

Article

XBusNet: Text-Guided Breast Ultrasound Segmentation via Multimodal Vision–Language Learning

Raja Mallina  and Bryar Shareef * 

Department of Computer Science, University of Nevada, Las Vegas, NV 89154, USA

* Correspondence: bryar.shareef@unlv.edu

Abstract

Background/Objectives: Precise breast ultrasound (BUS) segmentation supports reliable measurement, quantitative analysis, and downstream classification yet remains difficult for small or low-contrast lesions with fuzzy margins and speckle noise. Text prompts can add clinical context, but directly applying weakly localized text–image cues (e.g., CAM/CLIP-derived signals) tends to produce coarse, blob-like responses that smear boundaries unless additional mechanisms recover fine edges. **Methods:** We propose XBusNet, a novel dual-prompt, dual-branch multimodal model that combines image features with clinically grounded text. A global pathway based on a CLIP Vision Transformer encodes whole-image semantics conditioned on lesion size and location, while a local U-Net pathway emphasizes precise boundaries and is modulated by prompts that describe shape, margin, and Breast Imaging Reporting and Data System (BI-RADS) terms. Prompts are assembled automatically from structured metadata, requiring no manual clicks. We evaluate the model on the Breast Lesions USG (BLU) dataset using five-fold cross-validation. The primary metrics are Dice and Intersection over Union (IoU); we also conduct size-stratified analyses and ablations to assess the roles of the global and local paths and the text-driven modulation. **Results:** XBusNet achieves state-of-the-art performance on BLU, with a mean Dice of 0.8766 and IoU of 0.8150, outperforming six strong baselines. Small lesions show the largest gains, with fewer missed regions and fewer spurious activations. Ablation studies show complementary contributions of global context, local boundary modeling, and prompt-based modulation. **Conclusions:** A dual-prompt, dual-branch multimodal design that merges global semantics with local precision yields accurate BUS segmentation masks and improves robustness for small, low-contrast lesions.

Keywords: breast cancer; breast ultrasound; medical image analysis; deep learning; vision transformer (ViT); U-Net; prompt learning; BI-RADS; segmentation; vision–language models (VLMs)



Academic Editors: Wen-Shan Jian and Shuo-Chen Chien

Received: 1 September 2025

Revised: 1 November 2025

Accepted: 4 November 2025

Published: 11 November 2025

Citation: Mallina, R.; Shareef, B. XBusNet: Text-Guided Breast Ultrasound Segmentation via Multimodal Vision–Language Learning. *Diagnostics* **2025**, *15*, 2849. <https://doi.org/10.3390/diagnostics15222849>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Breast cancer is the most common cancer in U.S. women (about 30% of new cases in 2025), and roughly 42,170 women are expected to die from it [1]. Early detection and timely treatment remain central to lowering mortality [1]. Mammography, magnetic resonance imaging (MRI), and ultrasound are the main imaging modalities used for screening and diagnostic workups [2,3]. Mammography and MRI are accurate, but each has limitations: mammography loses sensitivity in women with dense breast tissue [4], and MRI, though highly sensitive, is costly and impractical for routine use [5]. Ultrasound is safe, painless, nonionizing, affordable, and widely available, which is valuable for younger patients

and in settings with limited resources [6]. Its interpretation, however, is highly operator-dependent and is often hindered by speckle noise, heterogeneous tissue appearance, and indistinct lesion boundaries, challenges shown in benchmark datasets such as BUSIS [7,8]. These factors motivate automated segmentation methods that are accurate, robust across imaging conditions, and interpretable to radiologists in routine practice.

Accurate tumor segmentation improves diagnostic precision by enabling measurements aligned with BI-RADS (size, shape, and margins), yielding more reliable radiomics features, keeping classifiers focused on the lesion, and supporting longitudinal tracking [9,10]. Automated segmentation improves reproducibility and can streamline reporting in routine ultrasound. Over the past decade, breast ultrasound (BUS) segmentation has progressed in three broad stages. Early studies used classical image processing—thresholding, edge detection, region growing, and active contours [11]. Next came machine learning pipelines that relied on hand-crafted texture and shape descriptors (e.g., GLCM, local binary patterns, wavelets) paired with supervised classifiers, such as SVMs and random forests, or with clustering and graph-based methods (fuzzy c-means, k-means, conditional random fields, graph cuts) [7]. With deep learning, U-Net and related encoder–decoder models became the standard [12,13]. Variants that add residual or dense connections and attention gates improved boundary quality [14–16]. Attention mechanisms and Transformer designs helped capture long-range context [17–19], and hybrid CNN–Transformer models reported strong results on BUS data, including a multitask (segmentation and classification) CNN–Transformer trained jointly for both tasks [20,21]. Despite this progress, BUS still presents challenging scenarios: small tumor sizes, low contrast, heterogeneous backgrounds, and variability across scanners and sites. These difficulties are compounded by limited labeled data, class imbalance, and annotation variability. Pixel only training also rarely uses explicit clinical descriptors (e.g., BI-RADS or radiomics), which can make outputs harder to interpret in routine reading [7,22]. Small tumor-aware architectures such as STAN and ESTAN directly target this failure mode [23,24]. In current BUS practice, lesion measurements are predominantly manual; automated, BI-RADS-aligned masks can standardize reporting and reduce inter-operator variability.

To address this gap, prior studies have injected domain knowledge into BUS segmentation in several ways. First, clinical descriptors and radiomics are used as auxiliary supervision or fused with visual features [10,25]. Second, anatomy-aware models and losses encode priors on shape, boundaries, and topology to keep masks plausible [26–30]. Third, ultrasound physics has been leveraged through speckle and attenuation models and quantitative ultrasound features [31–34]. These strategies improve clinical grounding but still face limits in scalability and generalization, which motivates frameworks that flexibly integrate descriptors while preserving robustness across imaging conditions.

Prompt-based learning treats segmentation as a conditional task steered by auxiliary inputs (“prompts”). In practice, prompts take several forms: (i) Spatial prompts. Image coordinates or regions such as points, bounding boxes, polygons/masks, and scribbles. The Segment Anything Model (SAM) is a canonical spatially promptable segmenter and supports these types at scale [35]. (ii) Textual prompts. Class names or free-form phrases that describe the target. CLIPSeg shows that text or image prompts can condition masks [36]. Pairing open-vocabulary detectors or vision–language models with SAM enables text-conditioned localization and segmentation; for example, pairing Grounding DINO with SAM (“Grounded-SAM”) uses free text to propose regions, and BLIP provides strong image–text pretraining [37–41]. (iii) Other side information. Categorical tags or numeric attributes (e.g., approximate lesion size or laterality) that can be encoded as tokens or embeddings and used alongside visual features.

Adapting these ideas to medical imaging is challenging because lesions often have subtle contrast, boundaries are weak, and domain shift across scanners and sites is common [42]. Several adaptations target these issues: BUSSAM introduces CNN–ViT adapters for breast ultrasound, MedSAM pretrains on large medical datasets, and SAMed applies low-rank adaptation for efficient tuning [43–45]. Vision–language approaches have also been explored directly for ultrasound, such as CLIP-TNseg [46]; causal interventions have been proposed to stabilize text guidance under noisy or ambiguous prompts [46,47]. These studies indicate benefits from both spatial and text prompts, yet most methods adopt a single prompt modality. Few approaches jointly exploit global semantic context and fine-grained clinical attributes in a manner consistent with radiologist reasoning. Furthermore, it is desirable that segmentation outputs are consistent with BI-RADS descriptors [9]. Common tools such as Grad-CAM and attention visualizations offer partial insight but often produce diffuse or inconsistent patterns that are difficult to map to clinical terms [48–50].

In this work, we present XBusNet, a dual-branch, dual-prompt vision and language framework for BUS segmentation. A global branch based on Contrastive Language Image Pretraining (CLIP) and a Vision Transformer (ViT) is conditioned by a Global Feature Context Prompt (GFCP) that encodes lesion size and centroids to provide scene-level semantics. A local branch based on U-Net with multi-head self-attention preserves boundary detail. The local decoder is modulated by a text conditioned scaling and shifting mechanism, Semantic Feature Adjustment (SFA), which is driven by a Local or Attribute Guided Prompt (LFP) that encodes shape, margin, and Breast Imaging Reporting and Data System (BI-RADS) terms. Prompts are assembled reproducibly from structured metadata, require no manual clicks, and integrate with standard training. We evaluate XBusNet on the BLU dataset using five-fold cross-validation, include ablations to isolate the role of each component, analyze performance across lesion sizes, and provide Grad-CAM visualizations as qualitative views of model focus relative to BI-RADS descriptors. We hypothesize that combining a Global Feature Context Prompt (GFCP; size/location) with a Local Feature Prompt (LFP; shape, margin, BI-RADS) guides the model toward plausible regions while preserving boundaries, improving BUS segmentation—especially for small or low-contrast lesions.

Our contributions are as follows:

- We propose XBusNet, a dual-branch, dual-prompt vision and language segmentation architecture that combines a CLIP ViT global branch conditioned by GFCP with a U-Net local branch with multi-head self-attention that is modulated via SFA driven by LFPs.
- We design a reproducible prompt pipeline that converts structured metadata into natural language prompts, including global prompts for lesion size and centroid and local prompts for shape, margin, and BI-RADS.
- We introduce a lightweight SFA mechanism in the local decoder to inject attribute-aware scaling and shifting, improving boundary focus while preserving fine detail.
- We provide a comprehensive evaluation protocol with five-fold cross-validation, size-stratified analysis, component ablations, and Grad-CAM overlays used as qualitative visualizations of model focus relative to BI-RADS descriptors.

2. Materials and Methods

2.1. Datasets

We used the Breast Lesions USG (BLU) dataset [51], which contains 256 breast ultrasound scans from 256 patients. Each scan includes expert verified annotations for benign and malignant lesions. Four images with multiple annotated lesions (two tumors in the same image) were excluded because they represent a very small fraction of the dataset (4/256, $\approx 1.6\%$), which would yield an imbalanced and hard-to-interpret evaluation for

per-image metrics. The remaining 252 images include 154 benign and 98 malignant tumors. This dataset is also referred to as ‘BrEaST’. All images and labels are publicly available and fully de-identified in the BLU (‘BrEaST’) release; no new data were collected and no identifiable information was accessed.

The dataset provides pixel-wise segmentation masks, patient-level clinical attributes such as age, breast tissue composition, signs, and symptoms, image-level BI-RADS labels, and tumor-level diagnostic information such as shape, margin, and histopathology. We considered the BUSI [52] dataset; however, it does not provide the structured clinical text fields (e.g., BI-RADS shape/margin) needed for our prompt construction. Extending XBusNet to cohorts without such fields (via automatic prompt inference) is planned.

2.2. Pre-Processing and Prompt Construction

All images were resized to 352×352 pixels (we therefore took $H = W = 352$ throughout). We chose 352 for compatibility with the pretrained ViT/CLIP encoder (patch size, 16; $352 = 22 \times 16$) and to align with the backbone stride ($\times 32$). Prompts were derived from structured metadata in the dataset.

The Global Feature Context Prompt encodes lesion size and location. Size is taken from the dataset metadata and discretized using training-set quantiles (small/medium/large). Location handling differs by split: for training images, location is the centroid of the reference segmentation mask computed from image moments,

$$C_x = \frac{M_{10}}{M_{00}}, \quad C_y = \frac{M_{01}}{M_{00}}, \quad (1)$$

where M_{00} is the mask area and M_{10} and M_{01} are spatial moments. For validation/test images, location is estimated without masks from a first-pass probability map, as detailed in Section 2.7. The Local Feature Prompt encodes clinically used attributes—shape, margin, and BI-RADS—verbalized directly from metadata. Prompts are tokenized and embedded with the CLIP text encoder, as described in the model subsections.

2.3. Training Setup

We used five-fold cross-validation. For each fold, we split the 252 annotated images into train and validation sets at an 80 to 20 ratio, ensuring non-overlapping images. Unless stated otherwise, training ran for 1000 iterations per fold with a batch size of 4, the AdamW optimizer, an initial learning rate of 10^{-4} , and a cosine schedule. Mixed-precision training was enabled with automatic mixed precision.

Hyperparameter Tuning and Overfitting Control

Within each cross-validation fold, we performed a small hyperparameter sweep restricted to the fold’s training partition (no images from the fold’s validation/test partitions were used for selection). We varied a compact set of parameters (learning rate, weight decay, batch size, loss weights) while keeping all other settings fixed. The best configuration was chosen by the mean Dice of an internal split of the training partition and then fixed across all folds and methods we implemented for comparability. To mitigate overfitting, we limited the sweep budget, used weight decay (AdamW), and monitored the train–validation gap; we did not use data augmentation or dropout in order to isolate architectural effects. We did not observe unstable divergence across folds. We acknowledge that nested cross-validation would further decouple tuning from evaluation and leave a fully nested protocol to future work.

2.4. Evaluation Metrics

We report Dice, Intersection over Union (IoU), the false positive rate (FPR), and the false negative rate (FNR). Let $M \in \{0, 1\}^{H \times W}$ be the ground-truth mask, where H and W are the resized image height and width, and $P \in [0, 1]^{H \times W}$ is the predicted probability map. We obtain a binary prediction by thresholding at $\tau_{\text{seg}} = 0.5$:

$$\hat{Y} = \mathbb{1}\{P \geq \tau_{\text{seg}}\} \in \{0, 1\}^{H \times W}.$$

The pixel-wise counts (sums over all pixels (i, j)) are

$$\text{TP} = \sum_{i,j} \hat{Y}_{ij} M_{ij}, \quad \text{FP} = \sum_{i,j} \hat{Y}_{ij} (1 - M_{ij}), \quad \text{FN} = \sum_{i,j} (1 - \hat{Y}_{ij}) M_{ij}, \quad \text{TN} = \sum_{i,j} (1 - \hat{Y}_{ij}) (1 - M_{ij}).$$

The per-image metrics are

$$\text{Dice} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}, \quad \text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}},$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

We computed metrics per image and the average within each fold; the fold means were then summarized across the $K = 5$ folds (mean $\pm$ SD). For numerical stability, we used $\varepsilon = 10^{-8}$ in denominators during implementation.

2.5. Implementation Details and Hardware

All experiments were implemented in PyTorch v 2.9.0 and run on a cluster with $8 \times$ NVIDIA A100 Tensor Core GPUs (NSF MRI (#2117941)) in SXM form factor. The CLIP encoders were frozen during training. The ResNet50 encoder was initialized from ImageNet weights and fine-tuned. For deployment, we provided an inference-profiling script and notes on memory / throughput; latency depended on hardware and batch size.

2.6. Overall Architecture

XBusNet is a text-guided medical image segmentation framework that integrates fine-grained local context from medical images with global semantic cues derived from natural language prompts. The model follows a dual-branch, dual-prompt vision and language design with a Global Feature Extractor (GFE) and a Local Feature Extractor (LFE) that run in parallel and exchange information through a Semantic Feature Adjustment (SFA) mechanism (see Figure 1). The GFE uses a CLIP-based Vision Transformer with a patch size of 16 to capture global semantic relationships conditioned on a high-level conditional prompt, p_c . The LFE uses a ResNet50 encoder with Transformer blocks at deep levels and a U-Net-style decoder to represent spatial detail under a phrase-level prompt, p_ℓ .

Let $I \in \mathbb{R}^{B \times 3 \times H \times W}$ denote the input image batch and let p_c and p_ℓ be the conditional and local prompts. (Here, H and W denote the resized input-image size, whereas h and w denote the feature-map size at the working resolution.)

The two branches produce

$$\mathbf{F}_g = \text{GFE}(I, p_c), \quad \mathbf{F}_\ell = \text{LFE}(I, p_\ell), \quad (2)$$

with $\mathbf{F}_g \in \mathbb{R}^{B \times C_g \times h \times w}$ and $\mathbf{F}_\ell \in \mathbb{R}^{B \times C_\ell \times h \times w}$. (Here, C_g and C_ℓ denote the channel widths; in our implementation, $C_g = 64$ and $C_\ell = 32$). SFA modulates features in both branches using scaling and shifting parameters computed from the corresponding prompt embedding:

$$\begin{aligned}\hat{\mathbf{F}}_g &= \gamma_g \odot \mathbf{F}_g + \beta_g, & [\gamma_g, \beta_g] &= \text{SFA}(\mathbf{F}_g, \mathbf{e}_c), \\ \hat{\mathbf{F}}_\ell &= \gamma_\ell \odot \mathbf{F}_\ell + \beta_\ell, & [\gamma_\ell, \beta_\ell] &= \text{SFA}(\mathbf{F}_\ell, \mathbf{e}_\ell),\end{aligned}\quad (3)$$

where $\odot$ denotes element-wise multiplication. For channel-wise FiLM, $\gamma_g, \beta_g \in \mathbb{R}^{B \times C_g \times 1 \times 1}$ and $\gamma_\ell, \beta_\ell \in \mathbb{R}^{B \times C_\ell \times 1 \times 1}$, broadcast over (h, w) (not pixel-wise). Prompt embeddings $\mathbf{e}_c, \mathbf{e}_\ell \in \mathbb{R}^{B \times e}$ (with e being the text-embedding dimension) are linearly projected to a fixed width, d , for conditioning; we use the same fixed d throughout, including Equation (9). A residual form can be used at implementation time as $\mathbf{F}_{\text{out}} = \hat{\mathbf{F}} + \mathbf{F}$.

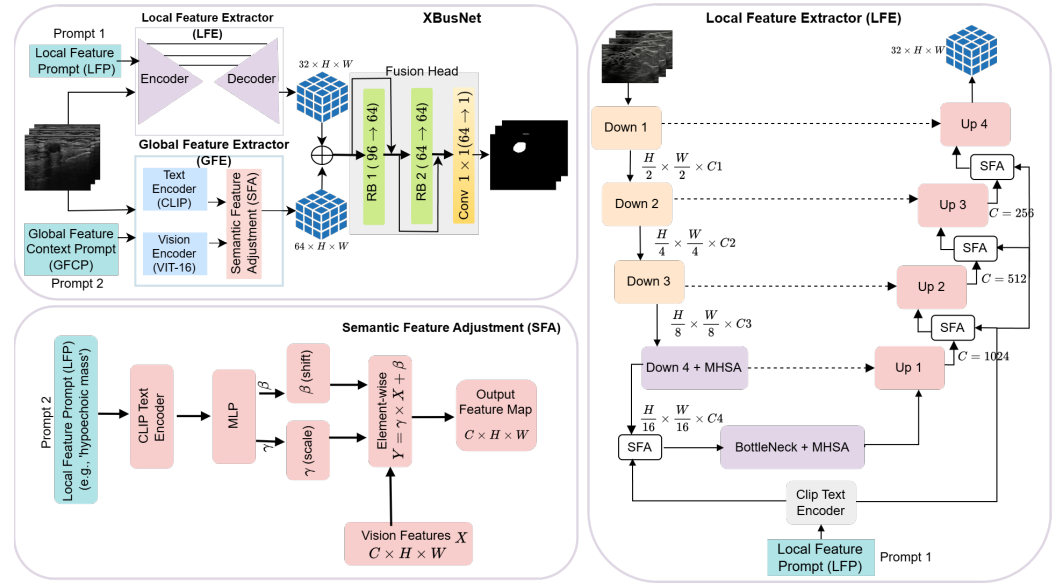


Figure 1. Overview of the proposed XBusNet architecture, showing the Local Feature Extractor (LFE), Global Feature Extractor (GFE), and Semantic Feature Adjustment (SFA) modules. RB = residual block; MHSA = multi-head self-attention.

The fusion head integrates the modulated features and prepares them for prediction:

$$\mathbf{F}_{\text{fused}} = \Phi_{\text{fuse}}([\hat{\mathbf{F}}_g, \hat{\mathbf{F}}_\ell]), \quad (4)$$

Here, Φ_{fuse} denotes the two residual blocks described in Section 2.10. This is followed by a 1×1 convolution that yields the segmentation logits:

$$\hat{\mathbf{Y}} = \text{Conv}_{1 \times 1}(\mathbf{F}_{\text{fused}}), \quad \hat{\mathbf{Y}} \in \mathbb{R}^{B \times 1 \times H \times W}. \quad (5)$$

This organization lets the network use language to steer scene-level context while the local pathway preserves boundaries and fine structure.

2.7. Global Feature Extractor (GFE)

2.7.1. Global Features and Prompts

Global features comprise size and location. Lesion size is read from the dataset meta-data and discretized using training-set quantiles (small/medium/large). Location handling differs by split: (i) Training. the location is the centroid (C_x, C_y) of the reference segmentation mask, computed from the image moments $C_x = M_{10}/M_{00}$, and $C_y = M_{01}/M_{00}$. (ii) Validation/Test. The location is estimated from a first-pass probability map, P , by

running the model without spatial text, thresholding at $\tau = 0.30$, retaining the largest connected component to form a coarse proposal, R , and computing its centroid, $(\hat{C}_x, \hat{C}_y)$. Centroids are mapped to breast quadrants (upper/lower $\times$ inner/outer) to produce a location token. Size and location tokens are verbalized to form the Global Feature Context Prompt p_c . Quantiles and thresholds are fixed a priori per fold using only the corresponding training split.

2.7.2. GFE Computation

The GFE computes a scene-level representation conditioned on p_c using a Vision Transformer. The prompt is embedded by the (frozen) text encoder to obtain $\mathbf{e}_c \in \mathbb{R}^{B \times e}$ (with e being the text-embedding dimension).

Selected Transformer layers provide token sequences, $\{\mathbf{Z}^{(j)}\}_{j \in \mathcal{L}}$, with $\mathbf{Z}^{(j)} \in \mathbb{R}^{B \times N \times D}$. Here, $L=|\mathcal{L}|$ is the number of selected layers, N the token count, and D the hidden token size. Tokens are linearly reduced to a common width and aggregated:

$$\mathbf{A} = \sum_{j \in \mathcal{L}} \mathbf{W}^{(j)} \mathbf{Z}_{[:,1:,:]}^{(j)} \in \mathbb{R}^{B \times (N-1) \times r}, \quad (6)$$

where the class token is removed. Here r is the reduced width, and $j \in \mathcal{L}$ indexes the selected layers with weights, $\mathbf{W}^{(j)}$. When attention logits, A , are used, we write $\tilde{A} = \text{softmax}(A)$. Prompt conditioning is applied as channel-wise scaling and shifting of the reduced tokens at a chosen Transformer depth:

$$\mathbf{c} = \mathbf{W}_c \mathbf{e}_c, \quad \tilde{\mathbf{A}} = (\mathbf{W}_{\text{mul}} \mathbf{c}) \odot \mathbf{A} + (\mathbf{W}_{\text{add}} \mathbf{c}). \quad (7)$$

Here c denotes the channel count of the reduced tokens; $\mathbf{W}_{\text{mul}}$ and $\mathbf{W}_{\text{add}}$ are learned linear maps that yield multiplicative/additive FiLM terms. The conditioned tokens are reshaped to a spatial grid and projected with a transposed convolution to obtain the global feature map

$$\mathbf{F}_g \in \mathbb{R}^{B \times 64 \times h \times w}. \quad (8)$$

SFA in the global branch follows the affine form in Equation (13) and produces $\hat{\mathbf{F}}_g$ as in Equation (3). The vision and text encoders are frozen; only the reduction, conditioning, and projection layers are trained.

2.8. Local Feature Extractor (LFE)

2.8.1. Local Features and Prompts

Local features comprise shape, margin, and BI-RADS. These attributes are read from the dataset metadata and verbalized to form the Local Feature Prompt p_ℓ (e.g., “irregular shape, microlobulated margin, BI-RADS 4”). The mapping from metadata to tokens is deterministic and applied identically across folds.

2.8.2. LFE Computation

The LFE recovers boundaries and fine structure under guidance from p_ℓ . We use a ResNet50 encoder with Transformer blocks at deep levels and a U-Net-style decoder. The prompt is embedded with the text encoder,

$$\mathbf{e}_\ell = \mathcal{T}(p_\ell) \in \mathbb{R}^{B \times e}, \quad (9)$$

and is linearly projected to the same fixed width, d , for SFA in the specified stages. Note: $\mathcal{T}(\cdot)$ denotes the (frozen) text encoder. Given $I \in \mathbb{R}^{B \times 3 \times H \times W}$, the encoder yields the

features $\{\mathbf{enc1}, \mathbf{enc2}, \mathbf{enc3}, \mathbf{enc4}, \mathbf{enc5}\}$. We add a Transformer block in the fourth encoder stage and at the bottleneck:

$$\mathbf{enc4}' = \text{TrEnc}(\mathbf{enc4}), \quad \mathbf{center}' = \text{TrEnc}(\mathbf{enc5}). \quad (10)$$

The decoder upsamples and merges skip connections using up blocks to produce $\mathbf{dec4}, \mathbf{dec3}, \mathbf{dec2}, \mathbf{dec1}$, followed by a final up projection to obtain a 32-channel map at the decoder output.

Semantic Feature Adjustment (SFA) uses the local prompt embedding in encoder stage four and after the first three decoder up blocks:

$$\begin{aligned} \mathbf{enc4}^* &= \text{SFA}(\mathbf{enc4}', \mathbf{e}_\ell), \\ \mathbf{dec4}^* &= \text{SFA}(\mathbf{dec4}, \mathbf{e}_\ell), \\ \mathbf{dec3}^* &= \text{SFA}(\mathbf{dec3}, \mathbf{e}_\ell), \\ \mathbf{dec2}^* &= \text{SFA}(\mathbf{dec2}, \mathbf{e}_\ell). \end{aligned} \quad (11)$$

The final local representation is taken after the last decoder stage:

$$\mathbf{F}_\ell \in \mathbb{R}^{B \times 32 \times h \times w}. \quad (12)$$

This configuration preserves boundary detail and small structures while injecting attribute cues in the stages listed above. Unless stated otherwise, the text encoder is frozen, and the LFE encoder–decoder parameters are trainable.

2.9. Semantic Feature Adjustment (SFA)

SFA injects prompt driven semantics into feature maps through channel-wise scaling and shifting. The module takes a visual tensor and a prompt embedding, predicts modulation parameters, and applies an affine transformation with an optional residual (see Figure 1).

Let $\mathbf{F} \in \mathbb{R}^{B \times C \times h \times w}$ be a feature map and let $\mathbf{e} \in \mathbb{R}^d$ be the embedding of the associated prompt. The core operation is

$$\begin{aligned} \hat{\mathbf{F}} &= \gamma \odot \mathbf{F} + \beta, \\ [\gamma, \beta] &= \Psi(\mathbf{e}), \end{aligned} \quad (13)$$

where $\odot$ denotes element-wise multiplication and $\gamma, \beta \in \mathbb{R}^{B \times C \times 1 \times 1}$ broadcast over spatial dimensions. Here Ψ denotes the per-stage projection network (two-layer MLP) that maps the prompt embedding to (γ, β) . Parameters are produced by a small projection network for each stage, k , with channel width, $C^{(k)}$, and are broadcast over spatial dimensions:

$$\begin{aligned} \mathbf{z}^{(k)} &= \text{MLP}^{(k)}(\mathbf{e}) \in \mathbb{R}^{2C^{(k)}}, \\ [\gamma^{(k)}, \beta^{(k)}] &= \text{split}(\mathbf{z}^{(k)}), \end{aligned} \quad (14)$$

Thus, $\gamma^{(k)}, \beta^{(k)} \in \mathbb{R}^{B \times C^{(k)} \times 1 \times 1}$ are per-sample, per-channel FiLM parameters broadcast over (h, w) .

In the global branch, we condition on $\mathbf{e}_c$ to obtain $\hat{\mathbf{F}}_g$; in the local decoder, we condition on $\mathbf{e}_\ell$ to obtain $\hat{\mathbf{F}}_\ell$, as summarized in Equation (3).

An additive skip may be used in implementation:

$$\mathbf{F}_{\text{out}} = \hat{\mathbf{F}} + \mathbf{F}. \quad (15)$$

2.10. Feature Fusion and Prediction Head

We concatenate the modulated global and local feature maps to obtain

$$\mathbf{F}_{\text{cat}} = \hat{\mathbf{F}}_g \parallel \hat{\mathbf{F}}_\ell \in \mathbb{R}^{B \times 96 \times h \times w}, \quad (16)$$

and refine with two residual blocks followed by a 1×1 convolution:

$$\mathbf{F}_1 = \text{RB}_1(\mathbf{F}_{\text{cat}}), \quad \mathbf{F}_{\text{fused}} = \text{RB}_2(\mathbf{F}_1), \quad \hat{\mathbf{Y}} = \text{Conv}_{1 \times 1}(\mathbf{F}_{\text{fused}}). \quad (17)$$

3. Results

We followed the training setup in Section 2.3 and the metrics in Section 2.4. All experiments used the implementation and hardware in Section 2.5.

3.1. Overall Performance

We compared XBusNet with representative non-prompt baselines including U-Net [12], U-ResNet, U_KAN [53], and AnatoSegNet [54], as well as prompt-guided baselines such as BUSSAM [43] and CLIP_TNseg [46]. All models were evaluated on BLU under the same five-fold cross-validation protocol (Section 2.3), the same pre-processing (Section 2.2), and a common operating threshold, $\tau_{\text{seg}} = 0.5$ (Section 2.4). We report the mean $\pm$ SD across folds for Dice, IoU, the FPR, and the FNR in Table 1 and provide a size-stratified analysis in Table 2.

Table 1. Overall comparison on BLU (five-fold CV). Dice/IoU cells show the fold mean $\pm$ SD at a fixed threshold, $\tau_{\text{seg}} = 0.5$, with 95% image-level bootstrap CIs in brackets. Lower is better for the FPR and FNR.

Method	Dice	IoU	FPR	FNR
U-Net [12]	0.604 $\pm$ 0.038 [0.564, 0.643]	0.500 $\pm$ 0.032 [0.463, 0.537]	0.019 $\pm$ 0.015	0.342 $\pm$ 0.069
U-ResNet	0.721 $\pm$ 0.040 [0.687, 0.754]	0.621 $\pm$ 0.049 [0.588, 0.654]	0.010 $\pm$ 0.007	0.264 $\pm$ 0.078
U-KAN [53]	0.752 $\pm$ 0.028 [0.739, 0.766]	0.614 $\pm$ 0.037 [0.599, 0.630]	0.009 $\pm$ 0.001	0.305 $\pm$ 0.036
AnatoSegNet [54]	0.733 $\pm$ 0.072 [0.702, 0.763]	0.627 $\pm$ 0.084 [0.596, 0.658]	0.015 $\pm$ 0.013	0.216 $\pm$ 0.034
BUSSAM [43]	0.833 $\pm$ 0.020 [0.815, 0.850]	0.735 $\pm$ 0.026 [0.713, 0.755]	0.014 $\pm$ 0.003	0.101 $\pm$ 0.024
CLIP-TNseg [46]	0.839 $\pm$ 0.019 [0.826, 0.852]	0.765 $\pm$ 0.020 [0.750, 0.778]	0.189 $\pm$ 0.002	0.134 $\pm$ 0.017
XBusNet (Ours)	0.876 $\pm$ 0.014 [0.863, 0.889]	0.815 $\pm$ 0.013 [0.800, 0.829]	0.099 $\pm$ 0.002	0.078 $\pm$ 0.012

Table 2. Dice and IoU scores (means with 95% CIs) across tumor-length intervals. Bold indicates the best per column.

Model/Length	0–110		111–250		250+	
	Dice	IoU	Dice	IoU	Dice	IoU
U-Net [12]	0.3668 [0.3211, 0.5105]	0.2750 [0.2467, 0.4135]	0.6461 [0.5992, 0.6948]	0.5251 [0.4928, 0.5843]	0.6739 [0.5776, 0.7310]	0.5625 [0.4779, 0.6246]
U-ResNet	0.6238 [0.5364, 0.7157]	0.5327 [0.4451, 0.6092]	0.7859 [0.7371, 0.8129]	0.6863 [0.6385, 0.7153]	0.7305 [0.5932, 0.7400]	0.6217 [0.4885, 0.6393]

Table 2. Cont.

Model/Length	0–110		111–250		250+	
	Dice	IoU	Dice	IoU	Dice	IoU
U-Kan [53]	0.7256 [0.7052, 0.7653]	0.5817 [0.5574, 0.6274]	0.7584 [0.7442, 0.7786]	0.6216 [0.6049, 0.6457]	0.7312 [0.7117, 0.7723]	0.5892 [0.5685, 0.6398]
AnatoSegNet [54]	0.6303 [0.5375, 0.7161]	0.5215 [0.4425, 0.6097]	0.7843 [0.7411, 0.8056]	0.6822 [0.6293, 0.7026]	0.7207 [0.6539, 0.7841]	0.6131 [0.5492, 0.6807]
BUSSAM [43]	0.7846 [0.7364, 0.8271]	0.6714 [0.6171, 0.7205]	0.8465 [0.8255, 0.8659]	0.7510 [0.7241, 0.7762]	0.8426 [0.8034, 0.8765]	0.7495 [0.7026, 0.7911]
CLIP-TNseg [46]	0.7689 [0.7375, 0.8133]	0.7026 [0.6725, 0.7450]	0.8587 [0.8462, 0.8750]	0.7867 [0.7732, 0.8057]	0.8447 [0.8169, 0.8605]	0.7621 [0.7255, 0.7777]
XBusNet (Ours)	0.8507 [0.8136, 0.8852]	0.7925 [0.7545, 0.8280]	0.8947 [0.8776, 0.9098]	0.8388 [0.8204, 0.8561]	0.8553 [0.8291, 0.8796]	0.7774 [0.7469, 0.8069]

3.1.1. Fold-Wise Performance

With five-fold cross-validation (Table 3), XBusNet attained consistently strong overlap across folds (Dice: 0.8583–0.8910; IoU: 0.7987–0.8302) with a stable FPR and FNR, indicating reliable behavior across splits.

Table 3. Fold-wise validation performance of XBusNet (latest run).

Fold	Dice	IoU	FPR	FNR
0	0.8846	0.8241	0.0984	0.0670
1	0.8910	0.8302	0.1280	0.0659
2	0.8583	0.7987	0.1077	0.0835
3	0.8649	0.8033	0.0857	0.0771
4	0.8836	0.8181	0.0771	0.0944
Mean	0.8766	0.8150	0.0994	0.0776

3.1.2. Comparison with Prior Methods

Table 1 summarizes mean cross-validation performance against U-Net variants, anatomy-aware models, and text-guided segmenters. XBusNet achieved the best Dice (0.8766) and IoU (0.8150) and the lowest FNR (0.0775), reflecting fewer missed lesion pixels. The FPR is comparable to that of the text-guided baseline and higher than that of some convolutional baselines that tend to under-segment; this trade-off favors recovering lesion extent while keeping spurious activations controlled.

3.1.3. Qualitative Assessment

Figure 2 shows typical error modes with false positives (blue) and false negatives (red). First, for high-contrast masses, U-Net variants exhibited small gaps at the posterior margin (missed pixels), whereas XBusNet produced continuous contours with fewer misses, consistent with the lower FNR in Table 1. Second, on small or faint lesions, several baselines fragmented the mask or missed the target entirely; XBusNet recovered a larger fraction of the lesion area, mirroring the gains in the 0–110 px bin in Table 2. Third, in low-contrast heterogeneous backgrounds, some text-guided baselines could add foreground in adjacent tissue, increasing the FPR; XBusNet suppressed these spurious activations while retaining the faint rim. Finally, for irregular echotexture, convolutional baselines scattered foreground outside the mass, while XBusNet better followed the irregular edge. Figure 3 shows an example of the exact prompts used in the proposed work.

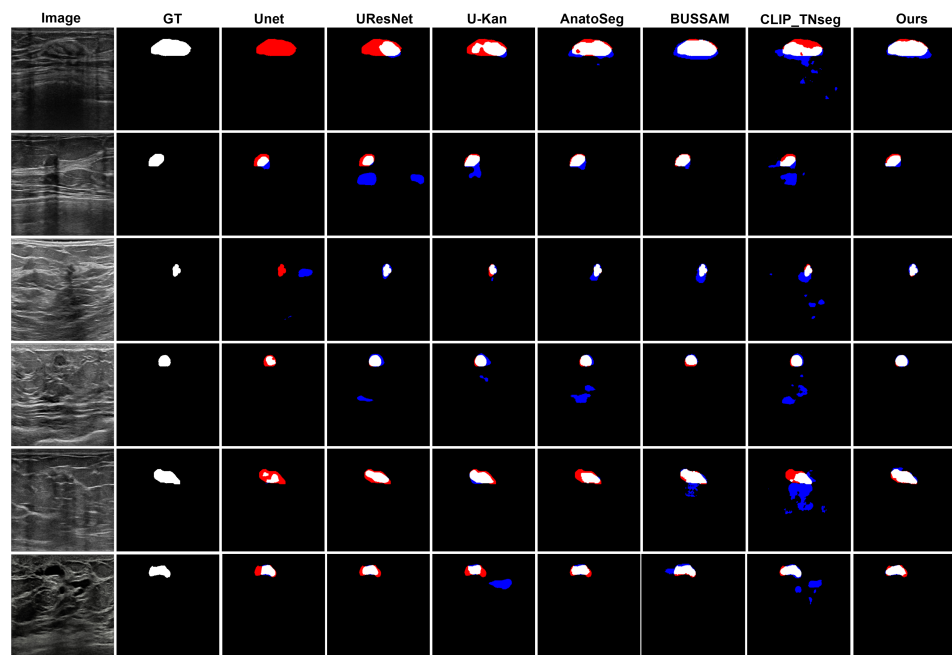


Figure 2. Qualitative comparison of breast ultrasound segmentations. False positives (FPs) are shown in blue and false negatives (FNs) in red.

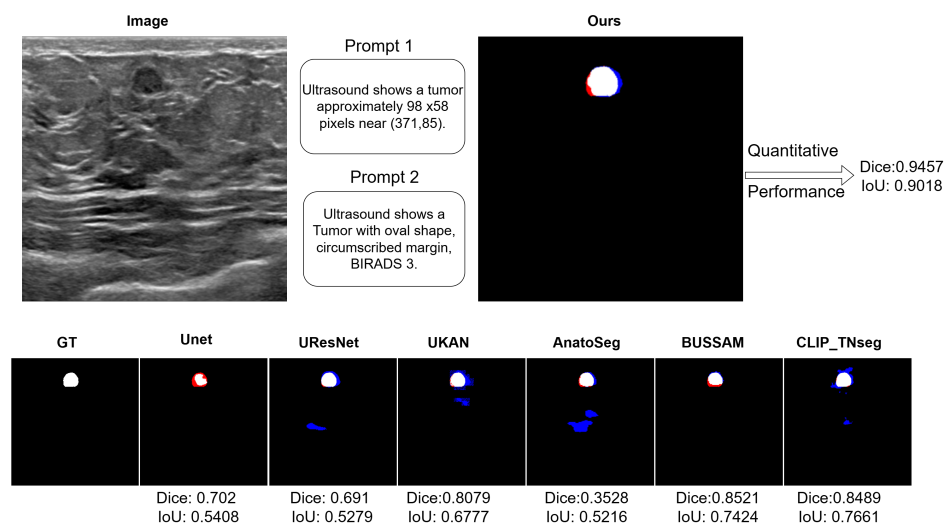


Figure 3. A qualitative example with the exact prompts used for a case. Top left: An input breast ultrasound (BUS) image showing the Global Feature Context Prompt (GFCP) and the Local Feature Prompt (LFP). Top right: Qualitative and quantitative results for XBusNet in this case. Bottom row: Segmentation overlays; ground truth is in white, false negatives are in red, and false positives are in blue.

3.1.4. Statistical Analysis and Comparison

All between-method comparisons were performed using image-level Dice and IoU using paired Wilcoxon signed-rank tests; when testing XBusNet against multiple baselines, p -values were adjusted via Benjamini–Hochberg FDR ($\alpha = 0.05$). Uncertainty was summarized with 95% bootstrap confidence intervals (10,000 image-level resamples) and is reported in Tables 1 and 2. Under this protocol, XBusNet significantly outperformed all baselines (adjusted two-sided $p < 10^{-4}$). Effect sizes (rank-biserial correlation) indicated large advantages (median $|r_{rb}| \approx 0.82$ for Dice; median $|r_{rb}| \approx 0.80$ for IoU).

3.2. Ablation Study

We quantified the contribution of the Local Feature Extractor (LFE), the Global Feature Extractor (GFE), and Semantic Feature Adjustment (SFA) by toggling each component. Table 4 shows that removing any single component degraded performance. Without the LFE, Dice fell from 0.8766 to 0.8572 and IoU from 0.8150 to 0.7865, underscoring the role of local detail and boundary focus. Without the GFE, Dice dropped to 0.8453 and IoU to 0.7772, highlighting the value of global semantic cues from the conditional prompt. Without SFA, Dice decreased to 0.8600 and IoU to 0.8068, showing the benefit of prompt-conditioned modulation that aligns visual features with clinical attributes. The full model with all three components yielded the best results, indicating complementary effects.

Table 4. Ablation study on effect of LFE, GFE, and SFA modules. The bold indicates the best per column.

LFE	GFE	SFA	Dice	IoU
Yes	Yes	Yes	0.8766	0.8150
No	Yes	No	0.8572	0.7865
Yes	No	Yes	0.8453	0.7772
Yes	Yes	No	0.8600	0.8068

3.3. Qualitative Visualizations

We employed Grad-CAM overlays to visualize image regions that most influenced the predicted masks (Figure 4). Among the compared methods, U-Net [12], U-ResNet, U-KAN [53], and AnatoSegNet [54] are non-prompt methods; BUSSAM [43] and CLIP-TNseg [46] are prompt-guided; and XBusNet uses dual prompts (global and attribute-conditioned). Non-prompt baselines frequently display diffuse responses in low-contrast tissue, particularly posterior to the lesion, corresponding to missed or fragmented masks. Prompt-guided methods improve spatial correspondence; CLIP-TNseg may still exhibit scattered hotspots around speckle, whereas XBusNet shows concentrated activation adjacent to lesion boundaries and within the mass. For very small lesions, the salient regions for XBusNet overlap the recovered mask more consistently, in line with the small-lesion gains reported in Table 2. These visualizations are qualitative and are not treated as formal explanations.

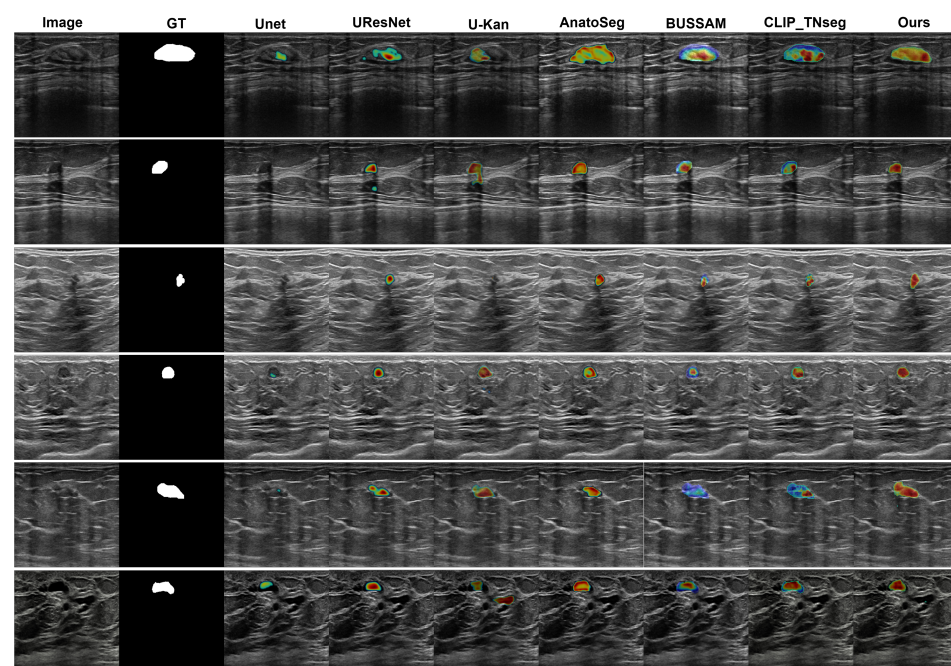


Figure 4. Grad-CAM comparison. Colors map low→high attribution (blue→red), normalized to [0, 1] with a shared scale across images/models.

4. Discussion

XBusNet showed strong performance on BLU. Across five folds, the model achieved a mean Dice of 0.8766 and an IoU of 0.8150, with narrow fold-to-fold variation, which suggests stable behavior under fold-wise splits. Compared with U-Net variants, anatomy-aware models, and text-guided baselines, XBusNet also produced a lower false negative rate, indicating fewer missed lesion pixels. The false positive rate was comparable to that of text-guided baselines but higher than that of some convolutional models that under-segment; this trade-off favors recovering lesion extent and can be tuned by adjusting the operating threshold when a stricter specificity is required. Findings align with prior BUS segmentation work where prompt signals improve localization under low contrast; remaining limits include single-cohort evaluation and reliance on structured text fields.

Size-stratified analysis supports these observations. The largest relative gains appear for the smallest lesions, where the model maintains competitive Dice and IoU, while performance remains strong for medium and large lesions. This pattern is consistent with the design: global prompt cues (size and approximate location) guide the search toward plausible regions, and local attribute cues (shape, margin, BI-RADS) help the decoder maintain sharp boundaries when pixel evidence is weak.

Ablation results point to complementary roles for both branches and for Semantic Feature Adjustment. Removing the local branch reduces Dice and IoU, highlighting the importance of boundary detail. Disabling the global branch leads to a larger drop, showing that scene-level cues improve localization and reduce misses. Turning off Semantic Feature Adjustment also degrades performance, which indicates that prompt-conditioned scaling and shifting adds value beyond the two-branch backbone. Together, these findings support the view that the full configuration is needed to realize the observed gains.

Qualitative panels mirror these trends. On high-contrast masses, several baselines show small gaps at the posterior margin, whereas XBusNet yields continuous contours, in line with the lower false negative rate. For small or faint lesions, baselines may fragment or miss the target, while XBusNet recovers a larger fraction of the lesion area. In heterogeneous backgrounds, some text-guided baselines spill into adjacent tissue; XBusNet restrains these activations while preserving the rim. With irregular echotexture, convolutional baselines scatter foreground outside the mass, whereas XBusNet follows the irregular edge more closely. Grad-CAM overlays are used as qualitative views of model focus and are not treated as formal explanations.

Clinical Relevance

There is no established minimal clinically important difference (MCID) for segmentation metrics in breast ultrasound. Nevertheless, our error profile—particularly improved coverage on small lesions and smoother boundaries—suggests potential benefits for measurement stability (e.g., size/margins aligned with BI-RADS), reduced inter-reader variability on follow-up, and more reliable pre-read triage. Future reader studies should translate Dice/IoU gains into task-level endpoints such as boundary error (mm), read time, and inter-reader variance to anchor a practical MCID. This study is limited by the availability of suitable public data. To our knowledge, there are no widely available breast ultrasound datasets that jointly provide pixel-level masks and rich clinical descriptors such as BI-RADS terms, structured lesion metadata (e.g., shape and margin), and histopathological outcomes. This is the main reason we did not combine BLU with a public external cohort for evaluation. As a practical recommendation for future dataset releases, we encourage curators to include not only classification labels and segmentation masks but also standardized BI-RADS descriptors, acquisition metadata, radiology reports, and, when possible, histopathology results. Such completeness would enable rigorous text-guided

segmentation research, cross-dataset evaluation, and clinically meaningful analyses. While structured BI-RADS fields facilitate prompting, entries can be missing or partial; ablations indicate graceful degradation without text cues, and future work will explore automatic prompt inference from images or reports. In practice, the local/global cues can be extracted from routine structured reporting (BI-RADS, shape, margin, size/quadrant). When such fields are absent, XBusNet remains functional with neutral prompts or image-only input; ablations indicate that Dice decreases from 0.8766 to 0.8453 without the GFE (no global prompt) and to 0.8600 without SFA (no prompt-driven modulation). As future work, we will address partially missing fields via automatic prompt inference (image/report-driven) and investigate radiomics-derived surrogates or pseudo-labels.

5. Conclusions

XBusNet fuses global prompts (size/location) with local prompts (shape, margin, BI-RADS) and applies Semantic Feature Adjustment to guide segmentation. On BLU, it achieves state-of-the-art Dice/IoU with the largest gains on small lesions, reducing missed pixels while keeping spurious activations controlled. Ablations show that the global branch, the local branch, and the prompt-conditioned modulation each contribute; together they yield balanced boundary quality and region coverage. These results indicate that simple, automatically assembled text cues can strengthen ultrasound segmentation without changes to current imaging practice. Integration options include (i) pre-read triage to flag small/low-contrast candidates and (ii) post-read standardization of lesion measurements; prompts can be auto-assembled from routine fields with minimal workflow impact.

Future Work

We will test cross-center generalization, add boundary-oriented metrics and calibration analysis, and explore coupling with detection and structured reporting.

Author Contributions: Conceptualization, B.S. and R.M.; methodology, B.S. and R.M.; software, R.M. validation, B.S. and R.M.; formal analysis, B.S. and R.M.; investigation, R.M.; resources, B.S.; data curation, R.M.; writing—original draft preparation, R.M. and B.S.; writing—review and editing, B.S. and R.M.; visualization, R.M.; supervision, B.S.; project administration, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to this study used a publicly available, fully de-identified dataset and involved no interaction with human subject.

Informed Consent Statement: Not applicable.

Data Availability Statement: Training and inference code, configurations, and evaluation scripts are available at github.com/AAR-UNLV/XBusNet, and the dataset is accessed on 9 May 2025 through <https://www.cancerimagingarchive.net/collection/breast-lesions-usg/>.

Conflicts of Interest: The authors declare no relevant conflicts of interest.

References

1. American Cancer Society. Key Statistics for Breast Cancer. 2025. Available online: <https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html> (accessed on 5 May 2025).
2. Nicholson, W.K.; Silverstein, M.; Wong, J.B.; Barry, M.J.; Chelmow, D.; Coker, T.R.; Davis, E.M.; Jaén, C.R.; Krousel-Wood, M.; Lee, S.A.; et al. Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2024**, *331*, 1918–1930. [CrossRef]

3. Sardanelli, F.; Boetes, C.; Borisch, B.; Decker, T.; Federico, M.; Gilbert, F.J.; Helbich, T.; Heywang-Köbrunner, S.H.; Kaiser, W.A.; Kerin, M.J.; et al. Magnetic resonance imaging of the breast: Recommendations from the EUSOMA working group. *Eur. J. Cancer* **2010**, *46*, 1296–1316. [\[CrossRef\]](#)
4. Carney, P.A.; Miglioretti, D.L.; Yankaskas, B.C.; Kerlikowske, K.; Rosenberg, R.; Rutter, C.M.; Geller, B.M.; Abraham, L.A.; Taplin, S.H.; Dignan, M.; et al. Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography. *Ann. Intern. Med.* **2003**, *138*, 168–175. [\[CrossRef\]](#)
5. Marcon, M.; Fuchsjäger, M.H.; Clauser, P.; Mann, R.M. ESR Essentials: Screening for breast cancer—General recommendations by EUSOBI. *Eur. Radiol.* **2024**, *34*, 6348–6357. [\[CrossRef\]](#)
6. Evans, A.; Trimboli, R.M.; Athanasiou, A.; Balleyguier, C.; Baltzer, P.A.; Bick, U.; Camps Herrero, J.; Clauser, P.; Colin, C.; Cornford, E.; et al. Breast ultrasound: Recommendations for information to women and referring physicians by the European Society of Breast Imaging. *Insights Imaging* **2018**, *9*, 449–461. [\[CrossRef\]](#)
7. Zhang, Y.; Xian, M.; Cheng, H.D.; Shareef, B.; Ding, J.; Xu, F.; Huang, K.; Zhang, B.; Ning, C.; Wang, Y. BUSIS: A Benchmark for Breast Ultrasound Image Segmentation. *Healthcare* **2022**, *10*, 729. [\[CrossRef\]](#)
8. Shareef, B.M.; Xian, M.; Sun, S.; Vakanski, A.; Ding, J.; Ning, C.; Cheng, H.D. A Benchmark for Breast Ultrasound Image Classification. SSRN Working Paper. 2023. Available online: <https://ssrn.com/abstract=4339660> (accessed on 20 May 2025).
9. D’Orsi, C.J.; Sickles, E.A.; Mendelson, E.B.; Morris, E.A. (Eds.) *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, 5th ed.; American College of Radiology: Reston, VA, USA, 2013.
10. Gu, J.; Jiang, T. Ultrasound radiomics in personalized breast management: Current status and future prospects. *Front. Oncol.* **2022**, *12*, 963612. [\[CrossRef\]](#)
11. Noble, J.A.; Boukerroui, D. Ultrasound image segmentation: A survey. *IEEE Trans. Med. Imaging* **2006**, *25*, 987–1010. [\[CrossRef\]](#)
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [\[CrossRef\]](#)
13. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the DLMIA/ML-CDS@MICCAI, Granada, Spain, 20 September 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. [\[CrossRef\]](#)
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; pp. 770–778. [\[CrossRef\]](#)
15. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; pp. 4700–4708. [\[CrossRef\]](#)
16. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999. [\[CrossRef\]](#)
17. Xie, Y.; Yang, B.; Guan, Q.; Zhang, J.; Wu, Q.; Xia, Y. Attention Mechanisms in Medical Image Segmentation: A Survey. *arXiv* **2023**, arXiv:2305.17937. [\[CrossRef\]](#)
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
19. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [\[CrossRef\]](#)
20. Rodriguez, J.; Huang, K.; Xu, M. Multi-Task Breast Ultrasound Image Classification and Segmentation Using Swin Transformer and VMamba Models. In *Proceedings of the 2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Hangzhou, China, 15–17 August 2024*; pp. 858–863. [\[CrossRef\]](#)
21. Shareef, B.; Xian, M.; Vakanski, A.; Wang, H. Breast Ultrasound Tumor Classification Using a Hybrid Multitask CNN-Transformer Network. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023, Proceedings of the 26th International Conference, Vancouver, BC, Canada, 8–12 October 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 344–353.
22. Guan, H.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [\[CrossRef\]](#)
23. Shareef, B.; Xian, M.; Vakanski, A. STAN: Small Tumor-Aware Network for Breast Ultrasound Image Segmentation. In *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020*; pp. 1–5. [\[CrossRef\]](#)
24. Shareef, B.; Vakanski, A.; Freer, P.E.; Xian, M. ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation. *Healthcare* **2022**, *10*, 2262. [\[CrossRef\]](#)

25. Hsieh, Y.-H.; Hsu, F.-R.; Dai, S.-T.; Huang, H.-Y.; Chen, D.-R.; Shia, W.-C. Incorporating the Breast Imaging Reporting and Data System Lexicon with a Fully Convolutional Network for Malignancy Detection on Breast Ultrasound. *Diagnostics* **2022**, *12*, 66. [\[CrossRef\]](#)
26. Oktay, O.; Ferrante, E.; Kamnitsas, K.; Heinrich, M.; Bai, W.; Caballero, J.; Cook, S.; de Marvao, A.; Dawes, T.; O'Regan, D.; et al. Anatomically Constrained Neural Networks (ACNN): Application to Cardiac Image Enhancement and Segmentation. *IEEE Trans. Med. Imaging* **2017**, *37*, 384–395. [\[CrossRef\]](#)
27. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ben Ayed, I. Boundary loss for highly unbalanced segmentation. *Med. Image Anal.* **2021**, *67*, 101851. [\[CrossRef\]](#)
28. Karimi, D.; Salcudean, S.E. Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. *arXiv* **2019**, arXiv:1904.10030. [\[CrossRef\]](#)
29. Hu, X.; Li, F.; Samaras, D.; Chen, C. Topology-Preserving Deep Image Segmentation. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
30. El Jurdi, R.; Petitjean, C.; Honeine, P.; Cheplygina, V.; Abdallah, F. High-level prior-based loss functions for medical image segmentation: A survey. *Comput. Vis. Image Underst.* **2021**, *210*, 103248. [\[CrossRef\]](#)
31. Yu, Y.; Acton, S.T. Speckle reducing anisotropic diffusion. *IEEE Trans. Image Process.* **2002**, *11*, 1260–1270. [\[CrossRef\]](#)
32. Destremes, F.; Cloutier, G. A Critical Review and Uniformized Representation of Statistical Distributions Modeling the Envelope of Ultrasonic Echoes. *Ultrasound Med. Biol.* **2010**, *36*, 1037–1051. [\[CrossRef\]](#)
33. Muhtadi, S.; Razzaque, R.R.; Chowdhury, A.; Garra, B.S.; Alam, S.K. Texture quantified from ultrasound Nakagami parametric images is diagnostically relevant for breast tumor characterization. *J. Med. Imaging* **2023**, *10* (Suppl. S2), S22410. [\[CrossRef\]](#)
34. Christensen, A.M.; Rosado-Méndez, I.M.; Hall, T.J. A systematized review of quantitative ultrasound based on first-order speckle statistics. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2024**, *71*, 872–886. [\[CrossRef\]](#)
35. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
36. Lüddecke, T.; Ecker, A.S. Image Segmentation Using Text and Image Prompts. *arXiv* **2021**, arXiv:2112.10003.
37. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv* **2023**, arXiv:2303.05499.
38. Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv* **2024**, arXiv:2401.14159. [\[CrossRef\]](#)
39. Li, J.; Li, D.; Hoi, S.C.H.; Xiong, C. BLIP: Bootstrapping Language–Image Pre-training for Unified Vision–Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 12888–12900.
40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755.
41. Yu, L.; Poirson, P.; Yang, S.; Berg, A.C.; Berg, T.L. Modeling Context in Referring Expressions. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; pp. 69–85.
42. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. Segment Anything Model for Medical Images? *arXiv* **2024**, arXiv:2304.14660. [\[CrossRef\]](#)
43. Tu, Z.; Gu, L.; Wang, X.; Jiang, B. Ultrasound SAM Adapter: Adapting SAM for Breast Lesion Segmentation in Ultrasound Images. *arXiv* **2024**, arXiv:2404.14837. [\[CrossRef\]](#)
44. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 1022. [\[CrossRef\]](#)
45. Zhang, K.; Liu, D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2304.13785. [\[CrossRef\]](#)
46. Sun, X.; Wei, B.; Jiang, Y.; Mao, L.; Zhao, Q. CLIP-TNseg: A Multi-Modal Hybrid Framework for Thyroid Nodule Segmentation in Ultrasound Images. *arXiv* **2024**, arXiv:2412.05530. [\[CrossRef\]](#)
47. Chen, Y.; Wei, M.; Zheng, Z.; Hu, J.; Shi, Y.; Xiong, S.; Zhu, X.X.; Mou, L. CausalCLIPSeg: Unlocking CLIP's Potential in Referring Medical Image Segmentation with Causal Intervention. *arXiv* **2025**, arXiv:2503.15949.
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [\[CrossRef\]](#)
49. Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556.
50. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018.

51. Pawłowska, A.; Ćwierz-Pieńkowska, A.; Domalik, A.; Jaguś, D.; Kasprzak, P.; Matkowski, R.; Fura, Ł.; Nowicki, A.; Żołek, N. Curated benchmark dataset for ultrasound based breast lesion analysis. *Sci. Data* **2024**, *11*, 148. [[CrossRef](#)] [[PubMed](#)]
52. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of breast ultrasound images. *Data Brief* **2020**, *28*, 104863. [[CrossRef](#)]
53. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; Yuan, Y. U-KAN makes strong backbone for medical image segmentation and generation. In Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; AAAI Press: Washington, DC, USA, 2025; AAAI'25/IAAI'25/EAAI'25. [[CrossRef](#)]
54. Xu, M.; Wang, Y.; Huang, K. Anatosegnet: Anatomy Based CNN-Transformer Network for Enhanced Breast Ultrasound Image Segmentation. In Proceedings of the 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA, 14–17 April 2025; pp. 1–5.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.