A Modality-Aware Mixture-of-Experts Framework for Clinical Multimodal Breast Ultrasound Diagnosis

Hao Suo[1], Zhongjun Zhu[1], Lingyu Chen[2], Qing Chang[3], Hongen Liao[1], Lin Jin[4(✉)], Cheng Li[5(✉)], and Fang Chen[1(✉)]

[1] Shanghai Jiao Tong University, Shanghai, China
chen-fang@sjtu.edu.cn
[2] Nanjing University of Aeronautics and Astronautics, Nanjing, China
[3] Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4] Guanghua Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China
jinlin205@163.com
[5] Department of Ultrasound, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
leecheng1018@126.com

Abstract. Current breast ultrasound CADx models are often insufficiently validated under real-world clinical heterogeneity, including spatial variability, long-tailed phenotypes, and multi-vendor acquisition variability. To bridge this gap, we construct a comprehensive multimodal benchmark capturing these complexities across pairwise combinations of B-mode with CDFI, Elastography, and CEUS. Furthermore, we propose MA-MoE (Modality-Aware Mixture-of-Experts) to adaptively integrate anatomical B-mode data with functional supplementary modalities. Our core innovation is a Shared-Specific MoE mechanism with constrained routing, which encourages modality-specific experts to preserve complementary functional cues and shared experts to model cross-modal pathological consensus. Evaluation on our clinical benchmark shows consistent and balanced improvements over representative fusion methods, suggesting that modality-aware routing is beneficial for paired multimodal breast ultrasound diagnosis. The code is available at https://github.com/StveHok/MA-MoE

Keywords: Breast Ultrasound · Mixture-of-Experts · Multimodal Fusion.

1 Introduction

Breast ultrasound is a primary, cost-effective screening modality that supports early detection and triage. In clinical breast ultrasound, B-mode imaging is commonly complemented by functional modalities such as Color Doppler/CDFI,

Hao Suo and Zhongjun Zhu contributed equally to this work.

Elastography, and CEUS, which provide additional vascular, stiffness, and perfusion information for lesion characterization. Prior studies have shown that targeted combinations built upon B-mode can improve diagnostic assessment and reduce false positives/benign biopsies compared with conventional ultrasound alone [8, 15]. Clinicians routinely integrate B-mode with functional imaging (e.g., CDFI, Elastography, or CEUS) to improve diagnostic confidence, yet most recent CADx advances remain B-mode-centric. As a result, despite strong benchmark performance, these unimodal models are still insufficiently validated under real-world clinical heterogeneity.

The complexity of real-world deployment arises from the interplay of endogenous pathological heterogeneity and exogenous acquisition variability, frequently causing a "performance gap" under distribution shifts [5, 18, 20]. To address this, we curate a multimodal benchmark capturing five real-world barriers often oversimplified in public datasets: (i) Clinical Uncertainty, where high-ambiguity cases render B-mode alone insufficient [19]; (ii) Spatial Variability, as lesions across all 12 clock-face positions exhibit appearance fluctuations due to quadrant-dependent structures; (iii) Long-tailed Phenotypes, where severe class imbalance impairs the recognition of rare cases [22]; (iv) Acquisition Shift, stemming from imaging-physics discrepancies across vendors (e.g., Mindray, GE, Toshiba) and operator variability; and (v) Modal Fragmentation, where clinically essential multimodal integration (B-mode paired with CDFI, Elastography, or CEUS) is inconsistently available and largely absent from existing benchmarks.

As shown in Fig. 1(a), we observe a clear generalization gap when transferring representative B-mode baselines from a public benchmark (BUSI) to our real-world clinical cohort (private). Although these methods achieve strong performance on BUSI, they consistently degrade on our cohort, ranging from mild to severe drops (e.g., UniFormer decreases by $\sim$20 points), suggesting that B-mode-only models validated on curated datasets may overestimate robustness under real clinical distribution shifts.

Fig. 1(b)–(d) summarizes key heterogeneity factors of our multimodal cohort that help interpret (a): Fig. 1(b) outlines our multi-modal acquisition (B-mode plus CDFI, Elastography, and CEUS) and multi-vendor collection, introducing complementary cues and device-dependent appearance variations; Fig. 1(c) shows diverse clinical sources and a broad lesion-size range, reflecting heterogeneous case mix and diagnostic difficulty; and Fig. 1(d) highlights non-uniform clock-face coverage and incomplete/long-tailed lesion-type annotations. Collectively, these acquisition-, cohort-, and annotation-level heterogeneities explain why public-benchmark performance may not translate to stable performance in our clinical cohort.

To address these limitations, we propose MA-MoE, a sample-adaptive fusion network grounded in a Shared-Specific Mixture-of-Experts mechanism. Evaluated on our clinical bimodal benchmark (B-mode paired with CDFI, Elastography, or CEUS), MA-MoE leverages CLIP and SAM for complementary feature extraction. By employing constrained routing to allocate tokens to modality-specific and shared experts, our design preserves unique modal characteristics

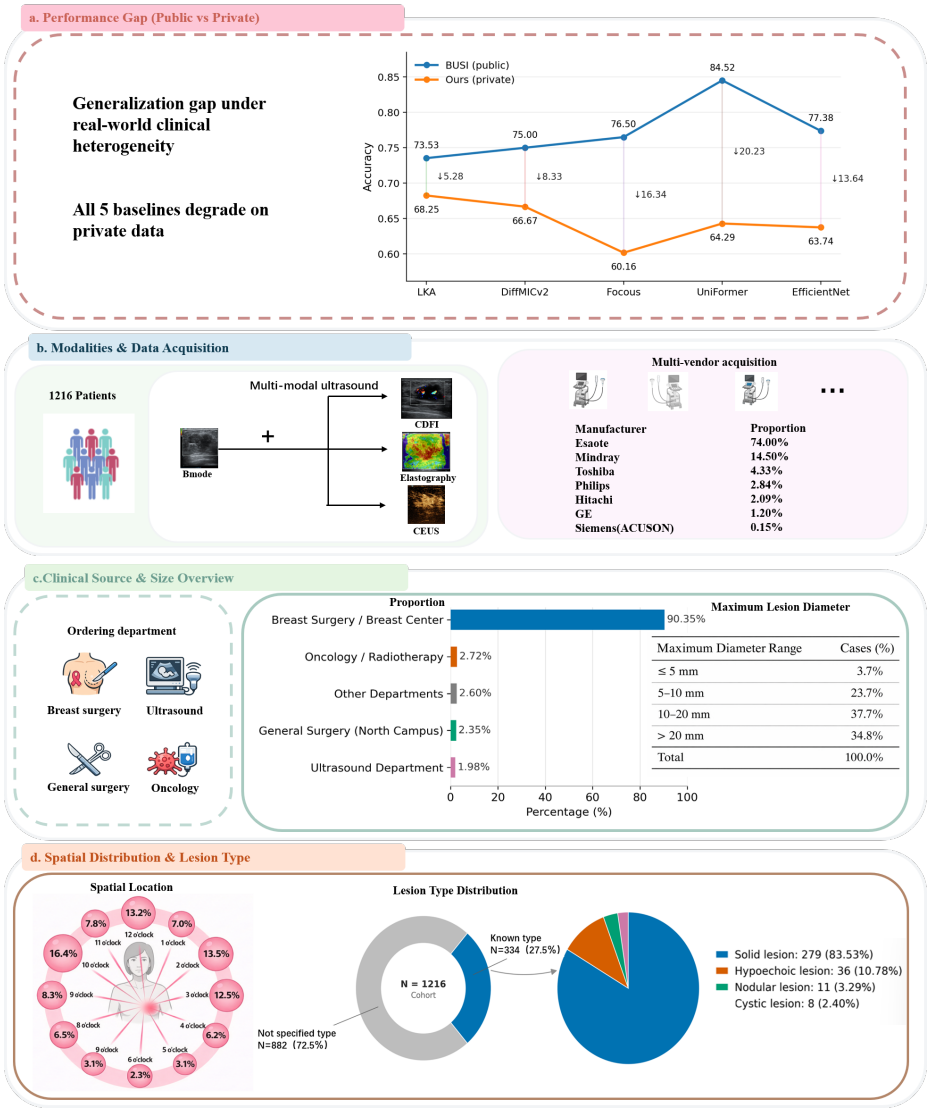

Fig. 1. Motivation and characterization of our multimodal breast ultrasound dataset. (a) B-mode baselines show severe performance drops (↓) when transferred from public (BUSI) to our private cohort. (b-d) Clinical heterogeneity of our private dataset, detailing paired modalities (B-mode with CDFI/Elasto/CEUS), multi-vendor acquisition, and diverse lesion distributions.

while capturing cross-modal dependencies, supporting more stable multimodal classification under clinical heterogeneity.

Our main contributions are as follows: (1) a paired multimodal clinical benchmark covering patient, pathological, and vendor heterogeneity; (2) a Shared-Specific MoE mechanism with constrained routing that encourages modality-specific and shared representation learning; and (3) consistent and balanced improvements over representative fusion baselines across three bimodal clinical settings (B+CDFI, B+Elasto, B+CEUS), suggesting the practical value of modality-aware fusion.

2 Related Work

Breast ultrasound datasets. Public benchmarks (e.g., BUSI [2], UDIAT [3, 12], BrEaST/BUS-Set [11,13]) have driven progress but are predominantly single-modality (BUS). In contrast, paired multimodal datasets (BUS with CDFI/CEUS) remain scarce and small-scale [17]. Large-scale cohorts that cover CEUS and real-world heterogeneity (e.g., vendor/protocol/operator variations) are still uncommon, limiting the validation of algorithms under clinical domain shifts and highlighting the need for heterogeneous multimodal data [1].

Medical image multimodal fusion with MoE. Fusion integrates anatomical (BUS) and functional (CDFI/CEUS) cues for improved diagnosis [8, 21]. Early CADx relied on concatenation [7], whereas recent methods (e.g., AINet [4, 10]) and foundation models (CLIP/SAM [9]) emphasize cross-modal alignment. However, most pipelines lack instance-wise reliability under multi-vendor quality fluctuations [14, 16]. MoE mitigates this via gated routing, e.g., modality-specific experts (M^4oE [16]) or sparse filtering (MExD [23]). Yet for paired ultrasound, modeling shared vs. modality-specific information remains limited, and weak routing constraints may induce redundancy and instability under heterogeneous acquisition conditions.

3 Method

3.1 Dataset

Public benchmarks such as BUSI and UDIAT are predominantly restricted to B-mode imaging. Even recent curated datasets (e.g., BrEaST, BUS-UCLM), despite improved curation and scale, still lack systematic functional pairings, as summarized in Table 1. In contrast, we curate a private paired multimodal clinical benchmark, collected from a collaborating hospital. Specifically, our dataset contains 1,216 paired B-mode–Elastography cases (stiffness), 1,199 paired B-mode–CDFI cases (vascularity), and 889 paired B-mode–CEUS cases (perfusion). This paired design enables rigorous multimodal fusion and supports evaluation under incomplete and non-uniform modality availability—a critical real-world challenge typically absent in existing unimodal benchmarks. All data were de-identified, organized at the patient level, and split without patient overlap.

Table 1. Comparison with public breast ultrasound benchmarks.

Dataset	Modality	Pairs	Notes
UDIAT (2017)	B-mode	—	163 images
OASBUD (2017)	B-mode	—	100 images
BUSI (2020)	B-mode	—	600 patients; 780 images
BrEaST (2023)	B-mode	—	256 images
US3M (2024)	B+Elasto+CDFI	303	248 patients; 1542 images
BUS-UCLM (2025)	B-mode	—	683 images
Ours	B+Elasto	1,216	1216 patients; 2432 images
	B+CDFI	1,199	1199 patients; 2398 images
	B+CEUS	889	889 patients; 1778 images

3.2 MA-MoE Framework

As illustrated in Fig. 2, MA-MoE is a sample-adaptive dual-stream framework designed to robustly integrate paired B-mode and supplementary modalities. For each modality image $\mathbf{I} \in \{\mathbf{I}_B, \mathbf{I}_S\}$, we extract representations using frozen CLIP ViT-B/16 and SAM ViT-B image encoders, where SAM is used as a dense structural token extractor rather than a segmentation module. CLIP produces semantic patch tokens $\mathbf{F}_{sem} \in \mathbb{R}^{L\times 768}$, while SAM yields a dense feature map flattened into structural tokens $\mathbf{F}_{str} \in \mathbb{R}^{4096\times 256}$. Both features are projected to a unified dimension D via $\tilde{\mathbf{F}}_{sem} = \mathbf{F}_{sem}\mathbf{W}_{sem}$ and $\tilde{\mathbf{F}}_{str} = \mathbf{F}_{str}\mathbf{W}_{str}$. To inject global structural context into semantic tokens, we apply a channel-wise gating fusion:

$$\mathbf{Z} = \mathbf{g} \odot \tilde{\mathbf{F}}_{sem} + (1 - \mathbf{g}) \odot \mathrm{Expand}\big(\mathrm{GAP}(\tilde{\mathbf{F}}_{str})\big), \tag{1}$$

where the gating weight is computed as $\mathbf{g} = \sigma(\mathrm{MLP}([\mathrm{GAP}(\tilde{\mathbf{F}}_{sem}) \,|\, \mathrm{GAP}(\tilde{\mathbf{F}}_{str})]))$. This process yields the final modality tokens $\mathbf{Z}_B, \mathbf{Z}_S \in \mathbb{R}^{L\times D}$.

3.3 Shared-Specific MoE Transformer

We concatenate both modality tokens as $\mathbf{Z}^{(0)} = [\mathbf{Z}_B; \mathbf{Z}_S] \in \mathbb{R}^{2L\times D}$, and feed them into N Transformer blocks. For the l-th block ($l = 1, \ldots, N$), the standard FFN is replaced by a sparse MoE layer, formulated as $\hat{\mathbf{Z}}^{(l)} = \mathbf{Z}^{(l-1)} + \mathrm{MSA}(\mathrm{LN}(\mathbf{Z}^{(l-1)}))$ and $\mathbf{Z}^{(l)} = \hat{\mathbf{Z}}^{(l)} + \mathrm{MoE}(\mathrm{LN}(\hat{\mathbf{Z}}^{(l)}))$.

We define an expert set $\mathcal{E} = \{E_1, \ldots, E_K\}$ explicitly partitioned as $\mathcal{E} = \mathcal{E}_B \cup \mathcal{E}_S \cup \mathcal{E}_{sh}$, where $\mathcal{E}_B$ and $\mathcal{E}_S$ denote modality-specific experts for B-mode and supplementary tokens, while $\mathcal{E}_{sh}$ denotes shared experts accessible by both modalities. The shared experts are designed to capture cross-modal pathological correspondences (e.g., whether suspicious morphology is consistent with stiffness, vascularity, or perfusion abnormalities).

In practice, we set $K = 4$ experts with a shared-specific ratio of $2 : 1 : 1$, where two experts are shared and one expert is assigned to each modality. This is the minimal symmetric configuration for Top-2 routing: each modality can

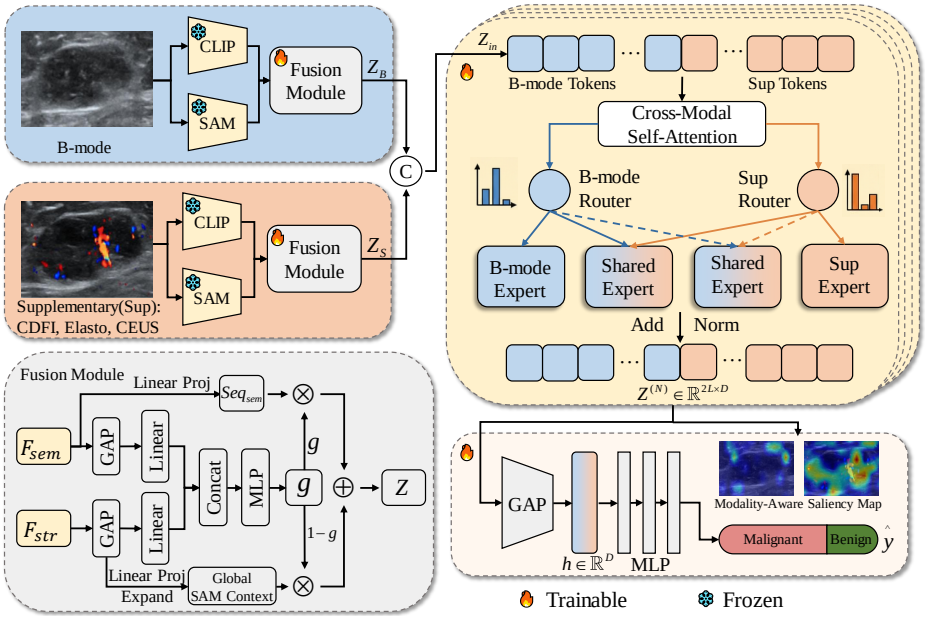

Fig. 2. Overall architecture of MA-MoE. Paired modalities are integrated via a constrained Shared-Specific MoE Transformer. The classification head yields both malignancy predictions (via GAP) and modality-aware saliency maps for spatial interpretability.

access at least two experts while retaining shared capacity for cross-modal consensus. Each expert adopts a standard ViT-style FFN (Linear–GELU–Linear) with hidden dimension $4D$.

To enforce the shared-specific constraint, we employ modality-aware masked routing. Specifically, two routers are used for B-mode and supplementary tokens, respectively. For a token $\mathbf{x} \in \mathbb{R}^D$, routing logits are computed as

$$\mathbf{l}(\mathbf{x}) = \mathbf{x}\mathbf{W}_g^{(m)}, \quad \mathbf{p}(\mathbf{x}) = \text{Softmax}\left(\mathbf{l}(\mathbf{x}) + \mathbf{M}^{(m)}\right), \quad m \in \{B, S\}, \tag{2}$$

where $\mathbf{W}_g^{(m)} \in \mathbb{R}^{D\times K}$ is the router weight. The fixed mask $\mathbf{M}^{(B)}$ disables experts in $\mathcal{E}_S$ for B-mode tokens, while $\mathbf{M}^{(S)}$ disables experts in $\mathcal{E}_B$ for supplementary tokens. Both masks allow access to $\mathcal{E}_{sh}$. This constraint reduces redundant cross-routing and encourages modality-aware specialization while retaining shared experts for cross-modal information exchange. Each token is sparsely dispatched to Top-k experts (we use $k = 2$), and aggregated by the routing probability:

$$\text{MoE}(\mathbf{x}) = \sum_{e\in \text{Top-}k(\mathbf{p}(\mathbf{x}))} p_e(\mathbf{x})\, E_e(\mathbf{x}). \tag{3}$$

To prevent expert collapse, we employ a Switch Transformer-style load-balancing auxiliary loss $\mathcal{L}_{aux}$, with $\lambda_{aux} = 0.01$. All routed tokens are processed by their assigned experts. When the supplementary modality is unavailable, we remove $\mathbf{Z}_S$

(or replace it with null tokens) and route only B-mode tokens through $\mathcal{E}_B \cup \mathcal{E}_{sh}$, enabling evaluation under incomplete modality availability without modifying network parameters.

We apply global average pooling on the final tokens and use a linear classifier: $\mathbf{h} = \mathrm{GAP}(\mathbf{Z}^{(N)})$ and $\hat{\mathbf{y}} = \mathrm{Softmax}(\mathbf{W}_{cls}\mathbf{h} + \mathbf{b}_{cls})$. The network is trained end-to-end with the standard loss $\mathcal{L} = \mathcal{L}_{CE} + \lambda_{aux}\mathcal{L}_{aux}$, where $\mathcal{L}_{CE}$ is the cross-entropy loss and $\mathcal{L}_{aux}$ is the MoE load-balancing regularizer to avoid expert collapse and encourage diverse expert utilization.

4 Experiments

We evaluate MA-MoE under three paired bimodal settings that reflect routine clinical usage: B-mode + CEUS (Contrast), B-mode + Elastography (Elastic), and B-mode + CDFI (Blood). Unless otherwise specified, all splits are performed at the patient level to avoid information leakage, and the same training pipeline is applied across settings.

4.1 Main Results on Our Clinical Paired Benchmark

Table 2. Comparison with bimodal fusion baselines (CAMVF [14], HVAN [6], XF-Mamba [24], M^{4}oE [16]) on our clinical benchmark. Best results in bold.

Setting	Method	Acc	Prec	Sens	Spec	F1	AUC	AP
B-mode + CEUS	CAMVF	0.6176	0.6158	0.5375	0.6889	0.5687	0.6306	0.6400
	HVAN	0.6069	0.6545	0.4865	**0.7324**	0.5581	0.6781	0.6586
	XFMamba	0.6345	0.6082	**0.7973**	0.4648	0.6901	0.6917	0.6931
	M^{4}oE	0.6069	0.6350	0.5410	0.6760	0.6060	0.6650	0.6325
	MA-MoE (Ours)	**0.6828**	**0.6750**	0.7297	0.6338	**0.7013**	**0.7298**	**0.7191**
B-mode + Elasto	CAMVF	0.6537	0.6854	0.6281	0.6819	0.6533	0.7034	0.6900
	HVAN	0.6911	0.6886	0.7324	0.6346	0.7324	**0.7752**	0.7349
	XFMamba	0.6776	0.6762	0.7396	0.6092	0.7065	0.6776	0.6745
	M^{4}oE	0.6504	**0.7260**	0.5380	**0.7760**	0.6480	0.7100	0.6977
	MA-MoE (Ours)	**0.7104**	0.6937	**0.8021**	0.6092	**0.7440**	0.7589	**0.7429**
B-mode + CDFI	CAMVF	0.6286	0.5786	0.8099	0.4636	0.6735	0.6745	0.6750
	HVAN	0.7238	**0.8235**	0.5957	**0.8621**	0.6914	**0.8325**	**0.8063**
	XFMamba	0.6519	0.6867	0.6064	0.7011	0.6441	0.6949	0.7291
	M^{4}oE	0.7333	0.6970	0.8550	0.6030	0.7270	0.8010	0.7623
	MA-MoE (Ours)	**0.8066**	0.7479	**0.9468**	0.6552	**0.8357**	0.7974	0.7577

As shown in Table 2, MA-MoE achieves the best Accuracy and F1 across all three paired-modality settings and remains competitive on AUC and AP. For B-mode + CEUS, it improves Accuracy by 4.8% over the strongest baseline (XFMamba) and also obtains the best AUC and AP. In B-mode + Elastography and B-mode + CDFI, some baselines achieve the best value on individual metrics, indicating that no single method uniformly dominates all operating points.

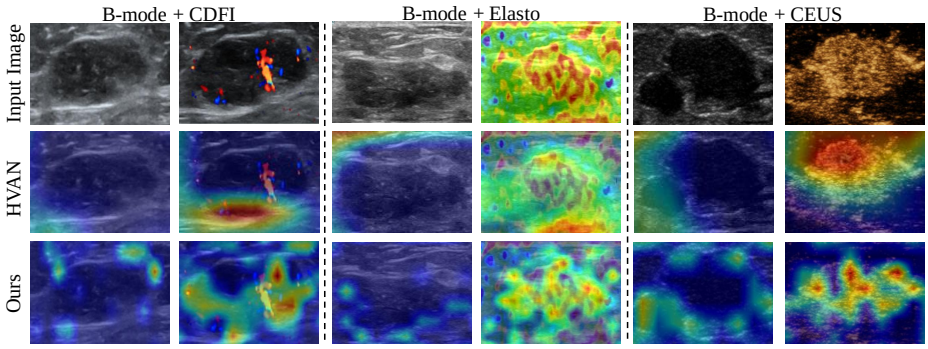

Fig. 3. Grad-CAM comparison across three bimodal cases. Compared with the diffuse activations in HVAN, MA-MoE yields more focused lesion responses and suggests complementary attention patterns across anatomical and functional cues: B-mode highlights anatomical margins, while functional branches emphasize vascularity (CDFI), stiffness (Elasto), and perfusion (CEUS).

Overall, MA-MoE provides a favorable balance across classification metrics, particularly Accuracy, Sensitivity, and F1, supporting the benefit of modality-aware routing for clinical multimodal fusion.

Qualitative Grad-CAM visualizations (Fig. 3) show that MA-MoE produces more focused lesion-related responses than HVAN. The modality-specific branches exhibit distinct attention patterns that are consistent with their clinical roles: the B-mode branch mainly highlights anatomical margins and lesion morphology, whereas the functional branches emphasize vascularity in CDFI, stiffness-related regions in Elastography, and heterogeneous perfusion in CEUS. These observations suggest that the shared-specific routing encourages complementary modality-aware attention, although we interpret the visualizations as qualitative evidence rather than definitive proof of disentanglement.

4.2 Ablation Study: Isolating the Contribution of Functional Modalities

To verify that the performance improvements of MA-MoE stem directly from the complementary functional information rather than the increased parameter capacity of a dual-stream architecture, we perform an input substitution ablation. Specifically, we maintain the exact MA-MoE network structure but duplicate the B-mode image into both the anatomical and supplementary streams (i.e., setting $\mathbf{I}_S = \mathbf{I}_B$). As shown in Table 3, replacing functional inputs with duplicated B-mode leads to consistent degradation across all subsets, with erratic sensitivity trade-offs (e.g., abnormally lowered sensitivity in the Elastography subset). This confirms that performance gains stem from genuinely exploiting complementary functional signals rather than increased parameter capacity.

Table 3. Input substitution ablation ($\mathbf{I}_S = \mathbf{I}_B$) versus true functional pairings. Best results are in bold.

Setting	Unimodal Ablation ($\mathbf{I}_S = \mathbf{I}_B$)				Bimodal MA-MoE (Ours)			
	Acc	Sens	F1	AUC	Acc	Sens	F1	AUC
B-mode + CEUS	0.6483	**0.7568**	0.6871	0.7195	**0.6828**	0.7297	**0.7013**	**0.7298**
B-mode + Elasto	0.6721	0.6146	0.6629	**0.7649**	**0.7104**	**0.8021**	**0.7440**	0.7589
B-mode + CDFI	0.6133	0.7979	0.6818	0.7050	**0.8066**	**0.9468**	**0.8357**	**0.7974**

5 Conclusion

We introduced a paired multimodal breast ultrasound benchmark combining B-mode with CDFI, Elastography, or CEUS, and proposed MA-MoE, a modality-aware fusion framework based on a Shared-Specific Mixture-of-Experts architecture. Through constrained dual-router masked routing, MA-MoE encourages modality-specific and shared representation learning while integrating complementary anatomical and functional cues. Experiments show balanced improvements across three bimodal settings, suggesting the practical value of modality-aware routing for paired multimodal breast ultrasound diagnosis. Broader external validation, larger multi-center cohorts, and more lightweight deployment designs remain important future directions.

Acknowledgments. This work was supported by the Science and Technology Commission of Shanghai Municipality (No. 24511104100), the National Key Research and Development Program of China (No. 2025YFC2426300), and the National Natural Science Foundation of China (Grant Nos. 82572314 and 62271246).

Disclosure of Interests. The authors declare no competing interests.

References

1. Afrin, H., Larson, N.B., Fatemi, M., Alizad, A.: Deep learning in different ultrasound methods for breast cancer, from diagnosis to prognosis: current trends, challenges, and an analysis. Cancers **15**(12), 3139 (2023)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., Andre, M.: Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. Biomedical signal processing and control **61**, 102027 (2020)
4. Chen, H., Li, Y., Zhang, J., Yang, L., Sun, Y., Chen, Y., Zhou, S., Li, Z., Qian, X., Xu, Q., et al.: An alignment and imputation network (ainet) for breast cancer diagnosis with multimodal multi-view ultrasound images. IEEE Transactions on Medical Imaging (2025)
5. Dan, Q., Xu, Z., Burrows, H., Bissram, J., Stringer, J.S., Li, Y.: Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. NPJ Precision Oncology **8**(1), 21 (2024)

6. Feng, Z., Fu, J., Zou, X., Ye, H., Wu, H., Zhou, J., Wang, Y.: Hybrid-view attention network for clinically significant prostate cancer classification in transrectal ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 260–269. Springer (2025)
7. Gadzicki, K., Khamsehashari, R., Zetzsche, C.: Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd international conference on information fusion (FUSION). pp. 1–6. IEEE (2020)
8. Kapetas, P., Clauser, P., Woitek, R., Wengert, G.J., Lazar, M., Pinker, K., Helbich, T.H., Baltzer, P.A.: Quantitative multiparametric breast ultrasound: application of contrast-enhanced ultrasound and elastography leads to an improved differentiation of benign and malignant lesions. Investigative radiology **54**(5), 257–264 (2019)
9. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-sam: Bridging text and image towards universal medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 643–653. Springer (2024)
10. Liu, Y., Li, J., Zhao, C., Zhang, Y., Chen, Q., Qin, J., Dong, L., Wang, T., Jiang, W., Lei, B.: Famf-net: Feature alignment mutual attention fusion with region awareness for breast cancer diagnosis via imbalanced data. IEEE Transactions on Medical Imaging (2024)
11. Pawłowska, A., Ćwierz-Pieńkowska, A., Domalik, A., Jaguś, D., Kasprzak, P., Matkowski, R., Fura, Ł., Nowicki, A., Żołek, N.: Curated benchmark dataset for ultrasound based breast lesion analysis. Scientific Data **11**(1), 148 (2024)
12. Thomas, C., Byra, M., Marti, R., Yap, M.H., Zwiggelaar, R.: Bus-set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets. Medical physics **50**(5), 3223–3243 (2023)
13. Vallez, N., Bueno, G., Deniz, O., Rienda, M.A., Pastor, C.: Bus-uclm: Breast ultrasound lesion segmentation dataset. Scientific Data **12**(1), 242 (2025)
14. Wan, P., Zhang, S., Shao, W., Zhao, J., Yang, Y., Kong, W., Xue, H., Zhang, D.: Correlation-adaptive multi-view ceus fusion for liver cancer diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 188–197. Springer (2024)
15. Watanabe, T., Yamaguchi, T., Okuno, T., Konno, S., Takaki, R., Sato, M., Tsuruoka, M., Shirai, H., Ogawa, Y., Ban, K., et al.: Utility of b-mode, color doppler and elastography in the diagnosis of breast cancer: results of the cd-confirm multicenter study of 1351 breast solid masses. Ultrasound in Medicine & Biology **47**(11), 3111–3121 (2021)
16. Wu, C., Shuai, Z., Tang, Z., Wang, L., Shen, L.: Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In: The Thirteenth International Conference on Learning Representations (2025)
17. Yan, L., Liang, Z., Zhang, H., Zhang, G., Zheng, W., Han, C., Yu, D., Zhang, H., Xie, X., Liu, C., et al.: A domain knowledge-based interpretable deep learning system for improving clinical breast ultrasound diagnosis. Communications Medicine **4**(1), 90 (2024)
18. Yoon, J.S., Oh, K., Shin, Y., Mazurowski, M.A., Suk, H.I.: Domain generalization for medical image analysis: A review. Proceedings of the IEEE (2024)
19. Yoon, J.H., Kim, M.H., Kim, E.K., Moon, H.J., Kwak, J.Y., Kim, M.J.: Interobserver variability of ultrasound elastography: how it affects the diagnosis of breast lesions. American Journal of Roentgenology **196**(3), 730–736 (2011)

20. Yu, A.C., Mohajer, B., Eng, J.: External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiology: Artificial Intelligence **4**(3), e210064 (2022)
21. Zhang, J., Wu, J., Zhou, X.S., Shi, F., Shen, D.: Recent advancements in artificial intelligence for breast cancer: Image augmentation, segmentation, diagnosis, and prognosis approaches. In: Seminars in Cancer Biology. vol. 96, pp. 11–25. Elsevier (2023)
22. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)
23. Zhao, J., Li, X., Yang, F., Zhai, Q., Luo, A., Zhao, Y., Cheng, H., Fu, H.: Mexd: An expert-infused diffusion model for whole-slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20789–20799 (2025)
24. Zheng, X., Chen, X., Gong, S., Griffin, X., Slabaugh, G.: Xfmamba: Cross-fusion mamba for multi-view medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 672–682. Springer (2025)