Research Square

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

A Decoder-Free Soft-Token Segmenter for Point-of-Care Breast Ultrasound

Chinmay Raut
Indian Institute of Technology Madras
Ekta Dhamija
All India Institute of Medical Sciences
Pradeeba Sridar
spradeeba@iitm.ac.in

Indian Institute of Technology Madras

Research Article

Keywords: Breast ultrasound, lesion segmentation, vision transformer, model compression, point-of-care imaging, imaging informatics

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11559512/v1

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

Additional Declarations: No competing interests reported.

A Decoder-Free Soft-Token Segmenter for Point-of-Care Breast Ultrasound

Chinmay Raut[1], Ekta Dhamija[2], Pradeeba Sridar[1*]

[1*]Department of Medical Science and Technology, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India.
[2]Department of Radiodiagnosis, All India Institute of Medical Sciences, New Delhi, Delhi, India.

*Corresponding author(s). E-mail(s): spradeeba@iitm.ac.in;

Abstract

Purpose: Automatic lesion localization can support point-of-care breast ultrasound (US), where operators often have limited training and scanners have limited compute and memory. Most segmentation models pair a pretrained encoder with a learned decoder, which adds parameters and computational cost. We propose a decoder-free soft-token segmenter that supervises each vision transformer (ViT) patch token with the fraction of lesion pixels in its patch.

Methods: A layer normalization and a single linear layer map each token to one logit, and the resulting grid is upsampled bilinearly to image resolution. We evaluated three encoder sizes and two initializations on the public BUS-BRA test set, both in-domain and out-of-domain, against five fully fine-tuned models with feature pyramid network (FPN) decoders.

Results: An oracle analysis showed that soft targets at patch size 8 raise the best achievable Dice to 0.970. The ViT-Tiny model, with 5.53 million parameters and a 577-parameter head, reached an in-domain Dice of 0.869 at 1,257 images/s, compared with 0.909 at 74 images/s for the most accurate FPN model DINOv2. Most of the remaining gap came from small lesions, where the model extended its mask into adjacent hypoechoic tissue, while low lesion contrast degraded all models similarly.

Conclusion: These results support decoder-free soft-token segmentation as an efficient lesion localizer for resource-constrained breast US.

Keywords: Breast ultrasound; lesion segmentation; vision transformer; model compression; point-of-care imaging; imaging informatics

ORCID. Chinmay Raut: 0000-0002-2694-8890; Ekta Dhamija: 0000-0001-8265-9674; Pradeeba Sridar: 0000-0003-1191-5188

Introduction

Clinical and deployment context

Breast cancer remains the most frequently diagnosed cancer among women worldwide and a leading cause of cancer mortality (Bray et al. 2024). Ultrasound is widely utilized for screening and evaluating palpable breast findings due to its affordability, real-time imaging capabilities, and broad availability, particularly in resource-constrained regions. However, ultrasound acquisition is highly operator-dependent, and accurate interpretation requires significant expertise. There is a critical need for automatic lesion segmentation to ensure consistency across different operators and to support reproducible measurement.

In remote and resource-limited settings, point-of-care breast ultrasounds are frequently performed and interpreted by front-line healthcare workers or non-specialists rather than trained radiologists. Because these operators often have minimal formal imaging training, manually reviewing scans in real-time can be challenging and increases the risk of overlooking malignant lesions. In these environments, a lightweight model that simply localizes a suspicious tumor, even without generating a perfectly refined segmentation mask, provides immediate clinical value. By acting as an automated triage mechanism, basic localization directly guides the attention of inexperienced operators to potential abnormalities. This targeted assistance reduces the time spent interpreting the scan and delivers critical decision support at the bedside, ensuring that patients with localized masses are promptly identified and referred to advanced diagnostic centers for definitive care.

Deploying these automatic segmentation models into clinical practice requires addressing the practical constraints of current hospital infrastructure. In many clinical settings, point-of-care ultrasound machines and standard hospital diagnostic workstations possess limited computational power and memory capacity. Moreover, to protect patient privacy, deployment on the machine is preferred over using a cloud-based service. Consequently, there is a pressing need for smaller, lightweight models designed specifically for deployment in these hardware-restricted environments.

Current segmentation models and encoder-decoder convention

Early research in automated segmentation was heavily dominated by Convolutional Neural Networks (CNNs), most notably the U-Net architecture, which established the foundational encoder-decoder pipeline. In this framework, an encoder processes the input image to extract abstract feature representations, and a decoder upsamples those features to recover a dense spatial mask. To capture global context, CNN encoders require significant depth, resulting in a very high parameter count.

More recently, vision transformers have been adopted for their inherent ability to capture long-range dependencies without deep convolutional layers. Building on this, domain-specific foundation models have emerged as powerful pretrained encoders. For ultrasound, models like UltraSAM (Meyer et al. 2026) and SAMUS (Lin et al. 2023) adapt the Segment Anything Model (SAM) paradigm to address the severe domain shift and image complexity of

sonographic data. Despite their utility, these ultrasound foundation models face limitations, including a frequent reliance on manual prompts, high computational memory footprints, and difficulties scaling to highly specialized clinical tasks due to the heterogeneity of their training data.

Furthermore, whether employing general vision transformers or specialized ultrasound foundation models, achieving dense pixel prediction typically requires coupling the pretrained encoder with a Feature Pyramid Network (FPN) or U-Net style decoder. Adding this decoder introduces substantial architectural complexity, significantly increases the overall parameter count, and demands considerably more training data.

A computationally cheap alternative to this heavy upsampling pipeline is treating segmentation directly as a token-level classification problem. Although frequently dismissed as too coarse for precise boundary delineation, this coarseness itself is a function of patch size. It is largely dictated by the chosen patch size and the reliance on hard binary labels rather than label softness. By reframing segmentation as a token classification problem instead of a dense spatial reconstruction task, the computational burden drops significantly. This simplification enables the network to achieve accurate results using an encoder with a much lower overall capacity.

Related work

Breast ultrasound lesion segmentation

Most deep learning methods for breast ultrasound lesion segmentation build on the U-Net encoder–decoder (Ronneberger et al. 2015), in which skip connections carry high-resolution encoder features to a symmetric decoder. Later CNN variants refine this template rather than replace it. Attention U-Net adds gates that suppress irrelevant background responses (Oktay et al. 2018), UNet++ redesigns the skip pathways as nested dense blocks (Zhou et al. 2018), and nnU-Net shows that a carefully configured plain U-Net remains a strong baseline across many segmentation tasks (Isensee et al. 2021). Breast-specific designs follow the same pattern. The selective-kernel U-Net adapts its receptive field to the wide range of mass sizes seen in breast ultrasound (Byra et al. 2020), and the global guidance network adds long-range context blocks and a boundary detection module to handle blurred lesion margins (Xue et al. 2021). Surveys of the field identify speckle noise, low contrast, posterior acoustic shadowing, and poorly defined margins as persistent challenges for automatic segmentation (Xian et al. 2018; Zhang et al. 2022; Thomas et al. 2023).

Transformer and hybrid models were introduced to model global context without very deep convolutional stacks. TransUNet places a transformer on top of a CNN encoder and keeps a U-Net-style decoder (Chen et al. 2021), and Swin-Unet uses shifted-window transformer blocks in both the encoder and the decoder (Cao et al. 2023). These models change how the encoder represents the image, but all of them keep a learned, multi-stage decoder to recover the full-resolution mask.

Ultrasound foundation models and self-supervised pretraining

Self-supervised pretraining learns image representations from unlabeled data and then transfers them to downstream tasks. Masked image modeling hides a large fraction of image patches and trains the network to reconstruct them. The masked autoencoder (MAE) does this

with a lightweight decoder that is discarded after pretraining, leaving only the encoder for transfer (He et al. 2022). The image-based joint-embedding predictive architecture (I-JEPA) predicts the representations of masked regions instead of their pixels, which steers learning toward semantic rather than textural features (Assran et al. 2023). Self-distillation methods take a third route. DINOv2 trains a vision transformer (ViT) (Dosovitskiy et al. 2021) to match the outputs of a momentum teacher across image views and, at scale, produces patch features that transfer to dense prediction with little adaptation (Oquab et al. 2024).

Ultrasound foundation models apply these ideas to sonographic data. USFM was pretrained on more than two million ultrasound images from multiple organs, centers, and devices, using a spatial–frequency dual masked image modeling objective designed for low-quality images (Jiao et al. 2024). USF-MAE applies MAE pretraining to roughly 370,000 images from 46 open-source ultrasound datasets and reports gains over natural-image initialization on downstream classification tasks, including BUS-BRA (Megahed et al. 2025). SAM-derived models form a separate, supervised branch. SAMUS adds a parallel CNN branch and feature and position adapters to SAM, which lets it run at a smaller input size (Lin et al. 2023), while UltraSam trains a SAM-style model on 43 public ultrasound segmentation datasets and accepts point and box prompts (Meyer et al. 2026).

These models are nearly always evaluated as encoders beneath a dense decoder or a prompt decoder. Less attention has gone to how much segmentation ability can be read directly from their patch tokens, and to how ultrasound-specific pretraining compares with supervised ImageNet initialization when the model is small. For this reason, the comparators in the present study include USFM, USF-MAE, DINOv2, Resnet50 and ViT-B encoders, each fine-tuned with an FPN decoder.

Use of decoders for segmentation

Transformer segmentation models differ mainly in how much work they assign to the decoder. SETR treats segmentation as sequence-to-sequence prediction on a plain ViT encoder and compares decoders ranging from simple upsampling to progressive and multi-level designs (Zheng et al. 2021). DPT reassembles tokens from several transformer layers into an image-like pyramid for dense prediction (Ranftl et al. 2021), and the simple feature pyramid of ViTDet builds multi-scale maps from a plain, single-scale ViT by deconvolution and pooling (Li et al. 2022). The Feature2Pyramid module used for the baselines in this study follows the same idea, and its output is fused by an FPN, which was first proposed for object detection (Lin et al. 2017). SegFormer takes the opposite direction. It pairs a hierarchical encoder with a lightweight all-MLP decoder and shows that a strong encoder reduces the need for a heavy decoder (Xie et al. 2021). Segmenter compares a mask-transformer decoder with a pointwise linear decoder applied directly to patch tokens and finds that the linear decoder is already competitive, with the mask transformer adding a consistent but modest gain (Strudel et al. 2021).

Linear decoders on patch tokens are also a standard diagnostic of representation quality. DINOv2, for example, reports semantic segmentation with a linear layer on frozen patch features, which tests how much spatial information the tokens carry without any learned upsampling (Oquab et al. 2024). The soft-token head in this study has the same form as such a probe, but it is trained end-to-end with the encoder and supervised with continuous per-patch targets. Soft targets have been studied separately as a way to keep partial-volume information

at object borders. SoftSeg trains with soft labels and soft outputs instead of binarized ones and reports that this preserves partial-volume information and improves segmentation compared with binary training (Gros et al. 2021).

Promptable models are often the most computationally expensive solution for automatic tumor segmentation. SAM combines a large ViT image encoder, a prompt encoder, and a mask decoder, and was trained on more than one billion masks (Kirillov et al. 2023). MedSAM fine-tunes SAM on a large collection of medical image–mask pairs and improves on the original model across medical imaging tasks (Ma et al. 2024). Systematic evaluations show that SAM's performance on medical images varies widely by modality and task and depends on the number and placement of prompts (Mazurowski et al. 2023). SAMUS and UltraSam narrow the domain gap for ultrasound (Meyer et al. 2026; Lin et al. 2023), but both expect a user prompt at inference unless an automatic prompt generator is added, as in the end-to-end variant of SAMUS (Lin et al. 2023).

Efficient and edge-deployable medical image analysis

Point-of-care breast ultrasound places the model close to the patient, often on portable hardware and in the hands of operators with limited training. Studies in Mexico showed that minimally trained operators using low-cost portable ultrasound, combined with computer-aided diagnosis, can triage women with palpable breast lumps (Love et al. 2018; Berg et al. 2023). In the more recent study, artificial intelligence applied to images from a portable scanner identified 96–98% of women with cancer and could have reduced specialist referrals by about half, even though the software had not been trained on images from that device (Berg et al. 2023). Both studies address lesion classification. Neither targets automatic lesion localization, which is the step that depends most on operator experience.

Efficient model design follows two main routes. The first uses architectures built for low compute, such as the inverted residual blocks with depthwise-separable convolutions in MobileNetV2 (Sandler et al. 2018) or the hybrid convolution–transformer blocks in MobileViT (Mehta and Rastegari 2022). The second shrinks a standard architecture. DeiT showed that small plain ViTs, including the Tiny and Small variants used in this study, can be trained on ImageNet alone, with optional distillation from a teacher network (Touvron et al. 2021). For promptable segmentation, MobileSAM distills SAM's image encoder into a much smaller one while keeping the original mask decoder (Zhang et al. 2023), and SAMUS lowers input resolution so that it runs on entry-level GPUs (Lin et al. 2023).

Much of this efficiency work compresses the encoder or the prompt pathway and leaves a learned decoder in place. Parameter count also gives an incomplete picture of deployment cost. Floating-point operations, peak memory, and latency on GPU and CPU can rank the same models differently, especially when the patch size or input resolution changes the number of tokens the encoder must process. The present study therefore reports all of these quantities for every architecture it evaluates.

Objectives and contributions

Vision transformers represent an image as a grid of patch tokens (Dosovitskiy et al. 2021). We hypothesized that each token can be supervised directly with a continuous target equal to the fraction of lesion pixels in its corresponding patch. Unlike a hard target that labels an entire

patch as foreground or background, this soft target preserves partial-volume information at lesion borders. A shared linear classifier then produces one logit per patch, and the resulting grid is interpolated to image resolution without a learned decoder. The design reduces trainable decoding components to a layer-normalization operation and a linear projection, but its spatial precision remains constrained by patch size and image resizing. In this study -

- We introduce a decoder-free soft-token segmentation method that supervises each image patch using the fraction of lesion pixels within the patch, preserving boundary information without a learned upsampling decoder.
- We conduct a controlled evaluation across three encoder capacities, two initialization strategies, and both in-domain and external-dataset settings, with comparisons against five fully fine-tuned FPN-based models.
- We characterize the accuracy–efficiency trade-off using parameter count, GFLOPs, peak memory, GPU and CPU latency, and throughput, while identifying small and low-contrast lesions as the principal sources of error.

Methods

Datasets and corpora

Table 1 Corpus inventory. Training manifest, eight public breast ultrasound sources. BUS-BRA is excluded by construction and used only as the held-out test population.

Source	Ref.	Images	Country	Lesion fraction (lesion-bearing only)[a]
BUS_UC	Iqbal and Sharif 2024	811	Netherlands	0.1395
BUSI	Al-Dhabyani et al. 2020	780	Egypt	0.0930
BUS-UCLM	Vallez et al. 2025	683	Spain	0.0758
BREAST	Pawłowska et al. 2024	252	Poland	0.0674
BUSC	Iqbal 2023	250	China	0.1708
BUID	Abbasian Ardakani et al. 2023	232	Iraq	0.1545
S1_OA_BUS	Guo et al. 2021	202	Taiwan	0.1481
UDIAT	Yap et al. 2018	163	Spain	0.0484
AIIMS	—[b]	410	India	0.1348
Total		**3783**	—	Average Lesion Fraction: 0.1147

[a] *Lesion fraction is the mean per-image tumour area fraction after the* 224^2 *NEAREST resize the trainer applies*

[b] Private dataset (AIIMS Delhi); not publicly available.

The development manifest contained 3,783 images from 3,382 patients across nine sources (Table 1). Source-specific identifiers were harmonized before splitting. Duplicate and corrupted images were removed. The BUS-BRA dataset (Gómez-Flores et al. 2024) was reserved for cross-dataset testing and was not used for training. All the datasets except AIIMS dataset are publicly available. Ethical approval was taken for the private AIIMS dataset sourced from AIIMS Delhi.

Preprocessing. Images were loaded as three-channel RGB, resized to 224 × 224 with bilinear interpolation, and normalised with the ImageNet channel means and standard

deviations. Reference masks were binarised first and then resized to 224 × 224 with nearest-neighbour interpolation, so that no interpolation ever runs on an unbinarised mask. Training augmentation was a horizontal flip applied jointly to image and mask with probability 0.5, followed by brightness and contrast scaling drawn from U(0.8, 1.2) and applied to the image only, also with probability 0.5. No histogram equalisation, denoising, or intensity clipping was applied.

Oracle analysis of token-level ceilings

Predicting one value per patch discards information before any model is trained, and the usual objection to token-level segmentation is that the resulting grid is too coarse. We quantified that loss directly, independently of learning, by decoding the reference mask through the token grid and scoring the best possible segmentation performance against each parameter choice.

Let g be the reference binary mask at 224 × 224 and p the patch size. The per-patch tumour fraction is $f = \mathrm{AvgPool}_p(g)$. Two decoders were evaluated, each the best any token head of that design could achieve:

- **Hard.** Predict $1[f \geq 0.5]$ and upsample by nearest neighbour. A patch is considered to be foreground when at least half its pixels belong to the lesion.
- **Soft.** Predict f, upsample bilinearly (align_corners=False), and threshold the result at 0.5.

Both were evaluated at $p \in \{16, 8, 4, 2\}$ on the BUS-BRA test set (n = 383, mean tumour fraction 0.0876). Because the predictions derive from the reference mask, these scores bound what any learned token head can reach at that patch size, and the gap between the two rules isolates the cost of hard patch labels from the cost of grid resolution.

Softness buys more than halving the patch (Table 2, Fig. 1). Moving from hard to soft at $p = 16$ gains 0.0671 Dice at an unchanged token count, while moving from $p = 16$ to $p = 8$ under hard labels gains 0.0810 at four times the tokens. Thus, we choose the patch size as $p = 8$ with soft targets to achieve the best balance between the number of tokens and performance ceiling.

Table 2 Oracle ceilings for token-level decoding. Best achievable score for *any* token head, BUS-BRA test (n = 383, mean tumour fraction 0.0876). **hard** = predict 1[f ≥ 0.5], nearest-upsample. **soft** = predict the tumour area fraction f, bilinear-upsample, threshold 0.5.

Patch	Grid	Tokens	Mode	Dice	IoU	HD95
16	14×14	196	hard	0.8355	0.7263	3.610
16	14×14	196	soft	0.9027	0.8383	1.529
8	28×28	784	hard	0.9165	0.8485	1.127
8	**28×28**	**784**	**soft**	**0.9702**	**0.9429**	**0.138**
4	56×56	3136	hard	0.9589	0.9217	0.261
4	56×56	3136	soft	0.9886	0.9776	0.000
2	112×112	12544	hard	0.9802	0.9613	0.016
2	112×112	12544	soft	0.9962	0.9923	0.000

IoU, intersection over union; HD95, 95th-percentile Hausdorff distance (pixels on the 224 × 224 grid)

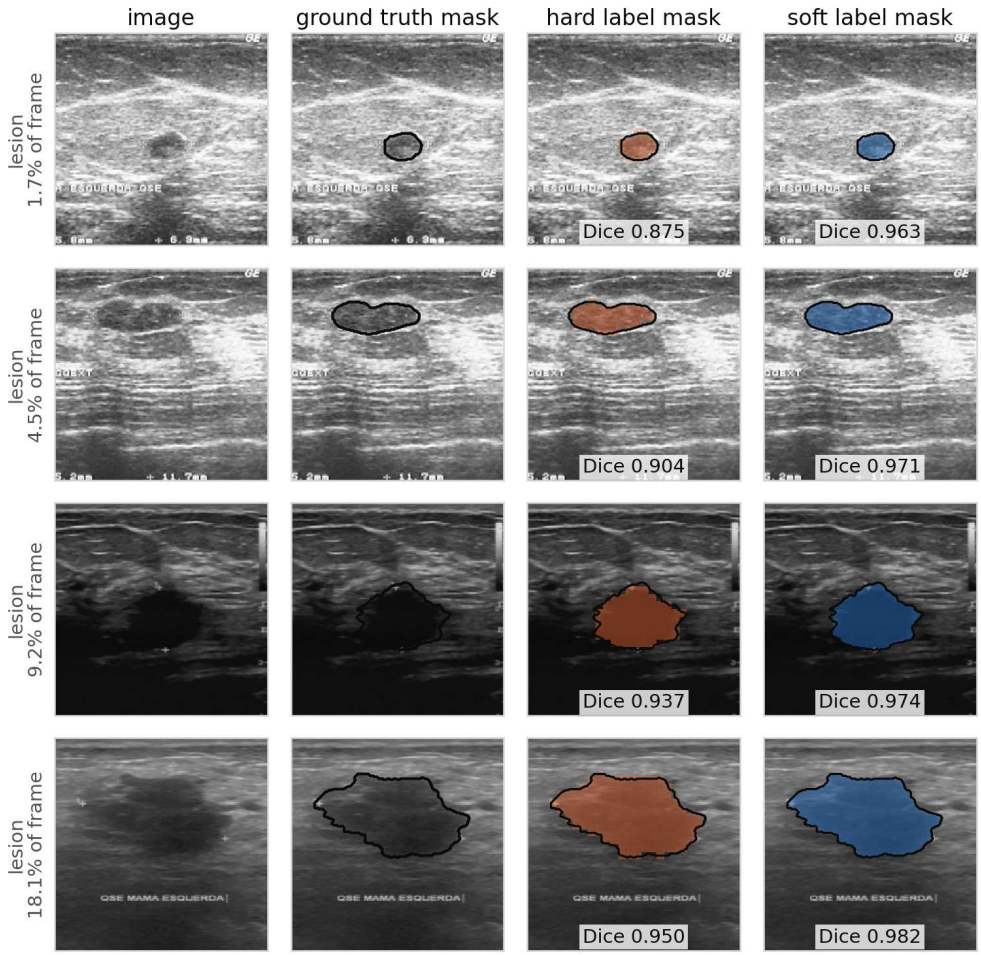


Fig. 1 Patch-level targets: a hard label discards the partial-volume information a soft one keeps. Oracle reconstructions from the ground truth mask alone at patch size 8 (28×28 tokens), the ceiling for any token head. Black outline is the ground truth lesion. hard = predict $\mathbb{1}[f \geq 0.5]$, nearest-upsample. soft = predict the tumour area fraction f, bilinear-upsample, threshold 0.5

Soft-token segmenter architecture

The segmenter is a vision transformer encoder followed by a layer normalisation and a single linear projection. A vision transformer was instantiated at 224×224 input with patch size 8, giving a 28×28 grid of 784 patch tokens. Classification and register tokens were discarded. Each remaining patch token passed through layer normalisation and a shared linear map to one logit per patch. At inference the logit grid was passed through a sigmoid, reshaped to 28×28, upsampled bilinearly to 224×224, and thresholded. No learned upsampling, feature pyramid, skip connection, or convolutional decoder was used. Figure 2 shows the full pipeline alongside the feature-pyramid decoder it replaces.

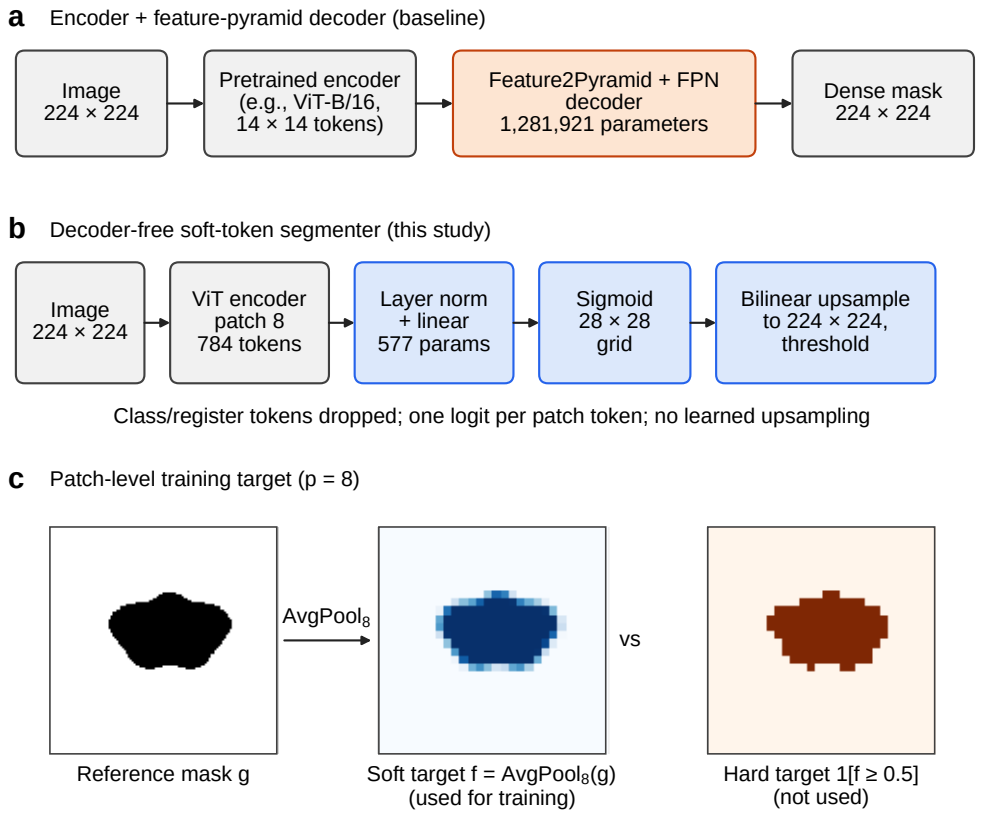


Fig. 2 Decoder-free soft-token segmenter compared with the feature-pyramid baseline. **(a)** Baseline: a pretrained encoder followed by a Feature2Pyramid module and FPN decoder (1,281,921 parameters) that recovers a dense 224 × 224 mask. **(b)** Proposed: a vision transformer at patch size 8 produces 784 patch tokens; class and register tokens are discarded, and a layer normalisation and shared linear map (577 parameters at ViT-Tiny) give one logit per patch, which is passed through a sigmoid, reshaped to 28 × 28, upsampled bilinearly to 224 × 224, and thresholded. **(c)** Training target: the reference mask g is average-pooled over each 8 × 8 patch to give the soft target f, the fraction of lesion pixels in each patch, which is used for training; the hard target $1[f \geq 0.5]$ is shown for comparison. The lesion shape in (c) is synthetic and for illustration only

For patch j containing N pixels, the target was $y_j = N^{-1} \sum_i m_i$, where $m_i \in \{0,1\}$ is the reference-mask value at pixel i. The target is therefore the continuous tumour area fraction rather than a binary class, which is what preserves the partial-volume information at lesion borders that the oracle analysis identifies as the dominant error term. The logit z_j was optimised with binary cross-entropy with logits, a proper scoring rule for probabilistic targets, averaged uniformly over all spatial tokens and all images. No token weighting, positive-class weighting, or auxiliary loss was used.

Lesion-free images were retained with an all-zero target rather than filtered out. Images with several annotated lesions were represented by the union of their lesion regions, since the reference masks are single-channel binary.

The segmentation head contains 577 parameters at ViT-Tiny, 384 in the layer normalisation and 193 in the linear map, against 1,281,921 for the feature-pyramid decoder in the baselines, a reduction of 2,222×. The head is 1,153 parameters at ViT-Small and 2,305 at ViT-Base. The complete ViT-Tiny/8 model contains 5,527,297 parameters, of which the head is 0.0104%.

Patch-16 to patch-8 checkpoint transfer

The ImageNet checkpoints were trained at patch 16, while the segmenter operates at patch 8, so the patch embedding and position embedding had to be adapted. The patch-embedding kernel was resized from 16×16 to 8×8 by bilinear interpolation, preserving the input and output channel dimensions. The position embedding grid was resized from 14×14 to 28×28 by bicubic interpolation, with the class token concatenated back unchanged. All transformer blocks were transferred verbatim. This resampling was performed by timm when the model was constructed at the target patch size.

Initialisations compared

Two prespecified initialisations were compared, with all other training and evaluation settings held constant so that initialisation is the only variable.

- **ImageNet.** Supervised ImageNet pretraining, loaded through timm at the target patch size.
- **Random.** No pretraining, used as the lower reference for the pretraining effect.

Training protocol

To isolate architecture and initialization as the sole independent variables, training procedures were strictly standardized across all soft-token arms. Models were optimized using AdamW with an initial learning rate of 3×10^{-4}, a weight decay of 0.05, and a batch size of 32. Training was conducted under automatic mixed precision for a maximum of 80 epochs, beginning with five warm-up epochs followed by a cosine decay schedule. Early stopping was enforced with a 15-epoch patience based on validation Dice scores computed at a fixed 0.5 threshold. The weights from the highest-scoring epoch were restored without utilizing checkpoint averaging. Following training, an optimal decision threshold was identified on the validation split via a grid search ranging from 0.20 to 0.80 in increments of 0.05. The threshold that maximized the validation Dice score, breaking ties toward the lower value, was frozen and applied exclusively to the test set.

The controlled experimental design evaluated combinations of encoder capacity (ViT-Tiny, ViT-Small, and ViT-Base) and initialization strategy (ImageNet and random) described in the Evaluation regimes subsection.

Baseline comparators comprised five fully fine-tuned models equipped with feature-pyramid decoders: DINOv2 ViT-B/14, USFM ViT-B/16, ImageNet ViT-B/16, ResNet50, and USF-MAE ViT-B/16. The baseline protocol employed the AdamW optimizer at an initial learning rate of 1×10^{-4} with layer-wise decay, a batch size of 16, and a maximum of 200 epochs with a 20-epoch patience on validation Dice. Furthermore, these comparators utilized a specific augmentation pipeline consisting of spatial flips, $\pm 30^{\circ}$ rotations, scale-and-crop operations, and both gamma and contrast jitter.

Evaluation regimes

In-domain. Training and testing both use BUS-BRA, through its own released split: 1,307 training, 185 validation and 383 test images from 1,875 total (69.7%, 9.9% and 20.4%). The split is patient-level on the Case identifier (744, 106 and 214 cases respectively). We verified that no case appears in more than one split. This regime is directly comparable to the published fine-tuning comparators, which use the same splits. The test split contains 264 benign and 119 malignant lesions acquired on three scanners (GE Logiq 5, GE Logiq 7, Toshiba Aplio 300).

External held-out dataset evaluation. Training uses the nine-source development manifest under an 80/20 split stratified on source, giving 3,114 training and 669 validation images, and testing uses the same locked BUS-BRA test split of 383 images. No BUS-BRA segmentation label is seen in training. This regime trains on *more* images than the in-domain regime, so any degradation measured between the two reflects distribution rather than sample size.

Metrics and statistical treatment

The primary evaluation endpoint was the Dice similarity coefficient (DSC). Secondary pixel-level metrics included sensitivity, specificity, and the 95th-percentile Hausdorff distance (HD95). All metrics were computed per image and subsequently aggregated rather than pooled across the test set. Because 16.4% of the development dataset was lesion-free, true negative results (empty prediction and reference masks) were assigned a DSC of 1.0. Cases presenting exactly one empty mask received a fixed HD95 penalty of 224, corresponding to the image side length. HD95 is reported in relative pixels on a 224 grid rather than in physical millimeters because the BUS-BRA dataset does not provide pixel spacing metadata.

Per-image Dice was averaged across training seeds before comparison, and all comparisons were paired on the same test images. Predictions and image-level metrics were computed separately for each image. Statistical inference was clustered at the patient level so that multiple views from the same patient were not treated as independent observations. For each comparison, the paired difference was defined as delta = $\text{Dice}_{\text{soft-token}} - \text{Dice}_{\text{baseline}}$. Non-inferiority was concluded when the one-sided 95% lower confidence bound for delta exceeded -0.02. Non-inferiority to each baseline was assessed at a margin of 0.02 Dice using a one-sided patient-level permutation test on the mean paired difference, with Holm correction across comparisons. The largest deficit consistent with the data was reported as the one-sided 95% upper confidence bound from a patient-clustered bootstrap. A 0.03 margin was examined as a post hoc sensitivity analysis. Failure to demonstrate non-inferiority does not imply inferiority.

Computational efficiency protocol

Inference performance was evaluated on a dedicated workstation equipped with an NVIDIA L40S GPU (46 GB VRAM, driver 580.126.09), dual Intel Xeon Silver 4314 processors (2.40 GHz), and 125 GB of system memory. All benchmarking procedures were executed using PyTorch and CUDA 12.4 in 32-bit floating-point precision, strictly within a torch.no_grad() context. To guarantee high-fidelity latency measurements, models were initialized with random weights at their respective target resolutions and evaluated on an otherwise idle GPU.

The assessment quantified both single-image latency at a batch size of 1 and processing throughput at a batch size of 32.

By capturing granular iteration times rather than aggregated block averages, the full latency distribution was modeled to accurately report the mean, median, interquartile range, and 95th percentile. The final evaluative suite additionally documents total parameter counts, fvcore-derived FLOPs, peak memory allocation, batch-1 latency across hardware platforms, and batched processing throughput.

Results

In-domain performance

Table 3 In-domain results (train on BUS-BRA, n_{train} = 1,307; test n = 383). Upper block is ours (decoder-free soft-token). Lower block is full fine-tuning with an FPN decoder from the same benchmark harness on the same splits.

Model	Params	GFLOPs	GPU ms	img/s	Dice	HD95	Sens	Spec
Soft-token (ours, decoder-free)								
ViT-Tiny/8	**5.53M**	4.2	3.3	1257	0.8694 ± 0.0044	5.32 ± 0.50	**0.901**	**0.989**
ViT-Small/8	21.67M	16.8	3.4	492	0.8828 ± 0.0013	4.48 ± 0.13	**0.910**	**0.989**
ViT-Base/8	85.81M	66.9	7.6	157	0.8870 ± 0.0036	4.18 ± 0.41	0.905	0.991
Full fine-tune + FPN decoder								
DINOv2 ViT-B/14 + FPN	87.86M	123.0	13.2	74	0.9088 ± 0.0032	4.06 ± 0.14	0.920	0.992
USFM ViT-B/16 + FPN	87.04M	18.4	6.7	644	0.8910 ± 0.0017	3.78 ± 0.15	0.913	0.990
ViT-B/16 + FPN	87.08M	17.7	4.5	690	0.8901 ± 0.0011	4.39 ± 0.35	0.920	0.989
ResNet50 + FPN	24.59M	4.9	4.8	917	0.8878 ± 0.0036	4.80 ± 0.10	0.919	0.990
USF-MAE ViT-B/16 + FPN	87.08M	17.7	4.5	690	0.8533 ± 0.0044	7.29 ± 0.22	0.859	0.991

Within the decoder-free soft-token family (Table 3), performance improved with encoder capacity, although the gains diminished as computational cost increased. Mean Dice rose from 0.8694 for ViT-Tiny/8 to 0.8828 for ViT-Small/8 and 0.8870 for ViT-Base/8. Most of the improvement occurred between Tiny and Small (+0.0134 Dice), whereas scaling from Small to Base added only 0.0042 Dice. Boundary accuracy followed a similar pattern, with HD95 decreasing from 5.32 to 4.48 and 4.18 pixels. Sensitivity did not increase monotonically, reaching its highest value with ViT-Small/8 (0.910), while specificity remained high across all three capacities (0.989-0.991). These incremental accuracy gains required approximately fourfold increases in parameters and FLOPs at each capacity step. Throughput declined from 1,257 images/s for ViT-Tiny/8 to 492 images/s for ViT-Small/8 and 157 images/s for ViT-Base/8.

The decoder-free ViT-Base/8 closely approached the corresponding ImageNet-initialized ViT-B/16 with an FPN decoder. Dice differed by only 0.0031 (0.8870 versus 0.8901), while the decoder-free model achieved a numerically lower HD95 (4.18 versus 4.39 pixels) and higher specificity (0.991 versus 0.989), although its sensitivity was lower (0.905 versus 0.920). Removing the FPN reduced the total parameter count from 87.08 million to 85.81 million because most parameters reside in the encoder. However, the patch-8 decoder-free model

required 66.9 GFLOPs, compared with 17.7 GFLOPs for the patch-16 FPN model, and consequently had lower throughput. Thus, eliminating the learned decoder substantially reduces head complexity but does not necessarily reduce total computation when the finer token grid is applied to a large encoder.

The ViT-Tiny/8 configuration provided the most favorable compact operating point. It achieved a Dice score of 0.8694 with 5.53 million parameters, approximately 16 times fewer than the ViT-B FPN models. Its Dice was 0.0207 below the ImageNet ViT-B/16 with FPN and 0.0216 below the USFM ViT-B/16 with FPN, while requiring approximately one-quarter of their FLOPs and providing greater measured throughput. ViT-Tiny/8 also remained within 0.0184 Dice of ResNet50 with FPN despite using approximately 4.4 times fewer parameters. These results show that much of the in-domain segmentation performance can be retained in a substantially smaller model, although the remaining difference is relevant when precise boundary delineation is required.

DINOv2 ViT-B/14 with FPN achieved the highest Dice in Table 3 (0.9088), together with sensitivity and specificity of 0.920 and 0.992. This improvement came with the greatest computational cost: 87.86 million parameters, 123.0 GFLOPs, 13.2 ms GPU latency, and a throughput of 74 images/s. Relative to ViT-Tiny/8, DINOv2 improved Dice by 0.0394 but used approximately 16 times more parameters and 29 times more FLOPs, with nearly fourfold higher GPU latency and approximately 17-fold lower batched throughput. DINOv2 therefore defines the highest-accuracy operating point in this comparison, whereas ViT-Tiny/8 provides a substantially lighter alternative for settings in which memory and computation are constrained.

In-domain, ViT-Tiny/8 (5.5M parameters) was non-inferior to USF-MAE+FPN at the prespecified 0.02 Dice margin (Holm-adjusted one-sided $p = 0.001$). Non-inferiority at 0.02 was not shown for ResNet50+FPN, ImageNet ViT-B/16+FPN, USFM+FPN or DINOv2+FPN. However, its Dice deficit was at most 0.024 relative to ResNet50+FPN, 0.025 relative to ImageNet ViT-B/16+FPN and 0.029 relative to USFM+FPN (one-sided 95% upper confidence bounds), and at most 0.034 across all four baselines other than DINOv2 simultaneously. It achieves this with 1/16 the parameters of the ViT-B baselines and 1/4 those of ResNet50+FPN. In a post hoc sensitivity analysis at a 0.03 margin, non-inferiority also held against ResNet50+FPN and ImageNet ViT-B/16+FPN.

Computational cost

The efficiency results (Table 4, Fig. 3) clarify that the practical value of decoder-free segmentation depends on pairing the soft-token head with a compact encoder. ViT-Tiny/8 occupies the most favorable operating point. It retains most of the segmentation performance reported in the in-domain results while requiring substantially fewer parameters, less memory, and lower latency than the FPN-based ViT models. Its 577-parameter head is effectively negligible relative to the encoder, making the complete segmenter suitable for settings in which inference must occur locally under restricted memory and compute budgets.

Increasing encoder capacity produced diminishing returns. The modest improvement from ViT-Tiny/8 to ViT-Base/8 was accompanied by a large reduction in throughput and a nearly tenfold increase in peak memory. The principal efficiency gain therefore arises from simplifying both components of the system - using a small encoder and avoiding a learned decoder - rather than from decoder removal in isolation.

Table 4 Computational cost. Inference cost of every architecture in the study

Architecture	Covers	Input	Params	Head/decoder params	GFLOPs	Peak mem (MB, b1)	GPU ms (b1)	CPU ms (b1)	Throughput (img/s, b32)
ViT-Tiny/8 + soft-token head	ours (main)	224^2	**5.53M**	577	**4.22**	37 ± 0	**3.31** $\pm$ 0.07	26.1 ± 1.1	**1257** $\pm$ 15
ViT-S/8 + soft-token head	ours	224^2	21.67M	1,153	16.77	107 ± 0	3.39 ± 0.22	58.5 ± 4.3	492 ± 8
ViT-B/8 + soft-token head	ours	224^2	85.81M	2,305	66.87	364 ± 0	7.59 ± 0.07	135.6 ± 18.2	157 ± 2
ResNet50 + FPN	ResNet	224^2	24.59M	1,084,545	4.93	121 ± 0	4.83 ± 0.04	32.4 ± 2.2	917 ± 1
ViT-B/16 + Feature2Pyramid + FPN	ImageNet ViT-B, USF-MAE	224^2	87.08M	1,281,921	17.71	360 ± 0	4.49 ± 0.05	48.1 ± 2.9	690 ± 8
USFM ViT-B/16 (rel-pos bias) + Feature2Pyramid + FPN	USFM	224^2	87.04M	1,281,921	18.43	363 ± 0	6.66 ± 0.05	57.5 ± 2.0	644 ± 5
DINOv2 ViT-B/14 + Feature2Pyramid + FPN	DINOv2	518^2	87.86M	1,281,921	123.00	455 ± 0	13.15 ± 0.08	217.3 ± 8.6	74 ± 0

Covers, baseline models in Tables 3 and 6 that share each architecture; b1, batch size 1; b32, batch size 32; Peak mem, peak GPU memory; GPU/CPU ms, latency per image

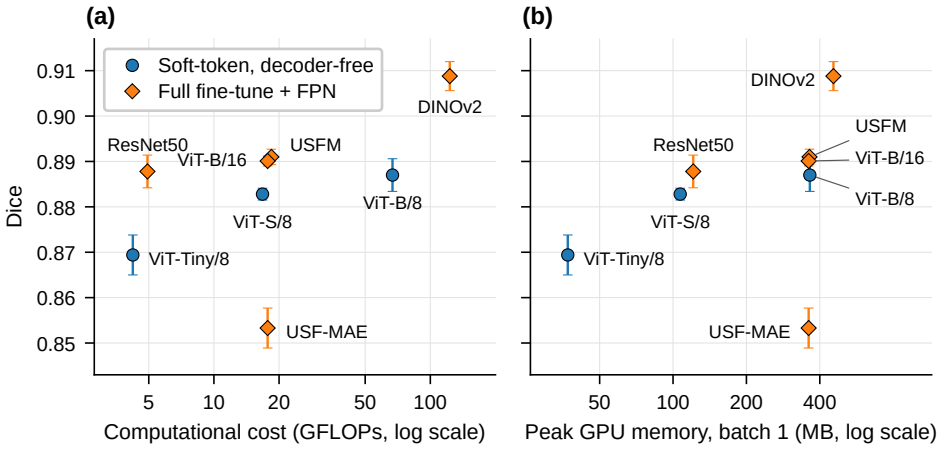


Fig. 3 Performance–efficiency comparison of the evaluated segmentation models. Dice score is plotted against **(a)** computational cost in GFLOPs and **(b)** peak GPU memory at batch size 1. Blue circles represent decoder-free soft-token models, while orange diamonds represent fully fine-tuned models with FPN decoders. Vertical error bars indicate the reported variation in Dice score. DINOv2 achieves the highest Dice score but has the greatest computational and memory requirements, whereas the soft-token models provide more resource-efficient alternatives

DINOv2 defined the opposite end of the trade-off. It achieved the highest Dice, but its larger input resolution resulted in the highest FLOP count, memory use, and latency, with approximately 17-fold lower throughput than ViT-Tiny/8. DINOv2 may therefore be appropriate where maximum overlap accuracy is prioritized and high-end computation is available, whereas ViT-Tiny/8 provides a more practical accuracy-efficiency balance for point-of-care and edge-oriented deployment.

Initialisation comparison

Table 5 Initialization comparison across encoder capacities. Random initialization was compared with ImageNet initialization for the decoder-free soft-token models (in-domain: train on BUS-BRA, test n = 383)

Model	Init	Dice	HD95	Sens	Spec
Soft-token (ours, decoder-free)					
ViT-Tiny/8	ImageNet	**0.8694 ± 0.0044**	**5.32 ± 0.50**	**0.901**	**0.989**
	Random	0.7075 ± 0.0110	18.09 ± 1.50	0.766	0.977
ViT-Small/8	ImageNet	**0.8828 ± 0.0013**	**4.48 ± 0.13**	**0.910**	**0.989**
	Random	0.6999 ± 0.0217	18.80 ± 2.69	0.774	0.975
ViT-Base/8	ImageNet	**0.8870 ± 0.0036**	**4.18 ± 0.41**	**0.905**	**0.991**
	Random	0.6709 ± 0.0089	22.96 ± 9.00	0.755	0.971

Bold marks the better initialization for each encoder capacity. ± is the standard deviation across seeds. HD95, 95th-percentile Hausdorff distance (pixels); Sens, sensitivity; Spec, specificity

ImageNet initialization substantially outperformed random initialization at every encoder capacity (Table 5). It improved Dice by 0.162 for ViT-Tiny, 0.183 for ViT-Small, and 0.216 for ViT-Base, while markedly reducing HD95. The widening advantage at higher capacity indicates that pretraining, rather than model size alone, is essential for effectively optimizing the decoder-free architecture.

External dataset performance

Table 6 External dataset results (train on 9-source manifest, n_{train} = 3,114; test on BUS-BRA, n = 383). No BUS-BRA segmentation label is seen in training. ± is the standard deviation across five seeds.

Model	Params	GFLOPs	Dice	HD95	Sens	Spec
Soft-token (ours, decoder-free)						
ViT-Tiny/8	**5.53M**	4.2	**0.7811 ± 0.0127**	17.39 ± 1.72	0.851	**0.982**
ViT-Small/8	21.67M	16.8	0.8071 ± 0.0086	14.25 ± 1.06	**0.865**	**0.984**
ViT-Base/8	85.81M	66.9	0.7973 ± 0.0161	14.83 ± 2.49	**0.850**	**0.985**
Full fine-tune + FPN decoder						
DINOv2 ViT-B/14 + FPN	87.86M	123.0	0.8715 ± 0.052	13.56 ± 2.64	0.922	0.984
USFM ViT-B/16 + FPN	87.04M	18.4	0.8382 ± 0.015	10.19 ± 2.30	0.875	0.986
ViT-B/16 + FPN	87.08M	17.7	0.8546 ± 0.0045	7.67 ± 0.79	0.894	0.987
ResNet50 + FPN	24.59M	4.9	0.8280 ± 0.0163	13.10 ± 3.41	0.868	0.985
USF-MAE ViT-B/16 + FPN	87.08M	17.7	0.7744 ± 0.0050	18.96 ± 2.71	0.791	0.991

External evaluation produced a consistent domain-shift penalty despite the larger and more diverse training set (Table 6). The decoder-free models lost approximately 0.08-0.09 Dice relative to their in-domain performance, with a concurrent increase in HD95. Increasing encoder size did not protect against this degradation: ViT-Small/8 achieved the strongest decoder-free result, whereas ViT-Base/8 was both less accurate and more variable across five seeds. The ordering therefore changed from the in-domain setting, indicating that additional capacity did not translate into greater robustness to unseen acquisition conditions.

The gap between decoder-free and FPN-based segmentation also widened externally. ViT-Base/8 had closely approached ViT-B/16 with FPN in-domain, but the FPN model retained a clearer advantage in both Dice and boundary distance on the external dataset. Nevertheless, ViT-Small/8 remained competitive with ResNet50 and USFM while operating at a lower computational cost, and ViT-Tiny/8 slightly outperformed the substantially larger USF-MAE model. DINOv2 achieved the highest mean Dice and sensitivity, but with the largest computational burden and considerable variation across five seeds. Conversely, ImageNet ViT-B/16 with FPN produced the lowest HD95, showing that overlap accuracy, boundary precision, computational efficiency, and training stability identified different preferred models.

These aggregate results establish that external degradation was not resolved by model scale alone, but they do not show whether performance declined uniformly across the test set. The accompanying increase in boundary error raises the possibility that the mean difference is driven by a smaller subset of severe failures. The following subsection therefore examines performance at the image level, separating lesion detection from boundary delineation and testing whether failure is associated with lesion size, localization, and image contrast.

Lesion-size, localization, and failure analysis

Performance differences between the decoder-free soft-token models and the FPN comparators were concentrated among the smallest lesions. When the BUS-BRA test set was divided into three categories by calculating tertiles using lesion area fraction, we found the cut off points to be 0.0408 and 0.1026 respectively for medium and large lesions. The mean Dice scores for the soft-token and FPN families were 0.8449 and 0.8590, respectively, for small lesions. In contrast, their scores were nearly identical for medium-sized lesions (0.8903 versus 0.8900) and large lesions (0.9060 versus 0.9062). The small-lesion stratum therefore accounted for essentially the entire aggregate Dice difference between the model families.

This difference primarily reflected boundary imprecision rather than failure to detect the lesion. For small lesions, the soft-token models produced fewer empty predictions than the FPN comparators (0.09% versus 0.52%), and no empty predictions were observed for medium or large lesions. However, small-lesion localization and boundary metrics favored the FPN models. HD95 was 5.56 versus 4.85 pixels for the soft-token and FPN families, respectively. This ordering did not persist for larger lesions. For medium and large lesions, the soft-token models achieved lower HD95 values than the comparators (3.90 versus 4.79 pixels and 4.45 versus 4.77 pixels, respectively). All distances are reported on the 224-pixel evaluation grid.

The characteristic small-lesion error was over-segmentation. Predictions exceeded the reference area by more than 25% in 24.5% of small-lesion cases for the soft-token models, compared with 18.8% for the FPN comparators. Rates of under-segmentation exceeding 20% were more similar (6.6% versus 5.0%). These errors were consistent with outward extension of the predicted mask into adjacent hypoechoic tissue.

Lesion-to-background contrast represented a second and largely independent source of error. Absolute local contrast was more strongly associated with mean Dice than lesion area fraction (Spearman rho = 0.433 and 0.383, respectively). Across increasing contrast-to-noise ratio quartiles, mean Dice was 0.829, 0.896, 0.905, and 0.904, indicating that degradation was concentrated in the lowest-contrast quartile. Both model families were affected similarly. Consequently, contrast explained substantial absolute error faced by all the models.

The qualitative examples in Fig. 4 illustrate these two failure modes. Rows 1 and 2 contain large, high-contrast lesions for which all eight models achieved Dice scores between 0.962 and 0.981. Row 3 (bus_0530-l) was the lowest-contrast image in the test set and was difficult for every architecture: two FPN comparators returned empty masks, while the soft-token models also produced low Dice scores. This case represents a shared failure under minimal lesion-to-background separation. In contrast, Row 4 (bus_0072-l) isolates a method-specific limitation. Although the lesion had moderate contrast, it occupied only 3.7% of the image. The soft-token mask extended into an adjacent hypoechoic band and achieved Dice scores of 0.398, whereas the FPN comparators achieved 0.829-0.920. Collectively, these findings support the decoder-free model as a reliable coarse localizer for lesions occupying more than approximately 4% of the image, while identifying precise delineation of smaller lesions as its principal remaining limitation.

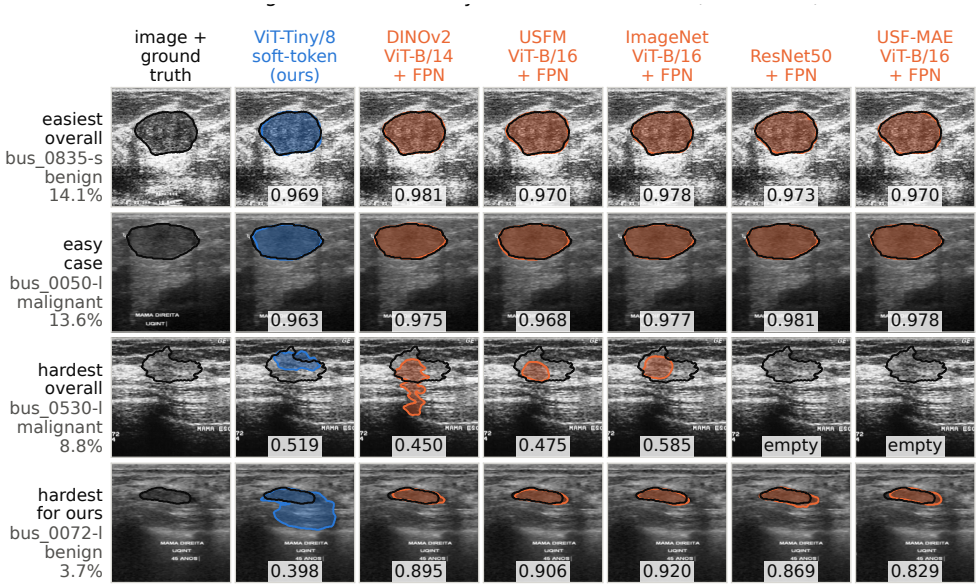


Fig. 4 Where the models agree and where they fail — BUS-BRA test set, in-domain, seed 42. Black outline is the ground truth lesion; coloured fill is the model's prediction. Blue = decoder-free soft-token head (577 parameters); orange = FPN decoder (1,281,921 parameters). Rows 1–2 are among the easiest images in the test set, row 3 the hardest for every model, row 4 the largest deficit of our arms against the comparators. DINOv2 predicts at 518^2 and is shown nearest-downsampled to 224^2

Discussion

Principal findings

This study establishes four principal findings regarding the design and deployment of breast ultrasound segmentation models. First, a learned decoder can be eliminated when the intended task is lesion localization rather than precise boundary characterization, enabling a substantially simpler model while retaining competitive segmentation performance. Second, performance degradation under cross-dataset domain shift is not distributed uniformly across the test set but is concentrated in a limited number of failure cases. Third, increasing encoder capacity from 5.5 million parameters in ViT-Tiny to more than 85 million parameters in ViT-Base produces diminishing returns, indicating that large architectures are not necessary for efficient point-of-care localization. Finally, pretrained initialization provides a clear advantage over training from scratch, with supervised ImageNet representations remaining effective even when transferred to a different patch stride.

Deployment implications

The documented efficiency metrics translate to profound differences in real-world clinical utility, particularly for edge deployment in resource-constrained environments. At point-of-care settings, the lack of expert radiologists necessitates assistive tools that operate with minimal memory and computational requirements. The proposed ViT-Tiny/8 configuration operates with only 5.5 million parameters and processes up to 1,200 images per second on

a standard workstation GPU, whereas heavily parameterized models like DINOv2 run at merely 6% of that speed. When mapped to native, low-power ultrasound machine compute, processing a standard Vision Transformer could take over a second per frame, resulting in an unusable 1 frame-per-second (FPS) video feed. Conversely, integrating this lightweight localization model enables native machines to operate smoothly at a real-time rate of 20 FPS. Localized, low-latency processing is vital in low-bandwidth environments because it circumvents the infrastructural bottlenecks of cloud computing or high-end hardware requirements. Furthermore, deploying multiple specialized, ultra-small models tailored to distinct clinical tasks can provide superior operational reliability on the edge compared to relying on a single, computationally expensive generalist model.

Localisation as a preprocessing step

Beyond standalone deployment, this highly efficient architecture offers significant utility as an automated preprocessing pipeline. High-end ultrasound segmentation frameworks, such as UltraSAM (Meyer et al. 2026) or SAMUS (Lin et al. 2023), often require manual prompting from users to initialize accurate dense pixel predictions. This manual interaction creates a severe bottleneck in high-throughput or non-specialist environments. Recent literature addressing foundation models in medical imaging emphasizes that automated prompting solutions, such as AutoProSAM for 3D scans (Li et al. 2025), are necessary to eliminate the reliance on domain experts. The proposed soft-token segmenter can serve as an exceptionally cheap preprocessing step to automatically generate coarse localizations, thereby replacing manual prompting with algorithmic initialization at a fraction of the computational cost.

Limitations

A primary limitation of the proposed architecture is that precise lesion boundaries are not densely replicated. Because the system predicts targets at the patch level rather than the exact pixel level, detailed morphologic characteristics of the tumor cannot be reliably interpreted solely by examining the resulting segmentation mask. Consequently, the system functions strictly as a highly efficient localization and triage model rather than a comprehensive, full-resolution segmentation tool intended for detailed boundary analysis.

Future work

Future research should focus on multi-organ validation to determine if the soft-token localization paradigm extends efficiently beyond breast sonography. Additionally, prospective, on-device clinical evaluations are required to measure real-time diagnostic impact and operator usability in authentic point-of-care settings. Finally, robust domain-coverage strategies must be developed to mitigate the isolated failure cases observed during cross-population deployment.

Conclusion

This study demonstrates that accurate breast ultrasound lesion localization does not necessarily require a large learned decoder. The decoder-free ViT-Tiny/8 retained most of the in-domain performance of substantially larger FPN-based systems while using only 5.53

million parameters, a 577-parameter segmentation head, and markedly less memory and computation. Its measured latency supports real-time inference, while its compact footprint makes integration into point-of-care ultrasound hardware plausible. Such a model could provide immediate visual guidance to non-expert operators, particularly in resource-constrained settings where specialist radiology support and high-performance computing infrastructure may be limited. It could also serve as an efficient first-stage localizer, supplying a region of interest or prompt to a larger segmentation or diagnostic model only when more precise analysis is required. Although external dataset shift and the delineation of small-lesion boundaries remain important challenges, the model continued to detect lesions reliably and offered a favorable balance between accuracy and computational cost. With prospective on-device and multicenter validation, decoder-free soft-token segmentation could enable practical, privacy-preserving, real-time decision support for breast ultrasound in settings where conventional high-capacity models are difficult to deploy.

Acknowledgements. This work was supported by the Indian Council of Medical Research (ICMR) under grant IIRP-2023-1815.

Statements and Declarations

Funding. This work was supported by the Indian Council of Medical Research (ICMR) under grant IIRP-2023-1815.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. The AIIMS New Delhi dataset was acquired under approval from the Institutional Ethics Committee of the All India Institute of Medical Sciences, New Delhi (approval number AIIMSA2312/07.10.2024), and all data were anonymized before analysis. The public datasets were used in accordance with their respective licenses and data-use terms. The study was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki and its later amendments.

Consent to participate. The requirement for informed consent was waived by the Institutional Ethics Committee because the study used retrospective, anonymized images.

Consent to publish. Not applicable. The manuscript contains no identifiable patient images or data.

Data availability. The public datasets analyzed in this study are available from the repositories cited in the manuscript. The private AIIMS dataset is not publicly available because of institutional and patient-privacy restrictions. Access requests may be directed to the corresponding author and are subject to ethics approval and institutional data-sharing requirements.

Author contributions. Chinmay Raut: conceptualization, methodology, software, experiments, formal analysis, and writing (original draft). Ekta Dhamija: clinical data acquisition and annotation of the AIIMS cohort, clinical interpretation, and writing (review and editing). Pradeeba Sridar: conceptualization, supervision, funding acquisition, and writing (review and editing).

References

Abbasian Ardakani A, Mohammadi A, Mirza-Aghazadeh-Attari M, Acharya UR. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Comput Biol Med. 2023;152:106438. https://doi.org/10.1016/j.compbiomed.2022.106438

Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863. https://doi.org/10.1016/j.dib.2019.104863

Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, LeCun Y, Ballas N. Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. pp. 15619–29. https://doi.org/10.1109/CVPR52729.2023.01499

Berg WA, López Aldrete AL, Jairaj A, Ledesma Parea JC, García CY, McClennan RC, Cen SY, Larsen LH, Soler de Lara MT, Love S. Toward AI-supported US triage of women with palpable breast lumps in a low-resource setting. Radiology. 2023;307:e223351. https://doi.org/10.1148/radiol.223351

Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63. https://doi.org/10.3322/caac.21834

Byra M, Jarosik P, Szubert A, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed Signal Process Control. 2020;61:102027. https://doi.org/10.1016/j.bspc.2020.102027

Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, vol 13803. Cham: Springer; 2023. pp. 205–18. https://doi.org/10.1007/978-3-031-25066-8_9

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306; 2021. https://doi.org/10.48550/arXiv.2102.04306

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR); 2021.

Gómez-Flores W, Gregorio-Calas MJ, Coelho de Albuquerque Pereira W. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Med Phys. 2024;51:3110–23. https://doi.org/10.1002/mp.16812

Gros C, Lemay A, Cohen-Adad J. SoftSeg: Advantages of soft versus binary training for image segmentation. Med Image Anal. 2021;71:102038. https://doi.org/10.1016/j.media.2021.102038

Guo Y, Duan X, Wang C, Guo H. Segmentation and recognition of breast ultrasound images based on an expanded U-Net. PLoS One. 2021;16:e0253202. https://doi.org/10.1371/journal.pone.0253202

He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. pp. 15979–88. https://doi.org/10.1109/CVPR52688.2022.01553

Iqbal A. BUSC dataset, version 1. Mendeley Data; 2023. https://doi.org/10.17632/vckdnhtw26.1

Iqbal A, Sharif M. Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. Eng Appl Artif Intell. 2024;127:107292. https://doi.org/10.1016/j.engappai.2023.107292

Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z

Jiao J, Zhou J, Li X, Xia M, Huang Y, Huang L, Wang N, Zhang X, Zhou S, Wang Y, Guo Y. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Med Image Anal. 2024;96:103202. https://doi.org/10.1016/j.media.2024.103202

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023. pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

Li C, Sultan RI, Khanduri P, Qiang Y, Chetty IJ, Zhu D. AutoProSAM: Automated prompting SAM for 3D multi-organ segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2025. pp. 3570–80. https://doi.org/10.1109/WACV61041.2025.00352

Li Y, Mao H, Girshick R, He K. Exploring plain vision transformer backbones for object detection. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol 13669. Cham: Springer; 2022. pp. 280–96. https://doi.org/10.1007/978-3-031-20077-9_17

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. pp. 936–44. https://doi.org/10.1109/CVPR.2017.106

Lin X, Xiang Y, Zhang L, Yang X, Yan Z, Yu L. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824; 2023. https://doi.org/10.48550/arXiv.2309.06824

Love SM, Berg WA, Podilchuk C, López Aldrete AL, Gaxiola Mascareño A, Pathicherikollamparambil K, Sankarasubramanian A, Eshraghi L, Mammone R. Palpable breast lump triage by minimally trained operators in Mexico using computer-assisted diagnosis and low-cost ultrasound. J Glob Oncol. 2018;4:1–9. https://doi.org/10.1200/JGO.17.00222

Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. 2024;15:654. https://doi.org/10.1038/s41467-024-44824-z

Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: An experimental study. Med Image Anal. 2023;89:102918. https://doi.org/10.1016/j.media.2023.102918

Megahed Y, Ducharme R, Walker M, Hawken S, Chan ADC. USF-MAE: Ultrasound self-supervised foundation model with masked autoencoding. arXiv preprint arXiv:2510.22990; 2025. https://doi.org/10.48550/arXiv.2510.22990

Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (ICLR); 2022.

Meyer A, Murali A, Zarin F, Mutter D, Padoy N. UltraSam: A foundation model for ultrasound using large open-access segmentation datasets. Int J Comput Assist Radiol Surg. 2026;21:93–102. https://doi.org/10.1007/s11548-025-03517-8

Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D. Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999; 2018. https://doi.org/10.48550/arXiv.1804.03999

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jegou H, Mairal J, Labatut P, Joulin A, Bojanowski P. DINOv2: Learning robust visual features without supervision. Trans Mach Learn Res. 2024.

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. https://doi.org/10.1038/s41597-024-02984-z

Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021. pp. 12159–68. https://doi.org/10.1109/ICCV48922.2021.01196

Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Cham: Springer; 2015. pp. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. pp. 4510–20. https://doi.org/10.1109/CVPR.2018.00474

Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021. pp. 7242–52. https://doi.org/10.1109/ICCV48922.2021.00717

Thomas C, Byra M, Martí R, Yap MH, Zwiggelaar R. BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets. Med Phys. 2023;50:3223–43. https://doi.org/10.1002/mp.16287

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning (ICML). PMLR 139; 2021. pp. 10347–57.

Vallez N, Bueno G, Deniz O, Rienda MA, Pastor C. BUS-UCLM: Breast ultrasound lesion segmentation dataset. Sci Data. 2025;12:242. https://doi.org/10.1038/s41597-025-04562-3

Xian M, Zhang Y, Cheng HD, Xu F, Zhang B, Ding J. Automatic breast ultrasound image segmentation: A survey. Pattern Recognit. 2018;79:340–55. https://doi.org/10.1016/j.patcog.2018.02.012

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.

Xue C, Zhu L, Fu H, Hu X, Li X, Zhang H, Heng PA. Global guidance network for breast lesion segmentation in ultrasound images. Med Image Anal. 2021;70:101989. https://doi.org/10.1016/j.media.2021.101989

Yap MH, Pons G, Martí J, Ganau S, Sentís M, Zwiggelaar R, Davison AK, Martí R. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform. 2018;22:1218–26. https://doi.org/10.1109/JBHI.2017.2731873

Zhang C, Han D, Qiao Y, Kim JU, Bae SH, Lee S, Hong CS. Faster segment anything: Towards lightweight SAM for mobile applications. arXiv preprint arXiv:2306.14289; 2023. https://doi.org/10.48550/arXiv.2306.14289

Zhang Y, Xian M, Cheng HD, Shareef B, Ding J, Xu F, Huang K, Zhang B, Ning C, Wang Y. BUSIS: A benchmark for breast ultrasound image segmentation. Healthcare (Basel). 2022;10:729. https://doi.org/10.3390/healthcare10040729

Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PHS, Zhang L. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. pp. 6877–86. https://doi.org/10.1109/CVPR46437.2021.00681

Zhou Z, Rahman Siddiquee MM, 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 (DLMIA/ML-CDS 2018). Lecture Notes in Computer Science, vol 11045. Cham: Springer; 2018. pp. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1