TRACE: Training-time Report-guided and Clinically Ordered Concept Editing

Wentao Yue[†]
Lanzhou University
Lanzhou, China
yuewt21@lzu.edu.cn

Tianyou Lai[†]
Lanzhou University
Lanzhou, China
laity21@lzu.edu.cn

Jiayu Luo
Lanzhou University
Lanzhou, China
luojy21@lzu.edu.cn

Qingyu Mao
Shenzhen University
Shenzhen, China
2150432008@email.szu.edu.cn

Ziying Wang
Southern Medical University
Guangzhou, China
ziyingwang@smu.edu.cn

Zhenyuan Ning[*]
Southern Medical University
Guangzhou, China
zhenyuan123@i.smu.edu.cn

Qilei Li[*]
Central China Normal University
Wuhan, China
qilei.li@ccnu.edu.cn

Abstract

Breast ultrasound diagnosis relies on clinically meaningful semantic concepts, yet most deep learning methods adopt end-to-end image-to-label paradigms that lack interpretability and robustness. While concept-based approaches offer a promising alternative, they often assume complete annotations or require multimodal inputs at inference, which significantly limits their real-world applicability. To tackle these issues, we propose Training-time Report-guided and Clinically Ordered Concept Editing (TRACE), a training-time report-guided framework that leverages structured radiology reports as privileged concept supervision while enabling image-only diagnosis at test time. TRACE refines image-derived concepts through a teacher-guided editing mechanism within a malignancy-aware ordered concept space. To address incomplete annotations, we introduce Strategic Concept Missing Training (SCMT) and train an image-only self-editor via edit distillation for autonomous concept refinement. Besides, we introduce BUSC, a concept-enriched benchmark linking images, labels, and structured attributes. Experiments across multiple datasets demonstrate that TRACE achieves superior performance and improved cross-domain robustness compared to existing methods. Code is available in our GitHub repository: `https://github.com/wentao-2/TRACE`

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Life and medical sciences**.

[*]Corresponding authors.
[†]These authors contributed equally to this work.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
`https://doi.org/10.1145/3767308.3836433`

Keywords

Concept Bottleneck Models; Missing Concept Training; Breast Ultrasound; Explainable Diagnosis; Asymmetric Testing

ACM Reference Format:
Wentao Yue, Tianyou Lai, Jiayu Luo, Qingyu Mao, Ziying Wang, Zhenyuan Ning, and Qilei Li. 2026. TRACE: Training-time Report-guided and Clinically Ordered Concept Editing. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 9 pages. `https://doi.org/10.1145/3767308.3836433`

1 Introduction

> *"The eye sees only what the mind is prepared to comprehend." — Robertson Davies*

In breast ultrasound diagnosis, reliable visual interpretation requires more than the detection of image cues. It also depends on clinically grounded concepts that organize visual evidence into diagnostic reasoning [5]. In clinical practice, radiologists do not make diagnoses solely from raw appearance patterns. Instead, they interpret lesions through structured semantic attributes such as shape, margin, orientation, posterior acoustic features, echogenicity, and calcification [16]. These concepts provide an intermediate abstraction between low-level image signals and final diagnostic decisions on benign or malignant status, which makes them highly valuable for robust and explainable breast ultrasound analysis [24].

Despite substantial progress in deep neural networks for breast ultrasound diagnosis, most existing methods still follow an end-to-end image-to-label paradigm. Strong image-only backbones such as ResNet50 [9] and ViT-B/16 [4] already achieve competitive performance, suggesting that current limitations may stem less from backbone capacity than from the lack of clinically grounded intermediate supervision [23]. Although such pipelines perform well in-domain, they often rely on shortcut visual correlations rather than clinically structured reasoning, limiting interpretability and robustness under distribution shifts or incomplete semantic supervision [13]. Concept-based learning offers a more clinically aligned

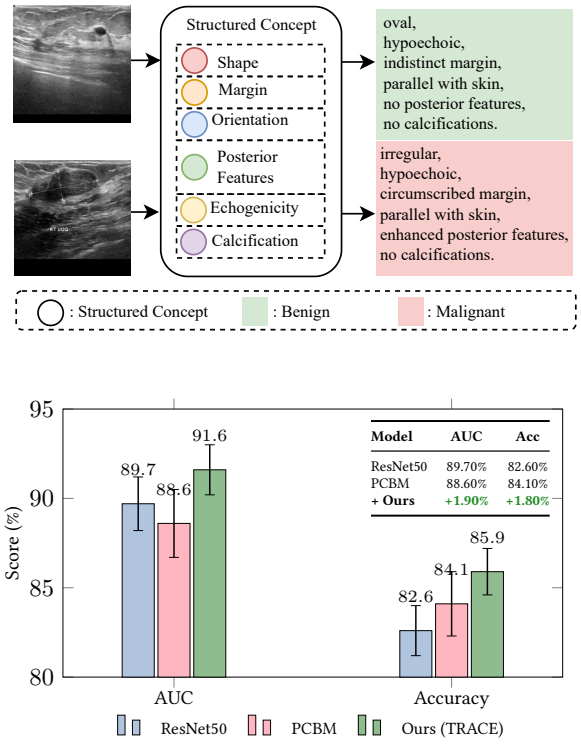

Figure 1: Overview of the structured concept representation used in BUSC-BUSBRA and performance comparison on the BUSC-BUSBRA dataset.

alternative by introducing semantically meaningful intermediate representations, as exemplified by VLG-CBM [19]. However, existing frameworks typically assume either complete concept supervision or multimodal inference with text available at test time. In practice, structured reports or concept annotations may be accessible during model development, yet are often unavailable, incomplete, or inconsistently documented in real deployment.

A related challenge is the lack of standardized breast ultrasound resources that align images, diagnostic labels, and structured clinical concepts. Most public datasets support image-level diagnosis or lesion segmentation, while concept annotations are often sparse, inconsistent, or unavailable, limiting systematic study of concept-guided diagnosis under incomplete concepts [2]. To address this gap, we construct the *Breast Ultrasound Structured Concept (BUSC)* dataset, a concept-enriched benchmark that associates breast ultrasound images with structured semantic attributes and diagnostic labels. BUSC includes two development subsets: *BUSC-BUSI647*, derived from the public BUSI [1] dataset by removing normal cases and retaining 647 lesion samples, and *BUSC-BUSBRA*, built upon the public BUSBRA [8] dataset. These subsets serve as the primary benchmarks for model development and in-domain evaluation.

These observations raise a practical question: *how can structured radiological reports be used during training to improve concept-level reasoning while supporting pure image-only diagnosis at test time?* A straightforward solution is to treat reports as an additional text modality and fuse them with image features [10, 26]. However, this formulation does not fully capture their clinical function. In breast ultrasound, report fields encode curated semantic judgments that can help correct imperfect image-derived concepts. We therefore treat structured reports as *privileged concept supervision* available only during training, rather than as an inference-time modality.

Inspired by these observations, we propose **TRACE**, a training-time report-guided and clinically ordered concept editing framework for robust breast ultrasound diagnosis under incomplete concepts. TRACE follows an asymmetric training–test paradigm: it uses breast ultrasound images together with report-derived concept annotations during training, but relies on images only at inference. Specifically, an image encoder first produces coarse concept predictions, after which a privileged concept teacher guides a teacher-driven editor to revise them. To better reflect clinical reasoning, TRACE organizes semantic attributes in a malignancy-aware ordered concept space, modeling diagnosis as clinically meaningful concept editing rather than flat independent classification.

To further handle incomplete concepts, TRACE introduces *Strategic Concept Missing Training (SCMT)*, which masks report concepts during training according to clinically informed missing patterns. This encourages compensatory reasoning across correlated attributes instead of over-reliance on any single concept. Since reports are unavailable at test time, TRACE further trains an image-only self-editor via edit distillation to mimic the privileged teacher editor, enabling concept refinement and fully image-only diagnosis.

We evaluate TRACE on the two BUSC development subsets, *BUSC-BUSBRA* and *BUSC-BUSI647*, and further assess zero-shot cross-domain generalization on three external datasets: *Ardakani*, *BUS_UC*, and *BrEaST*. Compared with standard visual backbones, concept bottleneck models, prototype-based methods, explainable medical classifiers, and vision-language models, TRACE consistently achieves superior in-domain performance and strong cross-domain robustness. These results show that learning to revise clinically meaningful concepts from training-time report supervision is more effective than relying on either end-to-end visual prediction or inference-time multimodal assistance.

In summary, the main contributions of this work are as follows:

- We construct the **Breast Ultrasound Structured Concept (BUSC)** dataset, a concept-enriched benchmark for breast ultrasound that explicitly links images, diagnostic labels, and structured clinical concepts. BUSC contains two development subsets, *BUSC-BUSI647* and *-BUSBRA*, and provides a practical basis for the study of concept-guided diagnosis and robustness under incomplete concept supervision.
- We propose **TRACE**, a training-time report-guided framework that treats structured radiology reports as a privileged concept teacher, thereby enabling stronger image-only diagnosis of breast ultrasound while requiring no reports at inference time.
- We introduce **clinically ordered concept editing**, which represents semantic attributes in a malignancy-aware ordered concept space and explicitly learns how report supervision refines coarse concepts derived from images.

- We develop **Strategic Concept Missing Training (SCMT)** together with an image-only self-editor and edit distillation, which improves robustness to incomplete concepts and supports generalizable image-only deployment across datasets.

2 Related Work

Structured Concepts in Breast Ultrasound. Breast ultrasound diagnosis is naturally grounded in structured clinical descriptions. BI-RADS-related attributes, such as shape, margin, orientation, posterior acoustic features, and echogenicity, are routinely used for lesion assessment and report writing [18]. Recent studies have incorporated such structured descriptions into deep learning models to improve interpretability and clinical readability. For instance, BI-RADS-Net [28] jointly predicts diagnostic labels and BI-RADS attributes via multi-task learning, while BI-RADS-Net-V2 [29] further integrates semantic and quantitative explanations for more interpretable computer-aided diagnosis. Other works also build unified pipelines for detection, description, and classification based on expert-annotated BI-RADS terminology, highlighting the clinical value of structured attribute supervision.

However, most existing methods assume structured attributes or report information are available during evaluation, or rely on complete concept annotations within the same domain [29] [10]. In practice, breast ultrasound datasets often provide only images and class labels, without standardized concept annotations. Recent multimodal approaches incorporate radiology reports through image–text fusion, but still require reports as inference-time inputs [22]. In contrast, we consider a setting where structured reports are only partially available during training, while testing and external generalization rely solely on images.

Concept Bottleneck Models. Concept Bottleneck Models (CBMs) introduce a human-interpretable concept layer between perception and prediction, enabling concept-based explanations and human intervention [17]. The seminal work [12] of Koh et al. establishes the CBM paradigm and shows that such models maintain competitive performance while supporting concept-level interpretability.

Recent medical interpretable models further improve concept learning from different perspectives. Explicd [7] learns interpretable concept representations by aligning images with textual diagnostic criteria provided by experts or large language models. VLG-CBM [19] enhances concept faithfulness through vision-language guidance and grounded detectors. MVP-CBM [21] improves concept bottleneck representations by modeling the preference of different concepts for features from different visual layers. Overall, these methods mainly focus on improving concept acquisition, concept–feature correspondence, or concept representation quality. In contrast, our goal is not simply to build a stronger CBM, but to address a practical breast ultrasound setting where structured reports are available only during training as privileged concept teachers, while diagnosis at test time relies on images alone.

Privileged Supervision. From a broader machine learning perspective, our problem is related to learning with additional information available during training but unavailable at test time. The Learning Using Privileged Information (LUPI) framework [20] proposed by Vapnik et al. shows that models can exploit extra teacher information during training while relying only on standard inputs at inference time. Subsequent work on generalized distillation further unifies knowledge distillation and privileged-information learning within a common framework.

This idea has also been explored in medical imaging [3]. For example, Yang et al. improve cross-modality image registration by introducing a third modality visible only during training [25], and Proto-Caps [6] combines prototype learning with privileged information for medical image classification. However, these methods typically treat extra information as an auxiliary source for distillation, a prototype-level constraint, or an extra modality. In contrast, we explicitly model structured breast ultrasound reports as concept-level teacher signals, aiming to learn a concept correction and diagnosis mechanism that transfers to image-only deployment.

3 Methodology

3.1 Problem Setting and Method Overview

We study an asymmetric clinical setting for breast ultrasound classification. Given a breast ultrasound image $x \in \mathcal{X}$, its class label is denoted by $y \in \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$ represents benign and malignant cases, respectively. For a subset of training samples, an expert-annotated structured concept vector is also available,

$$c = [c^{(1)}, c^{(2)}, \dots, c^{(M)}] \in \mathcal{C}, \tag{1}$$

where M denotes the number of concepts. Each concept dimension corresponds to a clinically meaningful BI-RADS-related attribute in breast ultrasound, such as shape, margin, orientation and calcification. Unlike multimodal methods that treat structured text as an inference-time input modality, we model it as a privileged concept teacher accessible only during training, while test-time diagnosis and external generalization rely on images alone.

Formally, the training set consists of two parts:

$$\mathcal{D}_s = \{(x_i, c_i, y_i)\}_{i=1}^{N_s}, \qquad \mathcal{D}_w = \{(x_j, y_j)\}_{j=1}^{N_w}, \tag{2}$$

where $\mathcal{D}_s$ denotes the sample set with structured concept supervision, and $\mathcal{D}_w$ denotes the weakly supervised sample set with only images and class labels. The test set is denoted by

$$\mathcal{D}_{\text{test}} = \{(x_k, y_k)\}_{k=1}^{N_t}, \tag{3}$$

where no structured concepts or report information are provided. Our goal is to learn a mapping

$$f : \mathcal{X} \to \mathcal{Y}, \tag{4}$$

which uses the limited structured concept supervision in $\mathcal{D}_s$ during training while still enabling robust classification and concept reasoning from images alone at deployment.

The core idea of TRACE is to learn concept correction rather than concept prediction alone. Specifically, the model extracts visual representations and produces initial image-derived concepts, then uses structured reports during training as concept-level teacher signals to learn a correction mapping from coarse to more reliable concepts. At test time, the learned correction behavior is applied using images only, enabling report-free deployment. The overall

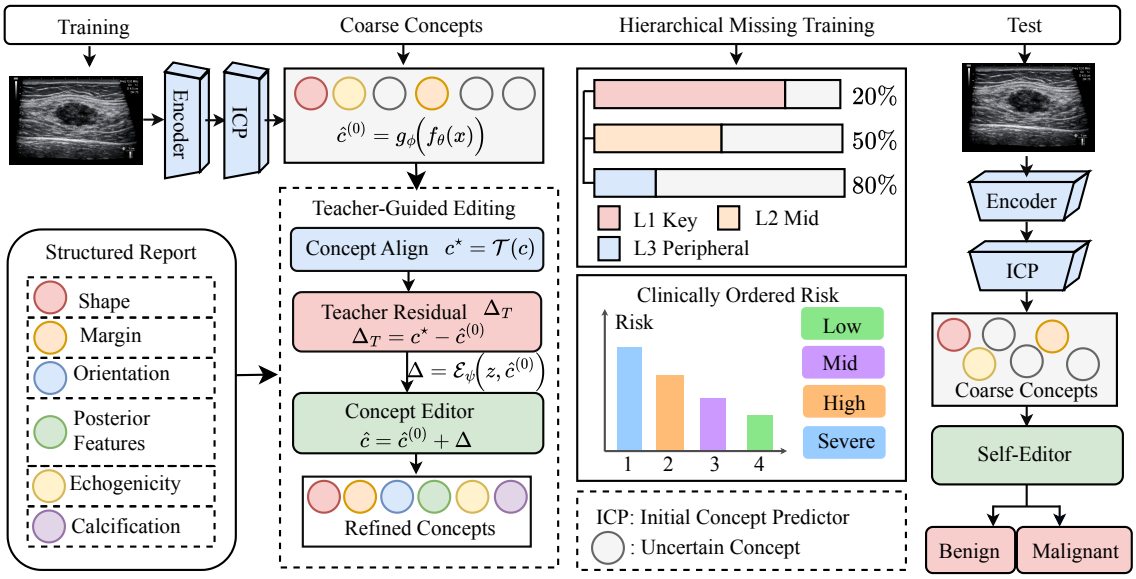

Figure 2: Overview of TRACE. During training, structured reports act as privileged concept teachers to correct coarse image-derived concepts through residual concept editing. TRACE further improves robustness with hierarchical concept missing training and a clinically ordered risk space. At test time, the report branch is removed, and diagnosis is made from images alone via self-editing concept refinement.

pipeline is formulated as

$$z = f_\theta(x), \qquad \hat{c}^{(0)} = g_\phi(z), \qquad \hat{c} = \mathcal{E}(z, \hat{c}^{(0)}), \qquad \hat{y} = h_\omega(\hat{c}), \tag{5}$$

where f_θ is the image encoder, g_ϕ is the initial concept predictor, $\mathcal{E}$ is the concept editor, and h_ω is the final classifier. During training, $\mathcal{E}$ learns concept correction under the guidance of the privileged concept teacher. At test time, $\mathcal{E}$ reduces to a self-editor that depends only on the image and the initial concepts.

3.2 Report-Guided Concept Editing

Given an input image x, the image encoder f_θ extracts a visual feature and the initial concept predictor g_ϕ produces a coarse concept representation:

$$\begin{aligned} z &= f_\theta(x) \in \mathbb{R}^d, \\ \hat{c}^{(0)} &= g_\phi(z) = \left[\hat{c}^{(0,1)}, \hat{c}^{(0,2)}, \ldots, \hat{c}^{(0,M)}\right]. \end{aligned} \tag{6}$$

For discrete concepts, $\hat{c}^{(0,m)}$ can be represented as class logits or probability distributions. For ordered concepts, $\hat{c}^{(0,m)}$ can also be represented as ordered real-valued scores or ordered embeddings. Because image appearance is affected by noise, device variation, and domain shift, concepts predicted directly by g_ϕ are often coarse and unstable, motivating subsequent concept-level correction.

For samples with structured annotations $(x, c, y) \in \mathcal{D}_s$, we do not treat c as a test-time input modality. Instead, we regard it as a concept-level teacher signal available only during training. We first transform the structured concept annotation into the unified concept space and compute the residual correction target:

$$c^T = \mathcal{T}(c), \qquad \Delta_T = c^T - \hat{c}^{(0)}, \tag{7}$$

where $\mathcal{T}(\cdot)$ maps the original structured concept annotation into the unified concept representation space. The teacher concept c^T specifies the desired semantic state, while Δ_T represents the correction direction and magnitude required to refine the initial image-derived concepts. Therefore, TRACE learns to estimate concept corrections from structured clinical knowledge rather than directly predicting a new set of concepts from scratch.

We introduce a concept editor $\mathcal{E}_\psi$, which takes the visual feature z and concept $\hat{c}^{(0)}$ as input and outputs a concept correction term:

$$\Delta_\psi = \mathcal{E}_\psi(z, \hat{c}^{(0)}), \qquad \hat{c} = \hat{c}^{(0)} + \Delta_\psi. \tag{8}$$

During training, for samples with structured concept supervision, we minimize the discrepancy between Δ_ψ and the teacher residual Δ_T, so that the editor learns how to correct rather than merely where to end up. Accordingly, the editing constraint is defined as

$$\mathcal{L}_{\text{edit}} = \frac{1}{M} \sum_{m=1}^{M} \ell_{\text{edit}}\left(\Delta_\psi^{(m)}, \Delta_T^{(m)}\right), \tag{9}$$

where $\ell_{\text{edit}}(\cdot, \cdot)$ denotes the concept-dependent editing loss.

Several concepts in breast ultrasound are not independent flat discrete labels, but follow clear clinical risk order. For example, concepts such as margin and shape correspond to different levels of malignancy risk. We therefore introduce clinically ordered constraints into the concept space so that the editing direction is more consistent with medical knowledge rather than arbitrary label switching. For the m-th ordered concept, let its ordinal level be

$$r^{(m)} \in \{1, 2, \ldots, K_m\}, \tag{10}$$

where K_m denotes the number of ordinal levels for that concept, and let the predicted ordinal score be $s^{(m)}$. We then define the ordinal loss as

$$\mathcal{L}_{\text{ord}} = \sum_{m \in \mathcal{M}_{\text{ord}}} \ell_{\text{ord}}\big(s^{(m)}, r^{(m)}\big), \tag{11}$$

where $\mathcal{M}_{\text{ord}}$ denotes the set of all concepts with clinical risk order. This constraint encourages concept correction to follow clinically plausible directions and reduces label transitions that are inconsistent with medical knowledge.

The final edited concept representation is denoted as $\hat{c}$ and used for downstream diagnosis:

$$\hat{y} = h_\omega(\hat{c}). \tag{12}$$

3.3 Hierarchical Concept Missing Training

Structured concepts in real clinical reports are unevenly available, and the missing probabilities vary substantially across attributes. Rather than simple random masking, we adopt a hierarchical concept missing training strategy to simulate supervision sparsity better reflecting real report patterns. Specifically, we divide all concepts into three levels by importance and clinical availability:

$$\mathcal{C} = \mathcal{C}_{L1} \cup \mathcal{C}_{L2} \cup \mathcal{C}_{L3}, \tag{13}$$

where $\mathcal{C}_{L1}$, $\mathcal{C}_{L2}$, and $\mathcal{C}_{L3}$ denote the key, mid-level, and peripheral concept sets, respectively. In our setting, the hierarchy is

$$\begin{aligned} \mathcal{C}_{L1} &= \{\text{margin}, \text{shape}\}, \\ \mathcal{C}_{L2} &= \{\text{orientation}, \text{posterior}, \text{echogenicity}\}, \\ \mathcal{C}_{L3} &= \{\text{calcification}\}. \end{aligned} \tag{14}$$

For concepts at level L, we define the missing rate as

$$\begin{aligned} r_L &= 0.2 + 0.3(L-1), \qquad L \in \{1,2,3\}, \\ r_1 &= 0.2, \qquad r_2 = 0.5, \qquad r_3 = 0.8. \end{aligned} \tag{15}$$

Thus, key concepts are more likely to be preserved, while peripheral concepts are more likely to be missing. For concept $c^{(m)}$, we define a visibility mask

$$b^{(m)} \sim \text{Bernoulli}\big(1 - r_{L_m}\big), \tag{16}$$

where L_m denotes the level of the m-th concept. The masked teacher concept is

$$\tilde{c}^{(m)} = b^{(m)} c^{(m)} + \big(1 - b^{(m)}\big)\emptyset, \tag{17}$$

where $\emptyset$ denotes a missing concept. With the full mask vector

$$b = [b^{(1)}, b^{(2)}, \ldots, b^{(M)}], \qquad \tilde{c} = b \odot c + (1-b) \odot \emptyset. \tag{18}$$

During training, the editor accesses only unmasked teacher signals, so the editing target becomes

$$\Delta_T^{\text{mask}} = b \odot \Delta_T. \tag{19}$$

The mask-aware editing loss is

$$\mathcal{L}_{\text{mask\_edit}} = \frac{1}{\sum_{m=1}^{M} b^{(m)} + \epsilon} \sum_{m=1}^{M} b^{(m)} \ell_{\text{edit}}\Big(\Delta_\psi^{(m)}, \Delta_T^{(m)}\Big), \tag{20}$$

where ϵ is a small constant to prevent division by zero. This design forces the model to learn cross-concept dependency and redundancy compensation under incomplete concept supervision, reducing the gap between training with visible structured reports and testing without reports.

3.4 Inference and Training Objectives

At test time, structured reports are unavailable and the teacher branch is removed. The model performs self-editing using only image features and initial concepts. Given a test image x, we compute

$$z = f_\theta(x), \qquad \hat{c}^{(0)} = g_\phi(z), \qquad \Delta_S = \mathcal{E}_\psi\big(z, \hat{c}^{(0)}\big), \tag{21}$$

which gives final concept representation and classification output

$$\hat{c} = \hat{c}^{(0)} + \Delta_S, \qquad \hat{y} = h_\omega(\hat{c}). \tag{22}$$

Thus, the goal of TRACE is not to recover structured reports at test time, but to transfer the correction behavior learned from the privileged concept teacher during training into report-free concept refinement.

For the main classification task, we use the supervised loss

$$\mathcal{L}_{\text{cls}} = \ell_{\text{cls}}(\hat{y}, y), \tag{23}$$

where ℓ_{cls} is typically the binary cross-entropy loss.

For training samples with concept annotations, the initial concept predictor is also trained to fit the teacher concepts. We define the initial concept supervision loss as

$$\mathcal{L}_{\text{init}} = \frac{1}{M} \sum_{m=1}^{M} \ell_{\text{con}}\Big(\hat{c}^{(0,m)},\, c^{\star(m)}\Big). \tag{24}$$

This term keeps the initial image-derived concepts close to the clinical concept space, providing a reasonable starting point for subsequent editing.

For samples with structured concept supervision, we use the previously defined $\mathcal{L}_{\text{mask\_edit}}$ to supervise the editing residual. To reduce the mismatch between teacher-guided editing during training and self-editing at test time, we also introduce a post-editing consistency constraint

$$\mathcal{L}_{\text{cons}} = \frac{1}{M} \sum_{m=1}^{M} \ell_{\text{con}}\Big(\hat{c}^{(m)},\, \tilde{c}^{(m)}\Big), \tag{25}$$

where the loss is computed only on currently visible teacher concept dimensions. This term encourages the self-edited concept state to approach the target concept state induced by teacher supervision.

Combining all terms above, the training objective of TRACE is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{init}} + \lambda_2 \mathcal{L}_{\text{mask\_edit}} + \lambda_3 \mathcal{L}_{\text{ord}} + \lambda_4 \mathcal{L}_{\text{cons}}, \tag{26}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ are balancing coefficients. For samples without concept supervision, namely $(x, y) \in \mathcal{D}_w$, the model optimizes only $\mathcal{L}_{\text{cls}}$. For samples with concept supervision, namely $(x, c, y) \in \mathcal{D}_s$, it jointly optimizes the concept-related terms. In this way, TRACE learns a transferable concept correction mechanism from limited structured concept supervision and enables robust breast ultrasound diagnosis from images alone at test time.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets. We evaluate TRACE on five breast ultrasound datasets, including two subsets from our **Breast Ultrasound Structured Concept (BUSC)** dataset and three external datasets for cross-domain evaluation. Specifically, we augment the public BUSBRA

Table 1: Main comparison on BUSC-BUSBRA and BUSC-BUSI647. We compare TRACE with eight representative baselines. Best and second-best results in each column are highlighted in red and blue, respectively.

Setting	Model	BUSC-BUSBRA			BUSC-BUSI647		
		AUC	Acc	F1	AUC	Acc	F1
Black-box	ResNet50 [9]	0.897±0.015	0.826±0.014	0.725±0.027	0.945±0.027	0.901±0.026	0.838±0.047
	ViT-B/16 [4]	0.876±0.018	0.821±0.025	0.711±0.042	0.944±0.021	0.886±0.035	0.809±0.068
Prototype	ProtoCaps [6]	0.727±0.037	0.718±0.024	0.391±0.075	0.803±0.024	0.807±0.014	0.658±0.052
CBM	PCBM [27]	0.886±0.019	0.841±0.018	0.738±0.031	0.951±0.022	0.901±0.025	0.837±0.047
	MVP-CBM [21]	0.889±0.020	0.830±0.021	0.749±0.027	0.931±0.027	0.878±0.045	0.806±0.069
	VLG-CBM [19]	0.883±0.018	0.837±0.024	0.733±0.037	0.941±0.028	0.910±0.018	0.857±0.029
Explainable	Explicd [7]	0.891±0.025	0.833±0.027	0.730±0.050	0.939±0.029	0.884±0.016	0.822±0.034
Foundation VLM	CLIP [15]	0.553±0.033	0.662±0.011	0.111±0.042	0.536±0.020	0.674±0.010	0.018±0.036
Ours	**TRACE**	0.916±0.014	0.859±0.013	0.776±0.026	0.970±0.030	0.938±0.057	0.902±0.083

[8] and BUSI [1] datasets with expert-annotated structured semantic concepts to construct BUSC, which aligns each ultrasound image with diagnostic labels and structured concepts. **BUSC-BUSBRA**, built upon BUSBRA, contains 1,872 cases annotated with six semantic concepts: *shape*, *margin*, *orientation*, *posterior acoustic features*, *echo pattern*, and *calcification*. **BUSC-BUSI647**, derived from BUSI after removing normal samples, contains 647 lesion cases annotated with the same concept schema. These subsets are used for in-domain evaluation.

For external validation, we adopt three public breast ultrasound datasets from different acquisition sources and populations. **Ardakani** [2] contains 232 pathologically confirmed lesions from Iran, including 123 malignant and 109 benign cases. **BUS_UC** [11] contains 811 images collected in Pakistan, including 358 benign and 453 malignant cases. **BrEaST** [14] is a curated dataset from Poland containing 256 scans with benign, malignant, and a small number of normal cases.

For cross-domain evaluation, we remove structured reports and concept annotations from **BUSC-BUSI647** and denote the resulting dataset as **BUSC-BUSI647***. Together with **Ardakani**, **BUS_UC**, and **BrEaST**, it is treated as a target domain without concept supervision. This setup reflects realistic deployment, where structured reports and concept annotations are available only during training, while inference follows an image-only protocol.

4.1.2 Compared Methods. We compare TRACE with eight representative baselines. ResNet50 [9] is used as a standard CNN baseline for image-based classification without concept supervision. ViT-B/16 [4] represents transformer-based image classification. ProtoCaps [6] is included as a prototype-based medical image classifier. PCBM [27] serves as a standard concept bottleneck baseline, while MVP-CBM [21] represents a stronger CBM variant with multi-view prototype learning. VLG-CBM [19] is adopted as a vision-language-guided concept bottleneck model, and Explicd [7] is included as an explainable medical image classifier. CLIP [15] serves as a representative foundation vision-language model based on contrastive pretraining.

4.1.3 Implementation Details. To ensure fair comparison, all BUSC images are processed with **zero-padding**, where black borders are added at the bottom to standardize spatial resolution while preserving the original aspect ratio. Unless otherwise specified, the image encoder is initialized with ImageNet-pretrained weights. For in-domain experiments, we adopt five-fold cross-validation and report the mean and standard deviation over the five folds. During training, TRACE uses ultrasound images with concept supervision derived from structured reports, while inference uses images only. The concept editor is optimized jointly with the image encoder and diagnosis head, and SCMT is applied during training to simulate incomplete concept conditions.

4.1.4 Evaluation Metrics. We report standard classification metrics for breast ultrasound diagnosis, including AUC, Accuracy, and F1-score. For in-domain evaluation, all methods are compared on BUSC-BUSBRA and -BUSI647 under the unified image-only inference setting. For external evaluation, models trained on the BUSC subsets are directly transferred to Ardakani, BUS_UC, BrEaST, and BUSI647* to assess zero-shot cross-domain generalization.

4.2 Main Results

Table 1 summarizes the comparison between TRACE and eight representative baselines on BUSC-BUSBRA and -BUSI647. All methods are evaluated under a unified **image-only** test protocol, where structured reports and concept labels are used only during training. TRACE achieves the best results on both datasets, demonstrating the effectiveness of using structured reports as privileged concept teachers for image-only breast ultrasound diagnosis.

On BUSC-BUSBRA, TRACE achieves the best results on AUC, Acc, and F1, outperforming both image-only baselines such as ResNet50 and concept-based methods including PCBM, MVP-CBM, and Explicd. The advantage of TRACE is further evident on BUSC-BUSI647, where it again achieves the best results on all three metrics with

Table 2: 5-fold cross-domain zero-shot performance (*mean*±*std*). Best and second-best results are highlighted in red and blue.

Model	Ardakani		BUS_UC		BrEaST		BUSC-BUSI647*	
	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc
ResNet50 [9]	0.857±0.013	0.794±0.025	0.525±0.074	0.470±0.053	0.508±0.075	0.515±0.125	0.873±0.007	0.819±0.016
ViT-B16 [4]	0.844±0.016	0.825±0.017	0.657±0.030	0.622±0.026	0.771±0.019	0.725±0.029	0.847±0.015	0.813±0.012
ProtoCaps [6]	0.756±0.015	0.792±0.016	0.538±0.025	0.518±0.024	0.671±0.023	0.655±0.008	0.764±0.015	0.764±0.010
PCBM [27]	0.851±0.015	0.808±0.051	0.653±0.042	0.589±0.051	0.779±0.111	0.691±0.096	0.864±0.010	0.826±0.015
MVP-CBM [21]	0.863±0.008	0.776±0.027	0.647±0.032	0.591±0.036	0.769±0.104	0.664±0.118	0.871±0.007	0.806±0.016
VLG-CBM [19]	0.845±0.016	0.800±0.040	0.651±0.012	0.593±0.016	0.779±0.058	0.711±0.071	0.867±0.010	0.825±0.010
Explicd [7]	0.850±0.012	0.793±0.020	0.635±0.044	0.587±0.037	0.785±0.100	0.715±0.121	0.872±0.007	0.815±0.009
CLIP [15]	0.621±0.032	0.733±0.013	0.472±0.021	0.445±0.055	0.661±0.094	0.621±0.046	0.5357±0.020	0.674±0.010
TRACE (Ours)	0.862±0.018	0.836±0.046	0.661±0.027	0.604±0.023	0.822±0.019	0.757±0.023	0.873±0.023	0.857±0.017

Table 3: Cross-domain comparison of different TRACE editors. Models are trained on BUSBRA and tested on external datasets without target-domain fine-tuning. Best and second-best results are highlighted in red and blue.

Target	Editor	AUC	Acc
Ardakani	TRACE-MLP	0.862±0.011	0.833±0.026
	TRACE-Att	0.862±0.018	0.836±0.046
	TRACE-Gate	0.861±0.014	0.853±0.017
BUS_UC	TRACE-MLP	0.660±0.011	0.614±0.008
	TRACE-Att	0.661±0.027	0.604±0.023
	TRACE-Gate	0.645±0.031	0.589±0.018
BrEaST	TRACE-MLP	0.813±0.015	0.741±0.031
	TRACE-Att	0.822±0.019	0.757±0.023
	TRACE-Gate	0.819±0.021	0.748±0.021
BUSI647*	TRACE-MLP	0.8736±0.017	0.8496±0.029
	TRACE-Att	0.8733±0.023	0.8574±0.017
	TRACE-Gate	0.8711±0.018	0.8651±0.030

clear improvements over competing methods. These results indicate that TRACE effectively exploits structured semantic supervision during training and transfers it into stronger image-only concept refinement at test time.

4.3 Cross-Domain Generalization

Table 2 reports zero-shot cross-domain results, where models are trained on BUSC-BUSBRA and directly evaluated on external target domains without fine-tuning. All target datasets, including BUSC-BUSI647*, Ardakani, BUS_UC, and BrEaST, are tested without structured reports, making this setting a stricter evaluation of transferability under realistic image-only deployment.

Overall, TRACE shows strong cross-domain generalization and achieves the best or near-best performance on multiple target datasets. In particular, TRACE obtains the best Accuracy on Ardakani, the best AUC on BUS_UC, and the best AUC and Accuracy on BrEaST. It also remains highly competitive on direct transfer from BUSC-BUSBRA to BUSC-BUSI647*, achieving the best Accuracy and one of the strongest AUC results. These results suggest that TRACE does not simply exploit report availability during training, but learns a transferable concept correction mechanism that remains effective without reports at test time. By contrast, pure image models can be competitive on individual targets, while existing concept bottleneck and multimodal methods do not show consistent advantages across domains. This indicates that standard concept modeling or image–text alignment is insufficient for the asymmetric clinical scenario where structured reports are available only during training. With privileged concept supervision, self-editing distillation, and concept-missing augmentation, TRACE achieves more robust generalization across shifts in country, device, and clinical center.

4.4 Ablation Study

We further analyze TRACE from three perspectives: editor comparison, supervision settings, and supervision scarcity.

First, Table 3 compares different editors under cross-domain evaluation. TRACE shows stable transferability across editor choices, with TRACE-Att achieving the best or tied-best results on multiple target domains. TRACE-MLP achieves higher AUC on some domains, while TRACE-Gate remains competitive in Accuracy, indicating different transfer preferences among editors. Attention provides the most reliable overall choice.

Second, Table 4 compares editors under Full and Missing settings. All editors perform well, suggesting that TRACE gains mainly come from teacher-guided concept editing rather than a specific architecture. Missing improves several metrics on BUSC-BUSBRA, while Full performs better on BUSC-BUSI647, indicating that missing perturbation is dataset-dependent.

Finally, Table 5 evaluates reduced supervision from 100% to 50% and 20%. TRACE remains effective with partial concept supervision. Attention and Gating show more stable trends, while MLP achieves a higher upper bound under full supervision.

Table 4: Core ablation of TRACE on BUSBRA and BUSC-BUSI647. We compare three editor designs under two training settings: *Full* (without concept missing) and *Missing* (with concept missing). Best and second-best results within each dataset and each setting are marked in red and blue, respectively.

Setting	Editor	BUSC-BUSBRA			BUSC-BUSI647		
		AUC	Acc	F1	AUC	Acc	F1
Full	Attention	0.924±0.007	0.874±0.016	0.794±0.023	0.945±0.023	0.893±0.024	0.822±0.071
	Gating	0.921±0.014	0.859±0.012	0.783±0.012	0.948±0.016	0.902±0.018	0.841±0.036
	MLP	0.916±0.014	0.859±0.013	0.776±0.026	0.970±0.030	0.938±0.057	0.902±0.083
Missing	Attention	0.924±0.013	0.874±0.019	0.802±0.018	0.873±0.023	0.857±0.017	0.759±0.025
	Gating	0.926±0.012	0.866±0.017	0.784±0.031	0.871±0.018	0.865±0.030	0.763±0.042
	MLP	0.921±0.010	0.872±0.006	0.797±0.013	0.874±0.017	0.850±0.029	0.745±0.036

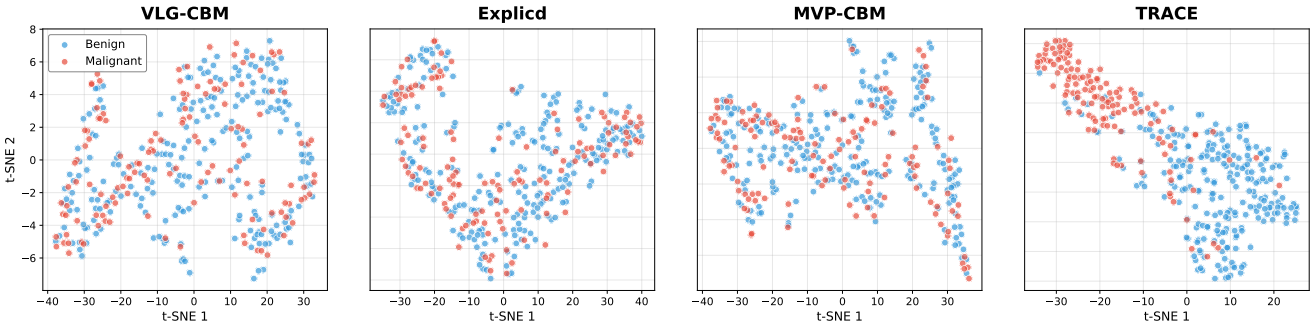

Figure 3: t-SNE visualization of feature embeddings learned by VLG-CBM, Explicd, MVP-CBM, and TRACE.

Table 5: Ablation on supervision ratio on BUSC-BUSBRA. We vary the proportion of training samples with structured concept supervision from 100% to 50% and 20%. Best and second-best results are highlighted in red and blue.

Editor	Ratio	AUC	Acc	F1
MLP	100%	0.9280±0.0078	0.8834±0.0173	0.8136±0.0150
	50%	0.9322±0.0188	0.8754±0.0126	0.7971±0.0318
	20%	0.9265±0.0192	0.8733±0.0178	0.7966±0.0206
Attention	100%	0.9263±0.0118	0.8807±0.0120	0.8070±0.0211
	50%	0.9304±0.0132	0.8770±0.0165	0.8035±0.0215
	20%	0.9272±0.0127	0.8770±0.0098	0.8033±0.0103
Gating	100%	0.9311±0.0182	0.8743±0.0190	0.7981±0.0261
	50%	0.9260±0.0139	0.8722±0.0230	0.7869±0.0307
	20%	0.9301±0.0123	0.8684±0.0148	0.7855±0.0276

4.5 t-SNE Visualization

To further examine representation quality, Fig. 3 visualizes the feature embeddings of VLG-CBM, Explicd, MVP-CBM, and TRACE using t-SNE, where blue and red points denote benign and malignant samples, respectively. TRACE shows clearer class separation, with more compact intra-class clusters and less overlap between the two classes. In contrast, VLG-CBM, Explicd, and MVP-CBM exhibit more cross-class mixing and scattered feature distributions. This suggests that TRACE learns more discriminative representations for breast ultrasound diagnosis.

This result provides representation-level support for TRACE. The improved separability indicates that teacher-guided concept editing transfers structured report supervision into more robust image-only representations, which is consistent with the superior AUC, Accuracy, and F1 performance in the main results.

5 Conclusion

In this paper, we presented TRACE, a training-time report-guided framework for robust breast ultrasound diagnosis under incomplete concepts. Rather than treating structured reports as an inference-time modality, TRACE uses them as privileged concept teachers during training to learn how coarse image-derived concepts should be refined. By integrating teacher-guided concept editing, a clinically ordered concept space, Strategic Concept Missing Training, and an image-only self-editor, TRACE enables report-free diagnosis at test time while preserving clinically meaningful concept refinement. We also introduced BUSC, a concept-enriched benchmark that aligns ultrasound images, diagnostic labels, and structured clinical concepts. Extensive experiments in both in-domain and zero-shot cross-domain settings show that TRACE consistently

outperforms strong baselines and supports robust image-only deployment across datasets. These findings indicate that learning concept correction from training-only structured supervision is a practical and effective approach to clinically grounded and generalizable breast ultrasound diagnosis.

Acknowledgments

This work is supported by the China Postdoctoral Science Foundation (Grant No. 2025M781597), the Hubei Provincial Natural Science Foundation of China (Grant No. 2026AFB700), the Fundamental Research Funds for the Central Universities, China (Grant No. XJ2026000901), and the Academy of Frontier Interdisciplinary Research at Central China Normal University.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data Brief* 28 (2020), 104863.

[2] Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, and U Rajendra Acharya. 2023. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine* 152 (2023), 106438.

[3] Junxin Chen, Zhiheng Ye, Renlong Zhang, Hao Li, Bo Fang, Li-bo Zhang, and Wei Wang. 2025. Medical image translation with deep learning: Advances, datasets and perspectives. *Medical Image Analysis* 103 (2025), 103605.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Mohammad Eghtedari, Alice Chong, Rebecca Rakow-Penner, and Haydee Ojeda-Fournier. 2021. Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. *American Journal of Roentgenology* 216, 4 (2021), 860–873.

[6] Luisa Gallée, Catharina Silvia Lisson, Timo Ropinski, Meinrad Beer, and Michael Götz. 2025. Proto-Caps: interpretable medical image classification using prototype learning and privileged information. *PeerJ Computer Science* 11 (2025), e2908.

[7] Yunhe Gao, Difei Gu, Mu Zhou, and Dimitris N. Metaxas. 2024. Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (Lecture Notes in Computer Science, Vol. 15010)*. Springer, 46–56.

[8] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. 2024. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics* 51, 4 (2024), 3110–3123.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[10] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3942–3951.

[11] Ahmed Iqbal and Muhammad Sharif. 2024. Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. *Engineering Applications of Artificial Intelligence* 127 (2024), 107292.

[12] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 5338–5348.

[13] Qiankun Li, Yimou Wang, Yani Zhang, Zhaoyu Zuo, Junxin Chen, and Wei Wang. 2025. Fuzzy-ViT: A Deep Neuro-Fuzzy System for Cross-Domain Transfer Learning From Large-Scale General Data to Medical Image. *IEEE Transactions on Fuzzy Systems* 33, 1 (2025), 231–241.

[14] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. 2024. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data* 11, 1 (2024), 148.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.

[16] Jean M. Seely and Mary B. Bissell. 2026. BI-RADS v2025: A Welcome Update. *American Journal of Roentgenology* (2026). Online ahead of print.

[17] Sanchit Sinha and Aidong Zhang. 2025. A Comprehensive Survey on the Risks and Limitations of Concept-based Models. *arXiv preprint arXiv:2506.04237* (2025).

[18] David Allen Spak, JS Plaxco, L Santiago, MJ Dryden, and BE Dogan. 2017. BI-RADS® fifth edition: A summary of changes. *Diagnostic and Interventional Imaging* 98, 3 (2017), 179–190.

[19] Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. 2024. VLG-CBM: Training Concept Bottleneck Models with Vision-Language Guidance. *Advances in Neural Information Processing Systems* 37 (2024), 79057–79094.

[20] Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: similarity control and knowledge transfer. *The Journal of Machine Learning Research* 16, 1 (2015), 2023–2049.

[21] Chunjiang Wang, Kun Zhang, Yandong Liu, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. 2025. MVP-CBM: multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 529–537.

[22] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 21372–21383.

[23] Peirong Xu, Luoqian Zhu, Jingkun Chen, Xin Qian, Yue Sun, Lingyun Bao, and Tao Tan. 2026. SAMASK-CLTR: A Spatial-Aware Mask Guided Learning Model for Benign and Malignant Tumor Classification in ABUS. In *Medical Image Computing and Computer Assisted Intervention, MICCAI 2025*, Vol. 15960. Springer, 567–577.

[24] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. 2023. Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models. *arXiv preprint arXiv:2310.03182* (2023).

[25] Qianye Yang, David Atkinson, Yunguan Fu, Tom Syer, Wen Yan, Shonit Punwani, Matthew J Clarkson, Dean C Barratt, Tom Vercauteren, and Yipeng Hu. 2022. Cross-modality image registration using a training-time privileged third modality. *IEEE Transactions on Medical Imaging* 41, 11 (2022), 3421–3431.

[26] Xiaoyan Yuan, Wei Wang, Junxin Chen, and Xiping Hu. 2025. Reading Between the Channels: Knowledge-Augmented Medical Time Series Classification. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 8978–8987.

[27] Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*.

[28] Boyu Zhang, Aleksandar Vakanski, and Min Xian. 2021. BI-RADS-Net: An Explainable Multitask Learning Approach for Cancer Diagnosis in Breast Ultrasound Images. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. 1–6.

[29] Boyu Zhang, Aleksandar Vakanski, and Min Xian. 2023. BI-RADS-NET-V2: a composite multi-task neural network for computer-aided diagnosis of breast cancer in ultrasound images with semantic and quantitative explanations. *IEEE Access* 11 (2023), 79480–79494.