RACA: Rule-Aligned Collaborative Agents for Evidence-Grounded Breast Ultrasound Classification

Lingyu Chen[1,5*], Yue Wang[2*], Cheng Li[3], Lin Jin[4], Daoqiang Zhang[1], Hongen Liao[2], Fang Chen[2†], and Qingjie Meng[5†]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4] Shanghai University of Traditional Chinese Medicine, Shanghai, China
[5] University of Birmingham, Birmingham, UK

lingyucher@163.com

Abstract. Breast cancer is highly prevalent worldwide. While breast ultrasound (BUS) serves as a primary screening modality, its reliable diagnostic interpretation remains inherently challenging. Most existing BUS classification methods prioritize direct image-to-label mapping and overall accuracy, neglecting the alignment with clinical diagnostic rules and decision traceability. Such limitations hinder their integration into rule-governed clinical workflows and diminish the trustworthiness of automated diagnostic decisions. Although agent-based frameworks offer a modular paradigm for organizing complex decision processes, their potential for rule-aligned and fully traceable BUS diagnosis remains largely unexplored. To address these challenges, we introduce RACA, a Rule-Aligned Collaborative Agent framework for evidence-grounded BUS classification. Our RACA introduces a novel, three-stage framework driven by explicit clinical rules. First, a perception module consolidates morphological and semantic evidence into structured diagnostic evidence. Then, a risk modeling agent aggregates the evidence under explicit clinical constraints to derive a rule-aligned risk estimate and an initial prediction. Finally, a critic-guided calibration agent refines predictions through memory-based retrieval. Extensive experiments on three BUS benchmarks demonstrate the superiority of RACA over the representative classification and medical agent-based frameworks. Comprehensive interpretability analyses and case demonstrations validates that its traceable decision pathway aligns with clinical diagnosis. The project page is available at: https://github.com/Swecamellia/RACA.

Keywords: Collaborative Agents · Diagnostic Rules · Evidence-Grounded · BUS Classification.

[*] Equal contribution
[†] Corresponding authors

1 Introduction

Breast ultrasound (BUS) is a cornerstone modality for breast lesion assessment [16, 18]. Nevertheless, interpretation remains challenging due to large speckle noise and acoustic artifacts that obscure critical morphology [14, 4]. In clinical practice, diagnosis relies on standardized guidelines [20, 25], most notably BI-RADS [24], which organizes defined imaging descriptors such as shape, margin, orientation, and posterior features into a rule-governed malignancy risk assessment [8, 6]. Crucially, this diagnostic process is inherently evidence-driven: clinical decisions are not based on isolated image features, but rather on the rule-aligned aggregation of fused evidence into calibrated risk estimates.

Recent advances in BUS classification [10, 2] have led to substantial performance gains, yet they fail to align with the rule-governed structure of clinical diagnosis. Most methods are optimized solely for image-level accuracy, and their predictions are rarely anchored to specific guideline-defined imaging descriptors [3]. This opacity weakens traceability and interpretability, constraining the their reliability within structured clinical workflows [29, 19]. This underscores the urgent need for BUS classification methods that explicitly align clinical rules to provide evidence-grounded predictions.

Agent-based frameworks in medical artificial intelligence serve as modular paradigms for modeling complex decision processes [27, 13]. By coordinating domain knowledge, clinical standards, and analytical components, they support structured information integration and explicit representation of diagnostic rules. However, the systematic adoption of agents for BUS remains largely underexplored due to a structural mismatch: existing agents are predominantly optimized for open-ended generative tasks that favor flexibility over strict clinical adherence [28], while BUS classification requires conclusions to be strictly governed by diagnostic criteria. Such clinical rules are rarely instantiated as explicit binding constraints within current agent architectures [22]. Furthermore, although referencing prior cases is crucial for refining diagnostic decisions [17], its principled integration under diagnostic rules constraints remains insufficiently formalized in existing models.

To address these limitations, we propose RACA, a Rule-Aligned Collaborative Agent framework that reformulates BUS classification into a structured decision process grounded in traceable imaging evidence. RACA operates in three stages: First, an evidence-grounded perception module employs three agents to jointly extract morphological measurements and semantic interpretations, generating structured diagnostic evidence. Second, a risk modeling agent emulates clinical risk stratification and maps the diagnostic evidence to a continuous malignancy risk, producing an initial rule-constrained classification prediction. Third, to resolve local evidence conflicts and emulate clinical prior-case verification, a critic-guided decision calibration agent leverages a memory retrieval mechanism to selectively refine the initial prediction, producing a more reliable final classification outcome. Our main contributions are as follows:

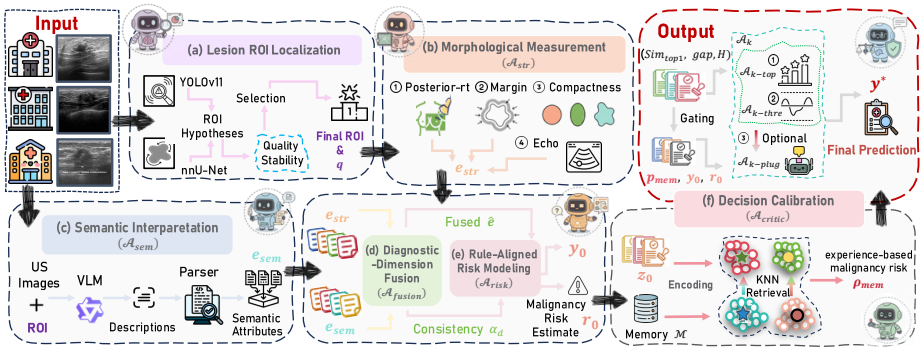


Fig. 1: Overview of the proposed RACA. (a) Lesion ROI is localized using a combination of existing detection and segmentation tools. (b-d) In the stage I, three perception agents extract and consolidate morphological and semantic evidence into structured diagnostic evidence. (e) In the stage II, diagnosis evidence is evaluated to derive a continuous malignancy risk estimate and an initial prediction via a rule-aligned risk modeling agent. (f) In stage III, a critic-guided decision calibration agent further refines the prediction through prior-case retrieval.

- We propose a rule-aligned collaborative AI agent framework for BUS classification. By anchoring the decision pathway in clinical guidelines, it enables a traceable risk assessment that aligns with the clinical diagnostic process.
- We develop a novel three-stage framework driven by clinical rules: learning structured diagnostic evidence, establishing an initial risk estimate, and performing critic-guided decision calibration with prior-case retrieval, thereby progressively mitigating evidence instability for a reliable final prediction.
- Beyond conventional accuracy, we deliver a rigorous analysis of diagnostic traceability, demonstrating that RACA forms stable and verifiable decision patterns that align with real-world clinical diagnosis.

2 Method

As demonstrated in Fig. 1, given a BUS image x, RACA formulates the diagnosis as a structured, three-stage decision process governed by standardized clinical rules. Stage I extracts clinically meaningful imaging evidence to form structured diagnostic evidence. Perception agents $\{\mathcal{A}_{str}, \mathcal{A}_{sem}, \mathcal{A}_{fusion}\}$ extract structural and semantic evidence from x and produce fused diagnostic evidence $\hat{\mathbf{e}}$, as depicted in Fig. 1(b)–(d). Stage II performs a risk modeling in Fig. 1(e). A rule-aligned risk modeling agent $\mathcal{A}_{risk}$ maps $\hat{\mathbf{e}}$ to an initial malignancy risk $r_0 \in [0, 1]$ and initial prediction y_0. Stage III conducts the decision calibration, as shown in the Fig. 1(f). A critic-guided decision calibration agent $\mathcal{A}_{critic}$ retrieves top-k prior cases to estimate a memory-based malignancy risk p_{mem}, which refines y_0 and generates the final prediction y^* through selective calibration.

Lesion ROI Localization. RACA initializes by extracting a validated region of interest (ROI) from the input BUS image. First, a hybrid localization strategy combining YOLO-based [11] and nnU-Net [9] is utilized to generate multiple ROI hypotheses. Then, we quantify each hypothese's quality (confidence and morphological plausibility) and stability (localization robustness). Finally, a soft selection outputs the final ROI and its normalized reliability score $q \in [0, 1]$ to the downstream agents, propagating q as a confidence signal in the stage II.

2.1 Evidence-Grounded Perception (EGP)

To enable rule-governed classification, EGP replaces implicit feature extraction with explicit, clinically aligned evidence construction in stage I. It anchors heterogeneous cues (e.g., morphologic measurements and semantic interpretations) to a unified diagnostic space: $\mathcal{D} = \{\text{shape}, \text{margin}, \text{echogenicity}, \text{posterior}\}$, which specifies standardized diagnostic attributes for rule-based modeling. The construction of structured evidence is realized by three collaborative agents: two independently extract quantitative morphological and semantic evidence, while the third fuses them into auditable diagnostic evidence, as shown in Fig. 1(b)-(d).

Morphological Measurement Agent ($\mathcal{A}_{str}$). Taking the ROI as input, $\mathcal{A}_{str}$ computes morphological and acoustic measurements, such as compactness, circularity, margin sharpness, echo contrast and posterior ratio, which forms quantifiable and traceable structural evidence $\mathbf{e}_{str} = \{e_d^{str}\}_{d\in\mathcal{D}}$.

Semantic Interpretation Agent ($\mathcal{A}_{sem}$). To complement structural evidence from $\mathcal{A}_{str}$, $\mathcal{A}_{sem}$ employs a vision-language model (VLM) to inject diagnostic semantics into the RACA. Through a rule-guided parsing function $\text{Parse}(\cdot)$, raw VLM outputs are normalized into a evidence-confidence pair over $\mathcal{D}$, producing semantic evidence $\mathbf{e}_{sem} = \{(v_d, c_d)\}_{d\in\mathcal{D}}$. Here, v_d denotes the discrete attribute value and c_d its confidence.

Auditable Diagnostic-attribute Fusion Agent ($\mathcal{A}_{fusion}$). To integrate $\mathbf{e}_{str}$ and $\mathbf{e}_{sem}$, $\mathcal{A}_{fusion}$ performs attribute-wise fusion over $\mathcal{D}$. For each attribute $d \in \mathcal{D}$, it first evaluates the structural-semantic consistency $\alpha_d = \mathcal{F}(e_d^{str}, v_d)$, where $\alpha_d \in \{\text{agree}, \text{weak-conflict}, \text{strong-conflict}\}$. Guided by α_d and ROI reliability q, a rule-conditioned operator $\hat{e}_d = \mathcal{G}(e_d^{str}, v_d, c_d, \alpha_d, q)$ produces fused diagnostic evidence $\hat{\mathbf{e}} = \{\hat{e}_d\}_{d\in\mathcal{D}}$. Formally, semantic evidence are retained under agreement or weak conflict; structural evidence is prioritized under strong conflict with sufficient reliability; otherwise, attribute-level uncertainty are preserved to avoid over-confident decisions. Thus, each $\hat{e}_d$ encapsulates all continuous, semantic, and auditing data (q, α_d) necessary for downstream modeling.

2.2 Rule-Aligned Risk Modeling Agent ($\mathcal{A}_{risk}$)

To mimic the risk-aware clinical assessment in BUS, $\mathcal{A}_{risk}$ translates the fused diagnostic evidence $\hat{\mathbf{e}}$ into a malignancy risk estimate, as illustrated in the Fig. 1(e).

First, each fused evidence $\hat{e}_d$ is mapped to a risk contribution $s_d = \mathcal{S}_d(\hat{e}_d)$, where $\mathcal{S}_d(\cdot)$ enforces clinically motivated risk weighting. Attributes associated with higher malignancy risk (e.g., spiculated margins) are assigned larger scores

than low-risk features, implicitly encoding attribute importance through rule-governed aggregation. Then an initial malignancy risk estimate is computed as:

$$r_0 = \sigma\left(\frac{\sum_{d\in\mathcal{D}} s_d - \lambda\,\psi(\{\alpha_d\})}{T_0 + \kappa P}\right), \tag{1}$$

where $\psi(\{\alpha_d\})$ models structural-semantic consistency across attributeal contributions, and P summarizes evidence instability propagated from degraded ROI quality. Larger P yields more conservative risk estimates. $T_0 > 0$ is a fixed baseline temperature hyper-parameter controlling risk sensitivity, $\kappa \geq 0$ scales instability modulation, and $\lambda \geq 0$ governs consistency penalization. Finally, a preliminary binary prediction is acquired as $y_0 = \mathbb{I}(r_0 \geq \tau_0)$, where τ_0 denotes a predefined rule threshold. In practice, the effective decision threshold may be adaptively adjusted according to internal decision statistics, while preserving rule consistency. It also compiles the decision state $\mathbf{z}_0 = \Big[r_0,\ \{\alpha_d\}_{d\in\mathcal{D}},\ \{\hat{e}_d\}_{d\in\mathcal{D}},\ P\Big]$ for subsequent calibration.

2.3 Critic-Guided Decision Calibration Agent ($\mathcal{A}_{critic}$)

To resolve uncertainty in the initial rule-aligned risk r_0, we introduce a critic-guided decision calibration agent $\mathcal{A}_{critic}$ with memory $\mathcal{M}$ for external consistency checking and selective calibration, directly emulating clinical prior-case retrieval and verification, as presented in the Fig. 1(f).

Memory Retrieval. $\mathcal{A}_{critic}$ first normalizes the decision state as $\tilde{\mathbf{z}}_0 = (\mathbf{z_0} - \mu)/\sigma$. Using cosine similarity $\mathrm{sim}_i = \cos(\tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_i)$, it retrieves top-$k$ neighbors $\mathcal{N}_k$ in $\mathcal{M}$. The memory-based malignancy risk p_{mem} is calculated as:

$$p_{\mathrm{mem}} = \sum_{i\in\mathcal{N}_k} \frac{\exp(\mathrm{sim}_i/\tau)}{\sum_{j\in\mathcal{N}_k}\exp(\mathrm{sim}_j/\tau)}\, y_i, \tag{2}$$

which reflects empirical malignancy proportion under similar decision states.

Reliability-Aware Memory Gating. To prevent unwarranted over-correction, the influence of p_{mem} is adaptively gated by the neighborhood reliability statistics $\mathcal{S} = (\mathrm{Sim}_{\mathrm{top1}}, \mathrm{gap}, H)$, which characterize similarity and label distribution stability. Here, $\mathrm{Sim}_{\mathrm{top1}}$ denotes the highest similarity score while gap is the difference between the top-2 neighbors. H is the entropy of neighbor labels. Only when neighborhood labels are consistent and H is low, p_{mem} is admitted for decision refinement; otherwise, its influence is suppressed to avoid over-correction. When admitted, p_{mem} modulates the subsequent selective calibration reactions.

Multi-Agent Selective Calibration. We construct calibration agents $\{\mathcal{A}_k\}$ that execute complementary strategies under rule constraints to ensure a highly reliable final prediction. Each agent generates a conditional candidate update:

$$\tilde{y}_k = \mathcal{A}_k(y_0, r_0, p_{\mathrm{mem}}, \mathcal{S}). \tag{3}$$

$\{\mathcal{A}_k\}$ includes (i) Dual-Threshold veto/rescue $\mathcal{A}_{k-thre}$, which applies asymmetric lower and upper bounds on p_{mem} to selectively veto overconfident malignant

Table 1: Quantitative comparison on BUSI, BUV and BUSBRA datasets.

	BUSI				BUV				BUSBRA			
Method	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
Fully-supervised												
LKA [32]	0.7353	0.6765	0.7667	0.7188	0.7647	0.7500	0.5000	0.6000	0.7979	0.7447	0.5738	0.6481
DiffMICv2 [30]	0.7500	0.7485	0.7518	0.7486	0.7647	0.7596	0.7045	0.7167	0.7606	0.7419	**0.6780**	**0.6915**
Semi-supervised												
OPENSSC [7]	0.7206	0.7540	0.7395	0.7191	0.7059	**0.8438**	0.5833	0.5503	0.7394	0.7188	0.6410	0.6500
PEAT [31]	0.6765	0.7381	0.6404	0.6211	0.6471	0.3235	0.5000	0.3929	0.7447	**0.7512**	0.6321	0.6382
Medical agent-based												
Medagents [27]	0.6119	0.6423	0.6297	0.6077	0.5294	0.5208	0.5227	0.5143	0.5260	0.5271	0.5312	0.5103
MDAgents [12]	0.5588	0.5000	**0.8000**	0.6154	0.7059	0.6000	0.5000	0.5455	0.6330	0.4524	0.6230	0.5241
Ours	**0.8088**	**0.7742**	**0.8000**	**0.7869**	**0.8235**	0.7143	**0.8333**	**0.7692**	**0.8085**	0.7273	0.6557	0.6897

predictions and rescue high-risk benign cases; (ii) Top-k rescue $\mathcal{A}_{k-top}$, which ranks gated benign predictions using a weighted score over p_{mem}, $\mathcal{S}$ and r_0, and then flips the top-k benign predictions as malignant to rescue likely false negatives; and (iii) an optional plug-in re-check $\mathcal{A}_{k-plug}$, which incorporates auxiliary confidence signals when permitted by rules. The final prediction y^* is determined by a rule-constrained selection policy $y^* = \Pi\big(y_0, \{\tilde{y}_k\}\big)$, where $\Pi(\cdot)$ integrates candidate updates according to rule priority and consistency constraints.

3 Experiment

Datasets and Implementation. We conduct experiments on three publicly available BUS datasets: BUSI [1], BUV [15], and BUSBRA [5]. We use AutoGPT [23] to build our agent framework. The calibration $\mathcal{A}_{k-plug}$ employs an EfficientNet [26] pretrained on BUSC [21], which is independent of the target datasets, ensuring no data leakage. Each dataset randomly divided into train, validation, and test sets with a 7:2:1 ratio. RACA only utilizes the validation split exclusively for hyperparameter selection and memory construction, while all baselines adopt identical data splits for fair comparison. Runtime is 31.40 s/image on one RTX 4090 24GB GPU, with most latency from VLM (31.38 s/image).

Evaluation Metrics. BUS classification performance is evaluated by standard metrics: Accuracy (ACC), Precision (PRE), Recall (REC), and F1-score (F1).

Comparison Results. We compare RACA against representative fully-supervised (LKA [32], DiffMICv2 [30]), semi-supervised (OPENSSC [7], PEAT [31]), and general medical agent frameworks (Medagents [27], MDAgents [12]) across three datasets. Table 1 shows that RACA achieves superior overall performance. When compared to the strongest fully-supervised baselines, RACA achieves substantial accuracy improvements, demonstrating that explicit rule-aligned decision modeling contributes to improved precision compared with conventional implicit feature learning. Against semi-supervised methods, RACA maintains a more balanced precision-recall trade-off. Crucially, RACA outperforms general

Table 2: Ablation study on BUSI and BUSBRA datasets.

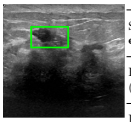
Modules					BUSI				BUSBRA			
EGP	$\mathcal{A}_{risk}$	$\mathcal{A}_{k-top}$	$\mathcal{A}_{k-thre}$	$\mathcal{A}_{k-plug}$	ACC	PRE	REC	F1	ACC	PRE	REC	F1
✓	✓				0.6176	0.5909	0.4333	0.5000	0.5212	0.3164	0.4098	0.3571
✓	✓	✓			0.7054	0.6744	0.6042	0.6374	0.6223	0.4151	0.3607	0.3860
✓	✓	✓	✓		0.7941	0.7667	0.7667	0.7667	0.7713	**0.7647**	0.4262	0.5474
✓	✓	✓	✓	✓	**0.8088**	**0.7742**	**0.8000**	**0.7869**	**0.8085**	0.7273	**0.6557**	**0.6897**

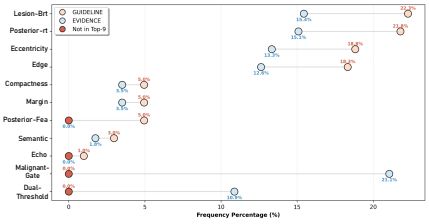
medical agents by massive margins exceeding 19% in accuracy, demonstrating the absolute necessity of explicitly integrating domain-specific diagnostic rules.

Ablation Studies. We evaluate the contribution of each component in RACA on the BUSI and BUSBRA. In Table 2, while baseline (EGP + $\mathcal{A}_{risk}$) establishes an evidence-grounded and rule-aligned foundation, sequentially integrating multiple calibration agents ($\mathcal{A}_{k\text{-}top}$, $\mathcal{A}_{k\text{-}thre}$, and $\mathcal{A}_{k\text{-}plug}$) achieves substantial, step-wise performance gains. This steady upward trajectory validates our core design philosophy: decision calibration with prior-case retrieval acts as a powerful, structured enhancement of rule-based inference rather than a replacement.

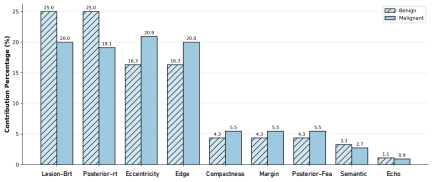
Diagnostic Interpretability. Moving beyond overall classification accuracy, we further assess whether the proposed RACA forms a stable, traceable, and clinically consistent decision-making process. We conduct a systematic experimental analysis of the structured feature representations within the evidence acquired by RACA. Specifically, we investigate the relationship between established BI-RADS guideline-based features and the explicit diagnostic evidence constructed by RACA. For each sample, we compute the normalized frequency of each feature within the top-9 ranked evidence, which reflects its relative contribution to the final decision-making process. Fig. 2(a) shows that the distribution of **guideline-based evidence** tightly aligns with the **overall structured evidence**, demonstrating that RACA strictly preserves clinical rules while robustly expanding its representational capacity to capture subtle diagnostic evidence. Fig. 2(b) shows a stark contrast: malignant predictions are heavily driven by invasive morphological features (e.g., Eccentricity, Edge), whereas benign predictions rely predominantly on acoustic intensity evidence (e.g., Lesion-Brt, Posterior-rt), proving the model naturally replicates real-world clinical diagno-

Table 3: Traceable rule-aligned inference process of RACA on a test sample. ↑ and ↓ indicate malignant and benign directional tendencies, respectively.

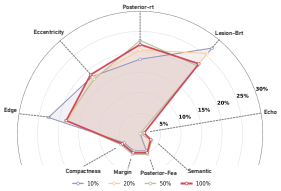
BUS Image	Module	Traceable Rule-Aligned Diagnostic Evidence (BUS)
	Morphological $\mathbf{e}_{str}$ ($\mathcal{A}_{str}$)	{Compactness:0.589 (Irregular ↑); AspectRatio (b/a):1.800 (Taller-than-wide ↑); Lesion-Brt:0.991 (Hypoechoic ↑); Edge:54.3 (Indistinct ↑)}
	Semantic $\mathbf{e}_{sem}$ ($\mathcal{A}_{sem}$)	{Margin: indeterminate (0.60; conflict: spiculated(0.75) vs indistinct) ↑; Compactness: irregular (0.65; uneven/jagged contour) ↑; Echo: hypoechoic (0.57; darker than parenchyma) ↑; Posterior-fea: shadowing (0.53; acoustic shadow band) ↑}
	Fused Evidence $\hat{\mathbf{e}}$ ($\mathcal{A}_{fusion} + \mathcal{A}_{risk}$)	{Aggregation: $S = \sum_i \text{Contrib}_i = 1.695$; StrongCue: Margin (strong malignant margin, +0.78); GateBias: activated (+0.30); OtherContributions: weak/gated evidence (e.g., ratio, strength, echogenicity); ⇒ **Probability:** $P(\text{malignant}) = 0.82$}
	Decision Gates ($\mathcal{A}_{critic}$)	{DecisionRouting: MalignancyGate: activated (1 strong, 3 supportive cues); Consistency: satisfied; Threshold: 0.60; **FinalDecision:** $P(\text{malignant}) > 0.60 \Rightarrow$ Malignant; (**GT: Malignant**)}



(a) Alignment between guideline-based evidence and overall structured evidence.



(b) Evidence contribution comparison between benign and malignant cases.



(c) Evidence contribution patterns under different sampling ratios.

Type	Evidence	Definition	Role
Continuous Structured Evidence	Lesion-Brt	$R_{lb} = \mu_{lesion}/\mu_{bg}$ (ROI mean intensity ratio)	Hypoechoic tendency ↑
	Posterior-rt	$R_{post} = \mu_{below}/\mu_{above}$ (posterior vs. anterior band)	Shadowing tendency ↑ (enhancement ↓)
	Eccentricity	$\epsilon = 1 - \frac{b}{a}$ (ellipse fit; major/minor axes a, b)	Elongated shape ↑
	Compactness	$C = \frac{4\pi A}{P^2}$ (A: area, P: perimeter of mask)	Irregular contour ↑
	Edge	$E = \frac{1}{\lvert\partial\Omega\rvert} \sum_{x \in \partial\Omega} \lVert\nabla I(x)\rVert$	Indistinct boundary ↑ (sharp/circumscribed ↓)
Categorical Semantic Evidence	Compactness	{round, oval, irregular}	Irregular ↑; round/oval ↓
	Margin	{circumscribed, indistinct, angular, microlobulated, spiculated}	Spiculated/angular/microlobulated ↑; circumscribed ↓
	Posterior-fea	{shadowing/attenuation, enhancement, none}	Shadowing/attenuation ↑; enhancement ↓
	Echo	{hypoechoic, isoechoic, hyperechoic, anechoic/cystic}	Anechoic/cystic ↓ (benign pattern cue)
	Semantic	{benign, malignant} (global summary)	Weak prior (down-weighted / gated)
Decision Gates	Malignant_Gate	Boolean rule on strong/consistent malignant cues	Enables malignant-path semantics; blocks weak overcalls
	Dual_Threshold	Dual-threshold routing using p and reliability	Determined vs. uncertain; conservative FP control

(d) Hierarchical layout of diagnosis evidence components.

Fig. 2: Comprehensive analysis of the proposed RACA across consistency, discrimination, and hierarchical representation on BUSI dataset. ↑ / ↓ indicates direction toward malignant/benign.

sis. Fig. 2(c) shows highly overlapping feature radar trajectories across different sampling ratios, indicating that RACA's diagnostic logic is exceptionally stable and anchored in fundamental rules rather than dataset size biases. Finally, Fig. 2(d) shows the hierarchical organization of perceptual components, explicitly mapping how low-level features aggregate into high-level decisions to guarantee that every decisions is fully traceable and backed by verifiable clinical evidence.

Case Demonstration. To explicitly demonstrate our framework's diagnostic traceability, we show the inference process of RACA using a randomly selected test sample. Table 3 presents the step-by-step intermediate output of each stage in RACA. The green box denotes the lesion ROI extracted in Fig. 1(a). In Stage I, $\mathbf{A}_{str}$ converts quantitative morphological measurements into clinically aligned diagnostic evidence. For example, AspectRatio $b/a = 1.800(> 1)$ exceeds the rule threshold of 1, indicating a taller-than-wide morphology (↑) under the predefined clinical rules, collectively supporting a malignant tendency. Similarly, $\mathbf{A}_{sem}$ projects raw outputs ("uneven contour" and "darker than parenchyma") of VLM into categorical evidence, where irregular (0.65) and hypoechoic (0.57) are interpreted as malignant-aligned evidence, with values indicating confidence. In Stage II, $\mathcal{A}_{risk}$ maps each fused evidence $\hat{e}_d$ to a attribute-wise contribution $s_d = \mathcal{S}_d(\hat{e}_d)$ and aggregating them into a malignancy risk estimate. Malignant margin evidence contribute +0.78, gate activation adds +0.30, and remaining

features provide smaller increments, resulting in an aggregated score $S = 1.695$, which is mapped through $\sigma(\cdot)$ to obtain $P(\text{malignant}) = 0.82$. In Stage III, $\mathcal{A}_{critic}$ verifies rule consistency and confirms malignancy gate activation (1 strong + 3 supportive evidence). As $P(\text{malignant}) > 0.60$ without rule-level conflict, the decision is routed to *Malignant*, consistent with the ground truth.

4 Conclusion

We propose RACA, a collaborative agent framework that grounds BUS classification in explicit rule-aligned clinical diagnostics. By orchestrating a traceable pipeline of evidence construction, risk modeling, and critic-guided calibration, RACA ensures transparent and calibrated predictions. Extensive evaluations across three BUS benchmarks demonstrate that our framework consistently outperforms representative classification and medical agent models. Future work will integrate chain-of-thought reasoning to further enrich interpretability and clinical workflow simulation.

Acknowledgments. This work was supported by National Nature Science Foundation of China Grants (62271246, 82572314), the Science and Technology Commission of Shanghai Municipality (Nos.24511104100). We acknowledge the computational resources provided by the Baskerville high-performance computing (HPC) facility at the University of Birmingham.

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)
2. Aumente-Maestro, C., Díez, 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)
3. Castelvecchi, D.: Can we open the black box of ai? Nature News **538**(7623), 20 (2016)
4. Chen, F., Chen, L., Han, H., Zhang, S., Zhang, D., Liao, H.: The ability of segmenting anything model (sam) to segment ultrasound images. BioScience Trends **17**(3), 211–218 (2023)
5. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W.: Busbra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical physics **51**(4), 3110–3123 (2024)
6. Gu, Y., Tian, J.W., Ran, H.T., Ren, W.D., Chang, C., Yuan, J.J., Kang, C.S., Deng, Y.B., Wang, H., Luo, B.M., et al.: The utility of the fifth edition of the bi-rads ultrasound lexicon in category 4 breast lesions: a prospective multicenter study in china. Academic radiology **29**, S26–S34 (2022)

7. He, A., Li, T., Zhao, Y., Zhao, J., Fu, H.: Open-set semi-supervised medical image classification with learnable prototypes and outlier filter. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 492–501. Springer (2024)
8. Hong, A.S., Rosen, E.L., Soo, M.S., Baker, J.A.: Bi-rads for sonography: positive and negative predictive values of sonographic features. American Journal of Roentgenology **184**(4), 1260–1265 (2005)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Jiang, T., Li, Y., Li, Y., Xing, W., Yu, M., Xie, F., Ta, D.: A segmentation knowledge-based global-local attention network for tumor classification in breast ultrasound images. Pattern Recognition p. 112152 (2025)
11. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
12. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)
13. Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.Q., Ma, W., et al.: Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957 (2024)
14. Li, M., Fang, Y., Shao, J., Jiang, Y., Xu, G., Cui, X.w., Wu, X.: Vision transformer-based multimodal fusion network for classification of tumor malignancy on breast ultrasound: A retrospective multicenter study. International Journal of Medical Informatics **196**, 105793 (2025)
15. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–623. Springer (2022)
16. Luo, L., Wang, X., Lin, Y., Ma, X., Tan, A., Chan, R., Vardhanabhuti, V., Chu, W.C., Cheng, K.T., Chen, H.: Deep learning in breast cancer imaging: A decade of progress and future directions. IEEE Reviews in Biomedical Engineering (2024)
17. Parola, M., Galatolo, F.A., La Mantia, G., Cimino, M.G., Campisi, G., Di Fede, O.: Towards explainable oral cancer recognition: Screening on imperfect images via informed deep learning and case-based reasoning. Computerized Medical Imaging and Graphics **117**, 102433 (2024)
18. Qian, X., Pei, J., Han, C., Liang, Z., Zhang, G., Chen, N., Zheng, W., Meng, F., Yu, D., Chen, Y., et al.: A multimodal machine learning model for the stratification of breast cancer risk. Nature Biomedical Engineering **9**(3), 356–370 (2025)
19. Rahman, M.A.: Hyformer-net: A synergistic cnn-transformer with interpretable multi-scale fusion for breast lesion segmentation and classification in ultrasound images. arXiv preprint arXiv:2511.01013 (2025)
20. Ren, W., Chen, M., Qiao, Y., Zhao, F.: Global guidelines for breast cancer screening: a systematic review. The Breast **64**, 85–99 (2022)
21. Rodrigues, P.S.: Breast ultrasound image. Mendeley Data **1**(10.17632) (2017)
22. Roeschl, T., Hoffmann, M., Hashemi, D., Rarreck, F., Hinrichs, N., Trippel, T.D., Gröschel, M.I., Unbehaun, A., Klein, C., Kempfert, J., et al.: Assessing the limitations of large language models in clinical practice guideline–concordant treatment decision-making on real-world data: retrospective study. Jmirx med **6**, e74899 (2025)

23. Significant Gravitas: Autogpt (2023), https://github.com/Significant-Gravitas/AutoGPT, accessed: 2025-02-27
24. Spak, D.A., Plaxco, J., Santiago, L., Dryden, M., Dogan, B.: Bi-rads® fifth edition: A summary of changes. Diagnostic and interventional imaging **98**(3), 179–190 (2017)
25. Stavros, A.T.: Breast ultrasound. Lippincott Williams & Wilkins (2004)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
27. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: Medagents: Large language models as collaborators for zero-shot medical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 599–621 (2024)
28. Xu, G., Li, X., Chen, Y., Duan, Y., Wu, S., Yu, A., Chiu, C.H., Ni, J., Tang, N., Li, T.J.J., et al.: A comprehensive survey of agentic ai in healthcare. Authorea Preprints (2025)
29. 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)
30. 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)
31. Zeng, Q., Xie, Y., Lu, Z., Xia, Y.: Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15671–15680 (2023)
32. Zhu, Z., Lu, S.Y., Huang, T., Liu, L., Liu, Z.: Lka: L arge k ernel a dapter for enhanced medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 394–404. Springer (2025)