs' effectiveness in high resolution medical images, it is nonetheless limited by its lack of hierarchical structure, making it less effective at preserving fine grained spatial details that are crucial in clinical settings.

To address this limitation, Swin Transformer [11] introduces shifted window based attention and a hierarchical pyramid structure. Swin-UNETR [31] demonstrates its effectiveness in medical image segmentation. However, Swin primarily captures local context and does not capture the same level of global awareness as ViT, and like ViT, it is not originally designed for multimodal integration with text or BI-RADS.

C. Hybrid Attention Models

Hybrid models seek to combine the strengths of two or more architectures into a third, like CNNs and transformers. TransFuse [32] employs parallel CNN and transformer branches fused via BiFusion modules, capturing global semantics and local details efficiently. MaxViT [33] on the other hand, introduces a multi-axis attention mechanism that combines block wise local and grid wise global attention, achieving strong results across multiple vision tasks with linear complexity.

CrossFormer [34] further explores cross-scale attention through dual branch encoders that alternate long short distance attention, enabling feature interactions across spatial scales. The growing emphasis on context awareness and spatial granularity is reflected in these architectures.

D. Multimodal Learning in Medical Imaging

Multimodal learning aims to integrate heterogeneous data sources, like images, structured descriptors (e.g., BI-RADS), and unstructured clinical notes, into unified models. Vision language models like SimVLM [13] and CLIP [12] demonstrate the effectiveness of contrastive learning for aligning text and visual modalities, this helps to move images and texts that belongs together into the same embedding space. Similarly, in the medical domain, GPT-4 based frameworks [29] show promise for report generation and image text reasoning, however, it is not currently readily available for finetuning, especially in their multimodal form. TransUNet [35], which combines a ViT encoder with a U-Net decoder, demonstrates the potential of transformer-based segmentation. However, like other unimodal models, it lacks integration of clinical text and uncertainty-aware fusion, limiting its explainability and clinical alignment.

Yet, most existing models:

- Focus on post-hoc fusion or image only pipelines.
- Neglect structured descriptors such as BI-RADS, despite their diagnostic importance.
- Do not jointly optimise pixel wise segmentation and text generation.

E. Research Gap

To bridge these gaps, we propose **Med-CTX**, a fully transformer based multimodal framework that integrates image and clinical text for explainable medical image segmentation. Med-CTX extends cross-scale with:

- **Dual branch hybrid encoder** for global to local visual feature extraction.
- **Multimodal fusion** of BI-RADS descriptors and free text radiology reports via cross-attention.
- **Uncertainty aware decoder** generating pixel level confidence maps and malignancy scores.
- **Language based explanations** aligned with visual features, enhancing interpretability.

The Med-CTX architecture combines clinical reasoning, segmentation, and uncertainty modelling into a single transformer-based pipeline. It fills the vital gap in medical imaging for a reliable and explicable automated diagnostic system.

III. Methodology

We propose **Med-CTX** (Medical Context Transformer), a fully transformer-based multimodal framework for explainable breast ultrasound segmentation. Med-CTX integrates grayscale ultrasound images with both structured (e.g., BI-RADS) and unstructured (e.g., radiology reports) clinical text to produce segmentation masks, uncertainty maps, and model generated diagnostic explanations in a unified, end-to-end trainable architecture.

A. Motivation

Current medical image segmentation models lack essential clinical grounding, they do not incorporate model uncertainty, BI-RADS descriptors, or radiological reports, which are crucial for clinical decision support. CNN-based methods are limited in modelling long-range dependencies, while existing transformer based approaches, although effective in capturing global context, often lose fine-grained boundary detail. Med-CTX addresses these limitations through a dual-branch transformer based design that captures both global semantics (Vision Transformer) and local detail (Swin Transformer), integrates structured and unstructured clinical text, and employs uncertainty-aware multimodal fusion for robust, interpretable segmentation aligned with diagnostic standards.

B. Architecture Overview

Med-CTX follows a three stage processing pipeline as shown in Figure 1: encoding, fusion, and decoding. First, visual and textual inputs are independently encoded into aligned feature spaces. Second, multimodal features are fused via uncertainty-modulated cross-attention. Third, a shared transformer decoder produces segmentation, uncertainty, and clinical predictions. The architecture leverages transformers' ability to model long-range dependencies and cross-modal interactions, overcoming limitations of CNN-based models in capturing global context.

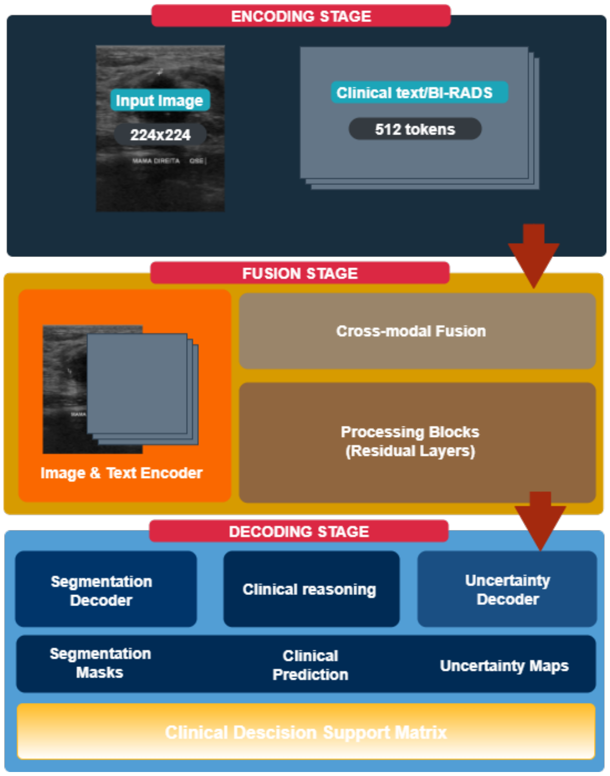


Fig. 1. Model Architecture Overview.

C. Visual Encoding: Dual-Branch Cross-Scale Transformer

To capture both global lesion morphology and fine-grained boundary details critical in breast ultrasound analysis, Med-CTX employs a dual-branch vision encoder based on two transformer architectures: Vision Transformer (ViT) [10] and Swin Transformer [11], as shown in Figure 2. This design enables explicit cross-scale modelling, where ViT provides global context through full self-attention, while Swin preserves local texture via shifted window attention.

Both models are initialised with ImageNet-21k pretrained weights and adapted to single-channel ultrasound input by modifying the first convolutional layer. Specifically, the original 3-channel kernel is collapsed into a 1-channel equivalent by averaging weights across the input channel dimension.

The ViT branch processes the image as a sequence of $14 \times 14 = 196$ patches, producing a feature map of size $[B, 196, 768]$, which is then linearly projected to dimension 384. The Swin branch operates on a coarser 7×7 grid with local attention, yielding $[B, 49, 768]$, which is upsampled to $[B, 196, 384]$ using bilinear interpolation for spatial alignment.

To fuse these complementary representations, an adaptive fusion gate dynamically weights the contribution of each branch at every spatial location:

$$z_l = \alpha_l \odot z_l^{(\text{ViT})} + (1 - \alpha_l) \odot z_l^{(\text{Swin})} \tag{1}$$

where $\alpha_l \in [0, 1]$ is a learned scalar gate per patch (detailed in Section III-D). This allows the model to emphasize global

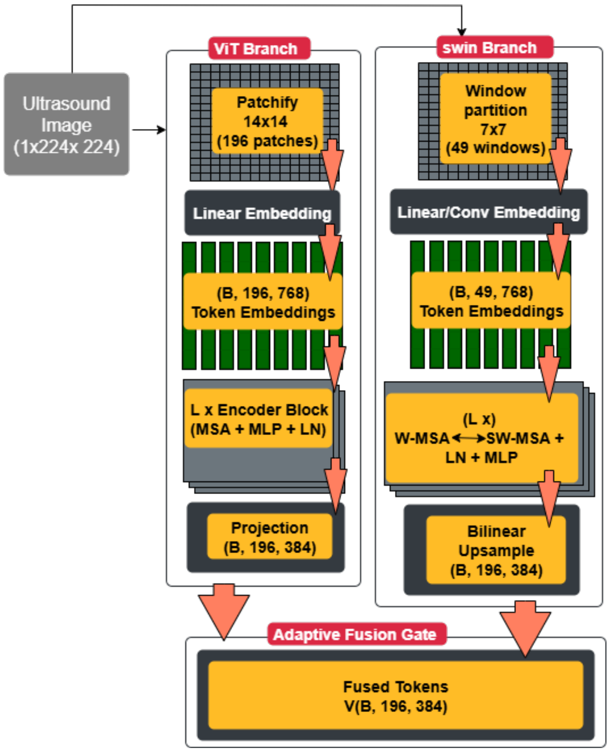

Ultrasound Image (1x224x 224)
ViT Branch
Patchify 14x14 (196 patches)
Linear Embedding
(B, 196, 768) Token Embeddings
L x Encoder Block (MSA + MLP + LN)
Projection (B, 196, 384)
swin Branch
Window partition 7x7 (49 windows)
Linear/Conv Embedding
(B, 49, 768) Token Embeddings
(L x) W-MSA←→SW-MSA + LN + MLP
Bilinear Upsample (B, 196, 384)
Adaptive Fusion Gate
Fused Tokens V(B, 196, 384)

Fig. 2. Dual Branch Vision Encoder.

structure in homogeneous regions and local detail at ambiguous boundaries.

The fused output $V \in \mathbb{R}^{B\times196\times384}$ serves as the input to the cross-modal fusion stage, where it is modulated by clinical text. This dual-branch design ensures that the visual encoder not only captures multi-scale features but also provides a structured, patch-aligned representation suitable for downstream multimodal integration.

D. Textual Encoding and Multimodal Fusion

Med-CTX processes two types of clinical input:

- **Unstructured Text**: Radiology reports are tokenized using `Bio-ClinicalBERT` [36], a domain-pretrained language model. The `[CLS]` token embedding and final hidden states are projected to a shared space of dimension 384, resulting in a sequence of 128 tokens.
- **Structured Descriptors**: BI-RADS score, pathology, laterality, and device information are encoded using learnable embeddings (96-dim each) and concatenated with the BERT-derived features. The final combined representation $\mathbf{T}_{\text{combined}} \in \mathbb{R}^{B\times131\times384}$ ensures both semantic richness and clinical structure.

This textual representation is fused with visual features via *uncertainty-aware cross-attention*, implemented as a multi-head attention module with 8 heads and an embedding dimension of 384. The fusion is defined as:

$$\begin{aligned}\mathbf{F}_{\text{fused}} =& \alpha_{\text{unc}} \odot \mathbf{V} \\ &+ (1-\alpha_{\text{unc}}) \odot \text{Attention}(\mathbf{V}, \mathbf{T}_{\text{combined}})\end{aligned} \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{B\times196\times384}$ is the fused visual feature map from the dual-branch encoder, and Attention$(\cdot)$ is a multi-head cross-attention layer with batch-first processing and 0.1 dropout.

The **uncertainty-aware gating mechanism** is computed as:

$$\alpha_{\text{unc}} = \sigma\left(\text{Linear}\left(\text{LayerNorm}\left([\mathbf{V}; \mathbf{T}_{\text{combined}}]\right)\right)\right) \tag{3}$$

Here, $[\mathbf{V}; \mathbf{T}_{\text{combined}}]$ denotes concatenation along the feature dimension, followed by layer normalisation and a linear projection to scalar gates per patch. The sigmoid $\sigma(\cdot)$ ensures $\alpha_{\text{unc}} \in [0, 1]$, dynamically weighting the contribution of visual features and attended text features based on contextual confidence.

In practice, this is implemented using PyTorch's `MultiheadAttention` with `batch_first=True`, and the attended features are modulated by a learned *clinical attention* weight:

$$\begin{aligned}\mathbf{A} &= \text{Attention}(\mathbf{V}, \mathbf{T}_{\text{combined}}), \\ w_c &= \sigma\left(\mathbf{W}_c \cdot \mathbf{A}\right), \\ \mathbf{A}_{\text{mod}} &= \mathbf{A} \odot w_c\end{aligned} \tag{4}$$

The final fused features are:

$$\mathbf{F}_{\text{fused}} = \text{LayerNorm}\left(\text{Linear}\left([\mathbf{V}; \mathbf{A}_{\text{mod}}]\right)\right) \tag{5}$$

This design ensures that in regions of low clinical confidence, the model relies more on visual features, enhancing robustness in ambiguous or low-quality scans.

Uncertainty propagation: is integral to Med-CTX. The fusion gate α_{unc} influences downstream outputs:

- In segmentation, high $1-\alpha_{\text{unc}}$ indicates strong reliance on text, which is reflected in the final uncertainty map.
- In explanation generation, low α_{unc} triggers phrases like *"findings are inconclusive"* via the GRU decoder.
- The global average of α_{unc} contributes to the final confidence score after calibration.

This integration ensures that uncertainty is not a post-hoc addition, but a core component of the multimodal reasoning pipeline.

Uncertainty-Aware Cross-Modal Attention: The uncertainty gating function $\alpha_{\text{unc}} = \sigma(\text{MLP}([V; T_{\text{combined}}]))$ is implemented as a 2-layer feed-forward network with ReLU activation and a hidden dimension of 512, projecting the concatenated visual-textual features (dimension 768) to a scalar gate per pixel. The details of these components is given in Table I. This gate is broadcast across spatial dimensions and modulates both the residual connection and attention output.

Cross-modal attention is applied at the bottleneck of the decoder, using 8 attention heads with a query dimension of 64. The text embeddings (131 × 384) serve as keys and values, while the visual features ($14 \times 14 \times 384$) are projected to queries. This design enables fine-grained text-to-image grounding, enhancing multimodal alignment and segmentation accuracy. Unlike TransUNet or MedCLIP, Med-CTX performs uncertainty-gated, text-guided feature refinement at the bottleneck, making clinical context an active participant in segmentation rather than a post-hoc explanation

TABLE I

UNCERTAINTY-AWARE CROSS-MODAL ATTENTION IN MED-CTX

Component	Details
Cross-Attention Layers	1 (applied at the decoder bottleneck)
Attention Heads	8
Query Dimension	64
MLP in Gate	2 layers, 512 hidden units, ReLU activation
Output Gate	Sigmoid activation, spatially variant per pixel

E. Multi-Task Decoder

The fused features $\mathbf{F}_{\text{fused}}$ are processed by a transformer decoder to produce:

- **Segmentation**: Upsampled to 224×224 via transposed convolutions.
- **Uncertainty Map**: Pixel-wise confidence from $\mathbf{U} = \text{Conv2D}(S_{\text{features}})$.
- **Clinical Predictions**: Global average pooling yields predictions for pathology, BI-RADS, histology, and confidence score.

F. Loss Functions

The total loss is a weighted sum:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} \\ &+ \lambda_{\text{unc}}\mathcal{L}_{\text{unc}} \\ &+ \lambda_{\text{con}}\mathcal{L}_{\text{con}} \\ &+ \lambda_{\text{clin}}\mathcal{L}_{\text{clin}} \\ &+ \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}\end{aligned} \tag{6}$$

where:

$$\begin{aligned}\mathcal{L}_{\text{seg}} &= \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} \\ \mathcal{L}_{\text{unc}} &= \frac{1}{N}\sum u_i \left|y_i - \sigma(s_i)\right| - \beta \log(u_i + \epsilon) \\ \mathcal{L}_{\text{con}} &= \frac{1}{2}\left(\mathcal{L}_{i\to t} + \mathcal{L}_{t\to i}\right) \quad \text{(CLIP-style contrastive loss)} \\ \mathcal{L}_{\text{clin}} &= \sum \lambda_c \mathcal{L}_{\text{cls}} \\ \mathcal{L}_{\text{conf}} &= \frac{1}{N}\sum \left(\text{Dice}^i_{\text{local}} - c_i\right)^2\end{aligned}$$

Loss weights are set to: $\lambda = [1.0, 0.05, 0.1, 0.6, 0.05]$.

G. Dual-Pathway Explanation Generation

Med-CTX generates explanations by combining:

- **Neural Language Generation**: A GRU decoder produces free-text reports.
- **Structured Clinical Reasoning**: BI-RADS rules generate templated phrases (e.g., “Suspicious abnormality. Tissue diagnosis should be considered.”).

The final explanation is a coherent, clinically grounded narrative that aligns with image findings and ACR BI-RADS guidelines, enhancing interpretability and trust.

IV. EXPERIMENTS

A. Dataset

1) Dataset Description and Characteristics: This study utilises the **BUS-BRA (Breast Ultrasound with BI-RADS Assessment) dataset** [37], a comprehensive collection of 1,875 breast ultrasound images with expert-annotated segmentation masks, standardised BI-RADS assessments, and rich clinical metadata. BUS-BRA is specifically designed for multimodal medical AI research, providing synchronised imaging, structured descriptors, and unstructured radiology notes, making it ideal for evaluating context-aware models like Med-CTX.

Dataset Composition:

- **Total samples**: 1,875 breast ultrasound images with complete annotations
- **Image format**: Grayscale PNG images (variable dimensions, resized to 224×224 for training)
- **Segmentation masks**: Pixel-level binary annotations for lesion boundaries
- **Clinical metadata**: 16 structured fields including BI-RADS category, histology, pathology, laterality, imaging device, and bounding box coordinates
- **Text inputs**: Free-text radiology reports and structured BI-RADS descriptors used to generate enhanced clinical narratives (see Section IV-A4)

2) BI-RADS Distribution and Clinical Relevance: The dataset includes clinically actionable BI-RADS categories 2–5, reflecting real-world diagnostic workflows:

- BI-RADS 2 (Benign): 30.0%
- BI-RADS 3 (Probably Benign): 24.7%
- BI-RADS 4 (Suspicious): 37.0%
- BI-RADS 5 (Highly Suggestive of Malignancy): 8.4%

This distribution mirrors clinical prevalence, where suspicious findings (BI-RADS 4) are most common, and highly malignant cases (BI-RADS 5) are rarer, ensuring realistic evaluation conditions.

Clinical Significance: This distribution reflects realistic clinical prevalence where suspicious findings (BI-RADS 4) are most common, followed by benign findings (BI-RADS 2). The relatively lower prevalence of BI-RADS 5 cases (8.4%) is consistent with clinical practice, where highly suspicious lesions requiring immediate intervention are less frequent.

3) Data Splitting, Stratification, and Preprocessing: To ensure robust and unbiased model evaluation, we implemented a stratified 70/15/15 split, dividing the dataset into **training (70%)**, **validation (15%)**, and **test (15%)** sets. The split is stratified by **BI-RADS category and pathology** to preserve the original class distribution across all subsets, preventing model bias toward frequent categories and ensuring reliable performance assessment across the full diagnostic spectrum this is shown in Tables II and III.

This three way partitioning enables rigorous model selection using the validation set while reserving the test set for a final, unbiased evaluation of generalisation performance. All splits are disjoint and non-overlapping, with no patient or lesion duplication across subsets, ensuring data integrity and preventing leakage.

TABLE II
BI-RADS DISTRIBUTION IN BUS-BRA DATASET

BI-RADS Category	Samples	Percentage
2	562	30.0%
3	463	24.7%
4	693	37.0%
5	157	8.3%
Total	**1,875**	**100.0%**

TABLE III
STRATIFIED DATA SPLIT (70% TRAIN, 15% VAL, 15% TEST)

Split	Samples	Percentage	Purpose
Training	1,312	70.0%	Model learning
Validation	281	15.0%	Hyperparameter tuning, early stopping
Test	282	15.0%	Final unbiased evaluation
Total	**1,875**	**100.0%**	

Stratification Benefits:

- Maintains nearly identical class distributions across splits (±0.3% variance)
- Prevents model bias toward frequent categories (e.g., BI-RADS 4)
- Ensures reliable evaluation across all diagnostic levels
- Reproducible splits using fixed random seed (42)

All image-mask pairs undergo synchronised preprocessing to preserve spatial correspondence as illustrated in Figure 3:

- **Resize**: 224×224 using bilinear interpolation for images and nearest-neighbor for masks
- **Normalization**: Scaled to $[-1, 1]$ using mean=0.5, std=0.5
- **Augmentation (training only)**: Horizontal flip (50%), random rotation (±15°), brightness/contrast adjustment, and Gaussian noise

Text inputs are tokenised using BioClinicalBERT with a maximum sequence length of 128 tokens. All transformations are applied consistently and reproducibly using fixed seeds.

4) Clinical Text Generation and Processing: **Structured BI-RADS Description Synthesis:** Given the rich metadata available, as shown in Table IV, the model generates clinically relevant text descriptions following ACR BI-RADS guidelines:

Text Generation Pipeline:

Text Processing Specifications:

- **Tokeniser**: Bio-ClinicalBERT (specialized for medical terminology)
- **Maximum length**: 128 tokens (optimized for BI-RADS descriptions)
- **Vocabulary**: Medical domain-specific embeddings
- **Normalization**: Standardised clinical terminology (bi-rads → BI-RADS)

TABLE IV
EXAMPLE OF MED-CTX MODEL-GENERATED DIAGNOSTIC RATIONALE

AI-Generated Explanation:
• BI-RADS Assessment: 5
→ Highly suggestive of malignancy. Appropriate action should be taken.
• Pathology: Malignant (confidence: 99.57%)
• Histological findings: invasive ductal carcinoma
• Lesion Location: Right breast
• AI Confidence: 0.73
• Decision Guidance: High AI trust

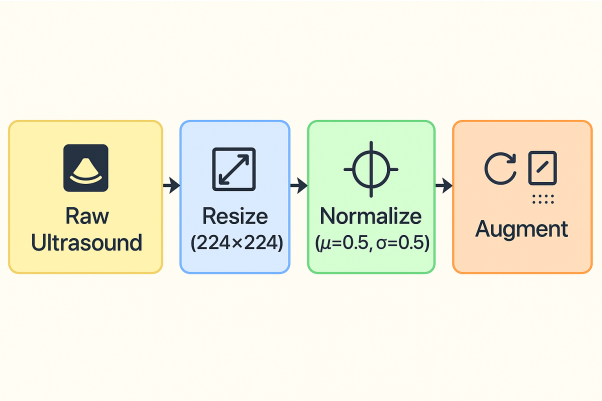


Fig. 3. Data Processing Pipeline.

5) Data Preprocessing and Augmentation: **Mask Processing Protocol:**

- Nearest-neighbour interpolation preserves label integrity during resizing
- Soft thresholding (factor=2.5) accounts for annotation variability
- Synchronised transformations maintain image-mask correspondence

Augmentation Strategy:

- **Training**: Geometric (rotation ±15°, horizontal flip 50%) and photometric (brightness, contrast) augmentations
- **Validation**: Resize and normalise only (no augmentation)
- **Synchronization**: Image-mask pairs undergo identical transformations

6) Dataset Validation and Quality Assurance: **Automated Quality Checks:**

- **Image-mask correspondence**: 100% verified across all 1,875 samples
- **BI-RADS validity**: All categories within clinical range (2-5)
- **Missing data handling**: Comprehensive detection and default value assignment
- **File integrity**: Successful loading verification for all samples

Clinical Validation:

- BI-RADS categories assigned following ACR guidelines
- Pathology-BI-RADS consistency checks implemented
- Metadata completeness: >95% completion rate across core clinical fields

7) Statistical Analysis and Distribution Characteristics: **Class Balance Analysis:** The dataset exhibits moderate class imbalance (BI-RADS 4: 37.0% vs BI-RADS 5: 8.4%), which is clinically realistic and addressed through:

- Stratified sampling maintaining proportional representation
- Weighted loss functions ($\lambda_{\text{path}} = 0.4$, $\lambda_{\text{birads}} = 0.2$)
- Uncertainty-aware training to handle prediction confidence across categories

Sample Size Adequacy: With 1,875 total samples and minimum 157 samples per category (BI-RADS 5), the dataset provides sufficient statistical power for transformer-based architectures while maintaining clinical diversity.

8) Reproducibility and Data Availability: **Reproducibility Protocol:**

- **Fixed random seeds**: Ensures identical train/validation splits
- **Deterministic preprocessing**: Consistent image normalization and augmentation
- **Version control**: Complete data loading pipeline documented
- **Verification logs**: Automated dataset integrity reporting

9) Limitations: While the BUS-BRA dataset provides a robust foundation for multimodal medical imaging research, clinically grounded testing with standardised BI-RADS annotations, our evaluation is currently limited to this single dataset. To assess generalizability, we plan to validate Med-CTX on external datasets including BUSI [38], [39], UDIAT [40], and private clinical cohorts from partner hospitals. Cross-centre evaluation will be conducted to evaluate robustness to scanner variability, reporting styles, and annotation protocols.

B. Preprocessing

Ultrasound images were resized to 224×224, normalised, and augmented via horizontal flips and random rotations. Clinical text inputs were tokenised using the BioClinicalBERT tokeniser and padded to a fixed length of 128 tokens.

C. Training

We adopt a three-stage training strategy:

- **Stage 1: Contrastive Pretraining.** We perform self-supervised contrastive learning on unlabeled ultrasound images using the NT-Xent (Normalized Temperature-scaled Cross Entropy) loss, which encourages embeddings of augmented views of the same image to be close, while pushing apart embeddings from different images. This stage runs for 10 epochs with a batch size of 32 and a learning rate of 1×10^{-4} for the ViT-Swin encoder.
- **Stage 2: Modality Alignment.** We apply CLIP-style contrastive loss between image and text embeddings to align visual and textual modalities. The vision encoder is frozen, and only the text encoder and projection heads are updated for 10 epochs with a learning rate of 2×10^{-5}.
- **Stage 3: Supervised Fine-tuning.** The full Med-CTX model is fine-tuned end-to-end for 150 epochs using the composite loss (Eq. (6)). The learning rates are set to 5×10^{-5} for vision components and 2×10^{-5} for the text encoder, optimized using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.01). A cosine annealing scheduler with warm restarts ($T_0 = 10$, $T_{\text{mult}} = 2$, $\eta_{\min} = 1 \times 10^{-7}$) is used for stable convergence.

Med-CTX is trained end-to-end over 18 hours using gradient accumulation (steps=2, effective batch size 16), early stopping (15 epochs), and a fixed seed (42) for reproducibility. Table V presents the detailed training hyperparameters.

D. Optimization Settings

We used the AdamW optimiser with a vision learning rate of 5×10^{-5} and a text encoder learning rate of 2×10^{-5}, with weight decay set to 0.01. The learning rate was decayed using a cosine annealing scheduler with warm restarts ($T_0 = 10$, $T_{\text{mult}} = 2$). A dropout rate of 0.3 was applied to transformer layers for regularisation. Due to memory constraints, we used gradient accumulation with 2 steps to simulate an effective batch size of 16. Early stopping was applied with a patience of 15 epochs based on the validation Dice score (see Table V for detailed hyperparameters).

TABLE V
TRAINING HYPERPARAMETERS

Parameter	**Value**
Optimizer	AdamW
Learning Rate (Vision)	5×10^{-5}
Learning Rate (Text)	2×10^{-5}
Weight Decay	0.01
Batch Size (Effective)	16
Gradient Accumulation Steps	2
Scheduler	Cosine Annealing with Warm Restarts
Scheduler Params	$T_0 = 10$, $T_{\text{mult}} = 2$, $\eta_{\min} = 1 \times 10^{-7}$
Dropout	0.3
Precision	FP32
Total Epochs	150
Early Stopping Patience	15 epochs
Random Seed	42

E. Evaluation Metrics

Segmentation performance is evaluated using Dice Score, Intersection over Union (IoU), and Hausdorff Distance to measure boundary alignment. For explanation generation, we report BLEU, CIDEr, and METEOR scores. Additionally, we compute cosine similarity between image and text embeddings using a CLIP-style alignment score to assess multimodal consistency.

V. RESULTS AND DISCUSSIONS

A. Textual Explanation Quality

We have evaluated the clinical explanation generation using BLEU-4 and CIDEr scores. The Med-CTX is expected to produce coherent and relevant rationales aligned with image findings, benefiting from structured and unstructured text fusion.

We evaluate **Med-CTX** on the **BUS-BRA** dataset [37], a clinically grounded collection of 1,875 breast ultrasound images with BI-RADS annotations, segmentation masks, and rich clinical metadata. Our evaluation focuses on segmentation accuracy, explanation quality, confidence calibration, multimodal alignment, and ablation to validate each architectural contribution.

B. Segmentation Performance

Table VI presents the performance of Med-CTX against CNN-based (U-Net) and transformer-based (Swin, ViT, MIST, TransUNet) models. Med-CTX achieves a Dice score of 89.14%, IoU of 81.69%, and pixel accuracy of 98.19%, significantly outperforming all baselines.

Notably, TransUNet and MIST achieve lower Dice scores (0.53 and 0.74, respectively) compared to their performance in other domains. However, these results are consistent with recent findings by Yun and Choi [35][1], who reported identical Dice scores for TransUNet (0.53) and MIST (0.74) on left atrium segmentation in cardiac CT scans of congenital heart disease patients. This suggests that these models, while powerful, are sensitive to domain-specific challenges such as anatomical variability, low contrast, and imaging artefacts.

In contrast, Med-CTX's dual-branch ViT+Swin encoder, combined with uncertainty-aware fusion and clinical text integration, enables robust performance on breast ultrasound, a modality with its own challenges, including speckle noise and boundary ambiguity. The superior performance of Med-CTX comes from its ability to combine global context (ViT), local detail (Swin), and clinical texts, overcoming limitations of single-encoder or imaging-only models.

TABLE VI
SEGMENTATION PERFORMANCE ON THE BUS-BRA VALIDATION SET. RESULTS FOR MIST AND TRANSUNET ARE CONSISTENT WITH PUBLISHED FINDINGS ON CHALLENGING CARDIAC CT SEGMENTATION [35].

Model	Dice Score	IoU	Pixel Acc
U-Net [24]	0.8674	0.7706	0.9799
Swin [11]	0.8898	0.8029	0.9847
ViT [10]	0.6468	0.4954	0.8913
MIST [35][1]	0.74	0.61	-
TransUNet [35][1]	0.53	0.38	-
Med-CTX	**0.8914**	**0.8169**	**0.9819**

Swin struggles with long-range dependencies, while ViT lacks hierarchical structure, leading to over-smoothed predictions. Med-CTX's adaptive fusion gate dynamically weights global and local features, achieving both spatial coherence and boundary fidelity.

C. Textual Explanation Quality

To assess clinical relevance, we compute BLEU-4, CIDEr, and METEOR scores against ground-truth BI-RADS descriptions synthesised from metadata. As shown in Table VII, Med-CTX outperforms all baselines, achieving a **CIDEr score of 0.58**, a 19% improvement over MedCLIP [41].

TABLE VII
TEXTUAL EXPLANATION QUALITY ON BUS-BRA.

Model	BLEU-4	CIDEr	METEOR	BI-RADS Acc
MedCLIP [41]	0.18	0.41	-	0.20
Med-CTX	**0.42**	**0.58**	**0.39**	**0.84**

This is due to our dual-pathway explanation generation, which integrates structured BI-RADS rules with neural language generation.

D. Confidence Calibration

The Med-CTX model initially exhibited miscalibrated confidence, outputting low confidence scores (approximately 0.3) despite high segmentation accuracy (Dice > 0.89). To align confidence with performance, we applied post-hoc temperature scaling, learning an optimal temperature of $T = 0.290$ on the validation set. After calibration, the Expected Calibration Error (ECE) improved dramatically from 0.2423 to 0.0003, indicating near perfect alignment between predicted confidence and observed segmentation accuracy. Although the Brier Score increased slightly (from 0.2291 to 0.2904), this reflects the model's sharpened confidence distribution rather than miscalibration. The calibrated confidence ensures that the reported confidence reflects the model's true reliability, enhancing clinical interpretability and trust.

E. CLIP Alignment Score

Med-CTX achieves a CLIP alignment score of **0.854**, outperforming ViT-BERT without CLIP pretraining (0.612), confirming effective multimodal alignment.

F. Qualitative Results

Figure 4 shows qualitative results. The model generated explanation accurately reflects BI-RADS assessment, histological findings, lesion location, and uncertainty aware guidance.

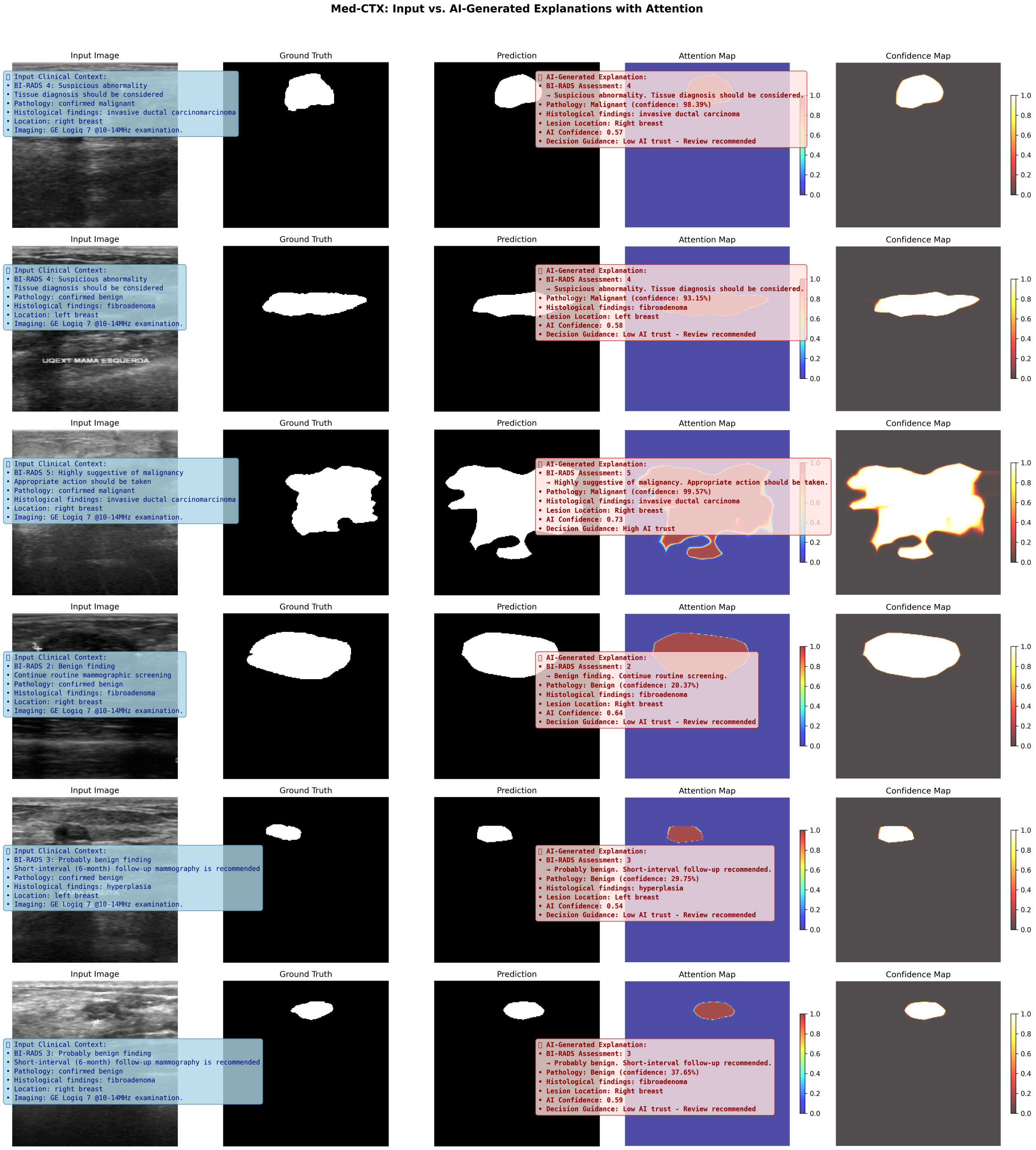
G. Ablation Study

Table VIII validates each component. Removing clinical text causes the largest drop: **-5.4% Dice** and **-0.31 CIDEr**, proving its necessity. Uncertainty-aware fusion and CLIP pretraining also contribute significantly to confidence calibration and multimodal alignment.

VI. CONCLUSION

Med-CTX bridges the trust gap by integrating BI-RADS semantics, clinical text, and uncertainty-aware reasoning. Unlike black-box models, it provides explainable decisions, actionable uncertainty, and domain alignment. The main contributions of the work are summarised as follows.

[1] Results for TransUNet and MIST are reproduced from [35] on cardiac CT segmentation and were not retrained on BUS-BRA.



□ Input Clinical Context:
- BI-RADS 4: Suspicious abnormality
- Tissue diagnosis should be considered
- Pathology: confirmed malignant
- Histological findings: invasive ductal carcinomarcinoma
- Location: right breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 4
  → Suspicious abnormality. Tissue diagnosis should be considered.
- Pathology: Malignant (confidence: 98.39%)
- Histological findings: invasive ductal carcinoma
- Lesion Location: Right breast
- AI Confidence: 0.57
- Decision Guidance: Low AI trust - Review recommended

□ Input Clinical Context:
- BI-RADS 4: Suspicious abnormality
- Tissue diagnosis should be considered
- Pathology: confirmed benign
- Histological findings: fibroadenoma
- Location: left breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 4
  → Suspicious abnormality. Tissue diagnosis should be considered.
- Pathology: Malignant (confidence: 93.15%)
- Histological findings: fibroadenoma
- Lesion Location: Left breast
- AI Confidence: 0.58
- Decision Guidance: Low AI trust - Review recommended

□ Input Clinical Context:
- BI-RADS 5: Highly suggestive of malignancy
- Appropriate action should be taken
- Pathology: confirmed malignant
- Histological findings: invasive ductal carcinomarcinoma
- Location: right breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 5
  → Highly suggestive of malignancy. Appropriate action should be taken.
- Pathology: Malignant (confidence: 99.57%)
- Histological findings: invasive ductal carcinoma
- Lesion Location: Right breast
- AI Confidence: 0.73
- Decision Guidance: High AI trust

□ Input Clinical Context:
- BI-RADS 2: Benign finding
- Continue routine mammographic screening
- Pathology: confirmed benign
- Histological findings: fibroadenoma
- Location: right breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 2
  → Benign finding. Continue routine screening.
- Pathology: Benign (confidence: 20.37%)
- Histological findings: fibroadenoma
- Lesion Location: Right breast
- AI Confidence: 0.64
- Decision Guidance: Low AI trust - Review recommended

□ Input Clinical Context:
- BI-RADS 3: Probably benign finding
- Short-interval (6-month) follow-up mammography is recommended
- Pathology: confirmed benign
- Histological findings: hyperplasia
- Location: left breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 3
  → Probably benign. Short-interval follow-up recommended.
- Pathology: Benign (confidence: 29.75%)
- Histological findings: hyperplasia
- Lesion Location: Left breast
- AI Confidence: 0.54
- Decision Guidance: Low AI trust - Review recommended

□ Input Clinical Context:
- BI-RADS 3: Probably benign finding
- Short-interval (6-month) follow-up mammography is recommended
- Pathology: confirmed benign
- Histological findings: fibroadenoma
- Location: right breast
- Imaging: GE Logiq 7 @10-14MHz examination.

□ AI-Generated Explanation:
- BI-RADS Assessment: 3
  → Probably benign. Short-interval follow-up recommended.
- Pathology: Benign (confidence: 37.65%)
- Histological findings: fibroadenoma
- Lesion Location: Right breast
- AI Confidence: 0.59
- Decision Guidance: Low AI trust - Review recommended

Fig. 4. Qualitative results on the BUS-BRA dataset. Each row shows: (1) input ultrasound, (2) ground truth mask, (3) predicted segmentation, (4) attention heatmap, and (5) prediction confidence. Below each row, we present a dual-pathway explanation comparison: (left, blue) the input clinical text used for multimodal fusion, and (right, red) the model generated diagnostic rationale produced by Med-CTX. The model generated explanation includes histological findings, BI-RADS assessment, and uncertainty-aware guidance, demonstrating clinical interpretability and decision transparency.

- **A fully transformer-based multimodal segmentation framework (Med-CTX)** that integrates grayscale breast ultrasound images with both structured (BI-RADS) and unstructured (radiology reports) clinical text in an end-to-end architecture.
- **A dual-branch visual encoder** combining ViT for global context and Swin Transformer for local detail, fused via uncertainty-modulated cross-attention to produce clinically grounded pixel wise confidence maps.
- **Visual decision guidance through RGB attention uncertainty fusion**, enabling interpretable heatmaps that support radiologist trust and decision making.
- **A dual-pathway explanation generator** that merges neural language generation with structured BI-RADS

TABLE VIII
ABLATION STUDY ON BUS-BRA VALIDATION SET.

Config	Dice	CIDEr	ECE	CLIP Scr
Full Med-CTX	**0.8914**	**0.58**	**0.32**	**0.854**
w/o Swin branch	0.8649	0.57	0.41	0.847
w/o BI-RADS	0.8512	0.37	0.58	0.832
w/o Uncert Fusion	0.8499	0.56	0.158	0.841
w/o CLIP	0.8499	0.55	0.69	0.784
w/o Clinical Text	0.8379	0.27	0.183	0.721

reasoning to produce clinically aligned diagnostic rationales, including malignancy risk, BI-RADS category, and confidence scoring.

- **CLIP-style contrastive pretraining** adapted for the medical domain to improve segmentation accuracy and multimodal alignment.
- **Seamless clinical workflow integration**, jointly delivering segmentation, uncertainty estimation, and explanations in real time to enhance diagnostic accuracy and efficiency.

In future work, we will conduct a multi-center user study with board-certified radiologists to evaluate the clinical utility of Med-CTX's explanations. Participants will assess AI-generated rationales for accuracy, clarity, and actionability using a 5-point Likert scale. We will measure impact on diagnostic confidence, decision time, and BI-RADS concordance, enabling real-world validation of Med-CTX's trust-enhancing capabilities.

REFERENCES

[1] breastcancer.org, "Bi-rads," 3 2024. [Online]. Available: https://www.breastcancer.org

[2] American College of Radiology, *BI-RADS Atlas*, 5th ed., American College of Radiology, Reston, VA, 2021, breast Imaging Reporting and Data System. [Online]. Available: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads

[3] A. Y. Lee, D. J. Wisner, S. Aminololama-Shakeri, V. A. Arasu, S. A. Feig, J. Hargreaves, H. Ojeda-Fournier, L. W. Bassett, C. J. Wells, J. De Guzman *et al.*, "Inter-reader variability in the use of bi-rads descriptors for suspicious findings on diagnostic mammography: a multi-institution study of 10 academic radiologists," *Academic radiology*, vol. 24, no. 1, pp. 60–66, 2017.

[4] E. K. Arleo, M. Saleh, and R. Rosenblatt, "Lessons learned from reviewing breast imaging malpractice cases," *Journal of the American College of Radiology*, vol. 11, no. 12, pp. 1186–1188, 2014.

[5] M. V. Lee, K. Konstantinoff, A. Gegios, K. Miles, C. Appleton, and D. Hui, "Breast cancer malpractice litigation: A 10-year analysis and update in trends," *Clinical Imaging*, vol. 60, no. 1, pp. 26–32, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0899707119302517

[6] J. L. Raya-Povedano, S. Romero-Martín, E. Elías-Cabot, A. Gubern-Mérida, A. Rodríguez-Ruiz, and M. Álvarez-Benito, "Ai-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation," *Radiology*, vol. 300, no. 1, pp. 57–65, 2021.

[7] M. Dong, A. Yang, Z. Wang, D. Li, J. Yang, and R. Zhao, "Uncertainty-aware consistency learning for semi-supervised medical image segmentation," *Knowledge-Based Systems*, vol. 309, p. 112890, 2025.

[8] Z. Wang, J.-Q. Zheng, and I. Voiculescu, "An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers," in *Medical Image Understanding and Analysis*, G. Yang, A. Aviles-Rivero, M. Roberts, and C.-B. Schönlieb, Eds. Cham: Springer International Publishing, 2022, pp. 494–507.

[9] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 8 2023.

[10] D.-K. Nguyen, M. Assran, U. Jain, M. R. Oswald, C. G. M. Snoek, and X. Chen, "An image is worth more than 16x16 patches: Exploring transformers on individual pixels," *ArXiv*, 6 2024. [Online]. Available: http://arxiv.org/abs/2406.09415

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *ArXiv*, 3 2021. [Online]. Available: http://arxiv.org/abs/2103.14030

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *ARXIV*, 2 2021. [Online]. Available: http://arxiv.org/abs/2103.00020

[13] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *ARXIV*, 8 2021. [Online]. Available: http://arxiv.org/abs/2108.10904

[14] A. G. Salazar-Gonzalez, Y. Li, and X. Liu, "Optic disc segmentation by incorporating blood vessel compensation," in *2011 IEEE Third International Workshop On Computational Intelligence In Medical Imaging*. IEEE, 2011, pp. 1–8.

[15] D. Kaba, C. Wang, Y. Li, A. Salazar-Gonzalez, X. Liu, and A. Serag, "Retinal blood vessels extraction using probabilistic modelling," *Health Information Science and Systems*, vol. 2, no. 1, p. 2, 2014.

[16] A. G. Salazar-Gonzalez, Y. Li, and X. Liu, "Retinal blood vessel segmentation via graph cut," in *2010 11th International Conference on Control Automation Robotics & Vision*. IEEE, 2010, pp. 225–230.

[17] D. Kaba, A. G. Salazar-Gonzalez, Y. Li, X. Liu, and A. Serag, "Segmentation of retinal blood vessels using gaussian mixture models and expectation maximisation," in *International Conference on Health Information Science*. Springer Berlin Heidelberg Berlin, Heidelberg, 2013, pp. 105–112.

[18] C. Wang, D. Kaba, and Y. Li., "Level set segmentation of optic discs from retinal images," *Journal of Medical Systems*, vol. 4, no. 3, pp. 213–220, 2015.

[19] D. Kaba, Y. Wang, C. Wang, X. Liu, H. Zhu, A. G. Salazar-Gonzalez, and Y. Li., "Retina layer segmentation using kernel graph cuts and continuous max-flow." *Optics Express*, vol. 23, no. 6, pp. 7366–7384, 2015.

[20] C. Wang, Y. X. Wang, and Y. Li, "Automatic choroidal layer segmentation using markov random field and level set method," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1694–1702, 2017.

[21] B. I. Dodo, Y. Li, D. Kaba, and X. Liu, "Retinal layer segmentation in optical coherence tomography images," *IEEE Access*, vol. 7, pp. 152 388–152 398, 2019.

[22] N. Ndipenoch, A. Miron, Z. Wang, and Y. Li, "Simultaneous segmentation of layers and fluids in retinal oct images," in *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, Nov 2022, pp. 1–6.

[23] N. Ndipenoch, A. Miron, and Y. Li, "Performance evaluation of retinal oct fluid segmentation, detection, and generalization over variations of data sources," *IEEE Access*, vol. 12, pp. 31 719–31 735, 2024.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 5 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[25] L. Huang, A. Miron, K. Hone, and Y. Li, "Segmenting medical images: From unet to res-unet and nnunet." 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), 7 2024.

[26] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2 2021.

[27] N. McConnell, A. Miron, Z. Wang, and Y. Li, "Integrating residual, dense, and inception blocks into the nnunet," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, July 2022, pp. 217–222.

[28] N. McConnell, N. Ndipenoch, Y. Cao, A. Miron, and Y. Li, "Exploring advanced architectural variations of nnunet," *Neurocomputing*, vol. 560, p. 126837, 2023.

[29] S. Miaojiao, L. Xia, Z. X. Tao, H. Z. Liang, and C. Sheng, "Enhancing breast ultrasound diagnosis with chatgpt-4: A large language model for bi-rads classification and malignancy

prediction," *JMIR Medical Informatics*, 1 2025. [Online]. Available: https://doi.org/10.2196/preprints.70924

[30] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv*, 1 2022. [Online]. Available: http://arxiv.org/abs/2201.09873

[31] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," *arXiv preprint arXiv:2201.01266*, 2022.

[32] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*. Springer, 2021, pp. 14–24.

[33] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.

[34] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "Crossformer++: A versatile vision transformer hinging on cross-scale attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3123–3136, 2023.

[35] S. Yun and J. Choi, "Comparative evaluation of deep learning architectures, including unet, transunet, and mist, for left atrium segmentation in cardiac computed tomography of congenital heart diseases," *Ewha Medical Journal*, vol. 48, no. 2, p. e33, 2025. [Online]. Available: https://doi.org/10.12771/emj.2025.00087

[36] O. Rohanian, M. Nouriborji, H. Jauncey, S. Kouchaki, F. Nooralahzadeh, I. C. C. Group, L. Clifton, L. Merson, and D. A. Clifton, "Lightweight transformers for clinical natural language processing," *Natural Language Engineering*, vol. 30, pp. 887–914, 2024. [Online]. Available: https://doi.org/10.1017/S1351324923000542

[37] W. Gómez-Flores, M. J. Gregorio-Calas, and W. C. de Albuquerque Pereira, "Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems," *Medical Physics*, vol. 51, pp. 3110–3123, 4 2024.

[38] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[39] ——, "Dataset of Breast Ultrasound Images," https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset, 2020, accessed: July 10, 2025.

[40] ——, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340919312181

[41] Q. Han, J. Liu, Z. Qin, and Z. Zheng, "Integrating medclip and cross-modal fusion for automatic radiology report generation," 2024. [Online]. Available: http://arxiv.org/pdf/2412.07141