Dual-Res Tandem Mamba-3D: Bilateral Breast Lesion Detection and Classification on Non-contrast Chest CT

Jiaheng Zhou[♥♦♠§*] **Wei Fang**[♠♣Ψ*] **Luyuan Xie**[Θ♠] **Yanfeng Zhou**[Ω]
Lianyan Xu[Π] **Minfeng Xu**[♠♣] **Ge Yang**[♥♦†] **Yuxing Tang**[♠†]

[♥]Institute of Automation, Chinese Academy of Sciences
[♦]School of Artificial Intelligence, University of Chinese Academy of Sciences
[♠]DAMO Academy, Alibaba Group [♣]Hupan Laboratory, Hangzhou, China
[Ψ]Zhejiang University [Θ]Peking University
[Ω]Shenzhen University [Π]ZhongShan Hospital, Fudan University
[§]Work done during an internship at the Medical AI Lab, DAMO Academy, Alibaba Group.
[*]Equal contribution. [†]Corresponding authors.
{zhoujiaheng2022, ge.yang}@ia.ac.cn {lucas.fw, yuxing.t}@alibaba-inc.com

Abstract

Breast cancer remains a leading cause of death among women, with early detection significantly improving prognosis. Non-contrast computed tomography (NCCT) scans of the chest, routinely acquired for thoracic assessments, often capture the breast region incidentally, presenting an underexplored opportunity for opportunistic breast lesion detection without additional imaging cost or radiation. However, the subtle appearance of lesions in NCCT and the difficulty of jointly modeling lesion detection and malignancy classification pose unique challenges. In this work, we propose **D**ual-**R**es **T**andem **M**amba-**3D** (**DRT-M3D**), a novel multitask framework for opportunistic breast cancer analysis on NCCT scans. DRT-M3D introduces a dual-resolution architecture, which captures fine-grained spatial details for segmentation-based lesion detection and global contextual features for breast-level cancer classification. It further incorporates a tandem input mechanism that models bilateral breast regions jointly through Mamba-3D blocks, enabling cross-breast feature interaction by leveraging subtle asymmetries between the two sides. Our approach achieves state-of-the-art performance in both tasks across multi-institutional NCCT datasets spanning four medical centers. Extensive experiments and ablation studies validate the effectiveness of each key component.

1 Introduction

Breast cancer remains one of the most prevalent and deadly diseases among women worldwide [8]. Early detection is critical for improving survival outcomes [74]. While screening techniques such as mammography, ultrasound, and magnetic resonance imaging (MRI) are well established, they require dedicated protocols and are not routinely performed during thoracic exams for unrelated conditions [13]. However, non-contrast chest computed tomography (chest NCCT) is widely available and frequently conducted for pulmonary evaluation, which naturally includes the breast region [61]. As illustrated in Fig. 1, this creates a valuable opportunity for breast cancer assessment without additional imaging cost or radiation exposure.

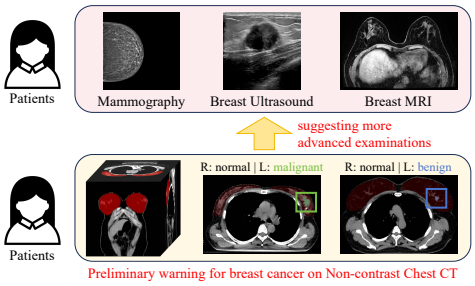


Figure 1: Non-contrast chest CTs enable opportunistic breast cancer analysis without additional cost. Leveraging bilateral information leads to more reliable detection and classification.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Despite this potential, breast cancer analysis on chest NCCT poses several technical challenges. First, the absence of contrast enhancement in NCCT hampers lesion conspicuity against surrounding tissue. Second, lesion detection and malignancy classification serve complementary clinical purposes: the former focuses on localization, the latter supports diagnosis. Yet they are often treated as separate tasks, limiting the opportunity to share relevant features. Third, most existing approaches process each breast separately, overlooking the bilateral context that radiologists routinely use to detect asymmetries and subtle lesions. These limitations underscore the need for a unified framework that jointly models both tasks and captures long-range interactions between bilateral breast regions.

However, building such a unified framework remains challenging [60]. Segmentation architectures like UNet [51] and its variants rely on convolutional backbones with limited capacity for global context [30, 52], while vision transformers (ViTs) [1, 14] enable long-range modeling but suffer from high computational cost or low spatial resolution [21], compromising fine-grained lesion localization/segmentation. A promising direction lies in selective state space models (SSMs), such as Mamba [10,17], which offer linear computational complexity with respect to sequence length, making it possible to use smaller patch sizes while preserving the ability to model long-range dependencies.

In this work, we propose Dual-Res Tandem Mamba-3D (DRT-M3D), a multitask framework that jointly performs segmentation-based lesion detection [1] and breast-level malignancy classification. Based on Mamba (S6) [10, 17], a selective state space model with linear complexity, DRT-M3D efficiently captures long-range dependencies across 3D volumes while preserving fine spatial granularity, which is crucial for detecting subtle lesions.

DRT-M3D introduces a dual-resolution architecture that separates voxel-level segmentation and breast-level classification into high-resolution and low-resolution paths, respectively. Mutual fusion modules align information across these two paths to improve representation quality without task interference. To model inter-breast context, we introduce a tandem input mechanism that processes bilateral breast volumes jointly, enabling cross-side interaction and the learning of asymmetry-aware representations, which are critical for detecting subtle signs of malignancy.

Experiments on three internal datasets and one external dataset demonstrate that DRT-M3D enables effective breast cancer analysis on chest NCCT scans, achieving state-of-the-art performance in both breast lesion segmentation and cancer classification tasks. Extensive ablation studies verify the effectiveness of each design choice, and we demonstrate that the tandem input strategy can also benefit vision transformer models when extended appropriately.

We summarize our contributions as follows:

- We propose a unified multi-task framework, DRT-M3D, for segmentation-based lesion detection and breast-level classification, enabling opportunistic breast cancer analysis on non-contrast chest CT scans.
- We design a dual-resolution architecture with mutual fusion to balance fine-grained spatial detail and global semantic context.
- We introduce a tandem input mechanism for bilateral modeling, enabling the network to leverage cross-side context as a structural prior.
- Evaluation on multi-institutional datasets shows that our model surpasses competitive baselines on both internal and external cohorts, demonstrating strong potential for clinical application.

2 Related Works

2.1 Breast Lesion Analysis in Medical Imaging

Breast lesion analysis has been widely studied across dedicated imaging modalities such as mammography [35, 53, 54], ultrasound [7, 23, 27, 65, 66], MRI [26, 71], and multi-modal scenarios [15, 28, 50], with deep learning models addressing tasks including lesion detection, segmentation, and malignancy classification. In contrast, research on breast cancer using non-contrast chest CT (NCCT) scans remains relatively underexplored [29, 33, 56]. U-Net variants dominate segmentation tasks

[1] Throughout this paper, *detection* refers to *segmentation-based lesion detection*, *i.e.*, identifying lesions through voxel-wise segmentation maps. We use *segmentation* and *detection* interchangeably in this context.

across modalities [3, 37, 68], while classification is typically performed via feature aggregation from segmentation backbones [39, 72] or using standalone CNN/ViT-based classifiers [16, 47].

Bilateral comparison is a routine practice in clinical mammography, and learning-based methods have explored dual-view fusion and symmetry-aware modeling [49, 67]. A few methods incorporate handcrafted symmetry features or pairwise comparisons [5, 12]. However, most approaches operate on 2D images and lack end-to-end bilateral modeling. In NCCT, where lesion visibility is often subtle due to low contrast, bilateral context is especially valuable yet remains underutilized, forming one of the key motivations of this study.

2.2 State Space Models for 3D Visual Data

State space models (SSM), such as structured state space sequence models (S4) [18–20], have emerged as linear-complexity alternatives to Transformers [59], alleviating the quadratic cost of attention for long-sequence processing. Mamba (S6) [10, 17] extends S4 with a selective mechanism, and has proven effective in capturing long-range dependencies in visual data, making it a compelling alternative to Vision Transformers [14], especially in 3D tasks. VideoMamba [36] extends bi-directional Mamba layers from 2D [73] to 3D (video) data. VMamba [44] introduces a tailored SSM scanning strategy for 2D images, while Mamba-ND [38] generalizes into Mamba-3D blocks, simplifying volumetric modeling without complicating the internal SSMs. E-ViM³ [70] further demonstrates that pre-training Mamba-3D as a masked autoencoder (MAE) [22, 58, 63] boosts downstream task performance, echoing MAE's success in ViTs [14].

In medical image analysis, U-Mamba [46] integrates Mamba into UNet [51] backbone, though it demonstrates no clear benefits over conventional segmentation models [31]. SegMamba [64] introduces tri-orientated Mamba modules to better capture 3D context. Swin-UMamba [42, 43] leverages VMamba [44] backbones pre-trained on ImageNet-1K [11] to improve performance on medical images, but remains limited to 2D analysis. EM-Net [6] combines Mamba with frequency-domain learning for multi-scale 3D features. More recently, Tri-Plane Mamba [62] incorporates SSMs into the Segment Anything Model (SAM) [32] for efficient interactive 3D segmentation.

Despite recent progress, existing methods are typically single-task, focusing solely on segmentation, without classification or bilateral context. In contrast, we integrate multi-task learning and bilateral-aware modeling into Mamba-3D, yielding a unified framework for breast cancer analysis on NCCT.

3 Methods

3.1 Overall Pipeline

Our pipeline for opportunistic breast cancer analysis on non-contrast chest CT (NCCT) scans is illustrated in Fig. 2. It takes the full NCCT 3D volume as input and produces both voxel-wise breast lesion segmentation and breast-level malignancy classification. The segmentation-based detection result highlights both benign and malignant lesions, guiding clinicians to focus on relevant areas. The classification result automatically identifies potentially malignant regions.

To focus on the breast regions within the entire chest CT, we introduce a Pre-Stage using a pre-trained segmentation network to perform coarse localization and cropping of each breast. The Main-Stage performs core analysis on the cropped breast regions, with all methods using consistent pre-processing and post-processing for fair evaluation. Implementation details are provided in Sec. C.

3.2 Dual-Res Mamba-3D

We first introduce the Dual-Res Mamba-3D (DR-M3D) network, which addresses the two interrelated yet distinct tasks: segmentation-based detection of breast lesions and malignancy classification on NCCT scans. DR-M3D operates independently on each breast region and also serves as a baseline for our study, in contrast to our full Dual-Res Tandem Mamba-3D (DRT-M3D) model.

3D data embedding Given a 3D CT image $\boldsymbol{X} \in \mathbb{R}^{1\times \mathrm{D}\times \mathrm{H}\times \mathrm{W}}$ (with 1 for one channel in grayscale CT, and $\mathrm{D}, \mathrm{H}, \mathrm{W}$ for depth, height, and width), we apply a 3D embedding layer to transforms $\boldsymbol{X}$ into the embedded feature map $\boldsymbol{Y} \in \mathbb{R}^{\mathrm{C}\times \mathrm{D_P}\times \mathrm{H_P}\times \mathrm{W_P}}$. Each token in $\boldsymbol{Y}$ with embedding dimension C

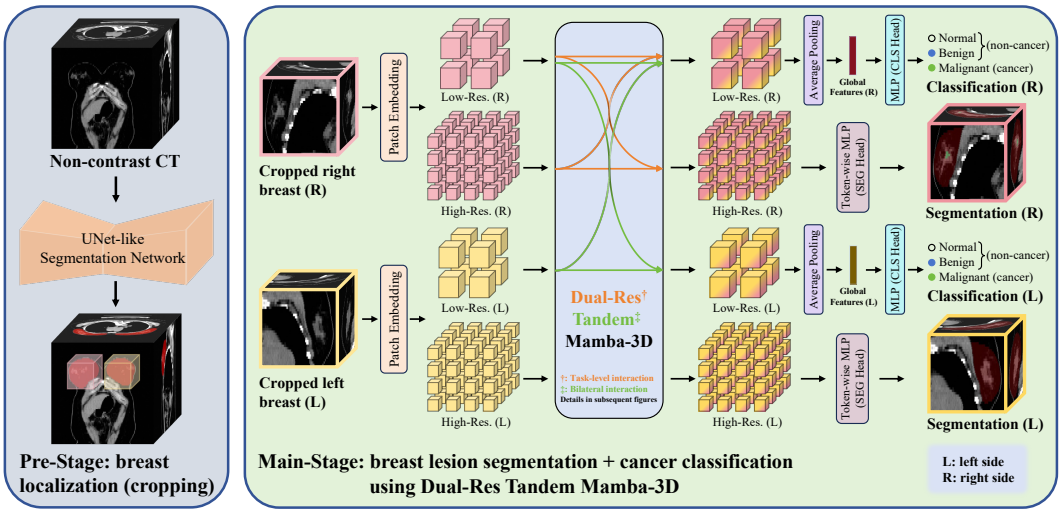


Figure 2: **The overall pipeline of the opportunistic breast cancer analysis approach.** The Pre-Stage employs a pre-trained segmentation network to perform coarse-grained localization and cropping of each breast region. The Main-Stage consists of the proposed Dual-Res Tandem Mamba-3D (DRT-M3D) network, which jointly performs segmentation-based detection of breast lesions and malignancy classification by leveraging dual-resolution features (intuitively represented by orange arrows) and facilitates cross-breast feature interaction between bilateral breast regions (green arrows). Please refer to Sec. 3.2, Sec. 3.3 and Fig. 3 for more details.

corresponds to a $p_d \times p_h \times p_w$ patch in the original image, where $p_d = \frac{D}{D_p}$, $p_h = \frac{H}{H_p}$, and $p_w = \frac{W}{W_p}$. We independently apply amplitude-learnable sinusoidal embeddings along each dimension and add them to equally split segments of the embedding space to help preserve spatial structure.

Dual-resolution design with mutual fusion To support both fine-grained lesion segmentation and breast-level malignancy classification, DR-M3D adopts a dual-resolution architecture. Specifically, the embedded feature map $\boldsymbol{Y} \in \mathbb{R}^{C \times D_p \times H_p \times W_p}$ is further downsampled via 3D MaxPooling to generate a lower-resolution feature map $\boldsymbol{Z} \in \mathbb{R}^{C' \times D'_p \times H'_p \times W'_p}$. Here, C' denotes the increased embedding dimension for a more powerful representation for the low-resolution path, and the downsampling factors are defined as $p'_d = \frac{D_p}{D'_p}$, $p'_h = \frac{H_p}{H'_p}$, and $p'_w = \frac{W_p}{W'_p}$. Thus, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are used as the initial inputs to the high-resolution (HR) and low-resolution (LR) paths, respectively, denoted as $\boldsymbol{Y}^0$ and $\boldsymbol{Z}^0$ in the following text.

As the DRT-M3D block shown in Fig. 3 (a), each DR-M3D block also contains an HR and an LR Mamba-3D sub-block, forming a dual-path structure. From a black-box view, the i-th DR-M3D block defines the mapping $(\boldsymbol{Y}^{i+1}, \boldsymbol{Z}^{i+1}) = \text{DR-M3D}^i(\boldsymbol{Y}^i, \boldsymbol{Z}^i)$ while preserving the spatial resolutions of both paths. The HR path retains spatial details for voxel-level segmentation, whereas the LR path, operating on shorter sequences, captures long-range dependencies crucial for classification.

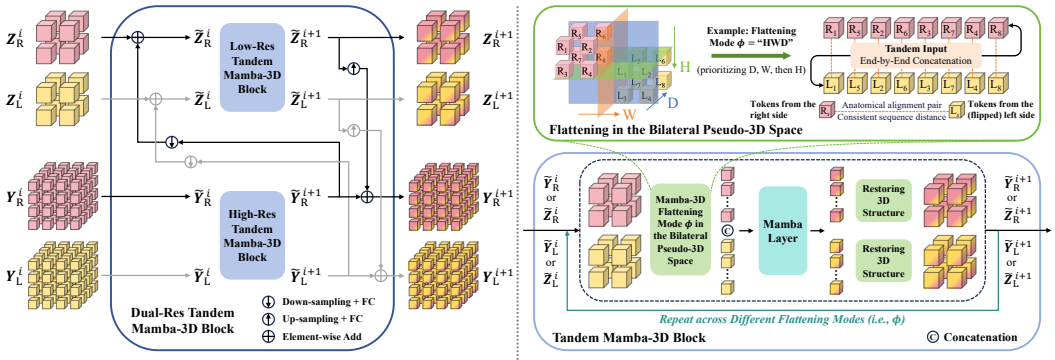
To facilitate cross-path information exchange, each DR-M3D block integrates a mutual fusion mechanism. Specifically, after the HR Mamba-3D sub-block processes $\boldsymbol{Y}^i$, the output $\widetilde{\boldsymbol{Y}}^{i+1}$ is downsampled and projected to match the LR resolution and channel dimension, allowing residual fusion with the LR path. Conversely, the LR output $\widetilde{\boldsymbol{Z}}^{i+1}$ is upsampled and projected to match the HR shape before being added to the input of the next HR block.

Formally, the internal update process within the i-th DR-M3D block can be expressed as:

$$\begin{cases} \widetilde{\boldsymbol{Y}}^i = \boldsymbol{Y}^i, & \widetilde{\boldsymbol{Z}}^i = \boldsymbol{Z}^i + f^i_{\text{down}}(\widetilde{\boldsymbol{Y}}^{i+1}) \\ \boldsymbol{Y}^{i+1} = \widetilde{\boldsymbol{Y}}^{i+1} + f^i_{\text{up}}(\widetilde{\boldsymbol{Z}}^{i+1}), & \boldsymbol{Z}^{i+1} = \widetilde{\boldsymbol{Z}}^{i+1} \end{cases} \tag{1}$$

where

$$\begin{cases} \widetilde{\boldsymbol{Y}}^{i+1} = \text{Mamba-3D}(\widetilde{\boldsymbol{Y}}^i) \\ \widetilde{\boldsymbol{Z}}^{i+1} = \text{Mamba-3D}(\widetilde{\boldsymbol{Z}}^i) \end{cases} \tag{2}$$



(a) Dual-Res Tandem Mamba-3D Block with Mutual Fusion

(b) Tandem Input of Mamba-3D Blocks

Figure 3: **Dual-Res Tandem Mamba-3D block.** (a) The high-resolution (HR) path captures fine-grained local features, while the low-resolution (LR) path focuses on global context. The upsampled LR output is fused back to the HR path, facilitating joint feature alignment. (b) The Tandem Mamba-3D block flattens both 3D inputs to 1D with specific dimension-priority flattening modes and concatenates them to form the Tandem Input. This prevents mixing of bilateral information while enabling cross-side feature interaction during selective scanning.

and the downsampling and upsampling transformations are defined as:

$$\begin{cases} f^i_{\text{down}}(\widetilde{\boldsymbol{Y}}^{i+1}) = \text{Linear}_{\text{C}\to\text{C}'}[\text{3D-MaxPooling}(\widetilde{\boldsymbol{Y}}^{i+1})] \\ f^i_{\text{up}}(\widetilde{\boldsymbol{Z}}^{i+1}) = \text{Linear}_{\text{C}'\to\text{C}}[\text{3D-Interpolation}(\widetilde{\boldsymbol{Z}}^{i+1})] \end{cases} \tag{3}$$

Our dual-resolution architecture assigns tasks to different paths, while residual coupling via mutual fusion encourages joint feature alignment. This setup can be seen as a form of soft parameter sharing [9] under a multi-task optimization framework, also emulating how radiologists integrate local and global cues in clinical reading.

3.3 Tandem Input for Mamba-3D Blocks

Building on the DR-M3D structure introduced in Sec. 3.2, we further seek to exploit the inherent correlation between bilateral breasts, aiming to mitigate the limited feature richness of NCCT images. To achieve this, we propose the tandem input mechanism as illustrated in Fig. 3 (b), which transforms each sub-block into a Tandem Mamba-3D block, forming the complete DRT-M3D architecture.

Specifically, we organize the tokens from the left and right breasts into a bilateral pseudo-3D space. The right breast occupies the coordinate range from $(1, 1, 1)$ to $(\text{D}_\text{R}, \text{H}_\text{R}, \text{W}_\text{R})$, while the left is positioned from $(\text{D}_\text{R} + 1, \text{H}_\text{R} + 1, \text{W}_\text{R} + 1)$ to $(\text{D}_\text{R} + \text{D}_\text{L}, \text{H}_\text{R} + \text{H}_\text{L}, \text{W}_\text{R} + \text{W}_\text{L})$. For anatomical alignment, the left breast is pre-flipped along the X-axis, so both regions share the same orientation and are juxtaposed within the extended space, without assuming direct connectivity along any particular axis. By flattening the pseudo-3D volume while skipping unoccupied regions, we obtain the tandem input sequence as shown in Fig. 3 (b), which is equivalent to flattening the left and right breast regions separately with the same mode and then concatenate them end-to-end. The internal different 3D flattening modes and scanning processes follow the Mamba-3D strategy [38, 70], as detailed in Sec. A and illustrated in Fig. 7.

The tandem input design not only ensures that Mamba-3D scans all voxels from one breast before moving to the other, but also guarantees a roughly fixed distance between anatomically symmetric locations in the sequence, allowing the model to better capture cross-breast relationships. The LR path further enhances cross-side modeling by offering a shorter sequence and higher embedding dimensionality. Moreover, the prioritized scanning strategy across different axes aligns with the multi-orientation analysis and comprehensive assessment performed in real clinical evaluation.

3.4 Training Procedure

Efficient self-supervised pre-training We adopt the efficient self-supervised pre-training strategy for Mamba-3D blocks proposed in [70], which improves downstream performance and accelerates convergence without extra data. We set the size of jointly masked regions in the HR path to match the downsampling factors $(\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w})$ from HR to LR, with a masking ratio p_{mask}. This ensures

that each LR token is computed only from unmasked HR tokens, avoiding information leakage and allowing for the removal of masked tokens during pre-training.

To construct the masked autoencoder [22, 58, 70], we append M additional DRT-M3D blocks as the decoder following the backbone network. The pretext task is to reconstruct the masked tokens from both the HR and LR outputs via token-to-patch linear projections. The pretraining objective is optimized using a mean squared error (MSE) loss, formulated as:

$$\mathcal{L}_{\mathrm{PRE}} = \frac{1}{|\mathrm{V}_{\mathrm{masked}}|} \sum_{j=1}^{|\mathrm{V}_{\mathrm{masked}}|} \left[\left(\left[\boldsymbol{X}_{\text{Y-pred}} \right]_j - \left[\boldsymbol{X} \right]_j \right)^2 + \left(\left[\boldsymbol{X}_{\text{Z-pred}} \right]_j - \left[\boldsymbol{X} \right]_j \right)^2 \right] \tag{4}$$

where $\mathrm{V}_{\mathrm{masked}}$ denotes the set of masked voxels.

Joint segmentation and classification fine-tuning We fine-tune the network jointly for voxel-wise lesion segmentation and breast-level malignancy classification, while the network structure enforces partial task decoupling and soft feature interaction through the dual-resolution paths.

For the segmentation task, supervision is applied to the HR path's output using a hybrid loss composed of the Dice loss [48] and voxel-level cross-entropy (CE) loss:

$$\mathcal{L}_{\mathrm{SEG}} = \mathrm{Dice}(\boldsymbol{M}_{\text{Y-pred}}, \boldsymbol{M}_{\mathrm{gt}}) + \frac{1}{|\mathrm{V}|} \sum_{j=1}^{|\mathrm{V}|} \mathrm{CE}\left(\left[\boldsymbol{M}_{\text{Y-pred}} \right]_j - \left[\boldsymbol{M}_{\mathrm{gt}} \right]_j \right)^2 \tag{5}$$

where $\boldsymbol{M}_{\text{Y-pred}}$ is the segmentation prediction obtained from the HR output $\boldsymbol{Y}^{N-1}$ via an MLP, $\boldsymbol{M}_{\mathrm{gt}}$ is the ground truth mask, and V denotes the set of all voxels in the cropped 3D-image.

For the classification task, supervision is applied to the LR path's output using standard image-level cross-entropy loss:

$$\mathcal{L}_{\mathrm{CLS}} = \mathrm{CE}(\boldsymbol{y}_{\text{Z-pred}}, \boldsymbol{y}_{\mathrm{gt}}) \tag{6}$$

where $\boldsymbol{y}_{\text{Z-pred}}$ is the predicted malignancy score obtained by applying global average pooling and an MLP to the LR output $\boldsymbol{Z}^{N-1}$, and $\boldsymbol{y}_{\mathrm{gt}}$ is the ground-truth breast-level malignancy label.

The overall fine-tuning loss is the weighted sum of the two task losses:

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{SEG}} + \lambda \mathcal{L}_{\mathrm{CLS}} \tag{7}$$

where λ balances the relative importance of the two tasks. We empirically set $\lambda = 0.1$, reflecting the relative simplicity and higher overfitting risk of the classification task.

4 Experiments

4.1 Experimental Setup

Datasets Following the medical image analysis paradigm, we use both internal and external datasets. The internal datasets include training and testing splits, while the external set is reserved for testing only, providing a distribution shift that enables a rigorous assessment of generalization.

The internal data comprises three cohorts collected from separate institutions, referred to as Inst. 1-3, with training and testing splits contain {341 / 315, 239 / 921, 141 / 113} and {82 / 78, 102 / 296, 34 / 28} *cancerous* / *non-cancerous* breast cases, respectively. The external dataset contains 214 breast samples as {105 / 109} *cancerous* / *non-cancerous*.

For all datasets in the Main-Stage, bilateral breasts from the same patient are always assigned to the same split to avoid data leakage. Each sample consists of a cropped NCCT image of the breast region, segmentation masks for the breast and lesions (if present), and a binary cancer label. More details regarding the datasets and the Pre-Stage process can be found in Sec. B.

Implementation Our model is implemented in PyTorch and trained on up to two NVIDIA A100 (80GB) GPUs. The training process consists of 500 epochs for self-supervised pre-training and 50 epochs for downstream fine-tuning, optimized using AdamW [45]. The base learning rate is

Table 1: **Quantitative results of the internal evaluation across three datasets.** The best and second-best results are **bolded** and underlined, respectively. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Method	Inst. 1 (Spec. = 0.9615)					Inst. 2 (Spec. = 0.9595)					Inst. 3 (Spec. = 0.9643)				
	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	0.6174	**0.9390**	0.8338	0.8049	0.9558	0.6413	0.9099	0.8806	0.7647	0.9501	0.5980	0.8611	0.8264	0.7353	0.9170
nnUNet-SEG	0.6231	0.9268	0.8760	-	-	0.6548	0.9099	0.8694	-	-	0.5862	0.8333	0.8194	-	-
nnUNet-CLS	-	-	-	0.7927	0.9622	-	-	-	0.7549	0.9540	-	-	-	0.7353	0.9275
VNet [48]	0.5623	0.9146	0.8333	0.7683	0.9500	0.5821	0.8649	0.8198	0.7451	0.9418	0.5116	0.8056	0.7431	0.7353	0.8797
swinUNETR [24]	0.5647	0.9146	0.8043	0.8171	0.9432	0.6009	0.9009	0.8423	0.7353	0.9423	0.5465	0.8333	0.7764	0.7059	0.8981
nnFormer [69]	0.5601	0.9268	0.8130	0.6707	0.9595	0.6122	0.9099	0.8709	0.7549	0.9444	0.5298	0.8333	0.7709	0.7353	0.9307
3D UX-Net [34]	0.5669	0.9268	0.7811	0.7683	0.9361	0.6015	0.9279	0.8288	0.7353	0.9382	0.5213	0.8333	0.7327	0.7647	0.9391
MedNeXt [52]	0.6175	0.9146	0.8719	0.7683	0.9580	0.6292	**0.9369**	0.8919	0.6863	0.9510	0.5604	0.8333	0.7848	0.7647	0.9076
U-Mamba-Bot [46]	0.6042	0.8659	0.8552	0.8049	0.9578	0.5989	0.7838	0.7928	0.7451	0.9596	0.5361	0.7500	0.7500	0.7941	0.9391
U-Mamba-Enc [46]	0.5708	0.8171	0.8171	0.7805	0.9548	0.5671	0.7838	0.7928	0.7451	0.9561	0.4901	0.6944	0.6944	0.7059	0.8792
EM-Net [6]	0.5803	0.9268	0.8059	0.7439	0.9467	0.6203	0.9189	0.8587	0.7549	0.9540	0.5461	0.8056	0.7361	0.6765	0.9160
SegMamba [64]	0.5852	0.9146	0.8509	0.7805	0.9555	0.6250	0.8919	0.8694	0.8039	0.9479	0.5109	0.7500	0.7152	0.7353	0.9074
Sun et al. 2025 [56]	0.6061	0.9268	0.8613	0.7561	0.9457	0.6317	0.9099	0.8649	0.7647	0.9437	0.5859	0.8889	0.7778	0.7059	0.9144
DR-M3D (ours)	0.6553 ±.0083	0.9309 ±.0070	0.9045 ±.0064	0.8780 ±.0122	0.9725 ±.0063	0.6744 ±.0075	0.9339 ±.0104	**0.9234** ±.0181	0.7549 ±.0098	0.9686 ±.0016	0.6003 ±.0078	0.8704 ±.0161	0.8264 ±.0184	0.7505 ±.0245	0.9226 ±.0097
DRT-M3D (ours)	**0.6608** ±.0018	0.9349 ±.0070	**0.9258** ±.0063	**0.9471** ±.0186	**0.9909** ±.0015	**0.6750** ±.0022	0.9279 ±.0090	0.9212 ±.0060	**0.8366** ±.0299	**0.9743** ±.0020	**0.6056** ±.0148	**0.8796** ±.0161	**0.8518** ±.0080	**0.8726** ±.0170	**0.9471** ±.0044

set to $1e{-3}$, with a linear warm-up phase followed by cosine annealing. Detailed experimental configurations and hyper-parameters are provided in Sec. C.

For baselines not compatible with MAE pre-training (mainly CNN-based models), we use supervised pre-training on segmentation followed by joint fine-tuning on segmentation and classification, which achieves better performance and enables a fairer comparison. For architectures that support MAE pre-training (*i.e.*, Transformer or Mamba-based models), we apply MAE pre-training (without extra data) and fine-tune with the same protocol as our method.

Evaluation metrics For segmentation-based detection, we evaluate performance using the Dice Similarity Coefficient (Dice), Hit-Rate (H.r.), and FROC-Score (F.s.) [41] derived from the Free-Response Receiver Operating Characteristic (FROC) curve [2]. Among these, Dice measures voxel-level segmentation quality, while Hit-Rate and FROC-Score assess lesion-level sensitivity and the ability to control false positives. For classification, we report sensitivity (Sens.) at a fixed high specificity (Spec.) threshold for different methods and the Area Under the Receiver Operating Characteristic Curve (AUC). Further details on evaluation metrics are available in Sec. D.

4.2 Main Results

Quantitative comparisons on internal datasets Tab. 1 presents quantitative results on three internal datasets, comparing our method against strong baselines and state-of-the-art approaches. For UNet-based methods, we incorporate multi-scale feature fusion [40, 72] for classification. To isolate the effect of task interaction, we include single-task nnUNet variants: nnUNet-SEG (segmentation-only) and nnUNet-CLS (classification-only). Our method, DRT-M3D, achieves leading performance across most metrics on all three datasets. In particular, it consistently improves segmentation-based detection by reducing false positives (as reflected in the FROC-score), and boosts cancer classification performance in terms of Sensitivity and AUC. We also note that our baseline DR-M3D model performs competitively, further validating the effectiveness of our dual-resolution design.

Quantitative comparisons on the external dataset The external dataset from a distinct institution (Inst. 4) poses a significant distribution shift, serving as a

Table 2: **Quantitative results of the external evaluation.** The best and second-best results are **bolded** and underlined, respectively. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Method	External (Spec. = 0.9083)				
	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	0.5668	0.8889	0.8173	0.6286	0.8889
nnUNet-SEG	0.5776	0.8632	0.8173	-	-
nnUNet-CLS	-	-	-	0.7238	0.9156
VNet [48]	0.5059	0.8120	0.7272	0.6571	0.8790
swinUNETR [24]	0.5364	0.8803	0.7639	0.6476	0.8818
nnFormer [69]	0.5270	0.8889	0.7568	0.6190	0.8806
3D UX-Net [34]	0.5217	0.8547	0.7589	0.6095	0.8777
MedNeXt [52]	0.5767	0.8974	0.8120	0.6571	0.8945
U-Mamba-Bot [46]	0.5415	0.7863	0.7799	0.7238	0.8974
U-Mamba-Enc [46]	0.5180	0.7436	0.7308	0.6095	0.8893
EM-Net [6]	0.5489	0.8547	0.7618	0.6000	0.8838
SegMamba [64]	0.5388	0.8462	0.7639	0.6381	0.9083
Sun et al. 2025 [56]	0.5624	0.8974	0.8246	0.6286	0.8827
DR-M3D (ours)	0.5909 ±.0038	0.8974 ±.0041	0.8632 ±.0064	0.7207 ±.0092	0.9082 ±.0091
DRT-M3D (ours)	**0.5948** ±.0027	**0.9117** ±.0082	**0.8766** ±.0057	**0.8762** ±.0080	**0.9371** ±.0046

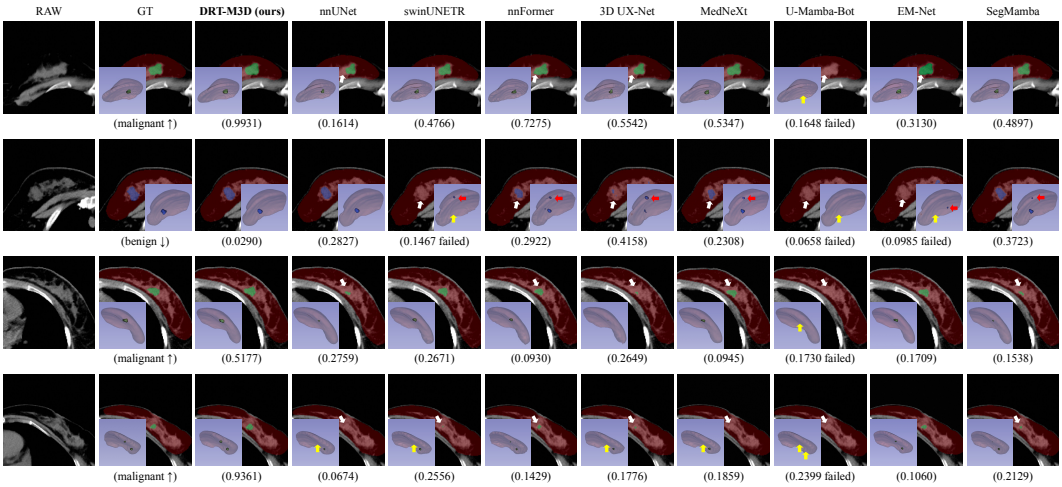


Figure 4: **Qualitative comparison of DRT-M3D with competing methods.** Examples from internal and external test sets are shown, including a representative 2D slice and a 3D view for each case. Red, green, and blue masks represent breast regions, malignant, and benign lesions, respectively. White, yellow, and red arrows mark pronounced segmentation errors in the slice, missed lesions, and segmentation of non-existent lesions, respectively. Predicted malignancy scores are shown for each case (higher for malignant, lower for benign). "Failed" denotes cases where none of the true lesions were localized (*i.e.*, a non-hit under the Hit-Rate metric).

Table 3: **Ablation study on the designs of the proposed DRT-M3D network.** DR, MF, and TI denote the use of Dual-Resolution, Mutual Fusion, and Tandem Input, respectively. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Variants				Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
DR	MF	TI		Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
			Vanilla M3D	0.6334	0.8908	0.8854	0.7936	0.9637	0.5788	0.8666	0.8359	0.7100	0.9053
✓				0.6449	0.9170	0.8879	0.7982	0.9574	0.5848	0.8879	0.8509	0.6620	0.8940
✓	✓		DR-M3D	0.6564	0.9228	0.9027	0.8073	0.9690	0.5909	0.8974	0.8632	0.7207	0.9082
		✓		0.6388	0.8952	0.8876	0.8211	0.9625	0.5898	0.8880	0.8565	0.8522	0.9184
✓		✓		0.6491	0.9214	0.9072	0.8303	0.9719	0.5925	0.9022	0.8698	**0.8762**	0.9264
✓	✓	✓	DRT-M3D	**0.6590**	**0.9229**	**0.9145**	**0.8578**	**0.9768**	**0.5948**	**0.9117**	**0.8766**	**0.8762**	**0.9371**

strong generalization benchmark. As shown in Tab. 2, DRT-M3D maintains leading performance across all metrics, outperforming all competing methods. The variant without tandem input (DR-M3D) also generalizes well, especially for segmentation, further supporting the robustness of the overall architecture.

For both DR-M3D and DRT-M3D, we report the mean and standard deviation of each metric over five runs with different random seeds, as reported in Tab. 1 and Tab. 2.

Visualization results for qualitative comparison Fig. 4 shows representative lesion detection and classification outputs. DRT-M3D delivers accurate localization and better differentiation between cancerous and non-cancerous cases. Competing methods more frequently miss lesions or misclassify malignant breasts.

4.3 Ablation Studies

Ablation on architectural designs We evaluate the impact of key components in DRT-M3D through ablation studies on Dual-Resolution (DR, w/o means using the HR path for both tasks), Mutual Fusion (MF), and Tandem Input (TI). The results in Tab. 3 demonstrate that the DR design enhances segmentation, while TI improves breast-level cancer classifica-

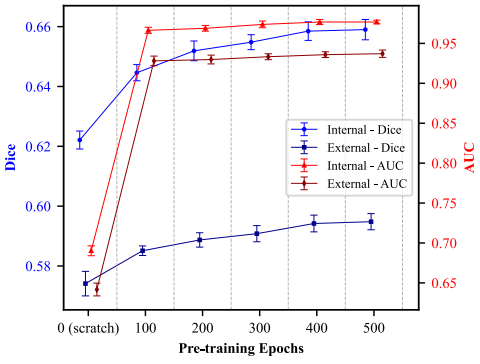


Figure 5: **Ablation study on the training strategy.** Error bars on the curves are obtained by repeating the experiments with five different random seeds.

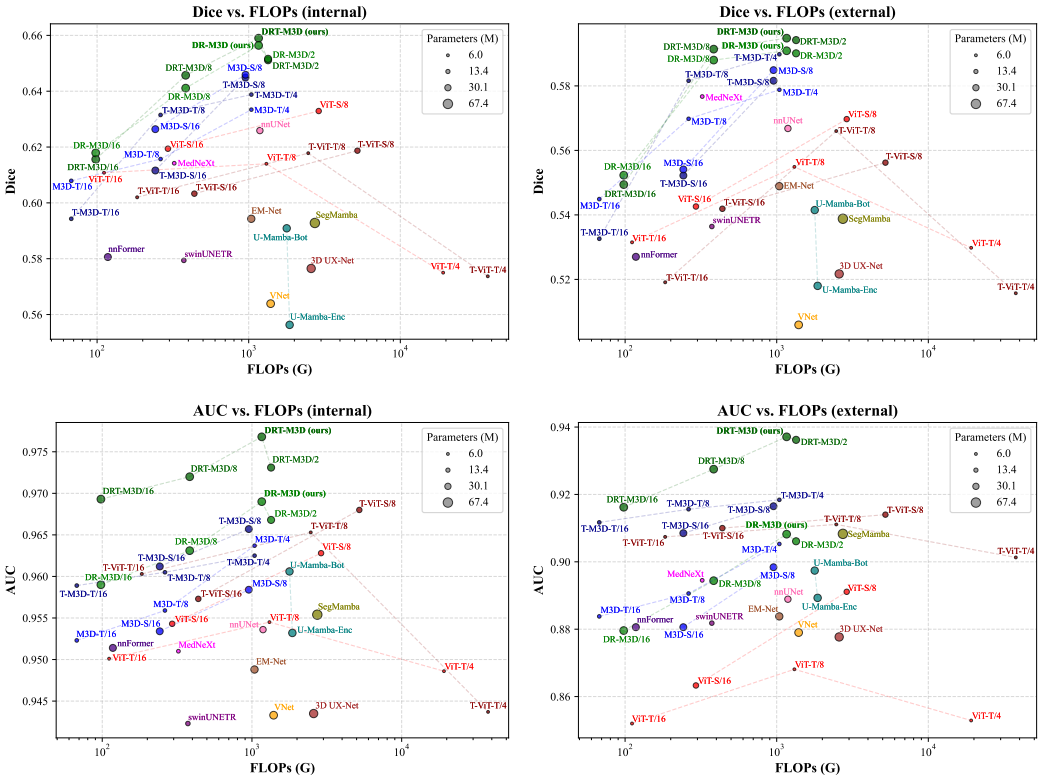


Figure 6: **Comparison of Dice and AUC vs. FLOPs on internal and external datasets.** Each FLOPs value is computed per bilateral input (or two unilateral inputs). ViT: vanilla Vision Transformer, M3D: vanilla Mamba-3D, T-~: tandem (bilateral) input form, ~/p: patch size of (2, p, p).

tion. Additionally, the MF mechanism facilitates information sharing between the Dual-Res paths, boosting performance across all metrics. For Tandem Input specifically, we further compare several variants in Sec. E.

Ablation on the training strategy MAE pre-training has been shown to benefit Mamba-3D on high-level semantic tasks [70]. Since our study also involves voxel-level segmentation, we further validate its effectiveness in this context. To fairly compare different pre-training durations, we adjust the fine-tuning epochs for full convergence. As shown in Fig. 5, the two-stage training strategy with adequate self-supervised pre-training also proves effective for the multi-task setting of segmentation and classification in this study.

Patch size tuning We further validate through experiments (see Fig. 6) that using a relatively smaller patch size in height (p_h) and width (p_w) of CT slices, such as (2, 4, 4) in our models, substantially improves performance. Due to the model's underlying linear computation design, this gain comes with minimal computational overhead. However, further reducing the patch size (*e.g.*, to (2, 2, 2)) leads to stagnation or even a drop in performance, presumably due to increased token fragmentation and reduced contextual capacity. Conversely, variants with larger patch sizes can significantly reduce computational costs while still outperforming competing methods. Detailed configurations of each variant are provided in Appendix Tab. 5 and Tab. 6. For experiments regarding the patch size in the depth dimension (p_d), please refer to Sec. E.

Bilateral input on vision transformers To further evaluate the Mamba-3D (M3D) structure and tandem input concept, we integrate bilateral input into the commonly used ViT [14] for comparison. We adopt ViT-T and ViT-S as two different scales, and denote their bilateral variants as T-ViT. In T-ViT, left-side and right-side encodings are added to the 3D positional encoding to distinguish patches. All ViT models follow the two-stage training pipeline as ours. As shown in Fig. 6, bilateral input improves cancer classification in ViT models but still fall short compared to our M3D-based models, particularly for ViT/4 with the same patch size (2, 4, 4) as DRT-M3D. The performance gap

is primarily due to the excessive number of tokens resulting from small patch sizes and bilateral input, which leads to peaked and less expressive attention distributions and thus hampers the learning of patch relationships [57]. For larger patch sizes, ViT's performance remains clearly inferior, especially in segmentation. Although larger patches reduce token count, they further decrease spatial resolution, and the lack of intrinsic local inductive bias in ViT prevents effective modeling of local structures.

5 Conclusion

In this paper, we propose DRT-M3D, a dual-resolution network for joint segmentation-based breast lesion detection and cancer classification on non-contrast chest CT scans. By disentangling the two tasks into resolution-specific pathways, DRT-M3D enables complementary learning between segmentation and classification, while tandem bilateral inputs enhance contextual understanding of subtle features across bilateral breasts. Experiments on multi-institutional datasets demonstrate consistent improvements over strong baselines, underscoring the clinical potential of DRT-M3D for robust and generalizable opportunistic breast cancer analysis.

Limitations and future work This study only focuses on breast cancer analysis using NCCT scans, but the idea of bilateral organs and the design of DRT-M3D may be applied to other organs such as lungs and kidneys. In addition, the data currently in use all come from one country and have not been extended to a broader population. We have already started to collect more diverse data from a wider range of regions for future research.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (grant 92354307), the National Key Research and Development Program of China (grant 2024YFF0729202), and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDA0460305). This work was also supported by Alibaba Group through Alibaba Research Intern Program.

References

[1] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons. A free response approach to the measurement and characterization of radiographic observer performance. In *Application of optical instrumentation in medicine VI*, volume 127, pages 124–135. SPIE, 1977.

[3] X. Cao, H. Chen, Y. Li, Y. Peng, Y. Zhou, L. Cheng, T. Liu, and D. Shen. Auto-denseunet: Searchable neural network architecture for mass segmentation in 3d automated breast ultrasound. *Medical image analysis*, 82:102589, 2022.

[4] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[5] P. Casti, A. Mencattini, M. Salmeri, and R. M. Rangayyan. Analysis of structural similarity in mammograms for detection of bilateral asymmetry. *IEEE transactions on medical imaging*, 34(2):662–671, 2014.

[6] A. Chang, J. Zeng, R. Huang, and D. Ni. Em-net: Efficient channel and frequency learning with mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–275. Springer, 2024.

[7] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, 6(1):24454, 2016.

[8] B. S. Chhikara, K. Parang, et al. Global cancer statistics 2022: the trends projection analysis. *Chemical Biology Letters*, 10(1):451–451, 2023.

[9] M. Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[10] T. Dao and A. Gu. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10041–10071, 2024.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] A. Dey, E. Ali, and S. Rajan. Bilateral symmetry-based abnormality detection in breast thermograms using textural features of hot regions. *IEEE Open Journal of Instrumentation and Measurement*, 2:1–14, 2023.

[13] C. D'Orsi, L. Bassett, and S. Feig. Breast imaging reporting and data system (bi-rads). *Oxford University Press, New York*, 2018.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[15] Y. Gao, S. Ventura-Diaz, X. Wang, M. He, Z. Xu, A. Weir, H.-Y. Zhou, T. Zhang, F. H. van Duijnhoven, L. Han, et al. An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. *Nature Communications*, 15(1):9613, 2024.

[16] B. Gheflati and H. Rivaz. Vision transformers for classification of breast ultrasound images. In *2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 480–483. IEEE, 2022.

[17] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[18] A. Gu, K. Goel, A. Gupta, and C. Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

[19] A. Gu, K. Goel, and C. Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[20] A. Gupta, A. Gu, and J. Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

[21] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[23] Q. He, Q. Yang, H. Su, and Y. Wang. Multi-task learning for segmentation and classification of breast tumors from ultrasound images. *Computers in Biology and Medicine*, 173:108319, 2024.

[24] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer, 2023.

[25] M. P. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel. Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE transactions on medical imaging*, 32(7):1239–1248, 2013.

[26] G. Holste, S. C. Partridge, H. Rahbar, D. Biswas, C. I. Lee, and A. M. Alessio. End-to-end learning of fused image and non-image features for improved breast cancer classification from mri. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3294–3303, 2021.

[27] Q. Huang, Y. Huang, Y. Luo, F. Yuan, and X. Li. Segmentation of breast ultrasound image with semantic classification of superpixels. *Medical Image Analysis*, 61:101657, 2020.

[28] R. Huang, Z. Lin, H. Dou, J. Wang, J. Miao, G. Zhou, X. Jia, W. Xu, Z. Mei, Y. Dong, et al. Aw3m: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. *Medical image analysis*, 72:102137, 2021.

[29] A. Hussain, A. Gordon-Dixon, H. Almusawy, P. Sinha, and A. Desai. The incidence and outcome of incidental breast lesions detected by computed tomography. *The Annals of The Royal College of Surgeons of England*, 92(2):124–126, 2010.

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

[31] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[33] J. Koh, Y. Yoon, S. Kim, K. Han, and E.-K. Kim. Deep learning for the detection of breast cancers on chest computed tomography. *Clinical breast cancer*, 22(1):26–31, 2022.

[34] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman. 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] H. Li, D. Chen, W. H. Nailon, M. E. Davies, and D. I. Laurenson. Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography. *IEEE Transactions on Medical Imaging*, 41(1):3–13, 2021.

[36] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024.

[37] M. Li, K. Sun, Y. Gu, K. Zhang, Y. Sun, Z. Li, and D. Shen. Developing large pre-trained model for breast tumor segmentation from ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–96. Springer, 2023.

[38] S. Li, H. Singh, and A. Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.

[39] Y. Li, Y. Shen, J. Zhang, S. Song, Z. Li, J. Ke, and D. Shen. A hierarchical graph v-net with semi-supervised pre-training for histological image based breast cancer classification. *IEEE Transactions on Medical Imaging*, 42(12):3907–3918, 2023.

[40] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[41] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):26286, 2016.

[42] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024.

[43] J. Liu, H. Yang, H.-Y. Zhou, L. Yu, Y. Liang, Y. Yu, S. Zhang, H. Zheng, and S. Wang. Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[44] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[45] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[46] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[47] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[48] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[49] T.-H. Nguyen, Q. H. Kha, T. N. T. Truong, B. T. Lam, B. H. Ngo, Q. V. Dinh, and N. Q. K. Le. Towards robust natural-looking mammography lesion synthesis on ipsilateral dual-views breast cancer analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2564–2573, 2023.

[50] X. Qian, J. Pei, C. Han, Z. Liang, G. Zhang, N. Chen, W. Zheng, F. Meng, D. Yu, Y. Chen, et al. A multimodal machine learning model for the stratification of breast cancer risk. *Nature Biomedical Engineering*, 9(3):356–370, 2025.

[51] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[52] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. H. Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.

[53] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019.

[54] V. K. Singh, H. A. Rashwan, S. Romani, F. Akram, N. Pandey, M. M. K. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig, et al. Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network. *Expert Systems with Applications*, 139:112855, 2020.

[55] J. T. Smith, A. Warrington, and S. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

[56] J. Sun, X. Xi, M. Wang, M. Liu, X. Zhang, H. Qiu, Y. Zhang, T. Fu, Y. Du, W. Ren, et al. A deep learning model based on chest ct to predict benign and malignant breast masses and axillary lymph node metastasis. *Biomolecules and Biomedicine*, 2025.

[57] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021.

[58] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] C. E. von Schacky, N. J. Wilhelm, V. S. Schäfer, Y. Leonhardt, F. G. Gassert, S. C. Foreman, F. T. Gassert, M. Jung, P. M. Jungmann, M. F. Russe, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology*, 301(2):398–406, 2021.

[61] F. Wang, D. Wang, Y. Xu, H. Jiang, Y. Liu, and J. Zhang. Potential of the non-contrast-enhanced chest ct radiomics to distinguish molecular subtypes of breast cancer: A retrospective study. *Frontiers In Oncology*, 12:848726, 2022.

[62] H. Wang, Y. Lin, X. Ding, and X. Li. Tri-plane mamba: Efficiently adapting segment anything model for 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646. Springer, 2024.

[63] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

[64] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[65] M. Xu, K. Huang, and X. Qi. A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification. *IEEE Access*, 11:5377–5392, 2023.

[66] C. Xue, L. Zhu, H. Fu, X. Hu, X. Li, H. Zhang, and P.-A. Heng. Global guidance network for breast lesion segmentation in ultrasound images. *Medical image analysis*, 70:101989, 2021.

[67] Z. Yang, Z. Cao, Y. Zhang, Y. Tang, X. Lin, R. Ouyang, M. Wu, M. Han, J. Xiao, L. Huang, et al. Momminet-v2: Mammographic multi-view mass identification networks. *Medical Image Analysis*, 73:102204, 2021.

[68] J. Zhang, J. Wu, X. S. Zhou, F. Shi, and D. Shen. Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. In *Seminars in Cancer Biology*, volume 96, pages 11–25. Elsevier, 2023.

[69] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu. nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

[70] J. Zhou, Y. Zhou, W. Fang, Y. Tang, L. Lu, and G. Yang. Mamba-3d as masked autoencoders for accurate and data-efficient analysis of medical ultrasound videos. *arXiv preprint arXiv:2503.20258*, 2025.

[71] L. Zhou, Y. Zhang, J. Zhang, X. Qian, C. Gong, K. Sun, Z. Ding, X. Wang, Z. Li, Z. Liu, et al. Prototype learning guided hybrid network for breast tumor segmentation in dce-mri. *IEEE Transactions on Medical Imaging*, 2024.

[72] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical image analysis*, 70:101918, 2021.

[73] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, pages 62429–62442. PMLR, 2024.

[74] N. Zielonke, A. Gini, E. E. Jansen, A. Anttila, N. Segnan, A. Ponti, P. Veerus, H. J. de Koning, N. T. van Ravesteyn, E. A. Heijnsdijk, et al. Evidence for reducing cancer-specific mortality due to screening for breast cancer in europe: A systematic review. *European journal of cancer*, 127:191–206, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clarified the research scope and main contributions of the paper in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have acknowledged and discussed the limitations of the work in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: As an application-oriented article on machine learning, the paper does not contain any content that requires theoretical assumptions and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the structure of the model, the parameter settings and the training strategy elaborately. For details, please refer to Sec. 3 and Sec. C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released once internal review and approval are completed. Due to privacy and institutional policy issues, the data cannot be directly and fully disclosed, but it will be accessible upon reasonable academic request with the institutional approval.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please refer to Sec. 4.1 and Sec. C as well as the ablation studies on main designs and hyper-parameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provided the error range of the main experimental results by repeating the training with different seeds for 5 times. Please refer to Tab. 1 and Tab. 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have reported the main hardware specifications in Sec. 4.1 and presented the computational cost (GFLOPs) and inference time of different methods in Fig. 6 and Fig. 9, respectively.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the possible broader impacts in Sec. F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve any content that can lead to a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For assets that do not need anonymization during the blind review period, we have already credited them in appropriate ways. Others will be further de-anonymized upon publication.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not involve any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. For research with human subjects, please refer to item 15 of the list.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The research participants are not exposed to any potential risks, as this study uses retrospective patient data and the research won't immediately change clinical pathways. The paper has obtained the IRB approval and it is stated in the appendix.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Preliminaries on Mamba-3D Blocks

A.1 Selective State Space Models

Selective state space models (S6), which are represented by Mamba [17], are essentially derived from the vanilla state space models (SSM) based on continuous linear time-invariant (LTI) systems.

SSMs utilize N-dimensional latent state vector $\mathbf{h}(t) \in \mathbb{R}^{\mathrm{N}}$ to model the transformation from a one-dimensional continuous input signal to its corresponding output $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$:

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}\mathbf{h}(t) + \mathbf{D}x(t) \end{aligned} \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ is the "state matrix", $\mathbf{B} \in \mathbb{R}^{\mathrm{N}\times 1}$ is the "input matrix", $\mathbf{C} \in \mathbb{R}^{1\times\mathrm{N}}$ is the "output matrix", and $\mathbf{D} \in \mathbb{R}^{1\times 1}$ is the "feed-through matrix". These are system matrices governing the evolution and output of the system.

To make these models applicable in deep learning settings with discrete inputs, structured state space models (S4) [18–20] discretize these equations using numerical methods such as the zero-order hold (ZOH) [55] approach, yielding the following discretization:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}\left[\exp(\Delta\mathbf{A}) - \mathbf{I}\right] \cdot \Delta\mathbf{B} \approx \Delta\mathbf{B} \end{aligned} \tag{9}$$

where Δ denotes the discretization step size. The larger the step size Δ, the faster the hidden state $\mathbf{h}$ changes, and the greater the impact of the current token x on the system.

In Mamba, $\mathbf{A}$ must be a diagonal matrix. Consequently, when entries of $\Delta\mathbf{A}$ are sufficiently small, the approximation $(\Delta\mathbf{A})^{-1}\left[\exp(\Delta\mathbf{A}) - \mathbf{I}\right] \approx \mathbf{I}$ holds true. This leads to the discrete-time S4 model for sequential input and output, as $x_t \in \mathbb{R} \rightarrow y_t \in \mathbb{R}$:

$$\begin{aligned} \mathbf{h}_t &= \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}x_t \approx \exp(\Delta\mathbf{A})\mathbf{h}_{t-1} + \Delta\mathbf{B}x_t \\ y_t &= \mathbf{C}\mathbf{h}_t + \mathbf{D}x_t \end{aligned} \tag{10}$$

For multi-channel inputs (*i.e.*, $\mathbf{x}_t \in \mathbb{R}^{\mathrm{C}}$) and the corresponding multi-channel outputs, the above operations are applied independently to each channel.

Building upon S4, the S6 structure enhances the model's expressiveness by making the parameters $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, $\mathbf{C}$ dynamically input-dependent across all channels of $\mathbf{x}_t \in \mathbb{R}^{\mathrm{C}}$, which is known as the selection mechanism:

$$\begin{aligned} \mathbf{B}_t &= \mathrm{Linear}_{\mathrm{C}\rightarrow\mathrm{N}}\left(\mathbf{x}_t\right) \\ \mathbf{C}_t &= \mathrm{Linear}_{\mathrm{C}\rightarrow\mathrm{N}}\left(\mathbf{x}_t\right) \\ \boldsymbol{\Delta}_t &= \mathrm{softplus}\left[\mathrm{Linear}_{\mathrm{C}\rightarrow\mathrm{C}}\left(\mathbf{x}_t\right)\right] \end{aligned} \tag{11}$$

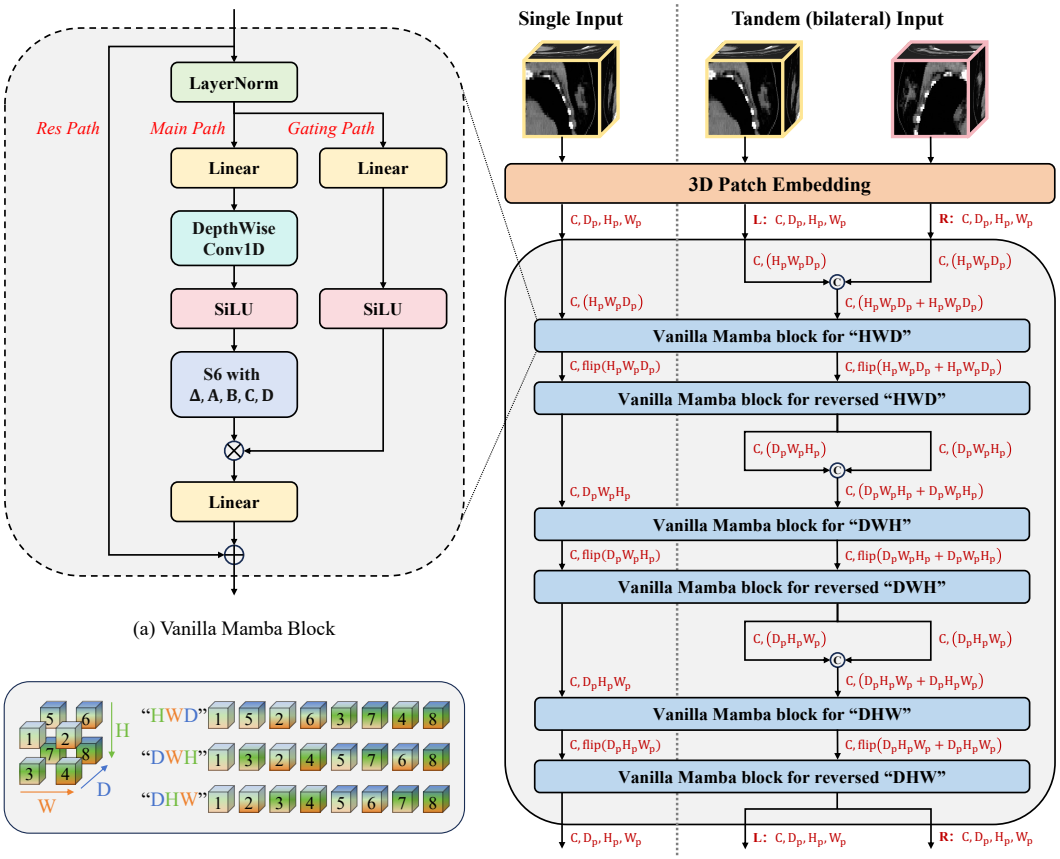
With the discretization in Eq. (9), $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, $\mathbf{C}$ are no longer fixed parameters in the inference phase, allowing the model to adapt to varying input signals and provides greater modeling capabilities than the original SSMs. Overall, S6 can be expressed as:

$$\begin{aligned} \mathbf{h}_t &= \exp(\boldsymbol{\Delta}_t\mathbf{A})\mathbf{h}_{t-1} + \boldsymbol{\Delta}_t\mathbf{B}_t x_t \\ y_t &= \mathbf{C}_t\mathbf{h}_t + \mathbf{D}x_t \end{aligned} \tag{12}$$

The input-dependent nature of S6 (Mamba) makes it particularly suitable for processing patchified visual data, where different spatial regions may require distinct dynamic responses, similar to words in natural language processing. The exponential transition term captures long-range dependencies, while the learned Δ_t modulates temporal sensitivity, making Mamba capable of adapting to both local and global features.

A.2 Mamba-3D Blocks

Unlike most visual Mamba works [36,44,73] that modify the internal structure of S6 blocks to support multi-directional scanning, Mamba-3D [38, 70] retains the original architecture of the Mamba block,



(a) Vanilla Mamba Block

(b) Different Flattening Modes in Mamba-3D

(c) Mamba-3D block with Single Input or Tandem (bilateral) Input

Figure 7: **The internal structure of the proposed Tandem Mamba-3D block.** (a) Vanilla Mamba block with selective state space module (S6). (b) Different sequence permutations used for the serial scanning processes in Mamba-3D. (c) Data flow inside one Mamba-3D block in the cases of Single Input and Tandem (bilateral) Input.

as depicted in Fig. 7 (a). This design choice allows Mamba-3D to maintain the same computational cost while enabling deeper networks that better capture directional visual features, yielding improved performance on 3D visual tasks compared to methods that introduce complexity into the S6 internals.

Mamba-3D achieves effective processing of 3D data through specific transformations of the sequence order among different Mamba blocks, as illustrated in Fig. 7 (b) and (c). Specifically, this involves separate forward and reverse Mamba operations across three permutations of the 3D axes: "HWD", "DWH", and "DHW", as three 3D-to-1D flattening modes. Each permutation ensures contiguity along the last axis after flattening, facilitating the modeling of local and non-local dependencies across all spatial dimensions. Furthermore, the standard **Single Input** form and the proposed **Tandem (bilateral) Input** are presented respectively in the left and right parts of Fig. 7 (c).

B Datasets

Further details on the datasets Our research utilizes three internal datasets from different institutions for training and evaluation, as well as an external dataset for independent evaluation. The data distribution and characteristics of the external dataset are prominently different from those of the internal datasets, which makes it more conducive to demonstrating the generalization capabilities of different methods. The number of samples in each dataset is shown in Tab. 4. The training and testing splits of internal datasets are designed to evenly distribute samples across different collection periods and varying lesion sizes.

The original data in each dataset consists entirely of non-contrast chest CT images with spatial resolution D $\times$ 512 $\times$ 512, where D denotes the number of slices per volume, typically ranging

Table 4: **Sample counts for the datasets used in this study.**

Dataset	Training				Testing			
	Patients	Breasts	Cancerous	Non-cancerous	Patients	Breasts	Cancerous	Non-cancerous
Internal Inst. 1	328	656	341	315	80	160	82	78
Internal Inst. 2	580	1160	239	921	199	398	102	296
Internal Inst. 3	127	254	141	113	31	62	34	28
External Inst.	-	-	-	-	107	214	105	109
Total	1035	2070	721	1349	417	834	323	511

from several dozen to several hundred depending on acquisition settings. To ensure the reliability and accuracy of the annotations, experienced radiologists with years of expertise delineated the boundaries of breast lesions based on contrast-enhanced CT images taken concurrently for each patient. These segmentation labels were then accurately registered to the non-contrast CT (NCCT) volumes using DEEDS [25]. All malignant cases were confirmed via pathological reports, while benign or normal cases were validated either through pathological examinations or two-year clinical follow-ups.

In Fig. 8, we present the visualizations of representative data samples with lesions from each dataset, from left to right including: (a) the representative NCCT 2D slice view with a window width of 300 and window level of 50, (b) the 2D view with segmentation labels for breast and lesion, (c) the registered CECT slices to highlight the challenge of using non-contrast CT instead of contrast-enhanced CT, (d) the 3D rendered image, and (e) the 3D rendered image with segmentation labels for breast and lesion. Red annotations represents the breast, while green and blue annotations respectively represent malignant or benign lesions. The samples from the four institutions are presented consecutively from top to bottom, with two samples from each institution.

The study received approval from the Institutional Review Board (IRB) with Approval No.: B2025-235R. All procedures adhered to established ethical standards and regulations. Confidentiality of participant data was stringently maintained, with all identifying information anonymized during data collection and analysis. Regular audits were conducted to verify compliance with ethical guidelines, and any unforeseen ethical concerns that arose during the study were promptly addressed in consultation with institutional review boards.

Pre-Stage: breast region cropping As described in Sec. 3.1, we use a pre-trained segmentation network to achieve coarse-grained cropping of each breast region in the Pre-Stage. Since this step does not require particularly precise segmentation results, we trained this coarse-grained segmentation model on a few hundred samples with breast masks using nnUNet [30], which achieved a high Dice score of 0.9598, demonstrating its accuracy and reliability.

To further minimize the risk of missing boundary lesions or subtle structures (*e.g.*, in obese patients or cases with blurred borders), we adopted a conservative cropping strategy: the bounding box of the predicted breast mask is expanded by 2 voxels along the D-axis and 16 voxels along the H- and W-axes. This ensures that small lesions near the periphery are retained within the cropped region. In future deployment, extremely rare failure cases of the Pre-Stage can be flagged for manual review or fallback processing.

Following this, we obtained the training sets actually used in the Main-Stage, where each sample contained two partial NCCT images corresponding to the left and right breast regions, along with the voxel-level lesion masks and breast-level cancer classification labels. For methods that do not use bilateral inputs, each side is treated as an independent sample. Additionally, we retained the breast region masks from this step as a third segmentation category, alongside the background and lesion masks, to guide effective learning in all networks, including ours and all competing methods.

C Implementation Details

This section elaborates on the implementation details of the Main-Stage in the overall pipeline (Fig. 2) of this study, while the the Pre-Stage has been introduced in Sec. B.

To ensure a fair comparison, all methods in this study—including those used for benchmarking and ablation studies—share identical pre-processing, post-processing, and data augmentation protocols. The only difference lies in the synchronization of randomness during data augmentation when bilateral inputs are used.

Pre-processing Firstly, we unify the three-dimensional spacings of CT images through resampling, with the target values being the median spacings of all samples in all training sets, which are (3 mm, 0.748 mm, 0.748 mm). Subsequently, we adopt the normalization scheme introduced by [30] to the Hounsfield Unit (HU) values of CT images. Let $\mathbf{x}$ denote the original Hounsfield Unit (HU) value of voxels in a CT image. For all foreground voxels (*i.e.*, breast and lesion regions), we first compute the 0.5th and 99.5th percentiles of the HU distribution, denoted as $q_{0.5}$ and $q_{99.5}$, respectively. The voxel intensities are then clipped and standardized as follows:

$$\begin{aligned} \mathbf{x}_{\text{clipped}} &= \min\big(\max(\mathbf{x}, q_{0.5}), q_{99.5}\big) \\ \mathbf{x}_{\text{norm}} &= \frac{\mathbf{x}_{\text{clipped}} - \mu}{\sigma} \end{aligned} \tag{13}$$

where $\mathbf{x}_{\text{clipped}}$ is the HU value after quantile-based clipping; μ and σ are the mean and standard deviation of the clipped HU values within the foreground; $\mathbf{x}_{\text{norm}}$ is the final normalized HU value. This normalization procedure helps suppress noise and outliers, while ensuring a standardized input range for downstream processing.

Given that the cropped breast regions vary in size, we further crop each single breast in the Main-Stage as nnUNet [30] did, so that it has a uniform size as $56 \times 160 \times 192$ when fed into our network and all the competitors, which is slightly smaller than the median size of all single-sided samples. In cases where the input no longer contains any lesion voxels after the uniformly sized cropping, we modify the breast-level label to "*non-cancerous*" by verifying that no lesion voxels are present in the cropped region during training.

Data augmentations To leverage the symmetric nature of breasts in alignment with our proposed Tandem Mamba-3D blocks, we flipped the left breast image along the X-axis (*i.e.*, the W-axis in the network's inputs and outputs) prior to applying data augmentations, thus treating it as a pseudo-right breast. During training, the cropped left and right breast images from the same patient undergo identical data augmentation procedures, which include a random combination of scaling, rotation, Gaussian noise, Gaussian blur, brightness adjustment, contrast adjustment, inversion, gamma adjustment, and three-dimensional mirroring, as proposed and planned by [30].

Post-processing for segmentation and classification As commonly used in medical image segmentation [4,30], we apply a sliding window approach during inference to handle variations in sample sizes. For segmentation, the softmax values of each voxel are restored to the original cropped breast region's size using Gaussian-weighted superposition and spacing resampling. For classification, the final score for malignancy (cancer) is taken as the highest score across all sliding windows.

In each window, mirroring is applied independently along all three dimensions, yielding eight different results. The averaged results are then used as the test-time augmentation. As mentioned earlier, the left breast is flipped to become a pseudo-right breast for bilateral inputs, and during the eight mirroring operations, the real-right and pseudo-right breasts are kept synchronized.

Detailed configurations and hyper-parameters Leveraging the linear complexity of Mamba with respect to sequence length and its efficient implementation, we adopt a relatively small patch size to enhance segmentation performance. Unless otherwise specified, the default main hyper-parameters are set as follows: the patch size for the high-resolution (HR) path is $\big(\mathrm{p_d}, \mathrm{p_h}, \mathrm{p_w}\big) = \big(2, 4, 4\big)$ with the embedding dimension of $\mathrm{C} = 192$; the downsampling factor from HR to low-resolution (LR) path is $\big(\mathrm{p_d'}, \mathrm{p_h'}, \mathrm{p_w'}\big) = \big(2, 4, 4\big)$ with the LR embedding dimension of $\mathrm{C'} = 384$; the number of backbone blocks is $N = 6$; the number of decoder blocks is $M = 2$; and the masking ratio for self-supervised pre-training is $p_{\text{mask}} = 0.8$. Each Mamba layer follows the default configurations as the original Mamba [17].

In Tab. 5 and Tab. 6, we respectively summarize the training configurations and hyper-parameters used in self-supervised pre-training and downstream fine-tuning for the proposed DR-M3D and DRT-M3D networks, including their variants used in the ablation study of different patch sizes.

During inference, all methods—ours and the competing ones—produce the exact same output format in the Main-Stage, including the voxel-leval segmentation softmax scores for each $\mathrm{D} \times \mathrm{H} \times \mathrm{W}$ sliding window, and window-level classification scores after applying the test-time augmentation (*i.e.*, mirroring and averaging as previously described).

Table 5: **Default configurations and hyper-parameters of DRT-M3D and its variants of different patch sizes for self-supervised pre-training.**

Configuration & Hyper-Parameter	DR-M3D series				DRT-M3D series			
	DR-M3D/16	DR-M3D/8	**DR-M3D**	DR-M3D/2	DRT-M3D/16	DRT-M3D/8	**DRT-M3D**	DRT-M3D-S
optimizer	AdamW [45] with $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1\text{e}{-15}$							
weight decay	$1\text{e}{-1}$							
learning rate schedule	cosine decay with linear warm-up							
basic learning rate	$1\text{e}{-3}$							
minimal learning rate	$1\text{e}{-5}$							
warm-up epochs	5 epochs							
total epochs	500 epochs							
total batch size	$32 = 4(\text{batch size}) \times 2(\text{ranks}) \times 4(\text{accumulation})$				$16 = 2(\text{batch size}) \times 2(\text{ranks}) \times 4(\text{accumulation})$			
max gradient norm	0.1							
mixed-precision	BFloat16							
loss function	Dual-Path MSE Loss as Eq. (4)							
Input Size ($1 \times \text{D} \times \text{H} \times \text{W}$)	$1 \times (1 \times 56 \times 160 \times 192)$				$2 \times (1 \times 56 \times 160 \times 192)$			
Patch Size ($\text{p}_\text{d}, \text{p}_\text{h}, \text{p}_\text{w}$)	(2, 16, 16)	(2, 8, 8)	(2, 4, 4)	(2, 2, 2)	(2, 16, 16)	(2, 8, 8)	(2, 4, 4)	(2, 2, 2)
Low-Res Pooling Size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$)	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)
High-Res Embedding Dimension (C)	192	192	192	96	192	192	192	96
Low-Res Embedding Dimension (C')	384							
Mutual Fusion DownSampling	3D MaxPooling with size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$) + Linear: $\text{C} \to \text{C}'$							
Mutual Fusion Upsampling	3D Nearest Interpolation with size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$) + Linear: $\text{C}' \to \text{C}$							
Backbone Mamba-3D Blocks (N)	6							
Decoder Mamba-3D Blocks (M)	2							
Masking Chain	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)
Masking Ratio (p_mask)	0.8							

Table 6: **Default configurations and hyper-parameters of DRT-M3D and its variants of different patch sizes for downstream fine-tuning.**

Configuration & Hyper-Parameter	DR-M3D series				DRT-M3D series			
	DR-M3D/16	DR-M3D/8	**DR-M3D**	DR-M3D/2	DRT-M3D/16	DRT-M3D/8	**DRT-M3D**	DRT-M3D-S
optimizer	AdamW [45] with $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1\text{e}{-15}$							
weight decay	$1\text{e}{-1}$							
learning rate schedule	cosine decay with linear warm-up							
basic learning rate	$1\text{e}{-3}$							
minimal learning rate	$1\text{e}{-5}$							
warm-up epochs	1 epoch							
total epochs	50 epochs							
total batch size	$32 = 4(\text{batch size}) \times 2(\text{ranks}) \times 4(\text{accumulation})$				$16 = 2(\text{batch size}) \times 2(\text{ranks}) \times 4(\text{accumulation})$			
max gradient norm	1.0							
mixed-precision	BFloat16							
loss function	$\mathcal{L}_\text{SEG} + \lambda\mathcal{L}_\text{CLS}$ as Eq. (7) with $\lambda = 0.1$							
Input Size ($1 \times \text{D} \times \text{H} \times \text{W}$)	$1 \times (1 \times 56 \times 160 \times 192)$				$2 \times (1 \times 56 \times 160 \times 192)$			
Patch Size ($\text{p}_\text{d}, \text{p}_\text{h}, \text{p}_\text{w}$)	(2, 16, 16)	(2, 8, 8)	(2, 4, 4)	(2, 2, 2)	(2, 16, 16)	(2, 8, 8)	(2, 4, 4)	(2, 2, 2)
Low-Res Pooling Size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$)	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)	(2, 1, 1)	(2, 2, 2)	(2, 4, 4)	(2, 8, 8)
High-Res Embedding Dimension (C)	192	192	192	96	192	192	192	96
Low-Res Embedding Dimension (C')	384							
Mutual Fusion DownSampling	3D MaxPooling with size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$) + Linear: $\text{C} \to \text{C}'$							
Mutual Fusion Upsampling	3D Nearest Interpolation with size ($\text{p}'_\text{d}, \text{p}'_\text{h}, \text{p}'_\text{w}$) + Linear: $\text{C}' \to \text{C}$							
Backbone Mamba-3D Blocks (N)	6							
Head for Segmentation	Linear: C from High-Res Path $\to (3 \times \text{p}_\text{d} \times \text{p}_\text{h} \times \text{p}_\text{w})$, "3" for (background, breast, lesion)							
Head for Classification	3D Global AvgPooling + MLP: C' from Low-Res Path $\to 512 \to 512 \to$ Malignant-Logit							

For the sliding window strategy, we use a maximum stride of 50% of the window size along each dimension to determine the number of required steps. If the sizes of the left and right breast regions differ, we use the larger of the two to compute the step count. For generating Gaussian importance weights in each window, we set $\sigma = 0.125$ when computing the weight tensor.

D Metrics

D.1 Lesion-Level (Segmentation-based Detection) Metrics

Dice similarity coefficient (Dice) We directly evaluate the voxel-level accuracy of the segmented parts using the dice similarity coefficient:

$$\text{Dice} = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{2 \times |s_{i,\text{pred}} \cap s_{i,\text{gt}}|}{|s_{i,\text{pred}}| + |s_{i,\text{gt}}|} \tag{14}$$

where $|S|$ denotes the total number of samples in the testing set; $s_{i,\text{pred}}$ is the foreground (lesion) voxel mask obtained by the model for the i-th sample; $s_{i,\text{gt}}$ is the corresponding ground truth. The dice similarity coefficient is inherently normalized, making it suitable for targets of varying sizes or scales, particularly for breast lesion regions in this study. Moreover, the lesion boundaries in breast CT images are often ambiguous. Consequently, employing the dice similarity coefficient that only considers the intersection and is more lenient towards minor deviations on the edges holds practical value.

Hit-rate (H.r.) We define a breast-level prediction as a "hit" if the Dice score for that sample is greater than zero, otherwise as a "non-hit". A "hit" indicates that the model has successfully segmented a lesion region that overlaps with the ground truth even if only partially, which can be sufficient to prompt further clinical attention. The Hit-rate is then defined as:

$$\text{H.r.} = \frac{\left|\{s_i \in S \mid \text{Dice}(s_{i,\text{pred}}, s_{i,\text{gt}}) > 0\}\right|}{|S|} \tag{15}$$

where the numerator $\left|\{s_i \in S \mid \text{Dice}(s_{i,\text{pred}}, s_{i,\text{gt}}) > 0\}\right|$ represents the number of samples where the dice similarity coefficient is greater than zero, indicating a "hit"; the denominator $|S|$ denotes the total number of test samples.

FROC-score (F.s.) For each connected region classified as the lesion class in the segmentation results, we compute its lesion probability t as the mean of the softmax scores of all voxels in that region. Subsequently, we obtained the FROC curve [2, 41] based on varying probability thresholds τ. The x-axis of the FROC curve represents the false positive per breast (FPPB) indicating the average number of false positive lesion regions per breast. The y-axis the FROC curve represents the detection sensitivity indicating the proportion of true lesion-containing breasts where at least one lesion is detected.

Unlike the standard ROC curve, the x-axis of the FROC curve spans $[0, \infty)$, making it unsuitable for traditional AUC computation. Therefore, based on the characteristics of the used datasets, we define the FROC-score as the average detection sensitivity at four predefined FPPBs (0.1, 0.2, 0.3, and 0.4):

$$\text{F.s.} = \frac{1}{4}\sum_{i=1}^{4} \text{LesionSensitivity}(R_{\text{pred}}, R_{\text{gt}} | \text{FPPB}_i) \tag{16}$$

where each FPPB_i is associated with a specific threshold τ_i used to determine R_{pred}, the set of predicted lesion regions across all samples; R_{gt} denotes the corresponding set of ground-truth lesion regions. Therefore, a higher FROC-score indicates better lesion detection performance with fewer false positive regions.

D.2 Breast-Level (Classification) Metrics

Sensitivity under fixed specificity (Sens.) For opportunistic breast cancer analysis, maintaining a relatively high specificity is crucial to prevent excessive false positive alerts that could unnecessarily strain medical resources. Consequently, we selected a specificity value α (relatively high, while also avoiding excessively extreme classification score thresholds) for each dataset to enable fair and effective comparisons of breast-level cancer classification across different methods:

$$\text{Sens.} = \frac{\text{TP}}{\text{TP} + \text{FN}}\bigg|_{\text{Specificity}=\alpha} \tag{17}$$

Under this setting, the sensitivity values may vary across methods, reflecting their ability to detect true positive cases at the same high specificity level. A higher sensitivity suggests more effective identification of true positives without increasing the false alarm rate.

Area Under the receiver operating characteristic Curve (AUC) Apart from the sensitivity under high specificity, we also evaluate the overall classification performance using the AUC metric, defined as:

$$\text{AUC} = \int_0^1 \text{Sensitivity}(\beta)\, d\beta \tag{18}$$

where $\beta = 1 - \text{Specificity}$ represents the false positive rate of breast-level cancer classification ranging from 0 to 1, which differs from the definition of FPPB used in lesion-level metrics.

E Additional Experiments

More hyper-parameter tuning In addition to the ablation studies in the main paper, we also conducted experiments on several other key hyper-parameters, including the patch size in the D (slice of CT scans) dimension ($\mathrm{p_d}$), the embedding dimensions (C, C'), and the depth of the backbone (N).

As shown in Tab. 7, a smaller patch size in the D dimension, which better aligns with the larger spacing (3 mm vs. 0.748 mm in H and W dimensions), yields improved segmentation performance. However, this also leads to an increase in computational cost. To keep the computational cost (FLOPs) similar to the strong baseline (nnUNet [30]), $\mathrm{p_d} = 2$ is used as the default. For the ablation study on patch size in the H and W dimensions, please refer to the main paper Sec. 4.3.

As demonstrated in Tab. 8, increasing the High-Resolution Path's embedding dimension C significantly improves the voxel-level metric (Dice), while increasing the Low-Resolution Path's embedding dimension C' can markedly enhance the breast-level metrics (Sens. and AUC). To balance parameter count and computational overhead, a moderate configuration ($\mathrm{C} = 192$ and $\mathrm{C}' = 384$) is adopted.

According to Tab. 9, Dice and FROC-score are relatively sensitive to the depth of the backbone (N), with significant improvements observed as the network deepens. In contrast, the classification part reaches a near-optimal trade-off between complexity and performance when $N = 6$. At this setting, the model's parameter count (45.71M) and computational load (1162.65G FLOPs) are comparable to those of the strong baseline nnUNet [30], which has 31.00M parameters and 1185.24G FLOPs.

Comparison of other 1D sequence construction strategies with tandem input As described in the main text, the tandem input strategy first places the left and right 3D regions into the pseudo-3D volume and then flattens them into a 1D token sequence. This approach is actually equivalent to independently flattening the tokens of the left and right breasts along the prioritized axis and then concatenating them to form the input sequence. To further validate the superiority of this strategy, we compare it with three alternative variants: 1. last-axis-concat: Concatenate along the last (prioritized) axis (*i.e.*, D-, H-, W-axis for HWD, DWH, DHW scanning, respectively) before flattening. 2. W-axis-concat: Always concatenate along the W-axis before flattening. 3. interleaving: Interleave tokens from right and left breasts (*i.e.*, R1, L1, R2, L2, ... ; Rx/Lx denotes the x-th token in the flattened right/left sequence). Results are shown in Tab. 10.

Notably, our tandem input design ensures that all tokens from one side are processed before those from the other within the Mamba layer, and also maintains an approximately constant distance between anatomically symmetric locations, regardless of scanning order. This facilitates effective learning and comparison of bilateral features. In contrast, all alternative variants introduce some degree of mixing between left and right tokens, which leads to information confusion—most notably in the Interleaving variant, where the alternation of tokens from both sides results in the greatest disruption.

Bilateral input on CNN-based UNet In Sec. 4.3 of the main paper, we have attempted to apply the bilateral input concept to Vision Transformers [14] and verified the effectiveness of this idea, yet DRT-M3D achieves substantially better results. Here, we further attempt to apply the bilateral input form to CNN-based UNet-like networks, as also shown in Tab. 10. The "UNet-W-axis-concat" method concats bilateral breasts along the W axis as in the original NCCT images; the "UNet-channel-concat" method uses bilateral breast inputs as two channels, with the left breast flipped as a pseudo-right breast; the "UNet-CLS-head-merge" method processes unilateral breasts in the UNet, with features for classification concatenated, thereby mixing the bilateral information in the classification part.

CNNs are generally incapable of modeling long-range features, as their local receptive fields require stacking many convolutional layers to capture larger spatial extents, making them less effective than our proposed DRT-M3D. Thus, simply concatenating left and right views along the W-axis ("UNet-W-axis-concat") is not effective for explicit bilateral modeling. Although stacking these views as channels ("UNet-channel-concat") may seem like a reasonable alternative, breast symmetry is not strictly voxel-to-voxel—due to natural anatomical differences, positional variation, and imaging variability—so this approach can mix features from non-corresponding regions and ultimately introduce negative effects, leading to the worst performance. Moreover, simply concatenating features from the two views at the network head ("UNet-CLS-head-merge") does not allow the model to effectively exploit their spatial correspondence.

Comparison of model efficiency Tab. 12 provides the GPU memory usage, training time, and inference time for all major competing methods. The reported memory usage is measured per training sample, corresponding to one breast region. For bilateral methods, although each training sample contains two breast regions by design, we set the batch size to half that of their unilateral counterparts; this ensures that memory usage per breast region is fairly comparable across all methods. Training time reflects the sum of both pre-training and downstream fine-tuning phases (using two NVIDIA A100 80GB GPUs). Inference time is measured per patient (using a single NVIDIA A100 80GB GPU).

In addition to comparing the performance and computational cost of each method using GLOPs in the main paper, we assess the practical deployment efficiency by replacing FLOPs with inference time for each patient. The results are shown in Fig. 9.

Additional visualization results for qualitative comparison In Fig. 10,we present additional qualitative comparisons between the proposed DRT-M3D and competing methods, which serve as a supplement to the results shown in Fig. 4 of the main paper.

Delta visualization for the Tandem Input mechanism To better illustrate the advantages introduced by the Tandem (bilateral) Input mechanism in DRT-M3D, we visualize the Delta values ($\mathbf{\Delta}$, as described in Sec. A.1) of the Mamba-3D blocks in both DRT-M3D and DR-M3D. This allows a direct comparison of their responsiveness to different regions in the chest NCCT images, particularly along the low-resolution (LR) path. A larger step $\mathbf{\Delta}_t$ indicates a faster change in the hidden state $\mathbf{h}_t$, implying a greater influence of the current token $\mathbf{x}_t$ on the overall system dynamics. In this sense, $\mathbf{\Delta}_t$ can be viewed as a soft analog to attention weights in Transformers [59].

To better observe responses across all layers, we use a simplified setup with $N = 1$ for both models, meaning each high-resolution (HR) and low-resolution (LR) path contains a single Mamba-3D block composed of six internal Mamba layers (as detailed in Sec. A.2). We generate the Delta visualization by averaging first over all channels and then across the six layers.

The resulting visualizations are presented in Fig. 11. The Delta responses on the HR path mainly reflect the model's focus on local textures and anatomical details, which aligns with its primary role for the segmentation task. Meanwhile, the LR path is designed to focus on capturing bilateral context with shorter sequence length when Tandem Input mechanism is used (*i.e.*, in DRT-M3D), so the Delta visualization of the LR path shows clear cross-breast clues. This bilateral awareness enhances cancer classification reliability. Notably, the model's ability to focus on corresponding regions across both breasts is learned rather than derived from explicit coordinate symmetry, which is often lacking due to anatomical variation.

Results on contrast-enhanced CT data Here, we conduct additional experiments on contrast-enhanced CT (CECT) as the approximate upper bound of performance using NCCT data for this task. The data sources are as described in Sec. B. Since the corresponding CECT has been registered with NCCT, the segmentation labels used by both are exactly the same, with only the images being different. The intuitive differences between NCCT and CECT in breast lesions are shown in Fig. 8. The experimental results on internal and external datasets are presented in Tab. 11.

F Discussion of Broader Impacts

As discussed in the main contributions of this paper, the proposed approach aims to enhance opportunistic breast cancer analysis techniques while minimizing economic costs and radiation risks encountered by patients. This advancement holds the potential to provide earlier diagnosis and treatment options to a larger population of potential breast cancer patients, thereby significantly impacting their health outcomes.

Regarding potential negative impacts, it is important to note that this proposed method has not yet been deployed in practical healthcare systems. Nonetheless, it is foreseeable that unavoidable false positive cases may impose an additional burden on healthcare systems. Despite possible concerns, the anticipated benefits of early detection, classification, further examination, and subsequent treatment are expected to outweigh these drawbacks, resulting in an overall positive impact.

Table 7: **Ablation study on the patch size in the D (slice of CT scans) dimension.** All the experiments were conducted based on the default DRT-M3D model, with modifications applied to $\mathrm{p_d}$. The gray row represents the default settings. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Variants	Complexity		Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
$\mathrm{p_d}$, $\mathrm{p_h}$, $\mathrm{p_w}$	Params (M)	FLOPs (G)	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	31.00	1185.24	0.6259	0.9127	0.8548	0.7844	0.9536	0.5668	0.8889	0.8173	0.6286	0.8889
4, 4, 4	45.74	582.65	0.6517	0.9214	0.9065	0.8242	0.9737	0.5867	0.8975	0.8652	0.8309	0.9347
2, 4, 4	45.71	1162.65	0.6590	0.9229	0.9145	0.8578	0.9768	0.5948	0.9117	0.8766	0.8762	0.9371
1, 4, 4	45.70	2322.66	0.6651	0.9243	0.9159	0.8654	0.9729	0.6143	0.9211	0.9142	0.9002	0.9439

Table 8: **Ablation study on the embedding dimensions of two resolution paths.** All the experiments were conducted based on the default DRT-M3D model, with modifications applied to C for the High-Resolution Path and C' for the Low-Resolution Path. The gray row represents the default settings. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Variants		Complexity		Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
C	C'	Params (M)	FLOPs (G)	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]		31.00	1185.24	0.6259	0.9127	0.8548	0.7844	0.9536	0.5668	0.8889	0.8173	0.6286	0.8889
96	192	12.51	339.74	0.6404	0.9214	0.8767	0.8073	0.9674	0.5842	0.9093	0.8644	0.8202	0.9215
192	192	19.37	1075.07	0.6584	0.9083	0.9028	0.8073	0.9683	0.5920	0.9093	0.8653	0.8282	0.9319
192	384	45.71	1162.65	0.6590	0.9229	0.9145	0.8578	0.9768	0.5948	0.9117	0.8766	0.8762	0.9371
192	768	148.15	1498.35	0.6636	0.9258	0.9094	0.8624	0.9757	0.6015	0.9164	0.8795	0.8789	0.9414
384	768	175.76	4258.06	0.6716	0.9170	0.9127	0.8670	0.9771	0.6036	0.9235	0.8955	0.8762	0.9374

Table 9: **Ablation study on the number of backbone blocks.** All the experiments were conducted based on the default DRT-M3D model, with modifications applied to N. The gray row represents the default settings. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Variants	Complexity		Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
N	Params (M)	FLOPs (G)	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	31.00	1185.24	0.6259	0.9127	0.8548	0.7844	0.9536	0.5668	0.8889	0.8173	0.6286	0.8889
1	8.47	196.28	0.6382	0.9214	0.8963	0.8165	0.9677	0.5769	0.9069	0.8674	0.8415	0.9252
3	23.37	582.83	0.6483	0.9214	0.8996	0.8318	0.9718	0.5925	0.9093	0.8760	0.8602	0.9331
6	45.71	1162.65	0.6590	0.9229	0.9145	0.8578	0.9768	0.5948	0.9117	0.8766	0.8762	0.9371
9	68.06	1742.48	0.6615	0.9228	0.9126	0.8486	0.9744	0.5960	0.9022	0.8809	0.8621	0.9357
12	90.40	2322.30	0.6659	0.9243	0.9120	0.8440	0.9738	0.5974	0.9046	0.8866	0.8629	0.9373

Table 10: **Evaluation of several DRT-M3D variants for Tandem Input effectiveness, together with results of attempts using the bilateral input form on the CNN-based UNet.** Representative unilateral (nnUNet, DR-M3D) models are also provided for comparison. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Method	Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	0.6259	0.9127	0.8548	0.7844	0.9536	0.5668	0.8889	0.8173	0.6286	0.8889
DR-M3D (ours)	0.6564	0.9228	0.9027	0.8073	0.9690	0.5909	0.8974	0.8632	0.7207	0.9082
UNet-W-axis-concat	0.6225	0.8940	0.8597	0.7431	0.9469	0.5696	0.8974	0.8120	0.6190	0.8936
UNet-channel-concat	0.5888	0.8958	0.8400	0.7798	0.9498	0.5220	0.8461	0.7760	0.5143	0.8839
UNet-CLS-head-merge	0.6208	0.9114	0.8575	0.8211	0.9559	0.5643	0.9059	0.8057	0.7333	0.9053
DRT-M3D-last-axis-concat	0.6536	0.9127	0.9083	0.8440	0.9693	0.5808	0.9060	0.8729	0.8190	0.9206
DRT-M3D-W-axis-concat	0.6551	0.9039	0.9062	0.8532	0.9732	0.5923	0.9060	0.8708	0.8476	0.9258
DRT-M3D-interleaving	0.6414	0.9083	0.9018	0.8211	0.9650	0.5749	0.8974	0.8644	0.7810	0.9162
DRT-M3D (ours)	**0.6590**	**0.9229**	**0.9145**	**0.8578**	**0.9768**	**0.5948**	**0.9117**	**0.8766**	**0.8762**	**0.9371**

Table 11: **Results on the corresponding contrast-enhanced CT (CECT) scans as the approximate upper bound of performance using NCCT scans for this task.**. (H.r.: Hit-Rate, F.s.: FROC-Score, Spec.: Specificity, Sens.: Sensitivity)

Method	Internal (Spec. = 0.9627)					External (Spec. = 0.9083)				
	Dice	H.r.	F.s.	Sens.	AUC	Dice	H.r.	F.s.	Sens.	AUC
nnUNet [30]	0.7078	0.9563	0.9323	0.8257	0.9724	0.6860	0.9487	0.9068	0.9238	0.9288
DR-M3D (ours)	0.7252	0.9476	0.9436	0.9129	0.9736	0.6878	0.9345	0.9146	0.9302	0.9376
DRT-M3D (ours)	0.7263	0.9534	0.9479	0.9480	0.9795	0.7018	0.9544	0.9337	0.9365	0.9497

Table 12: Training memory usage, training time, and inference time for all major competing methods. Memory usage is measured per breast region. Training time sums pre-training and downstream fine-tuning (2×A100 GPUs), and inference time is per patient (1×A100 GPU).

Method	**Training Memory (GB)**	**Training Time (h)**	**Inference Time (s)**
nnUNet [30]	3.826	31.0	0.126
VNet [48]	3.489	28.9	0.112
swinUNETR [24]	11.975	36.8	0.139
nnFormer [69]	2.063	10.9	0.029
3D UX-Net [34]	8.754	67.0	0.251
MedNeXt [52]	12.095	43.1	0.194
U-Mamba-Bot [46]	6.105	44.4	0.196
U-Mamba-Enc [46]	11.810	77.9	0.291
EM-Net [6]	6.218	34.9	0.116
SegMamba [64]	9.580	66.2	0.318
ViT-S/8 [14]	4.611	22.1	0.034
T-ViT-S/8	4.629	26.9	0.047
DR-M3D/8 (ours)	2.565	14.6	0.049
DRT-M3D/8 (ours)	2.581	15.7	0.059
DR-M3D (ours)	9.187	30.4	0.143
DRT-M3D (ours)	9.202	34.2	0.171

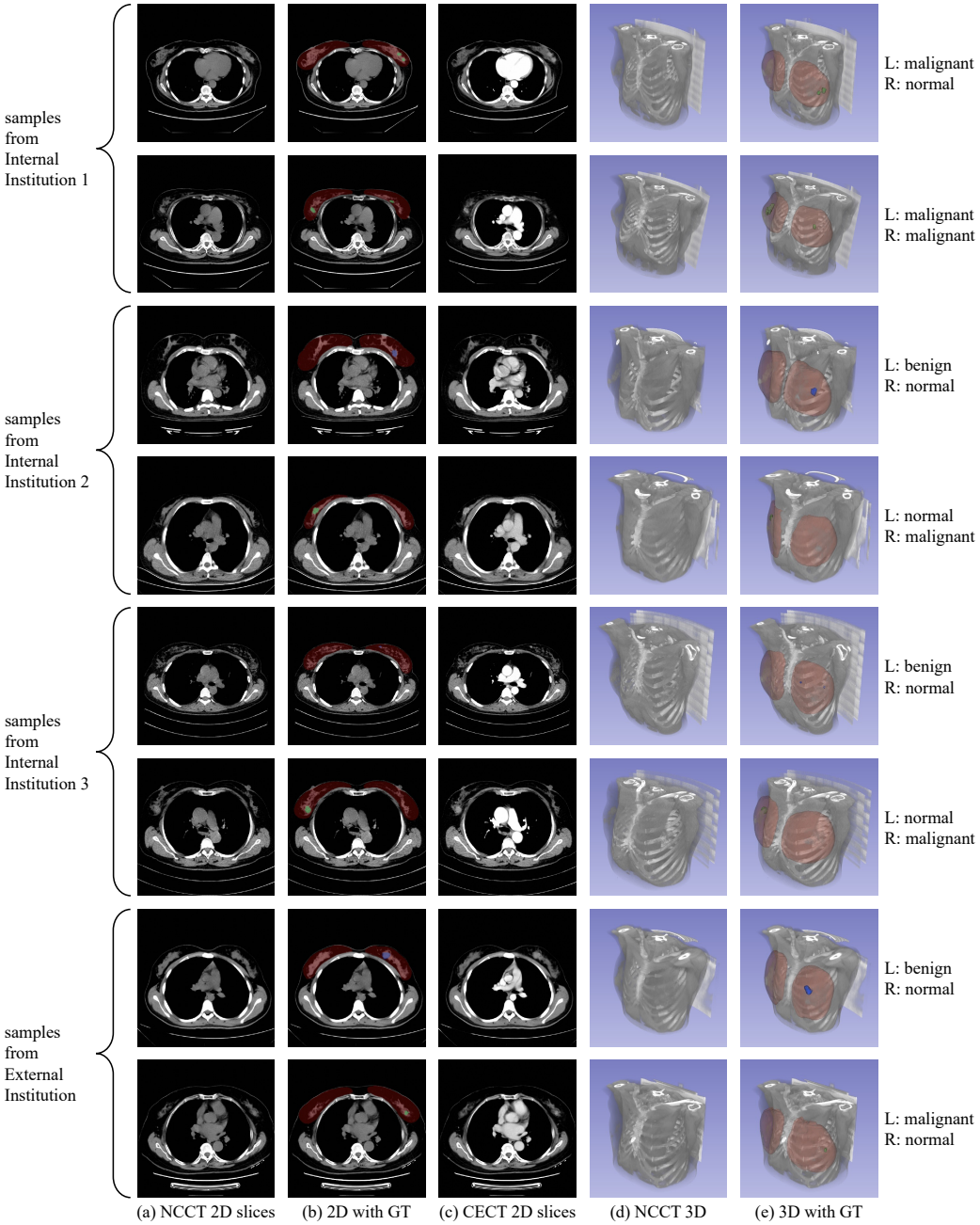


Figure 8: **Visualization of representative data samples from the four datasets used in this study.** (a), (b) and (c) respectively show typical NCCT 2D slices, their corresponding segmentation ground truth, and the corresponding CECT slices that have already been registered to NCCT. (d) and (e) present 3D renderings of the NCCT images and their segmentation ground truth. Green and blue indicate malignant and benign lesions, respectively. The breast-level classification labels are shown on the right side of each row. Note that the 2D slices are viewed in the foot-to-head direction, meaning the left side of each image corresponds to the right side of the human body.

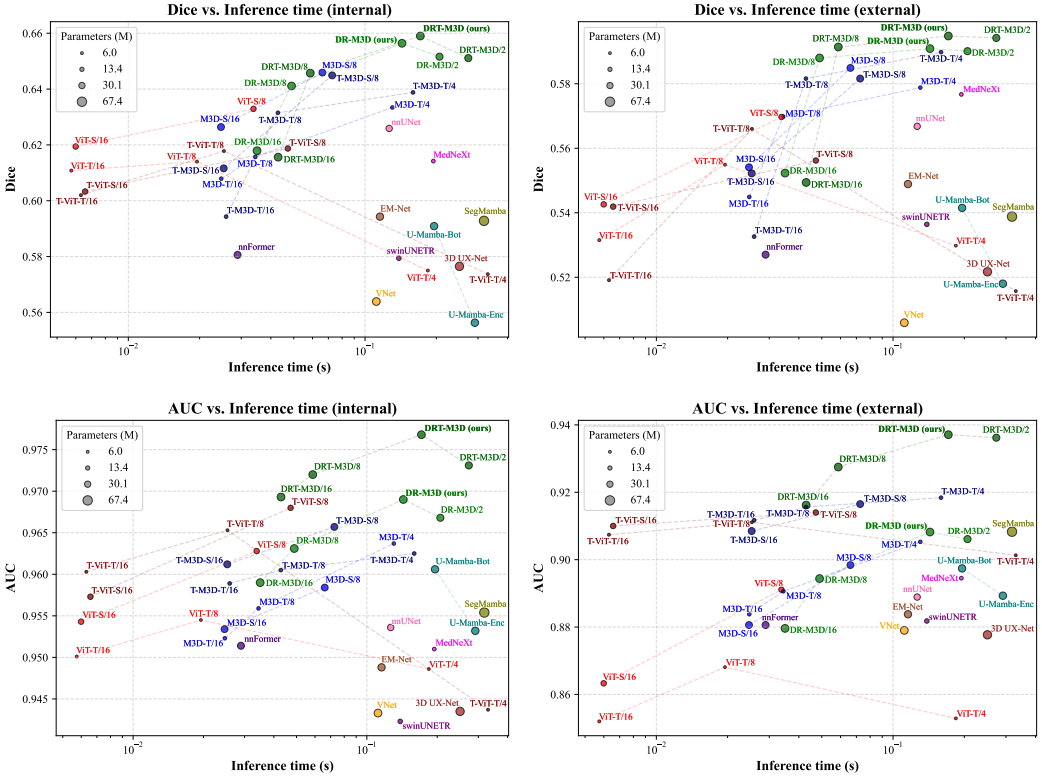


Figure 9: **Comparison of Dice and AUC vs. Inference time for internal and external datasets.** Each inference time is measured based on one bilateral (or two unilateral) input. ViT: vanilla Vision Transformer, M3D: vanilla Mamba-3D, T-~: tandem (bilateral) input form, ~/p: patch size of (2, p, p).

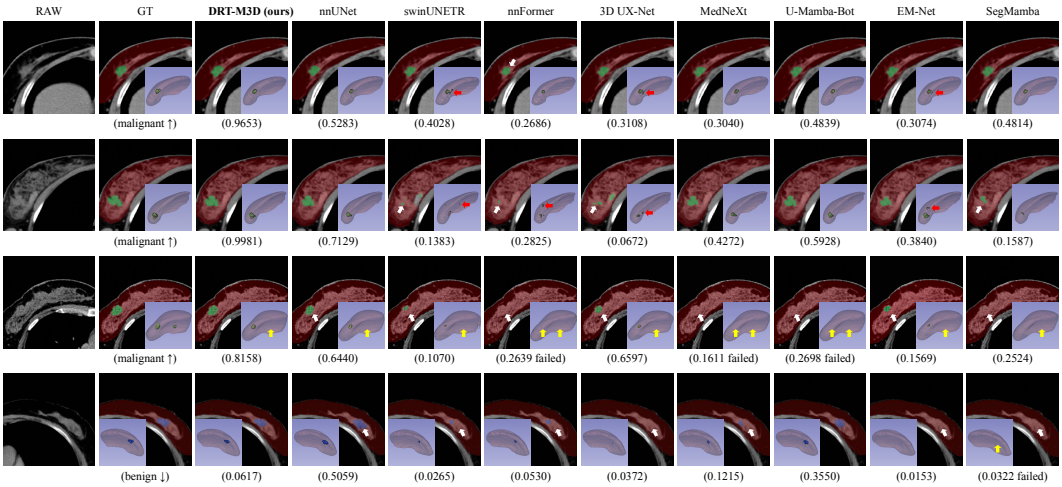


Figure 10: **More visualizations as a supplement to Fig. 4 in the main paper.** Red, green, and blue masks represent breast regions, malignant, and benign lesions, respectively. White, yellow, and red arrows mark pronounced segmentation errors in the slice, missed lesions, and segmentation of non-existent lesions, respectively. Predicted malignancy scores are shown for each case (higher for malignant, lower for benign). "Failed" denotes cases where none of the true lesions were localized (*i.e.*, a non-hit under the Hit-Rate metric).

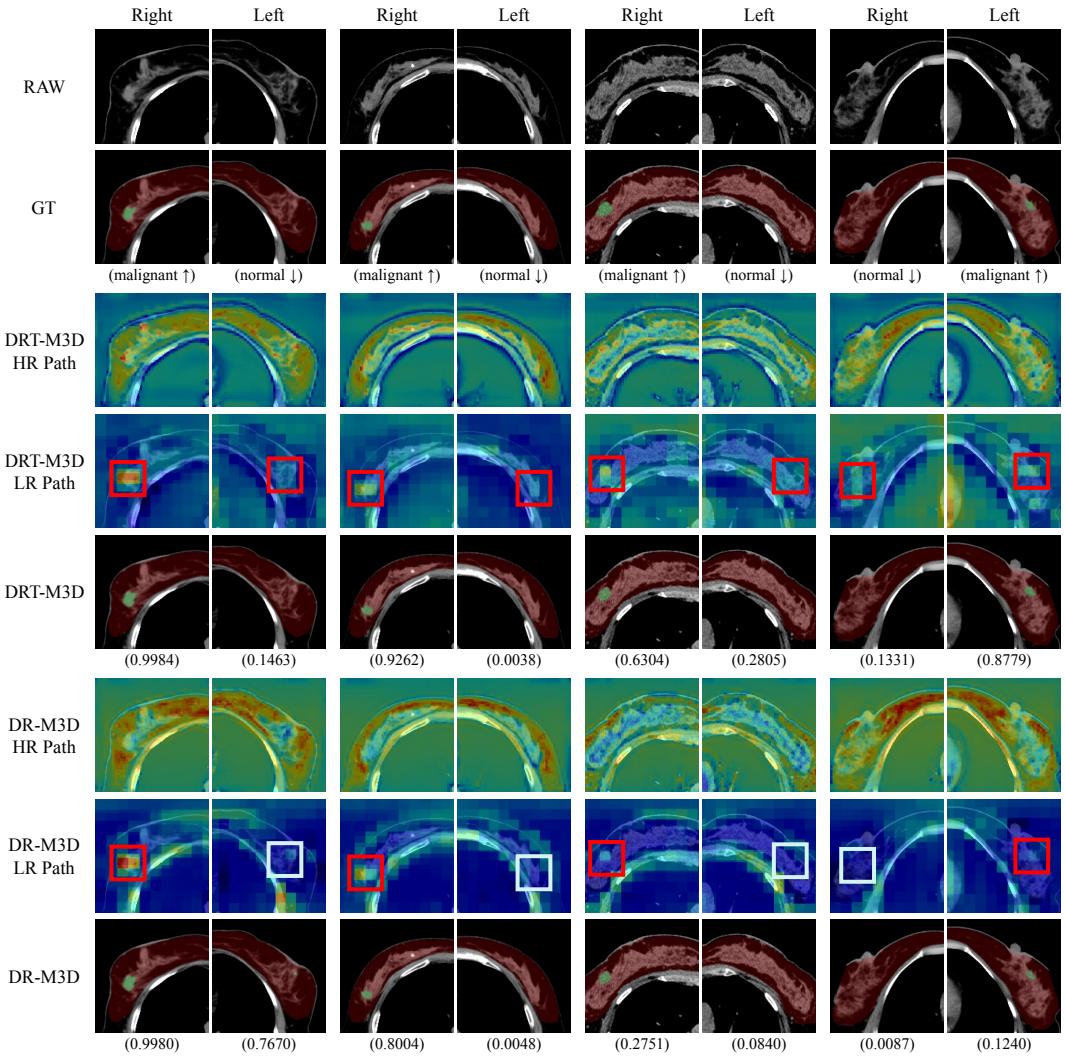


Figure 11: **Delta visualization of Mamba-3D blocks in DRT-M3D and DR-M3D.** Each column shows a test sample pair from three internal datasets and one external dataset; left and right images correspond to the right and left breast, respectively. Rows display (1) raw CT, (2) ground truth (red: breast, green: lesion, below: classification label), (3–4) Delta heatmaps of DRT-M3D (HR and LR paths), (5) DRT-M3D predictions, (6–7) Delta heatmaps of DR-M3D (HR and LR paths), and (8) DR-M3D predictions. The focus is on LR heatmaps: DRT-M3D, with the Tandem Input mechanism, highlights lesion areas and contralateral counterparts (red boxes), leading to superior classification results. In contrast, DR-M3D fundamentally lacks this ability to leverage contralateral information (light cyan boxes).