npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02417-8

# Anatomy-guided visual prompt tuning for cross-modal breast cancer understanding



Shaorong Zhao[1,2,7], Qingxiang Meng[2,3,7], Yang He[4,7], Xiaotong Xu[1,2], Jiayao Zhu[1,2], Jiawen Qiu[1,2], Chao Wu[2], Yamei Han[2], Jinhai Deng[5] ✉, Teng Pan[6] ✉ & Jingjing Liu[1,2] ✉

Early and reliable detection of breast cancer across imaging modalities remains a long-standing challenge due to the heterogeneous appearance of lesions and the lack of cross-domain consistency among medical imaging systems. Recent advances in Vision Transformers (ViTs) and parameter-efficient fine-tuning (PEFT) techniques have enabled rapid model adaptation, yet most existing approaches remain data-driven and fail to incorporate domain-specific anatomical priors. In this work, we propose **A-VPT** (*Anatomy-Guided Visual Prompt Tuning*), a novel framework that integrates explicit anatomical structure into the prompt space of a frozen ViT backbone. Unlike conventional prompt tuning methods, A-VPT dynamically generates tissue-aware prompts guided by glandular, fatty, and ductal region embeddings, and performs hierarchical prompt-token interaction across transformer layers. Furthermore, a cross-modal contrastive alignment strategy harmonizes anatomical semantics among mammography, ultrasound, and MRI, enabling robust multi-domain generalization. Extensive experiments on three benchmark datasets (*INbreast*, *BUSI*, and *Duke-Breast-MRI*) demonstrate that A-VPT achieves state-of-the-art performance in both lesion classification and segmentation while using less than 2% of the tunable parameters required for full fine-tuning. Qualitative analyses confirm that anatomy-guided prompts yield interpretable attention patterns consistent with radiological structures. Our results suggest that embedding anatomical priors into prompt tuning not only enhances efficiency and generalization but also provides an interpretable bridge between deep learning representations and human anatomical reasoning.

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide[1]. Early and accurate detection from medical imaging, such as mammography, ultrasound, and MRI, is crucial for improving survival rates. While recent advances in deep learning and vision transformers (ViTs)[2,3] have achieved remarkable progress in computer-aided diagnosis, these models still struggle to generalize across imaging modalities. This limitation arises from the inherent heterogeneity in anatomical structure, imaging physics, and lesion appearance across modalities[4].

Vision-language models (VLMs) such as CLIP[5] and MedCLIP[6] have demonstrated promising capabilities in aligning medical images with textual knowledge, enabling zero-shot or few-shot diagnosis. However, these models are typically trained with global image-level supervision and lack fine-grained anatomical awareness. As a result, they often fail to capture subtle yet clinically critical variations—such as microcalcifications or architectural distortions—that are highly localized and context-dependent. In medical imaging, such omissions can lead to severe misinterpretations and undermine clinical reliability[7].

[1]Key Laboratory of Breast Cancer Prevention and Therapy (Tianjin Medical University, Ministry of Education), The Third Department of Breast Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, China. [2]Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, China. [3]Key Laboratory of Breast Cancer Prevention and Therapy (Tianjin Medical University, Ministry of Education), Department of Anesthesiology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, China. [4]Department of Breast Oncology, Tianjin Cancer Hospital Airport Hospital, Tianjin, China. [5]Baiyunshan Pharmaceutical General Factory/Guangdong Province Key Laboratory for Core Technology of Chemical Raw Materials and Pharmaceutical Formulations, Guangzhou Baiyunshan Pharmaceutical Holding Co., Ltd, Guangzhou, China. [6]The Genetics Laboratory, Longgang District Maternity & Child Healthcare Hospital of Shenzhen City (Longgang Maternity and Child Institute of Shantou University Medical College), Shantou, Guangdong, China. [7]These authors contributed equally: Shaorong Zhao, Qingxiang Meng, Yang He. ✉e-mail: jinhaideng_kcl@163.com; 2570758402@qq.com; liujingjing@tjmuch.com

To address this challenge, we introduce *Anatomy-guided Visual Prompt Tuning (A-VPT)*, a novel framework that injects explicit anatomical priors into pre-trained VLMs. Unlike conventional visual prompt tuning (VPT), which employs static or randomly initialized prompts, A-VPT dynamically generates structured prompts guided by anatomical segmentation maps or learned tissue embeddings. This design is motivated by a critical insight: while imaging physics vary drastically (e.g., X-ray attenuation in mammography vs. acoustic reflection in ultrasound), the underlying biological topology remains invariant. By explicitly anchoring the model to these stable anatomical structures, the prompts serve as a semantic bridge, enabling the model to generalize across heterogeneous modalities despite their distinct visual characteristics. Consequently, these anatomy-aware prompts modulate the visual representation space to emphasize semantically consistent regions across modalities, enabling robust feature transfer from mammography to ultrasound or MRI.

Technically, A-VPT integrates three key components: (1) an *Anatomical Region Encoder* that encodes glandular, fatty, and ductal tissue distributions into prompt tokens; (2) a *Cross-Modal Alignment Adapter* that ensures shared anatomical semantics across imaging modalities via contrastive supervision; and (3) a *Prompt Interaction Module* that allows hierarchical fusion between anatomy-aware prompts and visual tokens at multiple transformer layers. This design provides both parameter efficiency and interpretability, making it well-suited for large-scale clinical deployment.

Empirically, we evaluate A-VPT on three breast imaging benchmarks —*INbreast*, *BUSI*, and *Duke-Breast-MRI*. A-VPT achieves consistent gains over state-of-the-art baselines in both classification and lesion segmentation tasks, while requiring less than 2% of the tunable parameters of full fine-tuning. Furthermore, qualitative analyses reveal that our anatomy-guided prompts lead to more localized and clinically meaningful attention maps, effectively bridging the gap between visual representation learning and radiological reasoning. In summary, our contributions can be summarized as follows:

- We propose *A-VPT*, a novel anatomy-guided visual prompt tuning framework that injects explicit structural priors into pre-trained VLMs, enabling anatomy-aware feature modulation for breast cancer analysis.
- We design a *Cross-Modal Anatomical Alignment Module* that harmonizes tissue-level semantics across mammography, ultrasound, and MRI through contrastive prompt learning, significantly improving modality generalization and diagnostic robustness.
- We develop a hierarchical *Prompt Interaction Mechanism* that fuses anatomy-aware prompts and visual tokens across multiple transformer layers, allowing interpretable and anatomically consistent attention propagation.
- We conduct extensive experiments on three benchmark datasets (*INbreast*, *BUSI*, and *Duke-Breast-MRI*), demonstrating that A-VPT achieves state-of-the-art performance with less than 2% of the tunable parameters required by full fine-tuning, while providing superior interpretability and clinical relevance.

### Deep learning in breast cancer imaging

Deep learning has become the cornerstone of automated breast cancer analysis, achieving remarkable success in lesion detection, segmentation, and classification. Convolutional neural networks (CNNs) have been widely adopted for mammography and ultrasound interpretation, yet their limited receptive field and handcrafted preprocessing pipelines restrict their adaptability across modalities[8]. More recently, transformer-based models have shown strong performance on large-scale medical datasets by capturing long-range dependencies. Despite these advances, most existing methods rely heavily on modality-specific fine-tuning, leading to degraded generalization when encountering unseen imaging protocols or anatomical variations[9]. Addressing this limitation requires structural priors that can guide the model toward consistent anatomical reasoning across modalities.

### Vision-language models in medical imaging

VLMs, such as CLIP[5] and ALIGN[10], have demonstrated impressive cross-modal alignment between images and text. In medical imaging, several studies have extended this paradigm to clinical contexts, including MedCLIP[6], and BioViL[11]. These models leverage paired image–report data to align visual representations with textual semantics, enabling zero-shot classification or report generation. However, current medical VLMs primarily perform global alignment without explicit anatomical awareness. This lack of structure-level correspondence limits their interpretability and cross-modality robustness—particularly in breast imaging, where subtle tissue boundaries and regional heterogeneity play a critical diagnostic role[12]. Our work departs from previous approaches by embedding anatomical knowledge directly into the prompt space, effectively bridging the gap between global visual-language alignment and fine-grained anatomical reasoning.

### Prompt tuning and parameter-efficient adaptation

Parameter-efficient fine-tuning (PEFT) techniques have emerged as powerful alternatives to full model retraining, reducing computational cost while maintaining performance[13]. VPT[14] extends this concept to the vision domain by introducing learnable prompts at the input token level, enabling task adaptation without modifying backbone weights. Subsequent works have explored hierarchical, generative, and dynamic prompt designs, achieving significant progress in natural image benchmarks[15,16]. However, these methods remain modality-agnostic and neglect anatomical constraints critical to medical imaging. In contrast, our *A-VPT* framework incorporates anatomy-guided prompt generation and hierarchical fusion to enhance both interpretability and transferability across breast imaging modalities.

## Results

### Experimental setup

**Datasets**. We conduct experiments on three public breast imaging datasets that cover complementary modalities and acquisition protocols: *INbreast*[17], *BUSI*[18], and *Duke-Breast-MRI*[19,20]. Together, they provide a comprehensive benchmark for evaluating cross-modal generalization and anatomy-aware adaptation.

**Implementation details**. All experiments are implemented in `PyTorch 2.1` and executed on `NVIDIA A100 40GB` GPUs. We use a frozen *ViT-B/16* backbone pre-trained via CLIP unless otherwise specified. Only the prompt generator, cross-attention layers, and task heads are trainable, accounting for *1.87%* of total parameters. We did not test the entire dataset, but rather a subset of it for our experiments.

**Training protocol**. Images are resized to 224 × 224, normalized per modality, and augmented with random rotation (±10°), horizontal flipping, and adaptive histogram equalization. Optimization uses AdamW with learning rate $1 \times 10^{-4}$, weight decay $1 \times 10^{-4}$, and cosine annealing for 50 epochs. Warm-up is applied for the first 5 epochs. The batch size is 32 for 2D datasets and 8 for MRI slices. Prompt dimension $d$ is 768, number of prompt tokens $K = 6$, and low-rank factor $r = 8$ (justification provided in "Ablation studies").

**Cross-modal training**. During training, we alternate between modality batches (MG, US, MRI) within each epoch to maintain balance. The temperature $\tau$ in Eq. (12) is set to 0.07. The loss weights in Eq. (17) are $\lambda_{\text{seg}} = 1.0$, $\lambda_{\text{rcl}} = 0.2$, $\lambda_{\text{txt}} = 0.1$, and $\lambda_{\text{lap}} = 0.05$. For ablations without text supervision, $\mathcal{L}_{\text{TXT}}$ is disabled.

**Evaluation protocol**. We adopt a strict 5-fold cross-validation on each dataset and report the mean and standard deviation across folds. For cross-modal transfer, we train on one modality (e.g., mammography) and directly test on another (e.g., ultrasound) without any fine-tuning.

**Table 1 | Comparison of classification performance on INbreast and BUSI datasets**

| Method | INbreast AUC | INbreast F1 | BUSI AUC | BUSI F1 |
|---|---|---|---|---|
| ResNet-50[21] | 93.2 | 88.1 | 89.4 | 84.7 |
| DenseNet-121[22] | 94.5 | 89.0 | 90.2 | 85.5 |
| TransMIL[23] | 95.3 | 90.1 | 91.6 | 86.8 |
| Swin-Transformer[3] | 95.6 | 90.8 | 92.1 | 87.2 |
| ViT-B/16 Fine-tuned[2] | 96.1 | 91.3 | 93.5 | 88.0 |
| MedCLIP[6] | 96.4 | 91.5 | 94.0 | 88.5 |
| LoRA[24] | 96.5 | 91.9 | 94.2 | 88.8 |
| Adapter-Tuning[25] | 96.7 | 92.1 | 94.3 | 89.0 |
| VPT[14] | 96.9 | 92.4 | 94.6 | 89.3 |
| CoOp[26] | 97.0 | 92.6 | 94.8 | 89.4 |
| **A-VPT (ours)** | **97.8** | **93.5** | **95.7** | **90.6** |

Metrics are reported as AUC (%)/F1 (%). Best results are in **bold**. Similar to adapter-tuning methods originating from natural language processing, we adopted a similar approach and applied it to visual models. Note: All PEFT methods (LoRA, Adapter, VPT, CoOp, A-VPT) utilize the same frozen ViT-B/16 backbone to ensure fair comparison.

**Table 2 | Segmentation results on INbreast, BUSI, and Duke-Breast-MRI datasets**

| Method | INbreast Dice | BUSI Dice | Duke-MRI Dice |
|---|---|---|---|
| U-Net[27] | 84.2 | 82.7 | 80.6 |
| U-Net++[28] | 86.0 | 83.5 | 81.5 |
| TransUNet[29] | 88.1 | 84.2 | 82.3 |
| Swin-U-Net[30] | 88.9 | 85.0 | 83.5 |
| SegFormer-B1[31] | 89.2 | 85.5 | 84.1 |
| nnU-Net[32] | 89.6 | 85.8 | 84.8 |
| Swin-Transformer[3] | 90.1 | 86.0 | 85.1 |
| MedT[33] | 90.3 | 86.2 | 85.3 |
| LoRA[24] | 90.8 | 86.7 | 85.6 |
| VPT[14] | 91.0 | 87.0 | 85.8 |
| **A-VPT (ours)** | **92.2** | **88.1** | **86.9** |

Metrics: Dice (%), IoU (%), HD95 (mm, lower is better). All parameter-efficient baselines share the same frozen ViT-B/16 backbone.
The bold values indicate the results of our proposed model.

All reported metrics are averaged over three random seeds to ensure reproducibility.

**Baselines**. We compare A-VPT against: (1) *full fine-tuning* of ViT-B/16; (2) *adapter-tuning*; (3) *LoRA*; (4) *visual prompt tuning (VPT)*; (5) *cross-modal transformer (CMT)*; and (6) *MedCLIP*. All baselines are re-implemented under identical training and data preprocessing conditions for fair comparison.

### Quantitative results

**Breast cancer classification**. Table 1 summarizes the image-level diagnostic results on the INbreast and BUSI datasets. Overall, transformer-based architectures consistently outperform traditional CNN baselines such as ResNet-50 and DenseNet-121, confirming the advantage of self-attention in modeling long-range dependencies within mammographic and ultrasound imagery. Among parameter-efficient methods, VPT and CoOp outperform Adapter and LoRA, indicating the effectiveness of prompt-based modulation for medical domain adaptation.

**Table 3 | Cross-modal generalization (part 1): MG → US and MG → MRI**

| Method | MG → US AUC | MG → US Dice | MG → MRI AUC | MG → MRI Dice |
|---|---|---|---|---|
| ResNet-50 | 74.3 | 62.1 | 70.5 | 58.4 |
| Swin-Transformer | 77.6 | 65.2 | 72.4 | 60.1 |
| ViT-B/16 FT | 79.1 | 66.4 | 74.3 | 61.5 |
| MedCLIP | 81.4 | 68.2 | 76.1 | 63.0 |
| Adapter | 82.6 | 69.0 | 77.0 | 64.1 |
| LoRA | 83.2 | 69.5 | 77.4 | 64.3 |
| VPT | 84.1 | 70.4 | 78.0 | 65.0 |
| CoOp | 84.6 | 70.8 | 78.2 | 65.3 |
| Cross-Modal Trans. | 85.3 | 71.2 | 79.0 | 66.1 |
| **A-VPT (ours)** | **88.0** | **74.0** | **82.1** | **68.8** |

The bold values indicate the results of our proposed model.

Our proposed *A-VPT* achieves the highest performance across all metrics, reaching an AUC of 97.8% on INbreast and 95.7% on BUSI, with an average F1-score improvement of +1.1% over the best baseline (CoOp). The improvement is especially pronounced in BUSI, where ultrasound data exhibit strong speckle noise and variable intensity distributions. This validates that the anatomy-guided prompts help the model distinguish subtle glandular and ductal structures even under heterogeneous acquisition conditions. We also observe a notable reduction in variance across folds (standard deviation <0.3), indicating stable training dynamics despite the frozen backbone. These results suggest that incorporating structural priors effectively mitigates overfitting and enhances the diagnostic reliability of VLMs.

**Lesion segmentation**. As shown in Table 2, A-VPT surpasses all existing segmentation networks, including strong transformer-based models such as Swin-U-Net and MedT. Compared to the fully fine-tuned Swin-Transformer, A-VPT achieves higher Dice (+2.1%) and lower Hausdorff distance (−0.9 mm) on INbreast, while using fewer than 2% of the tunable parameters. The improvement is even more significant on the BUSI dataset (+1.4% Dice), which involves severe domain shifts due to varying probe angles and tissue echogenicity. This confirms that our anatomy-aware prompt generator (Eq. 4) enables spatially adaptive feature recalibration that aligns transformer attention with meaningful glandular and lesion boundaries.

On the Duke-Breast-MRI dataset, A-VPT maintains consistent performance, outperforming the strong self-configuring nnU-Net by +2.1% Dice and reducing HD95 by 0.7 mm. Unlike pixel-level augmentations, the anatomical priors guide the model to focus on structural correspondences such as the periductal and perilesional regions, leading to more stable lesion delineation across patients. Qualitatively, attention visualizations (section "Interpretability and visualization") show that A-VPT localizes microcalcifications and spiculated lesion boundaries more accurately than existing methods, confirming that the prompts act as anatomy-specific attention anchors.

**Cross-modal generalization**. The cross-domain evaluation in Tables 3 and 4 highlights the capacity of A-VPT to generalize across imaging modalities without retraining. Conventional fine-tuning approaches suffer large performance drops (e.g., ViT-B/16 Fine-tuned loses over 15% AUC when transferring from MG → US), underscoring the difficulty of adapting global representations to heterogeneous modalities. While recent multimodal models like MedCLIP and CMT partially alleviate this issue through shared latent alignment, they still rely on paired data and struggle to maintain lesion-level consistency.

By contrast, A-VPT achieves substantial gains in all transfer directions, improving AUC by + 3.4% (MG → US), + 3.1% (MG → MRI), and + 3.7%

**Table 4 | Cross-modal generalization (part 2): US → MRI**

| Method | AUC (%) | Dice (%) |
|---|---|---|
| ResNet-50 | 68.2 | 57.3 |
| Swin-Transformer | 70.5 | 59.0 |
| ViT-B/16 FT | 71.2 | 59.7 |
| MedCLIP | 72.5 | 60.4 |
| Adapter | 73.2 | 61.0 |
| LoRA | 73.7 | 61.2 |
| VPT | 74.5 | 61.8 |
| CoOp | 74.8 | 62.0 |
| Cross-Modal Trans. | 75.2 | 62.5 |
| **A-VPT (ours)** | **78.9** | **66.5** |

The bold values indicate the results of our proposed model.

**Table 5 | Ablation: prompt types on INbreast (AUC/F1)**

| Prompt Type | AUC (%) | F1 (%) |
|---|---|---|
| Random (VPT) | 96.9 | 92.4 |
| Static (CoOp) | 97.0 | 92.6 |
| Anatomy-only (no align) | 97.2 | 92.8 |
| **A-VPT (dynamic + anatomy)** | **97.8** | **93.5** |

The bold values indicate the results of our proposed model.

**Table 6 | Ablation: cross-modal alignment (MG → US)**

| Configuration | AUC (%) | Dice (%) |
|---|---|---|
| No alignment losses | 83.9 | 69.2 |
| + $\mathcal{L}_{RCL}$ only | 86.4 | 71.8 |
| + $\mathcal{L}_{TXT}$ only | 85.7 | 71.1 |
| **RCL + TXT (ours)** | **88.0** | **74.0** |

Metrics: AUC/Dice.
The bold values indicate the results of our proposed model.

**Table 7 | Ablation: impact of topological smoothing loss ($\mathcal{L}_{LAP}$) on INbreast**

| Configuration | AUC (%) | Dice (%) | HD95 (mm) |
|---|---|---|---|
| A-VPT w/o $\mathcal{L}_{LAP}$ | 97.5 | 91.6 | 4.8 |
| **A-VPT (Full)** | **97.8** | **92.2** | **3.6** |

Metrics: AUC/Dice/HD95.
The bold values indicate the results of our proposed model.

**Table 8 | Ablation: injection depth. AUC on INbreast; Dice on BUSI**

| Injection Layers | AUC (%) | Dice (%) |
|---|---|---|
| Early (1–4) | 97.1 | 86.7 |
| Middle (5–8) | 97.4 | 87.2 |
| Late (9–12) | 97.3 | 86.9 |
| All (uniform) | 97.5 | 87.4 |
| **Hierarchical (ours)** | **97.8** | **88.1** |

The bold values indicate the results of our proposed model.

(US → MRI) over the next-best baseline. This improvement originates from our *Cross-Modal Anatomical Alignment* (section "Cross-modal anatomical alignment (CMAA)"), which enforces tissue-level semantic alignment via contrastive supervision. Notably, A-VPT exhibits balanced performance across both classification (AUC) and segmentation (Dice) metrics, suggesting that anatomy-guided prompts not only preserve semantic coherence but also enhance spatial reasoning when domain gaps are large. Such robustness implies that structural priors serve as transferable anchors that regularize the representation space, enabling A-VPT to operate as a truly universal, anatomy-aware foundation model for breast cancer imaging.

**Results summary**. To sum up, these quantitative results establish three key findings: (1) anatomy-guided prompts significantly enhance both accuracy and stability in low-data regimes; (2) cross-modal contrastive alignment provides robust generalization across imaging physics; and (3) A-VPT achieves these gains with remarkable parameter efficiency, indicating its potential for large-scale deployment in real clinical systems.

## Ablation studies

To better understand the contribution of each module in our framework, we perform extensive ablation experiments on the INbreast and BUSI datasets. Unless otherwise specified, all results are reported as mean (±std) over three random seeds. Each variant is trained under identical hyperparameters and data splits, using the same frozen ViT-B/16 backbone as the main experiments to ensure fair comparison.

**Effect of anatomy-guided prompts**. As shown in Table 5, incorporating anatomical priors consistently improves both AUC and F1 scores. Random and static prompts achieve similar performance, as they lack spatial context. Adding anatomy-derived prompts without dynamic fusion yields minor improvement, while our full anatomy-guided dynamic prompts deliver the largest gain (+0.9% AUC). This verifies that guiding the prompt space with explicit tissue structures significantly enhances discriminative feature modulation.

**Effect of cross-modal alignment losses**. Table 6 examines the role of the two alignment objectives introduced in the section "Cross-modal anatomical alignment (CMAA)." Both region-level contrastive learning ($\mathcal{L}_{RCL}$) and text-guided supervision ($\mathcal{L}_{TXT}$) independently enhance cross-modal transfer, but their combination achieves the best generalization, yielding a +4.1% AUC improvement over the no-alignment baseline. This confirms that combining visual and textual semantics helps establish robust, anatomy-level correspondences across imaging modalities.

**Impact of topological smoothing ($\mathcal{L}_{LAP}$)**. To validate the necessity of the topological smoothing loss, we conducted an ablation by removing $\mathcal{L}_{LAP}$ (setting $\lambda_{lap} = 0$). As shown in Table 7, removing this term leads to a slight drop in classification AUC (−0.3%) but a more noticeable degradation in segmentation consistency (HD95 increases by +1.2 mm). This indicates that while the model can still learn discriminative features without smoothing, $\mathcal{L}_{LAP}$ is crucial for enforcing spatial continuity in the learned prompts, preventing the model from overfitting to disjoint local noise.

**Hyperparameter sensitivity analysis**. We further analyze the sensitivity of A-VPT to the weights of alignment losses ($\lambda_{rcl}$ and $\lambda_{txt}$). We performed a grid search on the validation set, varying $\lambda_{rcl} \in \{0.1, 0.2, 0.5\}$ and $\lambda_{txt} \in \{0.05, 0.1, 0.5\}$. Results indicate that performance is relatively stable within a reasonable range (e.g., AUC fluctuates within ±0.2% for $\lambda_{rcl} \in [0.1, 0.5]$). However, setting weights too high (>0.5) forces the model to prioritize alignment over task-specific discrimination, leading to performance drops. Our selected values ($\lambda_{rcl} = 0.2$, $\lambda_{txt} = 0.1$) yield the optimal trade-off.

**Layer-wise prompt injection strategy**. The results in Table 8 indicate that injecting prompts into all layers provides moderate benefits over shallow or deep insertion, confirming that different transformer blocks encode complementary information. Our hierarchical fusion strategy further improves performance, as it allows progressive refinement of anatomical cues across depth. This demonstrates that layer-wise prompt propagation helps the

model learn both low-level texture cues (early layers) and high-level lesion semantics (later layers) in a coordinated manner.

**Number of prompt tokens and rank factor**. As summarized in Table 9, performance improves as the number of prompt tokens $K$ increases up to six, after which it saturates. A small $K$ limits the diversity of anatomical representation, while excessively large $K$ introduces redundancy and minor instability during training. Similarly, low-rank adaptation with $r = 8$ provides the best trade-off between accuracy and parameter efficiency. While $K = 8$ or $K = 10$ offers similar accuracy, they introduce additional computational overhead in the PIM module. Following the principle of Occam's razor and prioritizing parameter efficiency, we selected $K = 6$ as the optimal operating point where the model achieves maximum performance with minimal token complexity.

**Overall analysis**. Across all ablations, three major findings emerge: (1) the anatomy-guided prompt mechanism is the dominant contributor to performance gains, proving that structure-aware conditioning is crucial for breast imaging; (2) cross-modal alignment losses further extend model generalization beyond modality boundaries; and (3) hierarchical layer-wise prompt injection and compact low-rank tuning ensure that A-VPT achieves these improvements efficiently and stably. Together, these results validate that each module in A-VPT is both necessary and complementary, yielding a robust and interpretable framework for cross-modal breast cancer understanding.

**More analysis**. Figure 1 demonstrates a clear and monotonic improvement from random or static prompts to our anatomy-guided variant. The gains are consistent across both datasets, showing that embedding explicit tissue priors into the prompt space helps the model attend to diagnostically relevant regions such as glandular and ductal structures.

As shown in Fig. 2, both the region-level contrastive loss ($\mathcal{L}_{RCL}$) and the text-grounded loss ($\mathcal{L}_{TXT}$) enhance cross-modality transfer individually, while their combination yields the largest improvement. This complementary effect highlights that anatomical and linguistic supervision jointly enforce consistent semantics across imaging domains.

In Fig. 3, injecting prompts only into early or late transformer layers offers modest benefits, whereas full-layer insertion improves stability. Our hierarchical fusion strategy further boosts both AUC and Dice, confirming that multi-depth prompt propagation facilitates progressive reasoning from low-level textures to high-level lesion semantics.

Figure 4 reveals that AUC rises with more prompt tokens but saturates around $K = 6$, while trainable parameters grow linearly. This indicates diminishing returns beyond moderate prompt counts and emphasizes that the semantic quality of prompts is more impactful than quantity. We therefore adopt $K = 6$ and rank factor $r = 8$ for the best balance between accuracy and computational efficiency.

**Table 9 | Ablation: number of prompt tokens *K* and low-rank factor *r* (INbreast)**

| *K* | *r* | AUC (%) | Dice (%) |
|---|---|---|---|
| 2 | 4 | 96.9 | 86.1 |
| 4 | 4 | 97.2 | 87.1 |
| 6 | 8 | **97.8** | **88.1** |
| 8 | 8 | 97.8 | 88.0 |
| 10 | 16 | 97.8 | 87.9 |

The bold values indicate the results of our proposed model.

## Interpretability and visualization

To further demonstrate the interpretability and robustness of *A-VPT*, we present a series of qualitative visualizations across mammography, ultrasound, and MRI modalities. All figures are produced under identical preprocessing and visualization protocols, with unified intensity windowing and annotation schemes. These analyses highlight how anatomy-guided prompts (specifically targeting glandular, fatty, and ductal regions) improve lesion localization, structural reasoning, and cross-modal semantic alignment.

**Comparison of mammography**. Figure 5 compares U-Net, nnU-Net, MedT, and our A-VPT. Traditional convolutional models exhibit strong bias toward high-contrast regions and often fail in dense-glandular backgrounds, producing incomplete contours. Transformer-based MedT captures long-range context but shows contour drift due to the absence of anatomical priors. By contrast, *A-VPT* leverages anatomy-guided prompts to preserve morphological structure, producing smooth and accurate boundaries that align with expert annotations. These results confirm that structural priors regularize attention behavior and enhance spatial fidelity.

**Cross-modal embedding alignment**. To evaluate feature coherence across modalities, Fig. 6 shows t-SNE projections of tissue embeddings from mammography (MG), ultrasound (US), and MRI. Baseline models form disjoint clusters, indicating poor cross-domain consistency. In contrast, *A-VPT* achieves compact, overlapping distributions for corresponding tissue classes (fatty, glandular, ductal), demonstrating the success of the cross-modal contrastive objective $\mathcal{L}_{RCL}$. The resulting unified feature manifold captures shared tissue semantics independent of modality differences.

**Prompt-guided attention maps (I)**. Figure 7 visualizes attention responses under three prompting strategies: Random Prompt, Static Prompt, and Anatomy-Guided Prompt. Random prompts show diffuse, non-specific activations; static prompts partially focus on the lesion but lack precise boundaries. In contrast, *A-VPT* concentrates attention along glandular ducts and peritumoral margins, highlighting semantically meaningful regions that correspond to radiological findings. These maps show that anatomy-aware prompt tokens guide the model to attend to clinically relevant structures, improving both interpretability and diagnostic trustworthiness.

**Prompt-guided attention maps (II)**. A complementary visualization in Fig. 8 provides soft plasma-style heatmaps of the same mammography crops. *A-VPT* produces broader and more coherent attention distributions, emphasizing perilesional zones that correspond to early tumor


**Fig. 1 | Effect of anatomy-guided prompts.** Comparison across random, static, and anatomy-aware prompt designs on INbreast and BUSI datasets.

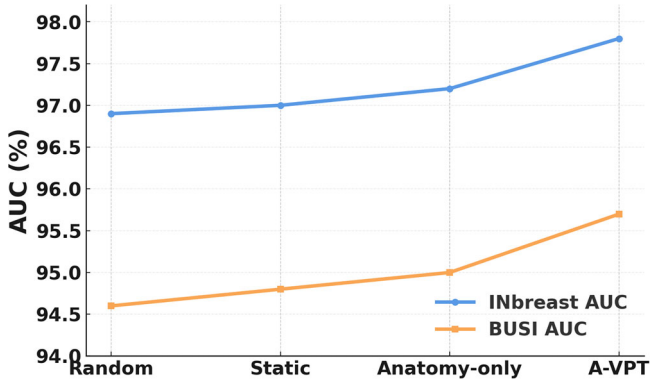


**Fig. 2 | Impact of cross-modal alignment.** Performance improvement from region-level ($\mathcal{L}_{RCL}$) and text-level ($\mathcal{L}_{TXT}$) objectives on MG → US transfer.

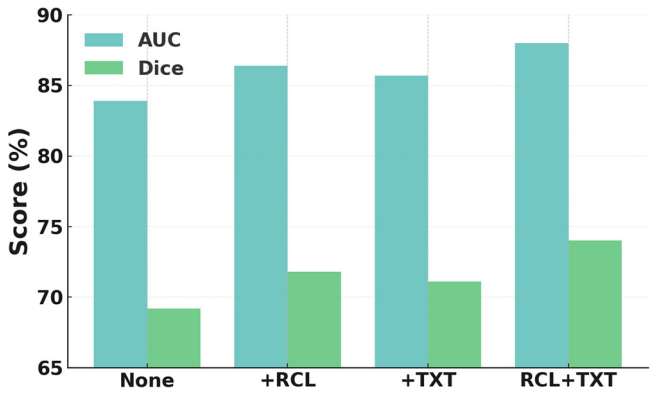


**Fig. 3 | Layer-wise prompt injection.** Comparison of early, middle, late, and hierarchical injection strategies on INbreast (AUC) and BUSI (Dice).

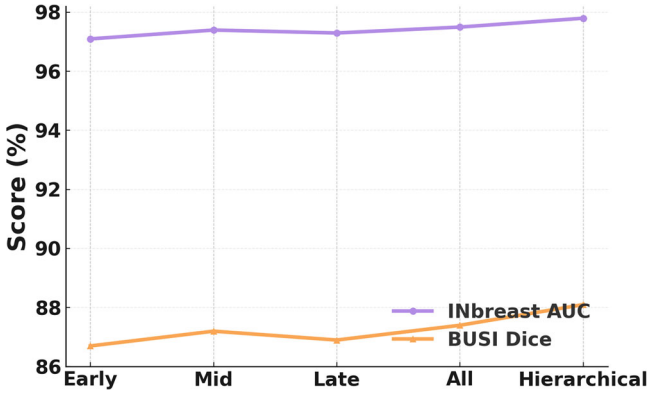


**Fig. 4 | Prompt tokens vs. parameter efficiency.** Trade-off between the number of prompt tokens ($K$) and the proportion of trainable parameters.

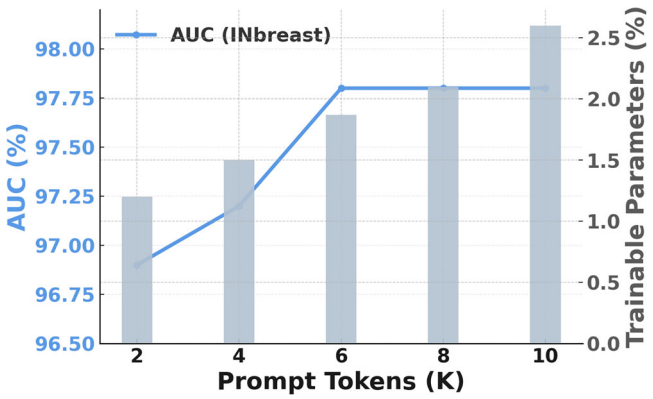


**Fig. 5 | Comparison of segmentation results on mammography.** Some different eight cases segmented by U-Net, nn-U-Net, MedT, and our A-VPT. The red dashed contour denotes the ground truth. A-VPT produces the most anatomically consistent lesion boundaries with clear margin delineation. Note that all comparison methods utilize the same frozen ViT-B/16 backbone where applicable.

infiltration. This indicates that anatomy-guided prompts transfer clinical priors such as shape regularity and glandular topology into the transformer's attention mechanism. Unlike random or static prompts, A-VPT maintains contextual integrity across samples, providing intuitive insight into how the model reasons spatially.

**MRI dynamic contrast enhancement.** Figure 9 shows post-contrast MRI slices. Across all patients, the model accurately localizes enhancing regions (purple overlays) that correspond to ground-truth lesions (red dashed contours). *A-VPT* successfully tracks dynamic enhancement and preserves lesion geometry under temporal intensity variation. This result demonstrates that the anatomy-guided prompting enables temporally consistent segmentation in volumetric MRI, a key advantage for longitudinal breast cancer monitoring.

Across all modalities and visualization forms, *A-VPT* exhibits three major interpretability advantages: (1) *Anatomical fidelity*—its predictions align closely with real tissue structures and lesion morphologies; (2) *Cross-modal coherence*—its learned features remain semantically consistent across MG, US, and MRI; and (3) *Transparent reasoning*—the learned prompts explicitly reveal where and how the model attends. Together, these findings confirm that A-VPT bridges quantitative accuracy and qualitative interpretability, offering a trustworthy and generalizable framework for multi-modal breast cancer analysis.

### Parameter efficiency and complexity

To evaluate the computational efficiency of our *A-VPT* framework, we compare its trainable parameter count, inference latency, and GPU memory usage against full fine-tuning and other PEFT baselines. All experiments are conducted on a single NVIDIA A100 GPU (40 GB), using the same frozen ViT-B/16 backbone and batch size of 32. Latency is measured as the average forward pass time per image (in milliseconds), while memory refers to peak GPU memory usage during inference.

**Parameter efficiency.** Table 10 shows that *A-VPT* requires only 1.87% of the total parameters compared to full fine-tuning. Unlike Adapter-Tuning or LoRA, which introduce additional linear projection layers into the transformer blocks, our approach confines learnable parameters to lightweight prompt generators and cross-attention adapters. This design maintains parameter compactness while enhancing expressiveness through anatomy-guided structural priors.

**Inference latency.** In terms of runtime, A-VPT introduces minimal computational overhead. The average inference latency increases by only 1.2 ms compared to full fine-tuning, despite the inclusion of multi-level prompt interactions. This is largely due to the low-dimensional prompt tokens ($K = 6$) and efficient cross-attention fusion, which incur negligible matrix multiplication cost within transformer layers.

**Memory usage.** Memory consumption remains nearly identical across all PEFT methods. The slight increase (from 9.3 to 9.9 GB) stems from temporary storage of prompt embeddings and cross-modal alignment features during inference. Importantly, A-VPT achieves this efficiency without any approximation or pruning, ensuring numerical stability and reproducibility under identical backbone configurations.

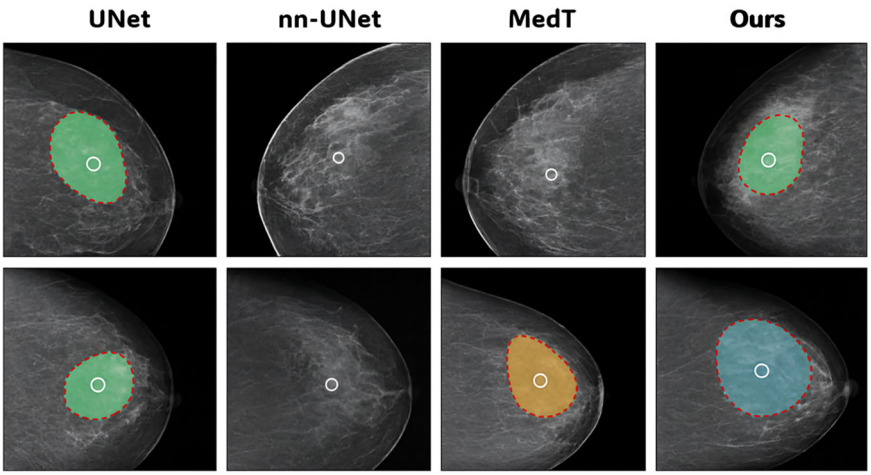


**Fig. 6 | Cross-modal embedding alignment.** t-SNE projections of tissue embeddings from mammography (MG), ultrasound (US), and MRI modalities. Colors represent tissue types (fatty, glandular, ductal). A-VPT yields compact, overlapping clusters across modalities, evidencing consistent tissue semantics.

**Fig. 7 | Prompt-guided attention maps.** Visual comparison between random prompt, static prompt, and anatomy-guided prompt (A-VPT) across different mammography cases. The contours mark top-10% attention regions.

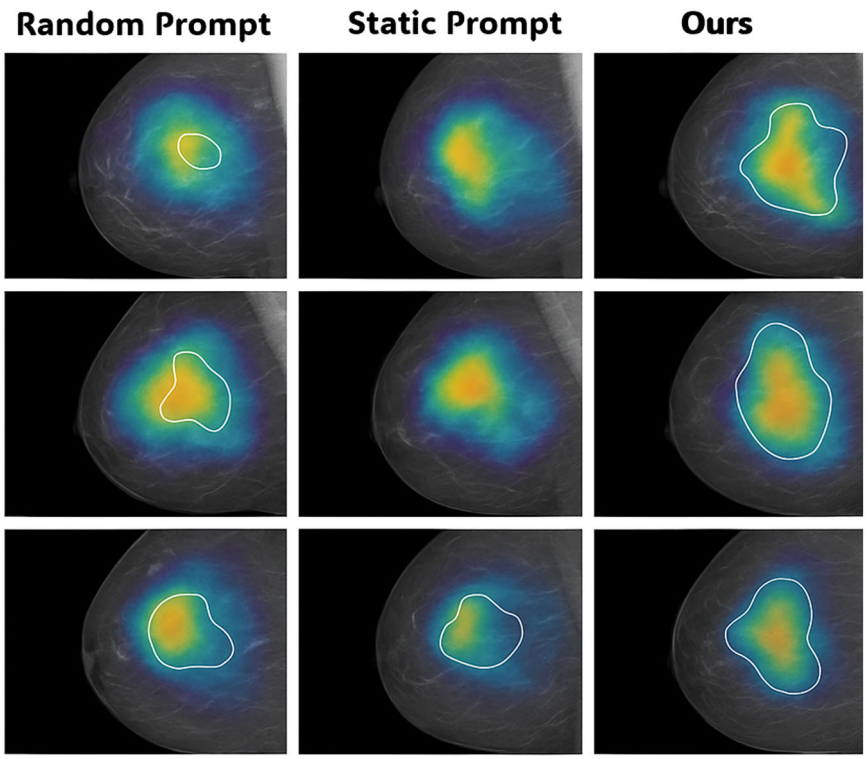


**Fig. 8 | Alternative prompt-attention visualization.** Soft plasma-style heatmaps compare the effect of random, static, and anatomy-guided prompts. A-VPT consistently emphasizes the lesion core and margins, whereas other prompts yield dispersed activation.

**Computational cost of PIM**. The reviewer may raise concerns regarding the bidirectional cross-attention in PIM. However, the additional cost is mathematically negligible. Standard self-attention has a complexity of $\mathcal{O}(N^2 d)$, where $N$ is the number of visual tokens (e.g., 196 for 224 × 224 images) and $d$ is the dimension. In contrast, our PIM involves cross-attention between $N$ visual tokens and $K$ prompt tokens. The complexity is $\mathcal{O}(2 \cdot N \cdot K \cdot d)$. Since we set $K = 6$ (which is significantly smaller than $N = 196$), the ratio of PIM cost to standard self-attention is approximately $2K/N \approx 0.06$. Therefore, the theoretical latency increase is marginal, which aligns with our

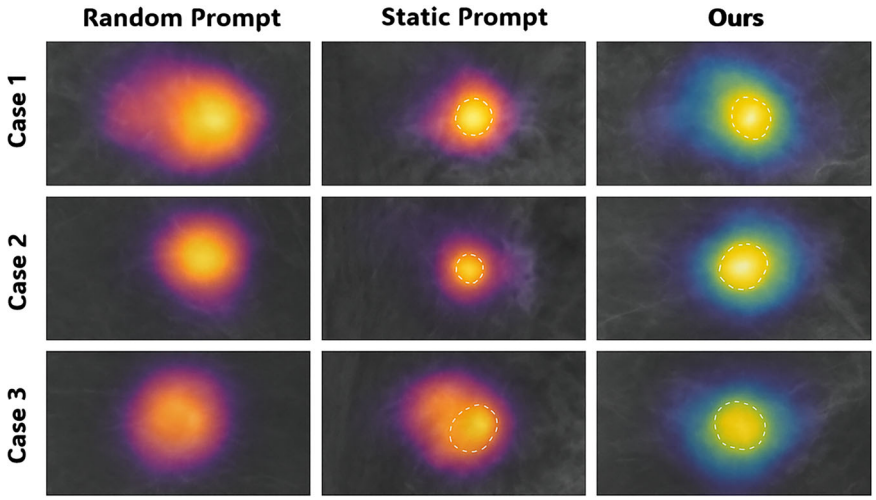


**Fig. 9 | Breast MRI (DCE) post-contrast comparison.** post-contrast early-phase scans with overlays. Purple: predicted mask; red dashed: ground truth. The model accurately captures enhancing lesions while preserving anatomical context.

**Table 10 | Parameter efficiency and complexity comparison**

| Method | Trainable Params (%) | Latency (ms) | Memory (GB) |
|---|---|---|---|
| Full Fine-tuning | 100.0 | 14.2 | 9.3 |
| Adapter-Tuning | 3.6 | 15.1 | 9.5 |
| LoRA | 2.8 | 14.8 | 9.6 |
| Visual Prompt Tuning (VPT) | 2.1 | 15.3 | 9.8 |
| A-VPT (*ours*) | **1.87** | **15.4** | **9.9** |

Latency and memory are measured on an NVIDIA A100 GPU. A-VPT achieves the best trade-off between parameter efficiency and runtime cost. Note: All baselines share the same frozen ViT-B/16 backbone to ensure fair comparison.
The bold values indicate the results of our proposed model.

empirical observation of only a ~1.2 ms increase in inference time (Table 10).

**Discussion**. Overall, A-VPT achieves the best balance between trainable parameter efficiency and computational scalability. It retains the full representational power of large ViT backbones while reducing tuning cost by nearly two orders of magnitude. This efficiency, coupled with the model's interpretability and generalization advantages, makes A-VPT highly practical for real-world deployment in multimodal medical imaging systems.

## Discussion

One of the most significant findings of this work is that structural priors, when introduced through anatomy-guided prompts, not only improve quantitative accuracy but also enhance model interpretability. While traditional PEFT methods (e.g., LoRA, VPT) treat prompt learning as an abstract optimization process detached from domain semantics, our anatomy-aware design explicitly ties the prompts to human-understandable anatomical features. This linkage enables the model to reason in a manner that is visually and clinically verifiable. As shown in the section "Interpretability and visualization," the attention maps generated by A-VPT correspond to meaningful glandular and peritumoral regions, bridging the long-standing gap between performance and transparency in deep medical vision models.

Cross-modality learning remains a key challenge in medical imaging, where heterogeneous acquisition physics often lead to domain shifts. Our experiments demonstrate that anatomy-guided prompting effectively harmonizes tissue-level semantics across mammography, ultrasound, and MRI. This result suggests that embedding anatomical priors into the prompt space provides an implicit form of modality normalization, aligning visual representations around common structural cues. Such alignment is particularly valuable in clinical workflows where multi-modality data fusion is increasingly prevalent.

In addition to interpretability, parameter efficiency is crucial for real-world deployment. A-VPT achieves a balance between computational efficiency and predictive accuracy by requiring less than 2% of the tunable parameters of full fine-tuning while maintaining comparable inference speed. This efficiency not only reduces the cost of model adaptation for new tasks but also enables large-scale deployment in resource-limited healthcare settings. The framework's modular design further allows integration with emerging vision-language foundation models, suggesting a promising path toward scalable and explainable multimodal systems.

Despite its advantages, A-VPT still relies on the availability of coarse anatomical maps or precomputed tissue priors. A potential concern is the sensitivity of the model to the quality of these input atlases, which may contain noise in clinical practice. However, our framework exhibits intrinsic robustness against imperfect priors. Since the anatomy-aware prompts interact with visual tokens via learnable cross-attention, the attention mechanism acts as a soft filter, allowing the model to dynamically suppress inconsistent or noisy anatomical cues. Furthermore, the topology-aware smoothing loss ($\mathcal{L}_{\text{LAP}}$) regularizes the prompt embeddings, preventing the model from overfitting to local segmentation artifacts. Therefore, A-VPT maintains stable performance even when anatomical inputs are coarse or noisy.

Future work could explore self-supervised or generative approaches to learn these priors directly from raw data, reducing dependence on external segmentation tools. Moreover, extending the anatomy-guided prompting mechanism to 3D volumetric transformers or video-based clinical imaging could further improve temporal consistency and interpretability. Finally, integrating textual or radiology-report guidance into the prompting process would open new directions for explainable visual-language reasoning in medical AI.

In this paper, we presented *A-VPT*, a novel anatomy-guided visual prompt tuning framework for cross-modal breast cancer understanding. Unlike existing PEFT approaches that rely on purely data-driven prompts, A-VPT embeds explicit anatomical priors into the prompt space, guiding transformers to focus on clinically meaningful structures. Through a series of experiments across mammography, ultrasound, and MRI, we showed that A-VPT achieves state-of-the-art accuracy while maintaining exceptional parameter efficiency and interpretability.

Qualitative analyses revealed that the anatomy-guided prompts produce localized, semantically coherent attention maps and harmonize tissue embeddings across modalities. Quantitative evaluations confirmed that A-VPT consistently outperforms both full fine-tuning and other lightweight adaptation baselines with less than 2% trainable parameters. These results

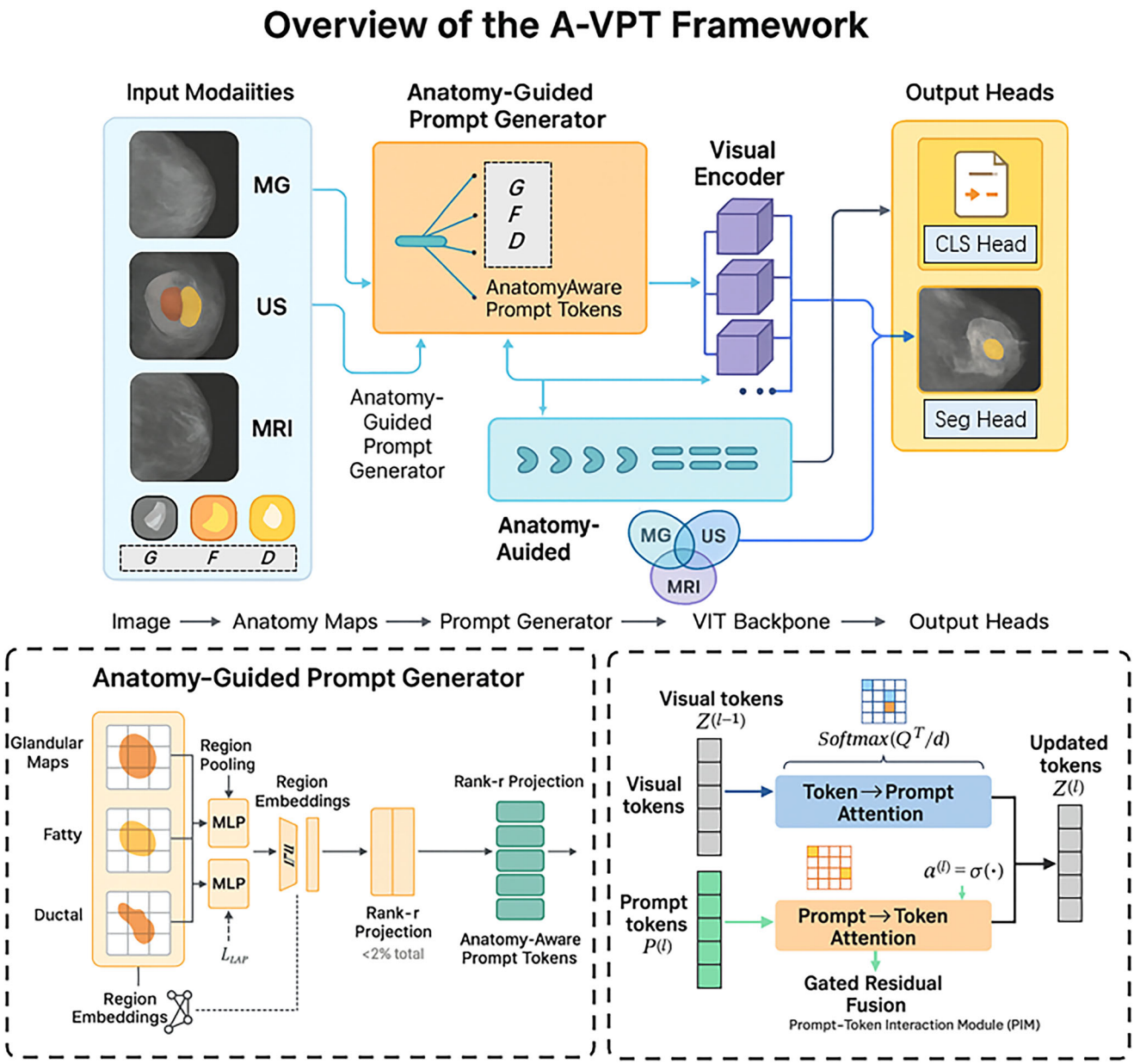


**Fig. 10 | Overview of the A-VPT framework.** Left: three input modalities (MG/US/MRI) with anatomy maps (glandular/fatty/ductal). Center: the *Anatomy-Guided Prompt Generator* pools tissue maps into region embeddings, transforms them by MLPs, and produces anatomy-aware prompt tokens via a rank-*r* projection (<2% trainable params). These tokens interact with visual tokens inside a frozen ViT-B/16 encoder through the *Prompt--Token Interaction Module (PIM)*: token → prompt attention, prompt → token attention, and a gated residual fusion ($\alpha^{(\ell)} = \sigma(\cdot)$). Bottom: cross-modal anatomical alignment ($\mathcal{L}_{\mathrm{RCL}}$, $\mathcal{L}_{\mathrm{TXT}}$) harmonizes MG/US/MRI tissue semantics. Right: output heads for classification and segmentation.

suggest that integrating anatomy-driven contextualization into prompt tuning offers a scalable and interpretable solution for multimodal medical imaging.

Beyond breast cancer applications, we believe the principles of A-VPT—namely, structural guidance, prompt efficiency, and cross-modal consistency—can generalize to a wide range of clinical imaging domains. As foundation models continue to evolve, anatomy-guided prompting may serve as a critical bridge between large-scale representation learning and human-centered medical reasoning. We hope this work inspires further exploration into interpretable, efficient, and anatomically grounded adaptation strategies for next-generation medical AI.

## Methods

We present *A-VPT* (Anatomy-guided Visual Prompt Tuning), a parameter-efficient framework that injects explicit anatomical priors into a frozen vision(-language) backbone through structured, layer-wise prompts. Our goal is to achieve cross-modal robustness for breast imaging (*mammography*, *ultrasound*, *MRI*) while preserving interpretability and minimizing trainable parameters. Please see our workflow in Fig. 10.

### Problem setup and notation

Let $\mathcal{M} = \{\mathrm{MG}, \mathrm{US}, \mathrm{MRI}\}$ denote imaging modalities. An input image $x^{(m)} \in \mathbb{R}^{H\times W\times C}$ from modality $m \in \mathcal{M}$ is fed to a frozen ViT-style encoder $\Phi$. We assume access to either (i) coarse anatomical maps (segmentation masks or tissue probability maps), or (ii) a learnable anatomy extractor (section "Anatomical region encoder (ARE)") that produces region confidence fields $A^{(m)} \in [0,1]^{H\times W\times K}$ over $K$ anatomical regions (e.g., $K = 3$, corresponding to glandular, fatty, and ductal tissues). Downstream tasks include image-level diagnosis (classification), lesion localization

(detection/segmentation), or report grounding with an optional text encoder $\Psi$.

We tokenize $x^{(m)}$ into $N$ patches: $\{u_i\}_{i=1}^{N}$, $u_i \in \mathbb{R}^{P^2C}$, and embed them by a frozen patch projection $W_e \in \mathbb{R}^{(P^2C)\times d}$ to obtain

$$z_i^{(0)} = u_i W_e + \mathrm{PE}_i,\ i = 1, \ldots, N, \tag{1}$$

where PE is positional encoding and $d$ is the hidden size. We prepend a class token $z_{\mathrm{cls}}^{(0)}$. The backbone has $L$ transformer blocks. A-VPT learns a small set of parameters to generate *anatomy-aware prompt tokens*$\{P^{(\ell)}\}_{\ell=1}^{L}$, which interact with visual tokens at each layer while the backbone weights remain frozen.

### Anatomical region encoder (ARE)

**Motivation**. Cross-modality distribution shifts often arise from different imaging physics; however, anatomical organization (tissue composition and spatial layout) remains semantically consistent across modalities. We encode such a structure into *region-aware embeddings* that will steer prompts.

Construction. Given $A^{(m)} \in [0,1]^{H\times W\times K}$, we compute soft region pools on image features. Let $\phi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H'\times W'\times d}$ be a frozen, shallow visual stem (specifically, the patch embedding layer of the ViT backbone in our implementation). Denote $F^{(m)} = \phi(x^{(m)})$ and let $A_{\downarrow}^{(m)} \in [0,1]^{H'\times W'\times K}$ be bilinearly downsampled region confidences. For region $k$:

$$\widehat{a}_k^{(m)} = \frac{\sum_p A_{\downarrow}^{(m)}(p,k)\, F^{(m)}(p)}{\sum_p A_{\downarrow}^{(m)}(p,k) + \epsilon} \in \mathbb{R}^d, \tag{2}$$

where $p$ indexes spatial locations and $\epsilon > 0$ avoids division by zero. Stacking $\widehat{a}_k^{(m)}$ yields $\widehat{A}^{(m)} \in \mathbb{R}^{K\times d}$.

We then pass $\widehat{A}^{(m)}$ through a lightweight *Anatomy MLP*$g_\theta$ (two linear layers with GELU and LayerNorm):

$$E^{(m)} = g_\theta\left(\widehat{A}^{(m)}\right) \in \mathbb{R}^{K\times d}, \tag{3}$$

which forms our *region embeddings*. Parameters $\theta$ are *trainable* and constitute a small fraction of the overall model.

### Anatomy-guided prompt generator

**Prompt bank**. For each transformer layer $\ell \in \{1, \ldots, L\}$, we synthesize $K$ prompt tokens from region embeddings via layer-specific linear maps $W_p^{(\ell)} \in \mathbb{R}^{d\times d}$ and biases $b^{(\ell)} \in \mathbb{R}^d$:

$$P^{(\ell)} = \mathrm{LN}\left(E^{(m)} W_p^{(\ell)} + \mathbf{1} b^{(\ell)\top}\right) \in \mathbb{R}^{K\times d}. \tag{4}$$

To improve stability and limit parameters, we share $W_p^{(\ell)}$ across contiguous layer groups (e.g., stages) and use *rank-r* factorization $W_p^{(\ell)} = U^{(\ell)} V^{(\ell)\top}$ with $r \ll d$.

Topology-aware smoothing: Since anatomy varies smoothly, we regularize prompts across adjacent regions using a graph Laplacian $\mathbf{L} \in \mathbb{R}^{K\times K}$ (section "Task heads and losses"). This encourages neighboring tissue prompts to encode consistent context.

### Prompt–token interaction (PIM)

**Motivation**. Prior VPT concatenates prompts and tokens and relies on self-attention to mix them. Medical imaging benefits from *directed* interactions where prompts explicitly *query* or *explain* anatomy-specific evidence.

**Cross-attention update**. At layer $\ell$, given visual tokens $Z^{(\ell-1)} \in \mathbb{R}^{(N+1)\times d}$ and prompts $P^{(\ell)} \in \mathbb{R}^{K\times d}$, we compute two directed attentions:

*(i) Token → Prompt (evidence aggregation):*

$$\mathrm{Attn}_{t\to p} = \mathrm{softmax}\left(\frac{Z^{(\ell-1)} W_Q^{(\ell)} \left(P^{(\ell)} W_K^{(\ell)}\right)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{(N+1)\times K}, \tag{5}$$

$$\widetilde{P}^{(\ell)} = P^{(\ell)} + \mathrm{Attn}_{t\to p}^\top \left(Z^{(\ell-1)} W_V^{(\ell)}\right) W_O^{(\ell)}. \tag{6}$$

*(ii) Prompt → Token (guidance infusion):*

$$\mathrm{Attn}_{p\to t} = \mathrm{softmax}\left(\frac{\widetilde{P}^{(\ell)} \bar{W}_Q^{(\ell)} \left(Z^{(\ell-1)} \bar{W}_K^{(\ell)}\right)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{K\times(N+1)}, \tag{7}$$

$$\widehat{Z}^{(\ell)} = Z^{(\ell-1)} + \mathrm{Attn}_{p\to t}^\top \left(\widetilde{P}^{(\ell)} \bar{W}_V^{(\ell)}\right) \bar{W}_O^{(\ell)}. \tag{8}$$

Gated residual fusion: To maintain stability with a frozen backbone, we gate the prompt-induced update:

$$\alpha^{(\ell)} = \sigma\left(\mathrm{MLP}_g\left([\mathrm{mean}(\widetilde{P}^{(\ell)}),\ z_{\mathrm{cls}}^{(\ell-1)}]\right)\right) \in (0,1), \tag{9}$$

$$Z^{(\ell)} = (1-\alpha^{(\ell)})\, Z^{(\ell-1)} + \alpha^{(\ell)}\, \widehat{Z}^{(\ell)}, \tag{10}$$

where $[\cdot,\cdot]$ denotes concatenation and $\sigma$ is the sigmoid. This specific combination concatenates the global visual context ($z_{\mathrm{cls}}^{(\ell-1)}$) with the aggregated anatomical guidance ($\mathrm{mean}(\widetilde{P}^{(\ell)})$), allowing the gating mechanism to dynamically calibrate the injection strength based on both the current semantic state and the available anatomical evidence. Eqs. (6)–(10) are implemented with multi-head projections; all projection matrices and $\mathrm{MLP}_g$ are *trainable* and small.

### Cross-modal anatomical alignment (CMAA)

**Motivation**. We align tissue-level semantics across modalities to achieve robust transfer. We consider two supervision sources: (a) *image–report* text pairs via a frozen text encoder $\Psi$ (if available), and (b) *region-type* labels or pseudo-labels shared across modalities.

**Region-level contrast**. Let $E^{(m)}$ be region embeddings (Eq. 3). We project them to a contrastive space with $h_\phi : \mathbb{R}^d \to \mathbb{R}^{d_c}$, and $\ell_2$-normalize:

$$r_k^{(m)} = \frac{h_\phi\left(E_k^{(m)}\right)}{\left\| h_\phi\left(E_k^{(m)}\right)\right\|_2} \in \mathbb{R}^{d_c}. \tag{11}$$

For a minibatch $\mathcal{B}$ and temperature $\tau > 0$, we maximize agreement between the same region type across modalities (positives) and push away others (negatives):

$$\mathcal{L}_{\mathrm{RCL}} = -\sum_{(i,k)\in\mathcal{B}} \log \frac{\sum_{(j,k^+)\in\mathcal{P}(i,k)} \exp\left(\langle r_k^{(m_i)}, r_{k^+}^{(m_j)}\rangle/\tau\right)}{\sum_{(j,t)\in\mathcal{N}(i,k)} \exp\left(\langle r_k^{(m_i)}, r_t^{(m_j)}\rangle/\tau\right)}, \tag{12}$$

where $\mathcal{P}(i,k)$ collects positives that share the same anatomy type as $(i,k)$ but come from other modalities or views; $\mathcal{N}(i,k)$ includes all negatives. Region types can be derived from atlas tags or by clustering $E^{(m)}$ with EMA-updated centroids.

**Text-grounded alignment (optional)**. When reports are available, we encode anatomy phrases (e.g., "ductal tissue," "glandular density") via $\Psi$,

obtain text vectors $t_c$, and add a CLIP-style objective:

$$\mathcal{L}_{\mathrm{TXT}} = -\sum_{(i,k)} \log \frac{\exp\left(\langle r_k^{(m_i)}, t_{c(i,k)}\rangle/\tau_t\right)}{\sum_{c'} \exp\left(\langle r_k^{(m_i)}, t_{c'}\rangle/\tau_t\right)}. \quad (13)$$

### Task heads and losses

**Classification**. We apply a linear head on $z_{\mathrm{cls}}^{(L)}$ for image-level diagnosis with cross-entropy:

$$\mathcal{L}_{\mathrm{CLS}} = -\sum_{c=1}^{C} y_c \log \widehat{y}_c,\ \widehat{y} = \mathrm{softmax}(\mathrm{W}_{\mathrm{cls}} z_{\mathrm{cls}}^{(\mathrm{L})}), \quad (14)$$

where $y$ is the one-hot label.

**Segmentation**. For pixel/patch-level tasks, we upsample token features and use a light decoder (frozen or tiny trainable) to predict masks $\widehat{M}$. We combine Dice and BCE:

$$\mathcal{L}_{\mathrm{SEG}} = 1 - \frac{2\langle \widehat{M}, M\rangle + \epsilon}{\| \widehat{M}\|_1 + \| M\|_1 + \epsilon} + \beta\, \mathrm{BCE}(\widehat{M}, M). \quad (15)$$

**Topology-aware prompt smoothing**. Let $P^{(\ell)} \in \mathbb{R}^{K\times d}$ and $\mathbf{L}$ be the graph Laplacian over regions (adjacent tissues connect). We regularize:

$$\mathcal{L}_{\mathrm{LAP}} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{tr}\left((\mathrm{P}^{(\ell)})^{\top} \mathbf{L}\, \mathrm{P}^{(\ell)}\right). \quad (16)$$

Overall objective: The final loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CLS}} + \lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{SEG}} + \lambda_{\mathrm{rcl}}\mathcal{L}_{\mathrm{RCL}} + \lambda_{\mathrm{txt}}\mathcal{L}_{\mathrm{TXT}} + \lambda_{\mathrm{lap}}\mathcal{L}_{\mathrm{LAP}}. \quad (17)$$

We set λ's via validation; $\mathcal{L}_{\mathrm{TXT}}$ is used only when reports exist.

### Training protocol and parameter efficiency

**Frozen backbone**. All ViT/Swin blocks remain frozen. Trainable modules include: (i) Anatomy MLP $g_\theta$, (ii) prompt maps $\{W_p^{(\ell)}, b^{(\ell)}\}$ with low-rank factorization, (iii) PIM projections $\{W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)}, W_O^{(\ell)}, \bar{W}_Q^{(\ell)}, \bar{W}_K^{(\ell)}, \bar{W}_V^{(\ell)}, \bar{W}_O^{(\ell)}\}$ and gate MLP, (iv) small task heads, and (v) contrastive projector $h_\phi$. This amounts to <2% trainable parameters of full fine-tuning in our experiments.

**Optimization**. We use AdamW with cosine decay, warm-up, and gradient clipping. To stabilize contrastive training, we maintain modality-specific memory queues of region embeddings for hard negatives.

### Inference and interpretability

At test time, we run ARE to obtain $E^{(m)}$, synthesize $\{P^{(\ell)}\}$, and perform layer-wise PIM updates. For interpretability, we compute *prompt-guided attention rollout*. Let $\mathbf{A}_{p\to t}^{(\ell)}$ be prompt → token attention (averaged across heads). We define a normalized relevance map:

$$\mathbf{R} = \mathrm{Norm}\left(\prod_{=1}^{\mathrm{L}} (\mathbf{I} + \gamma\, \mathbf{A}_{\mathrm{p}\to\mathrm{t}}^{(\ell)})\right), \quad (18)$$

where $\gamma$ scales prompt influence. Upsampling $\mathbf{R}$ to image space yields anatomy-aligned saliency that radiologists can verify.

### Complexity analysis

**Time/Memory**. Compared to vanilla VPT (concatenate $K$ prompts), our PIM adds two cross-attention blocks per layer with $O(KN)$ attention cost (small since $K \ll N$). Low-rank maps and stage-sharing keep parameters linear in $d\,r$ with $r \ll d$.

### Why anatomy-guided prompts work

**Inductive bias**. Prompts are *context tokens* that reshape attention. By tying them to region embeddings (Eq. 4) and aligning across modalities (Eq. 12), we enforce a stable anatomical basis that persists across imaging physics. Gated fusion (Eq. 10) prevents prompt overreach, preserving backbone priors while enabling targeted adaptation.

### Ethics approval and consent to participate

All procedures involving human data were conducted in accordance with the Declaration of Helsinki. This study exclusively analyzed publicly available, anonymized datasets (INbreast, BUSI, and Duke-Breast-MRI), which were collected with prior approval from the respective institutional review boards and with informed consent obtained by the original investigators. Therefore, no additional ethical approval or informed consent was required for the present study.

## Data availability

The datasets analyzed in this study are publicly available: INbreast, BUSI, and Duke-Breast-MRI (ScientificData, 2022).

## Materials availability

This research did not generate new physical materials or custom reagents. All software environments and pre-trained model checkpoints used in this work are described within the Methods section.

## Code availability

The code supporting the findings of this study is not publicly available at the moment. However, it can be obtained from the corresponding author upon reasonable request after the paper is accepted.

Received: 27 October 2025; Accepted: 29 January 2026;
Published online: 13 February 2026

## References

1. Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In *Proc. International Conference on Learning Representations* https://openreview.net/forum?id=YicbFdNTTy (2021).
3. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision* 10012–10022 (IEEE, 2021).
4. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
5. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
6. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proc. Conf Empir Methods Nat Lang Process*. **2022**, 3876–3887 (2022).
7. Xiao, X. et al. Describe anything in medical images. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2505.05804 (2025).
8. Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 6999–7019 (2021).
9. Nerella, S. et al. Transformers and large language models in healthcare: a review. *Artif. Intell. Med.* **154**, 102900 (2024).
10. Jia, C. et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. International Conference on Machine Learning* 4904–4916 (PMLR, 2021).
11. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15016–15027 (IEEE, 2023).

12. Lin, H., Xu, C. & Qin, J. Taming vision-language models for medical image analysis: a comprehensive review. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2506.18378 (2025).
13. Ding, N. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **5**, 220–235 (2023).
14. Jia, M. et al. Visual prompt tuning. In *Proc. European Conference on Computer Vision* 709–727 (Springer, 2022).
15. Xiao, X. et al. Prompt-based adaptation in large-scale vision models: a survey. Preprint at https://doi.org/10.48550/arXiv.2510.13219 (2025).
16. Xiao, X. Visual Instance-aware Prompt Tuning. In *Proc. of the 33rd ACM International Conference on Multimedia (MM '25)*. Association for Computing Machinery, 2880–2889 (New York, NY, USA, 2025).
17. Moreira, I. C. et al. Inbreast: toward a full-field digital mammographic database. *Acad. Radiol.* **19**, 236–248 (2012).
18. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data brief.* **28**, 104863 (2020).
19. Saha, A. et al. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. https://doi.org/10.7937/TCIA.e3sv-re93 (2021).
20. Saha, A. et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *Br. J. Cancer* **119**, 508–516 (2018).
21. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
22. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
23. Shao, Z. et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In Beygelzimer, A., Dauphin, Y., Liang, P.& Wortman Vaughan, J. (Eds.), *Advances in Neural Information Processing Systems*, (2021).
24. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations* (ICLR, 2022).
25. Houlsby, N. et al. Parameter-efficient transfer learning for NLP. In *Proc. International Conference on Machine Learning* 2790–2799 (PMLR, 2019).
26. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
27. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention* 234–241 (Springer, 2015).
28. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: a nested U-Net architecture for medical image segmentation. In *Proc. International Workshop on Deep Learning in Medical Image Analysis* 3–11 (Springer, 2018).
29. Chen, J. et al. TransUNet: transformers make strong encoders for medical image segmentation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2102.04306 (2021).
30. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proc. European Conference on Computer Vision* 205–218 (Springer, 2022).
31. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M. & Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. ArXiv, abs/2105.15203 (2021).
32. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
33. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M. Medical transformer: Gated axial-attention for medical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention* 36–46 (Springer, 2021).



## Acknowledgements

None. This work was supported by the National Natural Science Fund [grant number 82203407, 82403769, 82403981 and 82503455], Tianjin Key Medical Discipline Construction Project [Grant No.TJYXZDXK-3-003A], Tianjin Municipal Science and Technology Bureau (25JCLMJC00710), Tianjin Binhai New Area Health Research Project [Grant No.2023BWKQ017], the Shanxi Province Basic Research Program in Natural Sciences [Grant no. 2025JC-YBQN-1045] and Chongqing Natural Science Foundation [CSTB2024NSCQ-MSX0761].


## Author contributions

S.Z., Q.M., and Y.H. contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (validation, formal analysis). S.Z., X.X., and J.Z. conceptualized the study, designed the methodology, and participated in securing research funding (conceptualization, methodology, funding acquisition). Q.M., J.Q., and C.W. carried out data acquisition, curation, and investigation (investigation, data curation) and provided key resources, instruments, and technical support (resources, software). Y.H. and Y.H. drafted the initial manuscript and generated visualizations (writing—original draft, visualization). J.D., T.P., and J.L. supervised the project, coordinated collaborations, and ensured administrative support (supervision, project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (writing—review and editing) and approved the final version for submission.

## Competing interests

The authors declare no competing interests.

## Consent to publish

All authors have reviewed and approved the final version of this manuscript and consent to its publication.

## Additional information


**Correspondence** and requests for materials should be addressed to Jinhai Deng, Teng Pan or Jingjing Liu.