Boosting Ultrasound Image Classification via Attribute-Guided Dual-Branch Framework

Bo Zhao[1⋆], Yapeng Li[1⋆], Juhua Liu[1(✉)], and Bo Du[1]

School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China
liujuhua@whu.edu.cn

Abstract. Ultrasound image classification is essential for computer-aided diagnosis. However, current methods often neglect clinical priors, leading to poor generalization in challenging scenarios and a lack of interpretability that limits clinical adoption. To address these issues, we aim to develop a medical-prior module that can be seamlessly integrated into existing pipelines to enhance both diagnostic performance and interpretability. In this paper, we propose an attribute-guided dual-branch framework for ultrasound classification that introduces domain-agnostic medical attribute priors, improving generalization while offering interpretable evidence. Specifically, a baseline branch follows conventional architectures and predicts image categories via a fully connected classifier. An attribute-guided branch injects domain-agnostic attributes as priors and produces human-interpretable decision cues. Finally, an adaptive decision module fuses the two branches in a data-dependent manner to yield the final prediction. Experiments across diverse ultrasound classification tasks demonstrate that our approach can be integrated into multiple backbones and state-of-the-art methods with low overhead, consistently improving accuracy and interpretability. Code is available at: https://github.com/zhaobo253-crypto/AttrGuide.

Keywords: Ultrasound image classification · Attribute guidance · Plug-and-play

1 Introduction

Ultrasound imaging is one of the most widely used modalities in routine clinical practice, owing to its non-invasive nature, real-time acquisition, portability, and cost-effectiveness [4,1]. It plays an important role in a broad range of clinical workflows, including screening, follow-up examinations, and point-of-care assessment across fetal, breast and other organ systems [4,1,8,2]. Meanwhile, ultrasound images often exhibit distinctive characteristics such as speckle noise, low contrast, operator dependence, and substantial appearance variation across devices and acquisition settings, which have motivated extensive research on

⋆ Equal contribution. (✉) Corresponding author.

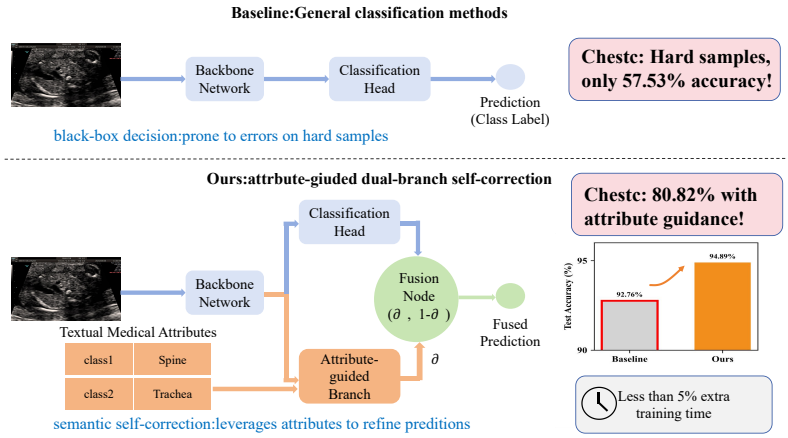


Fig. 1. Motivation. Traditional classifiers fail on hard samples, while AttrGuide adds an attribute-guided branch for semantic self-correction with negligible overhead.

deep learning-based analysis in medical imaging [16,20]. These factors make reliable interpretation challenging even for experienced clinicians, particularly in high-throughput settings, and raise concerns about model generalization under distribution shift [22]. Consequently, accurate ultrasound image classification has become a crucial component of computer-aided diagnosis.

Deep learning methods for ultrasound image classification broadly fall into two paradigms. Conventional representation-learning methods use transfer learning or self-supervised pretraining to extract discriminative features [3,22,23,8,2]. Interpretable and attribute-guided methods, including Concept Bottleneck Models (CBMs) and Vision-Language Models (VLMs), introduce semantic concepts to align learned representations with clinical reasoning [14,15,10,9,7,13,5]. However, both paradigms remain limited. Conventional "black-box" networks may rely on superficial textures rather than clinically meaningful attributes, causing errors on challenging cases (Fig. 1) and reducing clinical credibility. Attribute-guided models improve interpretability but often require dense annotations, task-specific designs, or costly prompt tuning [19,11,12], limiting lightweight adaptation. Therefore, achieving accurate and interpretable ultrasound classification without substantial annotation and computational burden remains challenging.

Motivated by the above observations, we aim to develop a plug-and-play module that can be seamlessly integrated into existing architectures [6] to concurrently bolster diagnostic accuracy and transparency. However, designing such a module entails two non-trivial challenges: (1) *Knowledge injection strategy*: how to effectively transform subjective, domain-agnostic medical priors into a structured representation that can guide the model's feature learning process; and (2) *Adaptive semantic integration*: how to dynamically reconcile the inherent global semantic features with specific attribute-based cues in a data-dependent manner to ensure robust performance across diverse clinical cases.

In this paper, we propose a versatile and effective attribute-guided dual-branch framework that can be integrated into existing methods with negligible overhead to enhance both generalization and interpretability. Specifically, to ensure the plug-and-play capability of the module, we treat the original classification architecture as a baseline branch. Building upon this, we design a parallel attribute-guided branch to explicitly inject clinical priors into the learning process, which bridges the gap between low-level visual features and high-level medical knowledge (Challenge 1). Furthermore, we introduce an adaptive fusion module to reconcile the global semantic predictions from the baseline branch with the interpretable, prior-informed cues from the attribute branch, thereby bolstering the overall diagnostic performance (Challenge 2). Since our dual-branch design requires minimal structural modifications and maintains a lightweight profile, the computational cost of our method remains low.

To summarize, the main contributions of this paper are:

- **Plug-and-play framework:** We propose an attribute-guided dual-branch framework that can be seamlessly integrated into existing models to enhance both robustness and interpretability.
- **Clinical-prior injection:** We design a dedicated attribute-guided branch that injects domain-agnostic medical priors to provide human-interpretable diagnostic evidence.
- **Adaptive decision fusion:** We introduce an adaptive fusion module that dynamically reconciles global and attribute-based decisions to yield superior classification performance.

Extensive experiments on diverse real-world datasets demonstrate that our module can be seamlessly integrated into various SOTA backbones, boosting performance and interpretability with negligible computational overhead.

2 Method

We consider a labeled ultrasound dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an image and $y_i \in \{1, \ldots, C\}$ is the class label. For each task, we use an LLM to generate task-discriminative medical attribute $\mathcal{A} = \{a_k\}_{k=1}^{K}$ and build a fixed class-attribute matrix $M \in \{0,1\}^{C \times K}$, where $M_{c,k} = 1$ if class c is expected to exhibit attribute k. Our goal is to learn a mapping $f_\theta : x \mapsto z_{\text{fus}} \in \mathbb{R}^C$, where θ denotes all learnable parameters and z_{fus} are fused logits that integrate a baseline branch with an attribute-guided branch before softmax. The architecture (Fig. 2) treats any existing encoder+FC classifier as the baseline branch producing logits z_{cls}, reuses its encoder features to drive an attribute-guided branch that produces attribute-based class logits z_{attr} via M, and finally combines z_{cls} and z_{attr} through a lightweight fusion block at the decision level.

2.1 Building the Medical Attribute Semantic Space

Ultrasound models face severe domain shifts from multi-center and cross-device variations, compromising generalization. Because clinical interpretability is crucial for deployment, achieving both at low cost remains a critical challenge.

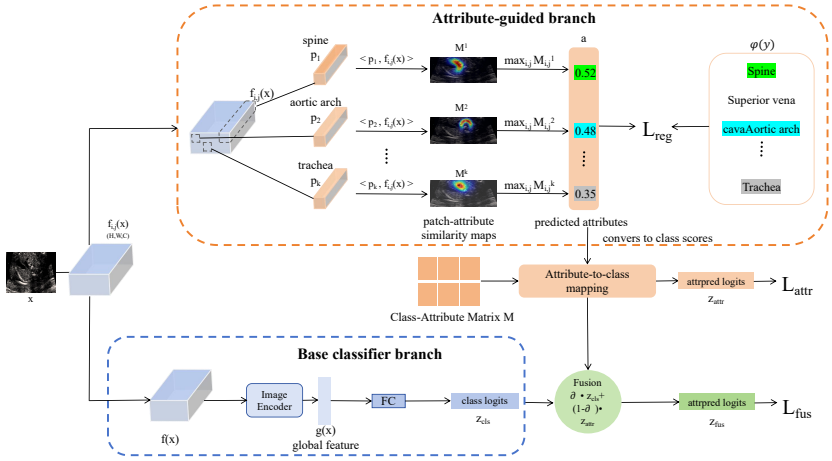


Fig. 2. Overview of AttrGuide. Our plug-in module integrates into existing ultrasound classifiers by reusing their encoder feature maps: a baseline branch follows the original encoder-classifier design, while an attribute-guided branch matches the same features with CLIP-derived attribute prototypes, aggregates them into attribute scores, maps them to class scores via a fixed class-attribute matrix, and is jointly optimized with the baseline prediction through classification and attribute-based regularization losses.

To address this, we construct a text-side attribute semantic space. Instead of expensive manual annotations, an LLM automatically generates scalable, task-discriminative, and domain-invariant medical attributes $\mathcal{A} = \{a_k\}_{k=1}^{K}$, i.e., visual cues less sensitive to data sources, centers, devices, and acquisition conditions, such as shape, margin, and texture. Unlike raw image features, these attributes provide a structured, interpretable prior for ultrasound recognition. Specifically, using a pre-trained CLIP text encoder $T(\cdot)$ [18], we encode ultrasound-specific prompts for each attribute a_k into semantic anchors $e_k = T(\texttt{prompt}(a_k)) \in \mathbb{R}^d$, stacking them offline into a fixed matrix $E = [e_1; \ldots; e_K] \in \mathbb{R}^{K \times d}$. Although CLIP is trained on natural images, shared visual primitives enable initial semantic grounding. Subsequently, learnable projection layers and supervised losses refine these embeddings, bridging the domain gap and calibrating them to clinical nuances.

On the image side, the backbone $B(\cdot)$ outputs local features v_i (ViT tokens or ResNet patches), which together with $\{e_k\}$ are projected into a shared space by learnable fully connected layers and ℓ_2-normalized as $\tilde{v}_i = \text{norm}(W_v v_i)$ and $\tilde{e}_k = \text{norm}(W_a e_k)$.

In this common space, we measure correlation between each patch and attribute via cosine similarity and aggregate patches to obtain per-attribute logits:

$$s_{k,i} = \cos(\tilde{v}_i, \tilde{e}_k), \quad w_{k,i} = \text{softmax}_i(\gamma \cdot s_{k,i}), \quad \ell_k = \sum_{i=1}^{N} w_{k,i}\, s_{k,i}. \tag{1}$$

where γ is a scaling factor and $\ell = [\ell_1, \ldots, \ell_K]$ is the attribute prediction vector. Given the class-attribute matrix M, we derive for each class y a binary target vector $m_y \in \{0,1\}^K$ that encodes which attributes are expected to be present, and supervise attribute prediction with:

$$\mathcal{L}_{\text{attr-pred}} = \text{BCEWithLogits}(\ell, m_y). \tag{2}$$

Beyond this attribute-prediction loss, we introduce a regularization term that pulls the predicted attribute activations towards the target vector m_y using an element-wise penalty and a directional constraint in the semantic space:

$$\mathcal{L}_{\text{reg}} = \text{MSE}(\sigma(\ell), m_y) + \big(1 - \cos(\sigma(\ell), m_y)\big). \tag{3}$$

2.2 Dual-Branch Adaptive Fusion Module

Baseline classification branch. The global feature g from the frozen or fine-tuned backbone is passed through the existing classifier head to obtain logits $z_{\text{cls}} = W_{\text{cls}} g$, which coincide with the baseline prediction (`cls_logits`).
Attribute-guided branch. The attribute logits ℓ are passed through a sigmoid to obtain scores $p = \sigma(\ell) \in [0,1]^K$, which we interpret as attribute activations for the image. These activations are aggregated into the class space using fixed matrix M. For class c, we compute a weighted average over its attribute set:

$$s_c = \frac{\sum_k M_{c,k}\, p_k \cdot p_k}{\sum_k M_{c,k}\, p_k + \epsilon}, \tag{4}$$

where $s = [s_1, \ldots, s_C]$ are class scores that are subsequently scaled to logits, yielding the attribute-guided output z_{attr}.
Adaptive fusion (low-cost self-correction). To combine the two decisions, we use a learnable fusion weight α and temperature τ that rescale and interpolate the two logits. The final fused logits are

$$z_{\text{fus}} = \alpha \cdot (z_{\text{cls}}/\tau) + (1 - \alpha) \cdot (z_{\text{attr}}/\tau). \tag{5}$$

During training, we always use the fused output z_{fus} as the sole classification head, computing the cross-entropy loss only on z_{fus}; the attribute prediction and regularization terms act as auxiliary supervision to shape the shared representation and fusion weights, as validated by the ablations in Table 2.

2.3 Joint Optimization Objective and Training Strategy

The training objective combines a cross-entropy loss on the fused logits, an attribute prediction loss, and a regularization term that aligns attribute predictions with the class-attribute matrix:

$$\mathcal{L} = \lambda_{\text{fus}}\, \mathcal{L}_{\text{fus}} + \lambda_{\text{reg}}\, \mathcal{L}_{\text{reg}} + \lambda_{\text{attr-pred}}\, \mathcal{L}_{\text{attr-pred}}. \tag{6}$$

where $\mathcal{L}_{\text{fus}}$ is the cross-entropy loss between the fused logits z_{fus} and the true class label, $\mathcal{L}_{\text{attr-pred}}$ is the BCEWithLogits loss on attribute logits ℓ against the binary target vector m_y, and $\mathcal{L}_{\text{reg}}$ is the regularization term in Eq. (3) that combines an MSE penalty and a cosine-similarity term between the predicted attribute activations (after sigmoid) and the same target vector m_y derived from M. Using both terms encourages the model to match each attribute probability numerically while also aligning the overall attribute direction in the semantic space. All modules, including the backbone, attribute branch and fusion weight, are optimized jointly using a single optimizer, while the primary supervision for classification is always applied to z_{fus}.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate on three benchmarks: BUSI [1] (Breast ultrasound), DDTI (Thyroid Ultrasound), and a Private dataset (Fetal ultrasound). The fetal task uses an internal three-center dataset with seven views: three-vessel, aortic arch, fetal bladder, transcerebellar plane, thorax, ductal arch, and fetal orbits.
Models and Training. We evaluate four ImageNet-pretrained backbones: ViT-B, ResNet50, Vim-s-16 (BU-Mamba [17]), and DiffMIC-v2 [21]. AttrGuide is attached to each backbone under the same Adam settings, and attribute tables are generated by an LLM from ultrasound guidelines and category descriptions.

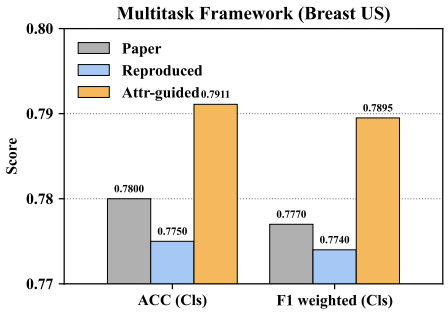
3.2 Results and Analysis

Comparison with existing methods. We integrate AttrGuide with the Vim-s-16 encoder of BU-Mamba [17] on BUSI and compare it with ViT and VMamba baselines. We also test a multi-task BUSI setting where classification is trained jointly with segmentation or auxiliary tasks [2]. Table 1 reports BUSI accuracy and the accompanying multi-task ACC/F1 plot, showing that AttrGuide achieves the best performance under both single-task and multi-task settings.
Ablation study. We examine component contributions to performance. Table 2 verifies that learnable fusion is crucial: the full model is best (+5.8 over Cls-only) and outperforms naive averaging. The attribute-only branch improves over the classification-only baseline, indicating that the medical attributes provide discriminative information rather than merely acting as auxiliary regularization.
Cross-backbone effectiveness. Table 3(a) shows that AttrGuide improves ResNet50, ViT-B, BU-Mamba, and DiffMIC-v2 by 1.8%, 4.3%, 0.8%, and 3.4%, respectively. These results demonstrate that the proposed method can be seamlessly integrated into diverse backbone architectures as a plug-and-play module to consistently enhance performance.
Task generalization. Table 3(b) confirms gains across datasets and backbones. On thyroid ultrasound, ResNet50 increases from 80.4% to 84.7%, demonstrating the task generalization of AttrGuide.

Table 1. Comparison on BUSI and multi-task BUSI setting. (a) BUSI test accuracy (mean ± std over 5 runs) for baselines and BU-Mamba with/without AttrGuide; † denotes results reported in BU-Mamba [17]. (b) Multi-task ACC and F1 on the joint segmentation-classification framework [2].

Encoder	ACC
ResNet50†	85.64 ± 2.72
VGG16†	85.47 ± 4.59
ViT-ti-16†	85.98 ± 4.48
ViT-s-16†	86.50 ± 4.44
ViT-s-32†	84.10 ± 3.81
ViT-b-16†	87.18 ± 2.70
ViT-b-32†	85.98 ± 1.39
BU-Mamba†	87.86 ± 2.72
AttrGuide	**88.72 ± 1.90**

(a) BUSI test accuracy.



(b) Multi-task ACC and F1.

Table 2. Component ablation on the private fetal 7-class task (ViT-B).

Cls branch	Attr branch	Learnable fusion	Test Acc (%)	Δ
✓			89.1	–
✓	✓	Avg	90.3	↑1.2
	✓		92.6	↑3.5
✓	✓	✓	**94.9**	↑5.8

Table 3. Model analysis. (a) Different backbones on the BUSI dataset. (b) Different datasets and backbones.

Backbone	Base	+Attr	Δ
ResNet50	81.5	**83.3**	+1.8
ViT-B	81.1	**85.4**	+4.3
BU-Mamba	87.9	**88.7**	+0.8
DiffMIC-v2	60.0	**63.4**	+3.4

(a) Different backbones on BUSI.

Dataset	Backbone	Base	+Attr	Δ
BUSI	ResNet50	81.5	**83.3**	+1.8
	ViT-B	81.1	**85.4**	+4.3
Fetal	ResNet50	92.2	**92.6**	+0.4
	ViT-B	92.8	**94.9**	+2.1
Thyroid	ResNet50	80.4	**84.7**	+4.3
	ViT-B	82.4	**86.1**	+3.7

(b) Different datasets and backbones.

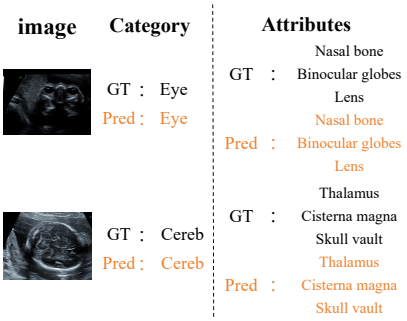
Computational cost. Table 4 shows that the plug-in branch adds marginal training-time and parameter overhead while improving all evaluated backbones.

Interpretability. Table 4 shows that the attribute-guided branch predicts clinically meaningful attributes with high accuracy (87.56% fetal, 83.29% BUSI), providing human-interpretable cues for verification. The visualization highlights regions associated with the selected attributes when positive and negative cues

Table 4. Computational cost and interpretability. (a) BUSI training cost, where time is minutes per fold and Params are in millions. (b) Attribute examples; pred-attr accuracy is 87.56% for fetal and 83.29% for BUSI.

Model	Set	Time	Param
ViT-B	Base	11.02	85.8
	+Attr	11.18	86.4
ResNet50	Base	10.35	23.5
	+Attr	11.20	24.8
Mamba	Base	73.98	25.4
	+Attr	75.91	25.9

(a) Training cost.



(b) Attribute examples.

are distinguishable; meanwhile, some hard cases remain imperfect, which suggests that attribute localization should be further strengthened in future work.

Semantic anchor analysis. We examine the construction of the attribute table and anchors. LLM-generated tables without expert annotation yield stable BUSI accuracies (85.4%, 85.6%, 84.9%). Manual expert selection improves this to 86.4%, indicating expert knowledge brings gains. Self-learned attribute embeddings achieve 82.8%; while useful, explicit attributes with language encoding provide stronger guidance. Replacing CLIP with MedCLIP improves accuracy from 85.4% to 86.3%, confirming medical text priors benefit attribute anchors.

4 Conclusion

We introduce an attribute-guided dual-branch framework to enhance ultrasound classification by bridging visual features with clinical priors. We demonstrate that our approach consistently boosts the performance of diverse backbones and recent methods, including ViT-B, ResNet50, DiffMIC-v2, and BU-Mamba, while remaining efficient with less than 5% additional training time. From the results, we conclude that: (1) the injection of domain-agnostic attributes provides essential evidence that bolsters both model robustness and clinical interpretability; (2) the adaptive fusion module is critical for synergistically integrating global and attribute-based semantics; and (3) the same lightweight plug-in generalizes from breast ultrasound to thyroid and fetal ultrasound tasks. We intend to extend this low-cost, plug-and-play enhancement to a broader range of medical imaging tasks and further improve attribute localization for hard cases.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC2705705, in part by the National Natural Science Foundation of China under Grants 62225113 and U23B2048, in part by the Innovative Research Group Project of Hubei Province under Grants

2024AFA017, in part by the Science and Technology Major Project of Hubei Province under Grants 2024BAB046 and 2025BCB026, and in part by the New Cornerstone Science Foundation through the XPLORER PRIZE. This work was also supported by WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Aumente-Maestro, C., Diez, J., Remeseiro, B.: A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. Computer Methods and Programs in Biomedicine **260**, 108540 (2025). https://doi.org/10.1016/j.cmpb.2024.108540
3. Azizi, S., Mustafa, B., Ryan, F., et al.: Big self-supervised models advance medical image classification. In: ICCV. pp. 3478–3488 (2021). https://doi.org/10.1109/ICCV48922.2021.00346
4. Burgos-Artizzu, X.P., Coronado-Gutierrez, D., Valenzuela-Alcaraz, B., et al.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**, 10200 (2020). https://doi.org/10.1038/s41598-020-67076-5
5. Chen, S., Hong, Z., Liu, Y., Xie, G.S., Sun, B., Li, H., You, X.: TransZero: Attribute-guided transformer for zero-shot learning. AAAI **36**(1), 330–338 (2022). https://doi.org/10.1609/aaai.v36i1.19909
6. Chen, Y., Zhao, S., Chen, B., Gustaf, M.: Clinically guided adaptive contrast adjustment for fetal plane classification: a modular plug-and-play solution. Frontiers in Physiology **16**, 1689936 (2025). https://doi.org/10.3389/fphys.2025.1689936
7. Fang, X., Lin, Y., Zhang, D., Cheng, K.T., Chen, H.: Aligning medical images with general knowledge from large language models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15010, pp. 57–67. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72117-5_6
8. Feng, Z., et al.: Hybrid-view attention network for clinically significant prostate cancer classification in transrectal ultrasound. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15960, pp. 260–269. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-04927-8_25
9. Gao, Y., Gu, D., Zhou, M., Metaxas, D.: Aligning human knowledge with visual concepts towards explainable medical image classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15010, pp. 46–56. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72117-5_5
10. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15012, pp. 632–642. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_59

11. Huang, Y., Cheng, P., Tam, R., Tang, X.: Fine-grained prompt tuning: A parameter and memory efficient transfer learning method for high-resolution medical image classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15012, pp. 120–130. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_12
12. Hussein, N., Shamshad, F., Naseer, M., Nandakumar, K.: PromptSmooth: Certifying robustness of medical vision-language models via prompt learning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15012, pp. 698–708. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_65
13. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 5338–5348. PMLR (2020)
14. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 453–465 (2014). https://doi.org/10.1109/TPAMI.2013.140
15. Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: CLIP-Lung: Textual knowledge-guided lung nodule malignancy prediction. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. LNCS, vol. 14226, pp. 403–412. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_38
16. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sanchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
17. Nasiri-Sarvi, A., Hosseini, M.S., Rivaz, H.: Vision mamba for classification of breast ultrasound images. In: Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 148–158. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77789-9_15
18. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
19. Shakeri, F., Huang, Y., Silva-Rodriguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15012, pp. 553–563. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_52
20. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging **35**(5), 1299–1312 (2016). https://doi.org/10.1109/TMI.2016.2535302
21. Yang, Y., Fu, H., Aviles-Rivero, A.I., Xing, Z., Zhu, L.: DiffMIC-v2: Medical image classification via improved diffusion network. IEEE Transactions on Medical Imaging **44**(5), 2244–2255 (2025). https://doi.org/10.1109/TMI.2025.3530399
22. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine **15**(11), e1002683 (2018). https://doi.org/10.1371/journal.pmed.1002683
23. Zheng, X., Chen, X., Gong, S., Griffin, X., Slabaugh, G.: XFMamba: Cross-fusion mamba for multi-view medical image classification. In: Medical Image Computing

and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15960, pp. 672–682. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-04927-8_64