Enhancing clinical safety in automated breast ultrasound segmentation through a sensitivity-prioritized detection and statistical fail-safe mechanism

Daffa Muhamad Azhar [a,1], Nanik Suciati [a,2,*]

[a] Department of Informatics, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS, Surabaya, 60111, Indonesia
[1] 6025241057@student.its.ac.id; [2] nanik@its.ac.id
* corresponding author

ARTICLE INFO

Article history
Received February 9, 2026
Revised April 4, 2026
Accepted April 9, 2026
Available online May 31, 2026

Keywords
Breast Ultrasound Segmentation
Segment Anything Model 2 (SAM 2)
Automated Prompting
Recall Optimization
Low-Rank Adaptation (LoRA)

ABSTRACT

Accurate breast ultrasound (BUS) lesion segmentation is critical for early diagnosis but is challenged by image artifacts and the reliance of foundation models on manual prompting. Existing automated frameworks often lack robust fail-safe mechanisms, leading to missed diagnoses. To address this reliability gap, this study proposes a novel, fully automated hybrid segmentation framework that synergistically integrates three key components: (1) a recall-optimized YOLOv9 detector tailored to minimize clinical false negatives; (2) a MedSAM2 foundation model efficiently fine-tuned via Low-Rank Adaptation (LoRA) for ultrasound specifics; and (3) a statistical fallback mechanism that acts as a crucial safety net to recover spatial prompts during detection failures. Evaluated on the public BUSI dataset, the recall-dominant detection module achieved a Recall of 0.8238. Supported by this robust prompting and fallback strategy, the segmentation module achieved a Dice coefficient of 0.8818 and an IoU of 0.8113. By effectively integrating specialized detection with adaptive segmentation and a statistical fail-safe, the proposed pipeline offers a highly reliable automated approach for computer-aided screening systems.

1. Introduction

Globally, breast carcinoma is the leading female malignancy and a primary cause of oncology-related deaths [1]. Because early detection and rapid intervention are vital for improving patient survival [2], [3], Breast Ultrasound (BUS) has become a crucial diagnostic modality favored for its affordability and radiation-free safety [4]. However, manual interpretation of BUS images is time-consuming and prone to inter-observer variability due to inherent image artifacts and low tissue contrast [5], [6], [7]. Consequently, Computer-Aided Diagnosis (CAD) systems have emerged as essential tools to assist radiologists [8].

While traditional Convolutional Neural Networks (CNNs) and transformer-based architectures have established strong benchmarks for breast lesion analysis [9], [10], a critical reliability gap persists in automated pipelines: the vulnerability to False Negatives (missed detections). Most existing detection and segmentation models are optimized for overall accuracy or Intersection-over-Union (IoU), rather than diagnostic safety (Recall). In a screening context, a missed detection is a catastrophic failure, yet standard models lack specific mechanisms to prevent this.

To handle complex ultrasound features, subsequent research has continuously refined deep learning models. Various studies have introduced specific attention mechanisms, such as dual-decoder networks, adaptive attention, and cross-attention, to suppress noise and highlight salient lesion features [11], [12], [13], [14], [15], [16], [17]. Other approaches have incorporated boundary-enhancement modules and cross-image dependency modeling to address irregular tumor shapes [18], [19], [20], [21]. Furthermore, multi-task learning frameworks have been developed to jointly optimize segmentation and diagnostic classification [22], [23], [24], [25], [26], [27]. Parallel to segmentation, automated lesion detection has also seen significant progress through the use of anatomical priors, multi-scale feature extraction, and anchor-free networks [28], [29], [30]. Despite these advancements, the aforementioned reliability gap regarding clinical safety remains largely unaddressed.

Recently, the Segment Anything Model (SAM) and its domain-specific adaptations, such as MedSAM2, have revolutionized medical image segmentation [31], [32], [33]. Through Parameter-Efficient Fine-Tuning (PEFT) strategies like Low-Rank Adaptation (LoRA), these models efficiently learn ultrasound-specific features without catastrophic forgetting [34], [35], [36], [37]. However, standard SAM architectures rely heavily on manual prompting or weakly-supervised methods [38], [39], [40], [41], [42]. To achieve fully automated workflows, recent "Auto-SAM" frameworks (e.g., Auto-BUSAM, Yolo-HLSAM, and APG-SAM) integrate object detectors to generate prompts autonomously [43], [44], [45].

Despite eliminating manual interaction, these Auto-SAM approaches inherit the vulnerability to False Negatives. If the autonomous detector fails to identify a lesion due to low contrast, the segmentation module receives no prompt, resulting in a silent diagnostic failure. Existing state-of-the-art Auto-SAM methods entirely lack a "fail-safe" mechanism to recover from such prompt-generation failures.

To address this critical vulnerability, this study proposes a novel, fully automated hybrid segmentation framework designed to prioritize diagnostic reliability and mitigate False Negatives. Unlike prior works, our pipeline synergistically integrates three key contributions:

- to employ a YOLOv9 variant trained to maximize Recall, ensuring valid prompts even for low-contrast lesions.
- to utilize Low-Rank Adaptation (LoRA) on MedSAM2 to ensure high-fidelity segmentation with minimal trainable parameters.
- to introduce a novel Mean Centroid Prior fallback mechanism. In rare cases of detection failure, this safety net triggers a statistical point prompt to recover a coarse segmentation.

2. Method

2.1. Dataset and Stratified Partitioning

Empirical validation in this study was conducted using the publicly accessible Breast Ultrasound Images (BUSI) repository [46], which aggregates clinical scans originating from 600 female subjects. The initial corpus consisted of sonographic images exported in PNG format, possessing a mean spatial dimension of approximately 500×500 pixels, and distributed across three diagnostic categories: normal, benign, and malignant. To ensure the study focuses strictly on pathological characterization, and because the segmentation task intrinsically requires the presence of a target object, the "normal" class was excluded, resulting in a curated dataset of 647 images, consisting of 437 benign and 210 malignant cases. For the automated localization task, these two categories were unified into a single "Lesion" class to facilitate binary object detection, shifting the model's objective toward structural morphology rather than pathological classification.

Data pre-processing involved several critical operations to optimize model convergence. First, multiple lesion masks within a single frame were consolidated into a unified binary ground truth using a logical "OR" operation. This procedure ensures the segmentation model effectively delineates all

clinically significant regions simultaneously (Fig. 1). Second, the dataset was partitioned using a Stratified 5-fold cross-validation scheme to preserve the benign-to-malignant ratio across all folds, thereby ensuring a consistent and unbiased class distribution during evaluation.

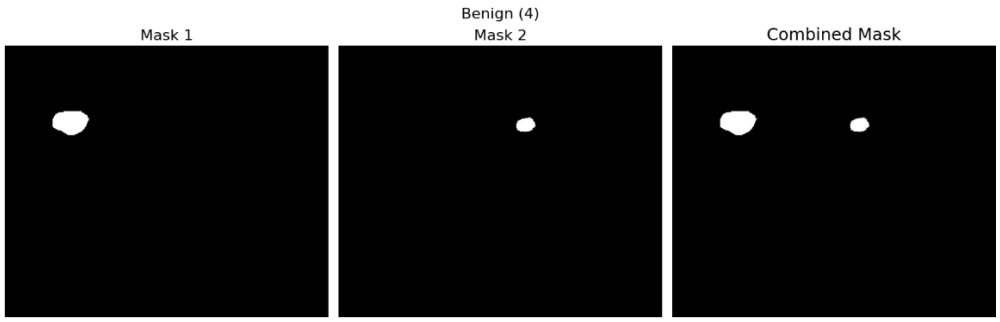

Fig. 1. Combined Ground Truth Mask.

Owing to the distinct architectural requirements of the detection and segmentation frameworks, images were resized using bilinear interpolation to two specific scales. The lesion localization stage utilized a 640×640 pixel input, incorporating default Ultralytics data augmentation and normalization to enhance detection robustness. Conversely, the segmentation stage employed a 512×512 pixel resolution, with MedSAM2 training utilizing standard ImageNet RGB normalization without additional augmentation. To address the inherent class imbalance, in which benign samples outnumber malignant cases by approximately 2:1, a Weighted Random Sampler was integrated into the MedSAM2 training pipeline. Malignant samples were assigned a weight of 2.0 to ensure a balanced representation in each training batch, thereby preventing the model from being biased toward the majority class. The comprehensive research methodology, encompassing the integration of YOLO-based detection, the statistical fallback mechanism, and MedSAM2-based segmentation, is visualized in Fig. 2. All computational experiments were conducted in a Kaggle environment powered by an NVIDIA Tesla P100 GPU.

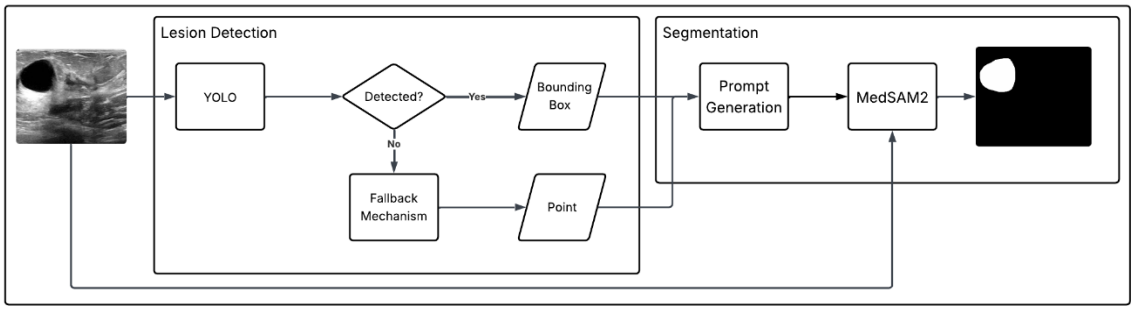

Fig. 2. Proposed research methodology for automated breast lesion segmentation.

2.2. Lesion Localization with Recall-Optimized YOLO

A fundamental contribution of this research involves Recall-Optimized Prompting, which strategically adapts the object detection training objective to ensure clinical safety. The YOLOv9 architecture was implemented to facilitate the automated localization of breast lesions. This specific variant was selected not only for its balanced performance between efficient parameterization and high inference speed, but also for its Programmable Gradient Information (PGI) architecture that effectively preserves deep features in medical images. During the detection stage, ultrasound images were resized to 640×640 pixels to predict bounding boxes encompassing suspected pathological regions.

While standard YOLO frameworks typically prioritize Mean Average Precision (mAP) within their fitness functions, the clinical context of breast cancer screening necessitates a higher emphasis on minimizing False Negatives to ensure patient safety. Consequently, the fitness function was redefined to strictly prioritize Recall (Sensitivity) over precision. The weights for the fitness metrics were empirically tuned through preliminary experiments according to (1).

$$Fitness = (0.1 \times P) + (0.7 \times R) + (0.1 \times mAP_{50}) + (0.1 \times mAP_{50-95}) \quad (1)$$

where P denotes Precision and R denotes Recall. By assigning a dominant weight of 0.7 to the Recall component, which was empirically yield the optimal balance between high sensitivity and acceptable false-positive rates, the optimization process encourages the model to identify the maximum number of potential lesions, thereby effectively reducing the false negative-rate.

The localization model was trained for 300 epochs with a batch size of 32. The training process used an automated optimizer selection strategy, which determined that AdamW was the most effective

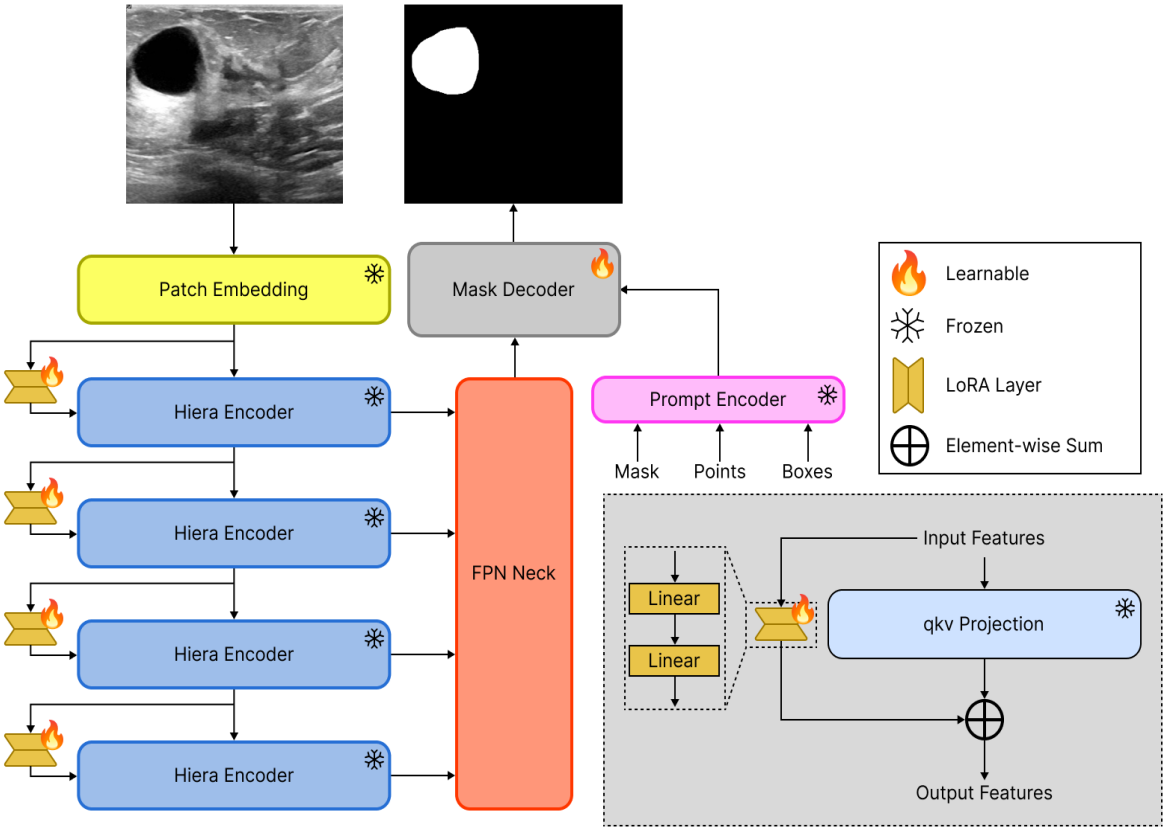

Fig. 3. The proposed MedSAM2 architecture with Low-Rank Adaptation (LoRA).

optimizer for this dataset. The identified optimal hyperparameters included a learning rate of 2×10^{-3}, a weight decay of 10^{-5}, and a patience of 75 epochs to mitigate overfitting while ensuring robust convergence. The bounding boxes generated from this stage serve as the automated spatial prompts to guide the subsequent MedSAM2 segmentation module.

2.3. MedSAM2 Architecture and LoRA Fine-Tuning

For the primary segmentation task, this research employs the MedSAM2 architecture, a specialized adaptation of the Segment Anything Model 2 (SAM 2.1) tailored for healthcare applications. Initialization was performed utilizing baseline MedSAM2 weights that were initially trained on three-dimensional medical imaging and volumetric video data. Within our proposed pipeline, the highly efficient Hiera-Tiny encoder was repurposed to analyze breast sonograms by handling them as discrete two-dimensional slices, standardized to a spatial dimension of 512×512 pixels.

To mitigate the risk of catastrophic forgetting while retraining on the relatively compact BUSI dataset, we integrated the Low-Rank Adaptation (LoRA) technique. As depicted in Fig. 3, most of the original encoder weights remained locked during training to retain their comprehensive medical feature extraction capabilities. We selectively introduced trainable LoRA modules into the Query-Key-Value (qkv) projection layers across all attention mechanisms in the visual encoder. Mathematically, LoRA

approximates the parameter update matrix ΔW through the multiplication of two lower-rank matrices, A and B, utilizing a rank dimension of r=16 alongside a scaling hyperparameter of α=32. Consequently, the forward computation for these modified linear layers operates according to (2).

$$h = W_0 x + \Delta W x = W_0 x + \frac{\alpha}{r}(BAx) \tag{2}$$

In this context, W_0 represents the static, pre-existing model parameters. The signal generated by the locked pathway is subsequently merged with the output of the LoRA pathway through element-wise addition. Following this, the extracted multi-scale representations are consolidated using a Feature Pyramid Network (FPN) architecture prior to their final evaluation by the Mask Decoder. Throughout the backpropagation phase, gradient updates were exclusively restricted to the Mask Decoder and the newly inserted LoRA adapters, keeping the rest of the network architecture frozen. To drive the learning process, the AdamW algorithm was selected as the optimizer, configured with a learning rate of 2×10^{-4} and a weight decay coefficient of 10^{-4}. Furthermore, an early stopping patience of 15 epochs was applied to guarantee steady model convergence and prevent overfitting.

2.4. Boundary-Aware Loss Function

To overcome the persistent issues of poor tissue contrast and ambiguous contour margins frequently encountered in sonographic scans, we integrated a hybrid boundary-sensitive optimization metric. This customized objective strictly penalizes spatial misalignments along the perimeters of the generated segmentation output relative to the actual clinical annotations. Consequently, the aggregate training penalty, denoted as L_{total}, is computed via a weighted aggregation of the regional Dice Loss (L_{dice}) and the perimeter-focused Edge Loss (L_{edge}), expressed mathematically in (3).

$$L_{total} = \alpha \cdot L_{dice} + \beta \cdot L_{edge} \tag{3}$$

The Dice Loss evaluates the global similarity between the predicted probability map P and the binary ground truth G, ensuring accurate regional localization according to (4).

$$L_{dice} = 1 - \left(\frac{2\sum_i P_i G_i + \epsilon}{\sum_i P_i + \sum_i G_i + \epsilon}\right) \tag{4}$$

Where ϵ represents a smoothing term utilized to prevent division by zero. To further refine contour precision, the Edge Loss utilizes Sobel operators to compute the gradient magnitude maps for both the prediction (∇P) and the ground truth (∇G). This loss is calculated as the Mean Absolute Error (MAE) between the normalized gradient maps, compelling the model to sharpen its predictions at the lesion contours as expressed in (5).

$$L_{edge} = \frac{1}{N}\sum_i |\nabla P_i - \nabla G_i| \tag{5}$$

Boundary precision was empirically prioritized by assigning a significantly higher weight to the edge term, specifically setting $\alpha = 1.0$ and $\beta = 10.0$. This configuration compels the network to prioritize the delineation of precise lesion margins, which is critical for clinical diagnostic accuracy. The impact of this boundary refinement directly correlates with the model's ability to achieve high Intersection over Union (IoU) scores, demonstrating precise spatial overlap.

2.5. Inference Strategy with Statistical Fallback

A hybrid inference strategy was implemented to integrate the detection and segmentation modules, operating under a fallback mechanism to mitigate the risk of False Negatives. The inference logic is primarily driven by the YOLOv9 detector. For every test image, a bounding box B_{pred} is predicted. If the detection confidence exceeds the empirically tuned threshold $\tau = 0.25$ (selected to maximize detection sensitivity without amplifying background noise), B_{pred} is utilized as the spatial prompt for the MedSAM2 encoder, representing the standard fully automated pipeline.

In cases where lesion detection fails, defined as $B_{pred} = \emptyset$ or a confidence score below τ, a fallback mechanism using a statistical point prompt is triggered to ensure the continuity of the segmentation process. To determine the optimal spatial prior, two distinct statistical approaches were investigated, with coordinates calculated strictly from the training set of each cross-validation fold to prevent data leakage. The first approach is the Mean Centroid (P_{mean}), computed by averaging the spatial (x, y) coordinate centers of all lesion bounding boxes, representing the geometric center of the lesion distribution. The second approach is the Mode Centroid (P_{mode}), determined via 2D spatial histogram binning to identify the coordinate with the highest frequency of occurrence within the training data. The prompt selection logic $Prompt(x)$ for an input image x is formalized as (6).

$$Prompt(x) = \begin{cases} Box(B_{pred}), if\ score(B_{pred}) > \tau \\ Point(P_{stat}), otherwise \end{cases} \quad (6)$$

Where $P_{stat} \in \{P_{mean}, P_{mode}\}$. This fallback assumes that breast lesions frequently occur in common anatomical clusters; however, it has the limitation that highly irregular or peripherally located lesions may receive suboptimal, coarse segmentations, since the statistical prior is not image-specific. An ablation study was conducted to compare these two priors and determine which statistical approach yields the most robust segmentation performance when automated object detection fails.

2.6. Evaluation Metrics

The empirical robustness of the developed hybrid framework was measured using established statistical criteria at the pixel level, specifically relying on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). During the initial localization phase, Sensitivity (Recall) serves as the principal indicator to gauge the system's proficiency in averting missed diagnoses, computed according to (7).

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

To validate the subsequent segmentation accuracy, the spatial concordance between the generated outputs and expert annotations is quantified via the Intersection over Union (IoU) in (8). Additionally, regional harmony is assessed utilizing the Dice Coefficient formulation in (9).

$$IoU = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP+FP+FN} \quad (8)$$

$$Dice = \frac{2|P \cap G|}{|P|+|G|} = \frac{2TP}{2TP+FP+FN} \quad (9)$$

Because it disproportionately rewards overlapping regions, the Dice metric is exceptionally well-suited for sonographic analysis where tissue borders often appear indistinct. Collectively, this metric suite delivers a holistic appraisal of the pipeline's diagnostic and delineative capabilities.

3. Results and Discussion

3.1. Evaluation of Recall-Optimized Lesion Localization

To establish a robust automated pipeline, the performance of various YOLO architectures was evaluated to identify the optimal prompt generator. A critical aspect of this evaluation was assessing the impact of the proposed recall-optimized fitness function, designed to minimize missed detections in a screening context. The comparative performance of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 is presented in Table 1.

The quantitative results in Table 1 validate the effectiveness of the proposed fitness function. Across all architectures, the optimized variants (denoted with "ᵃ") consistently demonstrated an improvement in Recall compared to their standard counterparts. YOLOv9ᵃ achieved the highest Recall of 0.8238 and the highest mAP@50-95 of 0.6389. Although YOLO11 registered the highest mAP@50 (0.8825), its Recall was notably lower (0.7886). The leading mAP@50-95 score of YOLOv9ᵃ indicates that it not only

detects more lesions but also provides bounding boxes that align more precisely with the ground truth across rigorous overlap thresholds.

Table 1. Comparative Performance of YOLO Variants.

Model	Recall	Precision	mAP@50	mAP@50-95
YOLOv8	0.7952±0.0510	0.8679±0.0352	0.8741±0.0330	0.6183±0.0309
YOLOv8[a]	0.8131±0.0174	0.8542±0.0324	0.8690±0.0207	0.6181±0.0268
YOLOv9	0.8166±0.0319	0.8768±0.0541	0.8751±0.0208	0.6350±0.0246
YOLOv9[a]	**0.8238±0.0393**	0.8699±0.0425	0.8763±0.0265	**0.6389±0.0252**
YOLOv10	0.7639±0.0320	0.8520±0.0533	0.8364±0.0360	0.5821±0.0337
YOLOv10[a]	0.7915±0.0544	0.8193±0.0514	0.8381±0.0312	0.5991±0.0279
YOLO11	0.7886±0.0369	**0.9040±0.0333**	**0.8825±0.0255**	0.6287±0.0264
YOLO11[a]	0.7932±0.0276	0.8773±0.0357	0.8771±0.0155	0.6252±0.0255

[a.] Trained using recall-optimized fitness function

Qualitative analysis further elucidates the practical implications of these metrics. Fig. 4 compares the detection capabilities of the standard YOLOv9 and the recall-optimized YOLOv9[a]. The enhanced sensitivity of YOLOv9[a] is evident in both benign and malignant cases where the standard model failed to produce any detection. Specifically, in malignant samples, YOLOv9[a] successfully identified lesions with confidence scores of 0.62, 0.85, and 0.54, respectively, preventing critical diagnostic failures.

It is observed that in certain difficult cases, such as Column 4, the optimized model predicts a slightly wider bounding box. While this inherently reduces Precision by introducing more background pixels into the bounding box, this trade-off is intentional and clinically justified. In the context of breast cancer screening, prioritizing Precision over Recall is dangerous; a False Positive may lead to an unnecessary biopsy, but a False Negative can delay life-saving treatment. Furthermore, in the context of this hybrid pipeline, this behavior is advantageous. A wider box ensures that the complete lesion context is captured and passed as a prompt to the MedSAM2 segmenter, whereas a missed

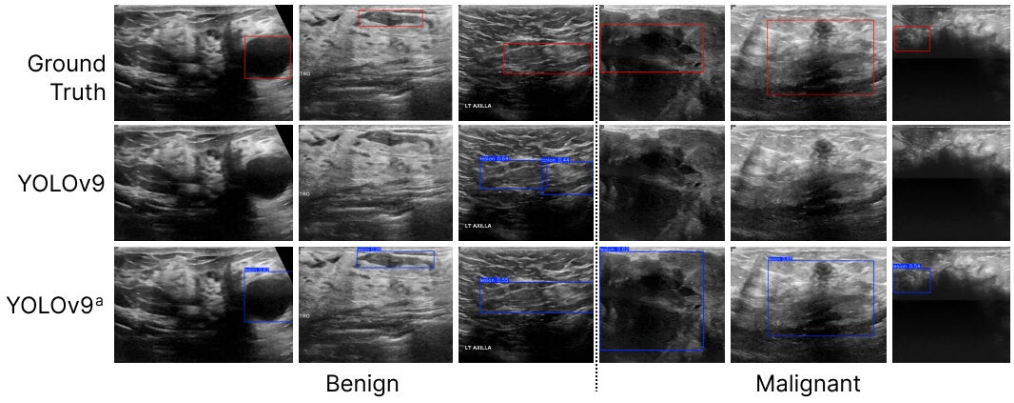

Fig. 4. Qualitative comparison of detection performance between standard YOLOv9 and recall-optimized YOLOv9[a] on benign and malignant samples.

detection or an overly tight box would result in information loss. The combination of the highest Recall and the highest mAP@50-95 confirms that YOLOv9a offers the best balance between high sensitivity and accurate localization, making it the ideal prompt generator for the proposed system.

3.2. Ablation Study on Segmentation Components

To validate the individual contributions of the proposed components, specifically the LoRA fine-tuning strategy, the recall-optimized detection prior, and the statistical fallback mechanism, an extensive ablation study was conducted. Table 2 summarizes the segmentation performance under various architectural configurations and prompting strategies.

Table 2. Ablation study of MedSAM2 configurations and fallback strategies.

Architecture	Dice	IoU	Recall	Accuracy	Prompt
MedSAM2-LoRA	0.9234±0.0030	0.8599±0.0051	0.9113±0.0103	0.9860±0.0014	Manual Box
MedSAM2	0.9177±0.0039	0.8506±0.0063	0.9128±0.0123	0.9839±0.0023	Manual Box
MedSAM2-LoRA[a]	**0.8818±0.0248**	**0.8113±0.0272**	**0.8924±0.0118**	**0.9765±0.0078**	YOLO + Mean
MedSAM2-LoRA	0.8793±0.0234	0.8087±0.0260	0.8826±0.0150	0.9764±0.0053	YOLO + Mean
MedSAM2	0.8727±0.0305	0.7987±0.0333	0.8854±0.0266	0.9747±0.0069	YOLO + Mean
MedSAM2-LoRA[a]	0.8803±0.0276	0.8101±0.0295	0.8906±0.0151	0.9761±0.0086	YOLO + Mode
MedSAM2-LoRA	0.8795±0.0240	0.8090±0.0257	0.8819±0.0150	0.9765±0.0060	YOLO + Mode
MedSAM2	0.8708±0.0343	0.7973±0.0360	0.8837±0.0294	0.9743±0.0077	YOLO + Mode
MedSAM2-LoRA[a]	0.8739±0.0383	0.8054±0.0373	0.8790±0.0335	0.9782±0.0053	YOLO
MedSAM2-LoRA	0.8673±0.0389	0.7997±0.0375	0.8666±0.0359	0.9775±0.0045	YOLO
MedSAM2	0.8668±0.0402	0.7944±0.0404	0.8788±0.0367	0.9754±0.0061	YOLO

[b.] Using YOLOv9 with recall-optimized fitness function.

The efficacy of Low-Rank Adaptation (LoRA) was evaluated using Manual (Ground Truth) prompts to isolate segmentation performance from detection errors. As shown in the first two rows of Table 2, MedSAM2-LoRA achieved a Dice score of 0.9234, outperforming the standard fine-tuned MedSAM2 (0.9177). This consistent empirical improvement suggests that injecting trainable adapters while freezing the encoder preserves robust pre-trained features more effectively than full fine-tuning, which minimizes the risk of overfitting on the limited BUSI dataset.

The ablation study confirms that the improvements in detection Recall (Section 3.1) translate directly into better segmentation outcomes. While manual bounding boxes naturally yield the highest segmentation performance, they inherently rely on continuous human intervention, which limits their utility in high-throughput clinical screening. In contrast, the proposed fully automated configuration (YOLO + Mean) completely eliminates the need for manual prompting while maintaining highly competitive accuracy. Architectures utilizing the recall-optimized YOLOv9a (denoted as MedSAM2-LoRA[a]) consistently achieved higher Dice and IoU scores compared to those using the standard YOLOv9. For instance, in the "YOLO + Mean" configuration, MedSAM2-LoRA[a] reached a Dice score of 0.8818, whereas the standard counterpart scored 0.8793, yielding a consistent empirical advantage across the cross-validation folds. Qualitative analysis in Fig. 5 further illustrates this impact. While the standard YOLOv9 configuration often results in fragmented or missed segmentations due to missed detections, the YOLOv9[a] column demonstrates enhanced stability. As visualized, the recall-optimized detector successfully captures the lesion within a valid bounding box, enabling MedSAM2 to generate a complete and coherent white segmentation mask that accurately delineates the lesion boundaries.

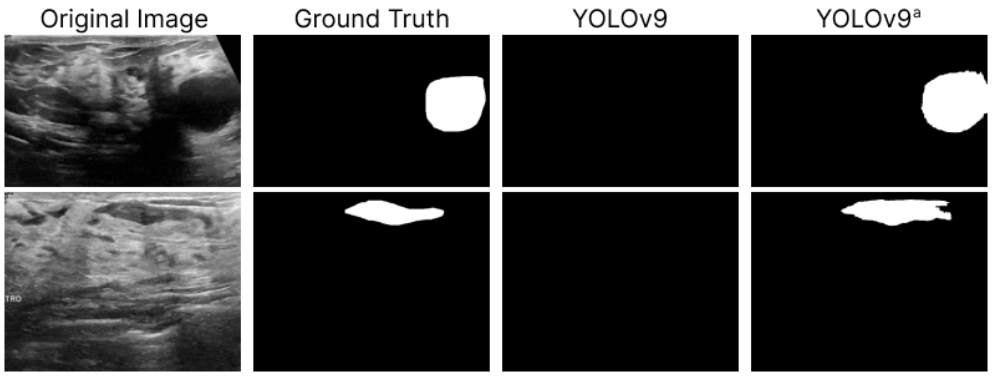

Fig. 5. Qualitative segmentation comparison.

The necessity of a fallback mechanism is evident when comparing the "YOLO Only" results against statistical fallback strategies. Relying solely on detection boxes yielded the lowest performance (Dice 0.8739) due to missing prompts. Integrating a fallback mechanism substantially mitigated this issue, with the Mean Centroid (P_{mean}) approach proving to be the most robust strategy (Dice 0.8818). Fig. 6 visually validates this "safety net" capability. In scenarios where the recall-optimized detector fails to

identify a lesion (Column 3), the segmentation output is typically lost, resulting in a blank mask. However, the Mean Centroid fallback (Column 4) successfully recovers the lesion location using the distribution prior, visualized as a red dot.

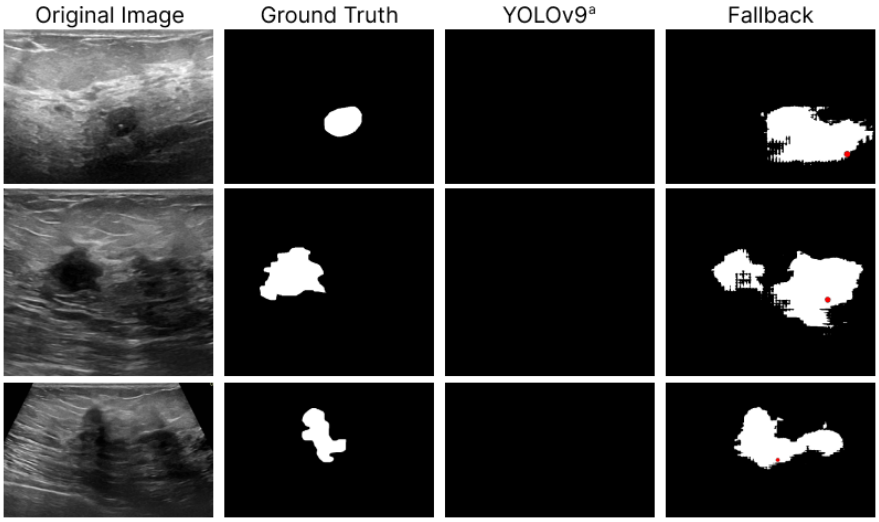

Fig. 6. Evaluation of the fallback mechanism.

While the quantitative increase in the overall Dice score provided by the fallback may appear incremental, the diagnostic relevance of this mechanism is paramount: recovering a missed detection prevents a catastrophic false negative, which is far more critical in a screening pipeline than marginal gains in boundary precision. However, failure case analysis reveals the inherent limitations of this statistical fallback. If a lesion is highly irregular or located at the extreme periphery of the breast tissue, far from the computed (P_{mean}), the generated point prompt may fall outside the actual lesion boundaries. In such cases, the resulting mask generated by MedSAM2 is often coarse, under-segmented, or captures adjacent shadowing artifacts instead of the true lesion. Despite this limitation, prioritizing detection recovery remains clinically advantageous, as it significantly minimizes the risk of the lesion going entirely undetected.

3.3. Comparison to state-of-the-art

To benchmark the effectiveness of the proposed hybrid framework, its performance was compared against established medical segmentation models (U-Net, Attention U-Net), the standard MedSAM foundation model, and recent SOTA methods specifically designed for breast ultrasound (MECA-SAM, Auto-BUSAM, APG-SAM). Table 3 presents the quantitative comparison of the BUSI dataset.

Table 3. SOTA Comparison.

Architecture	Dice	IoU	Recall	Accuracy
UNet	0.7823±0.0326	0.6719±0.0406	0.7757±0.0289	0.9596±0.0068
Attention Unet	0.7971±0.0365	0.6898±0.0434	0.7890±0.0308	0.9608±0.0084
MedSAM	0.8129±0.0058	0.6923±0.0075	0.7149±0.0077	0.9716±0.0041
MECA-SAM [42]	0.7063±0.0206	0.6150±0.0197	-	-
Auto-BUSAM [43]	0.8236	0.7443	0.8255	-
APG-SAM [45]	0.8370±0.0009	0.7190±0.0012	0.7851±0.0017	0.9702±0.0020
Proposed	**0.8818±0.0248**	**0.8113±0.0272**	**0.8924±0.0118**	**0.9765±0.0078**

The proposed method significantly outperformed traditional CNN architectures. Compared to UNet (Dice 0.7823) and Attention U-Net (Dice 0.7971), our approach achieved a Dice score of 0.8818, representing an improvement of approximately 10%. This performance gap highlights the limitation of training from scratch on small datasets like BUSI, whereas our method leverages the robust feature representations of the pre-trained SAM2 encoder. Compared with the standard MedSAM (Dice 0.8129),

our method demonstrated a clear advantage, validating the importance of the specific adaptations (LoRA fine-tuning and YOLO-based prompting) over a generic implementation. Furthermore, the proposed framework surpassed recent automated SAM variants. It outperformed Auto-BUSAM (Dice 0.8236) and the closest competitor, APG-SAM (Dice 0.8370).

To rigorously assess the reliability of these improvements against existing automated SAM frameworks, an independent two-sample t-test (Welch's t-test) was conducted using the reported summary statistics. The performance gap between the proposed method (0.8818±0.0248) and the closest competitor, APG-SAM (0.8370±0.0009), was proven to be statistically significant ($t = 4.037$, $p = 0.016$). This mathematical significance is further corroborated by the non-overlapping variances, which clearly indicate that the observed improvements offered by the proposed pipeline are robust and statistically stable.

A notable strength of the proposed method is the substantial improvement in IoU (0.8113), which is significantly higher than the best SOTA competitor (Auto-BUSAM at 0.7443). This indicates that the masks generated by our pipeline are not only accurate in location but also possess high boundary fidelity, adhering closely to the lesion shapes. Additionally, the method achieved the highest Recall (0.8924), surpassing APG-SAM (0.7851) and Auto-BUSAM (0.8255). This high sensitivity is a direct result of the recall-optimized detection prior (YOLOv9^{a}) and the fallback mechanism, ensuring that the system remains robust in identifying lesions.

Despite these promising outcomes, this study acknowledges certain limitations. The primary constraint is the reliance on a single, albeit widely recognized, public dataset (BUSI). While the stratified 5-fold cross-validation ensures internal validity and prevents data leakage, the model's generalizability to diverse, vendor-specific ultrasound machines or varying demographic populations remains to be verified. Future work will focus on validating the pipeline across multi-center datasets and refining the fallback mechanism to dynamically adapt to peripherally located lesions.

4. Conclusion

This study established a fully automated, highly reliable breast ultrasound segmentation pipeline that effectively addresses the reliability gap in existing Auto-SAM frameworks. Integrating a recall-optimized YOLOv9^{a} detector with a statistical fallback mechanism mitigated missed detections, a critical failure in medical screening. Fine-tuning MedSAM2 via Low-Rank Adaptation (LoRA) efficiently preserved robust pre-trained features. Evaluated on the BUSI dataset, the proposed method achieved a Dice score of 0.8818, a Recall of 0.8924, and an IoU of 0.8113, significantly outperforming traditional CNNs and recent state-of-the-art models like APG-SAM. Importantly, the fallback mechanism functioned as a vital safety net, recovering lesion localization during detection failures and substantially reducing critical false negatives.

Despite these promising outcomes, two primary limitations exist. First, relying solely on the BUSI dataset without multi-center external validation limits the generalizability of the findings across diverse ultrasound vendors and demographics. Second, the non-image-specific nature of the statistical fallback prior can yield coarse segmentations lacking fine boundary precision, particularly for highly irregular or peripherally located lesions.

Future research will address these constraints by incorporating image-adaptive priors or active contours to refine the fallback mechanism's boundary precision. Additionally, we plan to extend the inherently temporal MedSAM2 architecture to 3D Automated Breast Ultrasound and real-time video segmentation. Finally, extensive validation on larger, multi-institutional datasets will be prioritized to ensure the pipeline's robustness and diagnostic relevance in real-world scenarios.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. This work was funded by the Ministry of Higher Education, Science and Technology, Indonesia, and supported by the Department of Informatics, Institut Teknologi Sepuluh Nopember.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] A. Simon and K. Robb, "Cancer: breast," in *Cambridge Handbook of Psychology, Health and Medicine*, Cambridge University Press, 2001, pp. 577–580. doi: 10.1017/CBO9780511543579.131.

[2] J. S. Ahn *et al.*, "Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine," *J. Breast Cancer*, vol. 26, no. 5, p. 405, Oct. 2023, doi: 10.4048/jbc.2023.26.e45.

[3] Y. Shin *et al.*, "Breast Cancer Risk After Hysterectomy: A Health Insurance Database-Based Analysis," *J. Breast Cancer*, vol. 28, no. 4, p. 215, Aug. 2025, doi: 10.4048/jbc.2025.0031.

[4] M. A. Aslam, A. Naveed, N. Ahmed, and Z. Ke, "A hybrid attention network for accurate breast tumor segmentation in ultrasound images," *Sci. Rep.*, vol. 15, no. 1, p. 39633, Nov. 2025, doi: 10.1038/s41598-025-23213-6.

[5] G. Zhao, X. Zhu, X. Wang, F. Yan, and M. Guo, "Syn-Net: A Synchronous Frequency-Perception Fusion Network for Breast Tumor Segmentation in Ultrasound Images," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 3, pp. 2113–2124, Mar. 2025, doi: 10.1109/JBHI.2024.3514134.

[6] J. Huang *et al.*, "EMGANet: Edge-Aware Multi-Scale Group-Mix Attention Network for Breast Cancer Ultrasound Image Segmentation," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 8, pp. 5631–5641, Aug. 2025, doi: 10.1109/JBHI.2025.3546345.

[7] F. A. Hermawati, H. Tjandrasa, and N. Suciati, "Phase-based thresholding schemes for segmentation of fetal thigh cross-sectional region in ultrasound images," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4448–4460, Jul. 2022, doi: 10.1016/j.jksuci.2021.02.004.

[8] M. H. A. M. H. Himel, P. Chowdhury, and M. A. M. Hasan, "A robust encoder decoder based weighted segmentation and dual staged feature fusion based meta classification for breast cancer utilizing ultrasound imaging," *Intell. Syst. with Appl.*, vol. 22, no. June, p. 200367, Jun. 2024, doi: 10.1016/j.iswa.2024.200367.

[9] Z. Ji *et al.*, "BGRD-TransUNet: A Novel TransUNet-Based Model for Ultrasound Breast Lesion Segmentation," *IEEE Access*, vol. 12, pp. 31182–31196, 2024, doi: 10.1109/ACCESS.2024.3368170.

[10] K. Hu, X. Zhang, D. Lee, D. Xiong, Y. Zhang, and X. Gao, "Boundary-Guided and Region-Aware Network With Global Scale-Adaptive for Accurate Segmentation of Breast Tumors in Ultrasound Images," *IEEE J. Biomed. Heal. Informatics*, vol. 27, no. 9, pp. 4421–4432, Sep. 2023, doi: 10.1109/JBHI.2023.3285789.

[11] A. Sulaiman *et al.*, "Attention based UNet model for breast cancer segmentation using BUSI dataset," *Sci. Rep.*, vol. 14, no. 1, p. 22422, Sep. 2024, doi: 10.1038/s41598-024-72712-5.

[12] G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap, "AAU-Net: An Adaptive Attention U-Net for Breast Lesions Segmentation in Ultrasound Images," *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1289–1300, May 2023, doi: 10.1109/TMI.2022.3226268.

[13] A. A. Hekal, A. Elnakib, H. E.-D. Moustafa, and H. M. Amer, "Breast Cancer Segmentation From Ultrasound Images Using Deep Dual-Decoder Technology With Attention Network," *IEEE Access*, vol. 12, pp. 10087–10101, 2024, doi: 10.1109/ACCESS.2024.3351564.

[14] J. Wang, J. Liang, Y. Xiao, J. T. Zhou, Z. Fang, and F. Yang, "TaiChiNet: Negative-Positive Cross-Attention Network for Breast Lesion Segmentation in Ultrasound Images," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 3, pp. 1516–1527, Mar. 2024, doi: 10.1109/JBHI.2024.3352984.

[15] Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang, "SMU-Net: Saliency-Guided Morphology-Aware U-Net for Breast Lesion Segmentation in Ultrasound Image," *IEEE Trans. Med. Imaging*, vol. 41, no. 2, pp. 476–490, Feb. 2022, doi: 10.1109/TMI.2021.3116087.

[16] M. J. Umer, M. I. Sharif, and J. Kim, "Breast Cancer Segmentation From Ultrasound Images Using Multiscale Cascaded Convolution With Residual Attention-Based Double Decoder Network," *IEEE Access*, vol. 12, pp. 107888–107902, 2024, doi: 10.1109/ACCESS.2024.3429386.

[17] J. Sun *et al.*, "DDRA-Net: Dual-Channel Deep Residual Attention UPerNet for Breast Lesions Segmentation in Ultrasound Images," *IEEE Access*, vol. 12, pp. 43691–43703, 2024, doi: 10.1109/ACCESS.2024.3373551.

[18] Q. Qin *et al.*, "MBE-UNet: Multi-Branch Boundary Enhanced U-Net for Ultrasound Segmentation," *IEEE J. Biomed. Heal. Informatics*, vol. 30, no. 1, pp. 575–585, Jan. 2026, doi: 10.1109/JBHI.2025.3589293.

[19] S. Abuowaida, H. A. Owida, D. M. Alsekait, N. Alshdaifat, D. S. AbdElminaam, and M. Alshinwan, "UltraSegNet: A Hybrid Deep Learning Framework for Enhanced Breast Cancer Segmentation and Classification on Ultrasound Images," *Comput. Mater. Contin.*, vol. 83, no. 2, pp. 3303–3333, Apr. 2025, doi: 10.32604/cmc.2025.063470.

[20] J. Huang *et al.*, "UltraMamba: Mamba-Based Multimodal Ultrasound Image Adaptive Fusion for Breast Lesion Segmentation," *IEEE Trans. Med. Imaging*, vol. 45, no. 5, pp. 2360–2372, May 2026, doi: 10.1109/TMI.2026.3653779.

[21] H. Wu, X. Huang, X. Guo, Z. Wen, and J. Qin, "Cross-Image Dependency Modeling for Breast Ultrasound Segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 6, pp. 1619–1631, Jun. 2023, doi: 10.1109/TMI.2022.3233648.

[22] S. Hossain *et al.*, "Automated breast tumor ultrasound image segmentation with hybrid UNet and classification using fine-tuned CNN model," *Heliyon*, vol. 9, no. 11, p. e21369, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21369.

[23] M. R. Islam *et al.*, "Enhancing breast cancer segmentation and classification: An Ensemble Deep Convolutional Neural Network and U-net approach on ultrasound images," *Mach. Learn. with Appl.*, vol. 16, no. June, p. 100555, Jun. 2024, doi: 10.1016/j.mlwa.2024.100555.

[24] A. Su *et al.*, "Multi-task learning for multi-scale breast cancer ultrasound image segmentation and classification based on visual perception," *Biomed. Signal Process. Control*, vol. 110, no. December, p. 108212, Dec. 2025, doi: 10.1016/j.bspc.2025.108212.

[25] Z. Zhuang, Z. Yang, A. N. J. Raj, C. Wei, P. Jin, and S. Zhuang, "Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion," *Comput. Methods Programs Biomed.*, vol. 208, no. September, p. 106221, Sep. 2021, doi: 10.1016/j.cmpb.2021.106221.

[26] F. Taheri and K. Rahbar, "Improving breast cancer classification in fine-grain ultrasound images through feature discrimination and a transfer learning approach," *Biomed. Signal Process. Control*, vol. 106, no. August, p. 107690, Aug. 2025, doi: 10.1016/j.bspc.2025.107690.

[27] P. Zhang, Y. Dong, J. Li, L. Jiang, M. Hu, and Y. Ping, "MSSM-MFP: Medical semantic segmentation model based on multiscale fusion perception," *Biomed. Signal Process. Control*, vol. 112, no. February, p. 108481, Feb. 2026, doi: 10.1016/J.BSPC.2025.108481.

[28] L. Wu *et al.*, "Breast tumor detection in ultrasound images with anatomical prior knowledge," *Image Vis. Comput.*, vol. 162, no. October, p. 105724, Oct. 2025, doi: 10.1016/J.IMAVIS.2025.105724.

[29] Y. Wang and Y. Yao, "Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement," *Sci. Rep.*, vol. 12, no. 1, p. 14720, Aug. 2022, doi: 10.1038/s41598-022-18747-y.

[30] S. Jiang, K. Gan, Q. Wang, L. Gong, W. Kong, and J. Yuan, "Advancing breast nodule detection in ultrasound images with multi-scale feature extraction," *Biomed. Signal Process. Control*, vol. 110, no. December, p. 108196, Dec. 2025, doi: 10.1016/j.bspc.2025.108196.

[31] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," in *13th International Conference on Learning Representations, ICLR 2025*, International Conference on Learning Representations, ICLR, Oct. 2024, pp. 41175–41218. Accessed: Jun. 19, 2026. [Online]. Available: http://arxiv.org/abs/2408.00714

[32] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, "Medical SAM 2: Segment medical images as video via Segment Anything Model 2," pp. 1–15, Dec. 2024, Accessed: Jun. 19, 2026. [Online]. Available: http://arxiv.org/abs/2408.00874

[33] J. Ma *et al.*, "MedSAM2: Segment Anything in 3D Medical Images and Videos," pp. 1–13, Apr. 2025, Accessed: Jun. 19, 2026. [Online]. Available: https://arxiv.org/pdf/2504.03600

[34] E. Chukwujindu, K. Faiz, A. De Sequeira, S. Chidom, and H. Faiz, "Improving medical image segmentation with SAM2: analyzing the impact of object characteristics and finetuning on multi-planar datasets.," *Eur. J. Radiol. Artif. Intell.*, vol. 3, no. September, p. 100034, Sep. 2025, doi: 10.1016/j.ejrai.2025.100034.

[35] Y. Zhou *et al.*, "Semiautomated segmentation of breast tumor on automatic breast ultrasound image using a large-scale model with customized modules," *Sci. Rep.*, vol. 15, no. 1, p. 17329, May 2025, doi: 10.1038/s41598-025-97098-w.

[36] J. Wu *et al.*, "Medical SAM adapter: Adapting segment anything model for medical image segmentation," *Med. Image Anal.*, vol. 102, no. May, p. 103547, May 2025, doi: 10.1016/j.media.2025.103547.

[37] Z. Zhong, Z. Tang, T. He, H. Fang, and C. Yuan, "Convolution Meets LoRA: Parameter Efficient Finetuning for Segment Anything Model," in *12th International Conference on Learning Representations, ICLR 2024*, International Conference on Learning Representations, ICLR, Jan. 2024, pp. 1–25. Accessed: Jun. 19, 2026. [Online]. Available: https://arxiv.org/pdf/2401.17868

[38] S. N. Gowda and D. A. Clifton, "CC-SAM: SAM with Cross-Feature Attention and Context for Ultrasound Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15103 LNCS, Springer Science and Business Media Deutschland GmbH, 2025, pp. 108–124. doi: 10.1007/978-3-031-72995-9_7.

[39] Z. Tu, L. Gu, X. Wang, and B. Jiang, "Ultrasound SAM Adapter: Adapting SAM for Breast Lesion Segmentation in Ultrasound Images," pp. 1–9, Apr. 2024, Accessed: Jun. 19, 2026. [Online]. Available: http://arxiv.org/abs/2404.14837

[40] T. Jiang *et al.*, "A segmentation knowledge-based global-local attention network for tumor classification in breast ultrasound images," *Pattern Recognit.*, vol. 171, no. March, p. 112152, Mar. 2026, doi: 10.1016/j.patcog.2025.112152.

[41] L. Guo, H. Zhang, and C. Ma, "ESAM2-BLS: Enhanced segment anything model 2 for efficient breast lesion segmentation in ultrasound imaging," *Comput. Med. Imaging Graph.*, vol. 126, no. December, p. 102654, Dec. 2025, doi: 10.1016/j.compmedimag.2025.102654.

[42] X. Yue *et al.*, "Morphology-enhanced CAM-guided SAM for weakly supervised breast lesion segmentation," *Biomed. Signal Process. Control*, vol. 116, no. May, p. 109509, May 2026, doi: 10.1016/j.bspc.2026.109509.

[43] X. Liu, M. Hussain, J. Huang, Q. Li, M. T. Khan, and H. Wu, "Auto-BUSAM: Auto-segmentation of attention-diverted low-contrast breast ultrasound images," *Displays*, vol. 92, no. April, p. 103314, Apr. 2026, doi: 10.1016/j.displa.2025.103314.

[44] B. Liu, H. Chen, T. Zhu, Z. Ye, H. Cui, and K. Wang, "Yolo-HLSAM: Adapting foundation segment anything model for semi-automatic detection and segmentation of breast cancer microcalcification clusters," *Biomed. Signal Process. Control*, vol. 111, no. January, p. 108300, Jan. 2026, doi: 10.1016/j.bspc.2025.108300.

[45] D. Yin, Q. Zheng, L. Chen, Y. Hu, and Q. Wang, "APG-SAM: Automatic prompt generation for SAM-based breast lesion segmentation with boundary-aware optimization," *Expert Syst. Appl.*, vol. 276, no. June, p. 127048, Jun. 2025, doi: 10.1016/j.eswa.2025.127048.

[46] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, no. February, p. 104863, Feb. 2020, doi: 10.1016/j.dib.2019.104863.