



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Latent-CURE: Interpretable Breast Cancer Diagnosis via Dual-Asymmetric Chain-of-Thought

Weiyi Zhao<sup>1</sup>, Xiaoyu Tan<sup>2</sup>, Lu Gan<sup>3</sup>, Liang Liu<sup>3</sup>, and Xihe Qiu<sup>1,✉</sup>

<sup>1</sup> School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China  
qiuxihe1993@gmail.com

<sup>2</sup> Tencent YouTu Lab, Shanghai 200232, China

<sup>3</sup> Integrative Clinical Research Ward, Clinical Medicine Research Institute, Zhongshan Hospital, Fudan University, Shanghai 200032, China

**Abstract.** Multimodal Large Models have significantly advanced automated breast ultrasound diagnosis. However, most existing frameworks utilize opaque, end-to-end paradigms prioritizing global statistical correlations over structured clinical reasoning. Consequently, these models remain susceptible to shortcut learning amid extreme real-world epidemiological imbalances, often bypassing rare but decisive malignant indicators for dominant benign patterns. To address this disconnect, we propose Latent-CURE, a novel diagnostic framework driven by asymmetric weighted chain-of-thought methodology grounded in latent space reasoning. Unlike traditional approaches, our framework constructs an implicit reasoning trajectory forcing the model to sequentially infer standardized BI-RADS morphological descriptors before converging on a final diagnosis. Furthermore, to combat the extreme scarcity of critical malignant features, we couple this architecture with a dual-asymmetric optimization strategy. By dynamically adjusting margins and weights, this strategy safeguards high-specificity malignant descriptors from being overshadowed by common benign priors. Comprehensive evaluations demonstrate that our knowledge-injected approach provides transparent clinical evidence while achieving robust, accurate diagnostic performance in imbalanced medical cohorts.

**Keywords:** Breast Ultrasound · Multimodal Large Models · Latent Reasoning

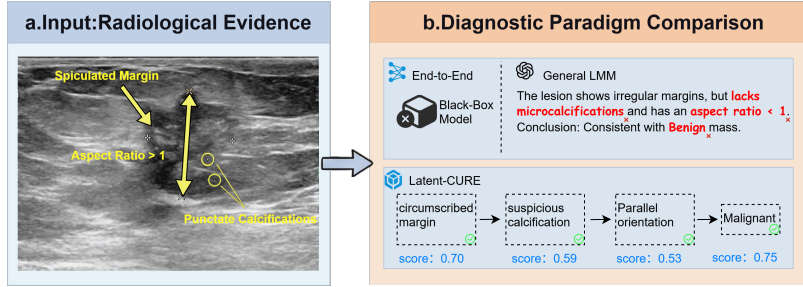
## 1 Introduction

Breast cancer detection relies heavily on breast ultrasound (BUS), a non-invasive modality requiring radiologists to meticulously evaluate morphological features, including margins, echogenicity, and orientation, according to the standardized Breast Imaging-Reporting and Data System (BI-RADS) [3, 25, 26]. To reduce diagnostic variability and operator dependence, clinical artificial intelligence has

---

✉ Corresponding author.

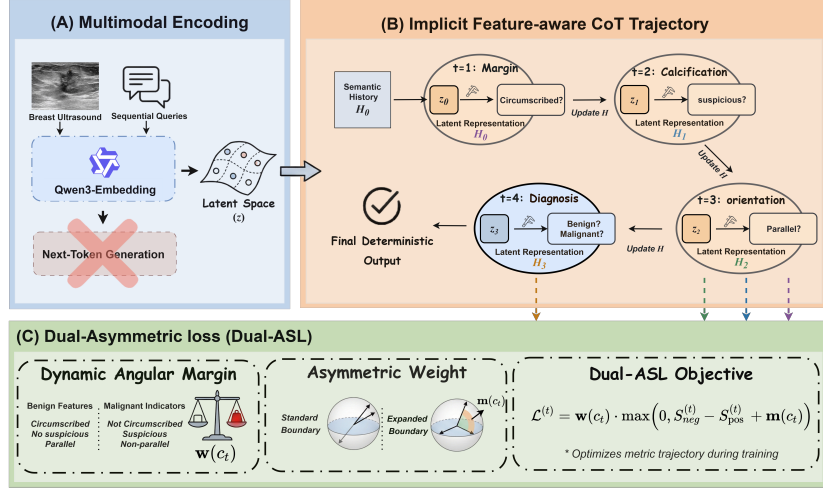
rapidly transitioned toward the adoption of Multimodal Large Models (MLMs) [18, 23, 1]. Pre-trained on massive datasets, these foundational architectures exhibit remarkable representational capabilities, driving state-of-the-art performance in automated lesion classification and segmentation [12, 7].



**Fig. 1.** Diagnostic challenge and our proposed paradigm. Compared to black-box LMMs prone to shortcut learning, Latent-CURE ensures transparent diagnosis via step-wise morphological reasoning.

Despite the representational power of these systems, existing LMM-based diagnostic frameworks predominantly operate under an opaque end-to-end paradigm that maps deep visual embeddings directly to diagnostic labels, prioritizing global statistical correlations over structured clinical reasoning [29, 13, 24]. This creates a critical vulnerability rooted in the extreme epidemiological imbalance of real-world medical cohorts, where benign lesions vastly outnumber malignant cases [15]. Consequently, even advanced foundation models are highly susceptible to shortcut learning. Rather than identifying statistically rare but decisive malignant indicators, such as non-parallel growth or subtle microcalcifications, these models tend to bypass them in favor of dominant, easily recognizable benign patterns, as illustrated in Fig. 1. This failure to navigate clinical imbalance fundamentally severs the link between the latent decisions of the model and rigorous expert logic [11, 9, 4].

To address this disconnect, we propose an asymmetric weighted chain-of-thought methodology grounded in latent space reasoning (illustrated in Fig. 2). Unlike traditional end-to-end methods, our framework exploits the autoregressive properties of advanced embedding spaces to construct an implicit reasoning trajectory [6, 16, 19, 20], forcing the model to sequentially infer key clinical BI-RADS descriptors before converging on a final diagnosis. To combat the extreme scarcity of critical malignant features, we couple this architecture with a novel dual-asymmetric loss (Dual-ASL) function. This strategy dynamically adjusts learning priorities, safeguarding high-specificity malignant descriptors from being overshadowed by common benign patterns to ensure transparent and robust diagnostic performance.



**Fig. 2.** Overview of the proposed Latent-CURE architecture. The framework consists of multimodal encoding, an implicit feature-aware CoT trajectory, and a Dual-ASL strategy.

**Contributions.** The primary contributions are three-fold: (1) **Knowledge-Injected Latent Reasoning**: an implicit Chain-of-Thought (CoT) approach embedding expert descriptors into the latent process; (2) **Explainability via Professional Descriptors**: a transparent alternative to end-to-end models with evidence-supported diagnosis; (3) **Dual-Asymmetric Optimization**: a novel loss strategy using dynamic margins to protect rare, high-specificity malignant features.

## 2 Methodology

### 2.1 Problem Formulation

In the context of MLMs, the conventional automated diagnostic process is typically formulated as an autoregressive generation task. Given a BUS image  $\mathbf{X}_v \in \mathbb{R}^{H \times W \times C}$  and a textual prompt  $\mathbf{X}_p$ , traditional MLMs generate a sequence of textual tokens  $\mathbf{Y} = (y_1, y_2, \dots, y_N)$  by maximizing the conditional probability  $P(\mathbf{Y}|\mathbf{X}_v, \mathbf{X}_p) = \prod_{i=1}^N p(y_i|\mathbf{X}_v, \mathbf{X}_p, y_{<i})$ . Although versatile, this unconstrained next-token prediction paradigm inevitably introduces semantic hallucinations [28, 30]. This vulnerability renders the approach unsafe for rigorous medical decision-making, a domain that strictly requires deterministic outputs.

To eliminate hallucinations at the architectural level, we reformulate the multimodal diagnostic process from an open-ended generation task to a deterministic metric learning problem within a normalized latent space. Let  $f_\theta(\cdot)$  denote the unified multimodal encoder parameterized by  $\theta$ , which is capable of projecting

multimodal contexts and unimodal text into a shared  $d$ -dimensional hypersphere. For a specific diagnostic sub-task  $t$ , we define a set of mutually exclusive clinical semantic anchors  $\mathcal{A}_t = \{a_{pos}^{(t)}, a_{neg}^{(t)}\}$ . The objective is to retrieve the target anchor that maximizes the cosine similarity with the given multimodal context:

$$\hat{a}^{(t)} = \arg \max_{a \in \mathcal{A}_t} \cos(f_\theta(\mathbf{X}_v, \mathbf{X}_p), f_\theta(a)) \quad (1)$$

By replacing the probabilistic decoding head with a deterministic retrieval-based objective, the output space is strictly bounded by predefined expert knowledge, which theoretically guarantees a zero-hallucination diagnostic trajectory.

## 2.2 Implicit Feature-aware Chain-of-Thought Modeling

Although the metric learning formulation presented in Sec. 2.1 guarantees deterministic outputs, a direct mapping from the image  $\mathbf{X}_v$  to the final diagnosis bypasses the intermediate morphological reasoning required by the BI-RADS standard. To bridge this gap, we propose an implicit feature-aware CoT module, which forces the latent representations to align sequentially with predefined clinical descriptors prior to the synthesis of the final decision.

We model the diagnostic process as a discrete Markovian decision sequence of length  $T$ . Let  $\mathcal{T} = \{1, 2, \dots, T\}$  denote the set of ordered reasoning steps. Based on established radiological cognitive pathways, we set  $T = 4$ . This sequence reflects a diagnostic progression from macro-structures to micro-signals, culminating in a holistic assessment: Margin ( $t = 1$ )  $\rightarrow$  Calcification ( $t = 2$ )  $\rightarrow$  Orientation ( $t = 3$ )  $\rightarrow$  Final Diagnosis ( $t = 4$ ).

At each step  $t \in \mathcal{T}$ , the model resolves a specific clinical query  $\mathbf{q}_t$ . To preserve the autoregressive nature of the reasoning process without requiring explicit text generation, the input at step  $t$  is conditioned on the visual context  $\mathbf{X}_v$  and the accumulated semantic history  $\mathbf{H}_{t-1}$  from all preceding steps. This history is constructed by concatenating deterministically selected anchors:

$$\mathbf{H}_{t-1} = \bigoplus_{i=1}^{t-1} \hat{a}^{(i)}, \quad \text{with } \mathbf{H}_0 = \emptyset \quad (2)$$

where  $\oplus$  denotes textual sequence concatenation.

The context-aware latent representation  $\mathbf{z}_t \in \mathbb{R}^d$  for the  $t$ -th reasoning step is extracted by the multimodal representation encoder  $f_\theta$  as follows:

$$\mathbf{z}_t = f_\theta(\mathbf{X}_v \oplus \mathbf{q}_t \oplus \mathbf{H}_{t-1}) \quad (3)$$

The optimal diagnostic anchor at step  $t$  is determined through nearest neighbor retrieval in the latent space:

$$\hat{a}^{(t)} = \arg \max_{a \in \mathcal{A}_t} \cos(\mathbf{z}_t, f_\theta(a)) \quad (4)$$

This sequential formulation ensures that the embedding vector  $\mathbf{z}_4$  utilized for the final classification of malignancy is intrinsically conditioned on explicit

anatomical findings. Embedding this causal chain directly into the latent space achieves a highly interpretable diagnostic trajectory wherein the ultimate prediction remains deterministically anchored to explicit, intermediate BI-RADS evidence.

### 2.3 Dual-Asymmetric Optimization for Long-tailed Clinical Features

Although the implicit CoT reasoning integrates clinical pathways, optimizing this trajectory under real-world epidemiological distributions presents fundamental challenges. Medical datasets are notoriously long-tailed; benign lesions consistently outnumber malignant ones, and specific malignant morphological indicators, such as non-parallel orientation or microcalcifications, remain statistically scarce. Under standard symmetric optimization objectives, the gradient landscape is overwhelmingly dominated by the abundant benign visual priors [14]. This imbalance invariably causes gradient starvation for rare features, which collapses the latent space toward majority classes and exacerbates the shortcut learning phenomenon. To mitigate this majority collapse, we propose a Dual-ASL tailored for the metric-based reasoning steps. This strategy injects asymmetry into two orthogonal dimensions: margin adaptation for boundary strictness and weight penalty for gradient magnitude, ensuring that scarce but critical malignant anchors dictate the optimization trajectory. For a given reasoning step  $t \in \mathcal{T}$  with the extracted latent state  $\mathbf{z}_t$ , let  $S_{pos}^{(t)} = \cos(\mathbf{z}_t, f_\theta(a_{pos}^{(t)}))$  and  $S_{neg}^{(t)} = \cos(\mathbf{z}_t, f_\theta(a_{neg}^{(t)}))$  denote the cosine similarities between the visual embedding and the corresponding textual anchors. To prevent majority collapse, the penalty is conditioned not solely on the step  $t$ , but strictly on the specific ground-truth clinical feature  $c_t$ . We formulate the step-wise Dual-ASL objective as a dynamically weighted cosine triplet margin loss:

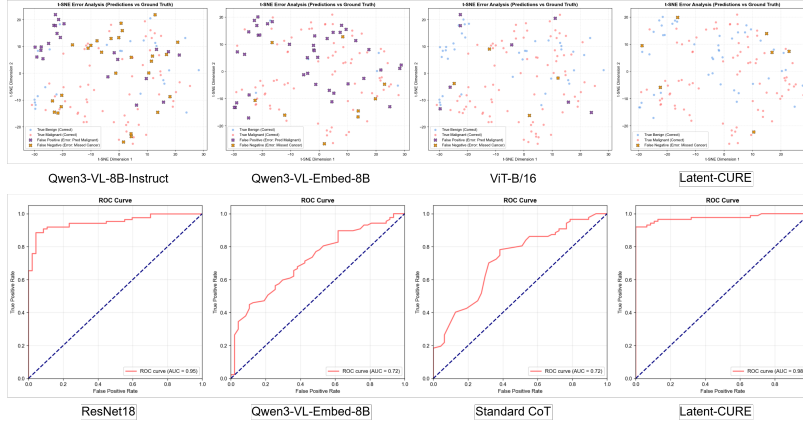
$$\mathcal{L}_{\text{Dual-ASL}}^{(t)} = \mathbf{w}(c_t) \cdot \max\left(0, S_{neg}^{(t)} - S_{pos}^{(t)} + \mathbf{m}(c_t)\right) \quad (5)$$

where  $\mathbf{w}(c_t)$  acts as the feature-specific dynamic sample weight and  $\mathbf{m}(c_t)$  defines the adaptive angular margin.

The values of  $\mathbf{w}(c_t)$  and  $\mathbf{m}(c_t)$  are dynamically conditioned on the epidemiological prior and the diagnostic specificity of the target anchor  $a_{pos}^{(t)}$ . The margin  $\mathbf{m}(c_t)$  forces a stricter latent separation boundary for malignant-associated features, compelling the encoder  $f_\theta$  to learn a discriminative manifold that resists benign feature entanglement. Simultaneously, the weight  $\mathbf{w}(c_t)$  aggressively up-weights structurally rare anchors. By amplifying the gradient signals of these high-specificity, low-frequency indicators, the proposed formulation effectively shields them from being overshadowed.

The global optimization objective is the summation of the asymmetric losses across all implicit reasoning steps:

$$\mathcal{L}_{total} = \sum_{t=1}^T \mathcal{L}_{\text{Dual-ASL}}^{(t)} \quad (6)$$



**Fig. 3.** Qualitative and quantitative analysis of diagnostic models.

This architecture ensures that the model cannot achieve a low global loss by merely defaulting to the prediction of benign outcomes; it must actively align with the rare pathological clues.

### 3 Experiments

#### 3.1 Experimental Setup

We utilized an IRB-approved multicenter dataset of 666 breast ultrasound cases, partitioned 8:2 for training and testing, with all diagnostic labels strictly derived from the surgical histopathology gold standard. To eliminate irrelevant background interference and prevent shortcut learning, physicians manually cropped the images to isolate regions of interest. Following consultations with multidisciplinary physicians, we distilled BI-RADS descriptors into three core dimensions: Margin, Calcification, and Orientation. To combat epidemiological imbalance and prevent gradient starvation on rare malignant features, the model was optimized via our Dual-ASL framework. Comprehensive implementation details are summarized in Table 1.

#### 3.2 Quantitative Baselines and Efficiency Analysis

As summarized in Table 2 and visualized in Fig. 3 and Fig. 4, Latent-CURE establishes a new state of the art, achieving 93.78% accuracy and an 87.07% Matthews correlation coefficient (MCC), systematically outperforming pure vision baselines such as ViT-B/16 and ResNet18. The evaluation simultaneously exposes severe limitations across general foundation models. Both large-scale proprietary frameworks and open adapted variants demonstrate a high susceptibility to shortcut learning and autoregressive hallucinations, overwhelmingly

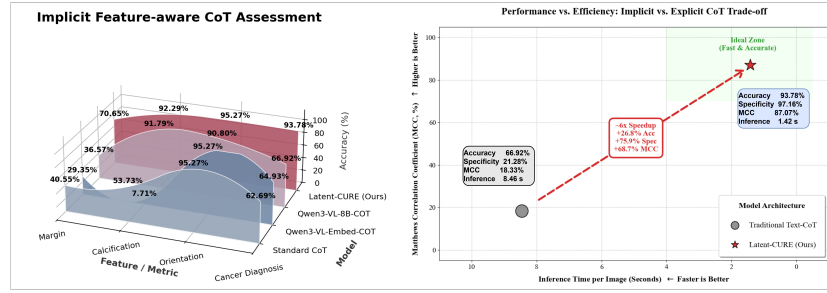


Fig. 4. Implicit CoT Assessment: Accuracy Breakdown and Efficiency Trade-off

Table 1. Comprehensive Implementation Details: Global Hyperparameters and Feature-Specific Dual-ASL Configurations.

| Part I: Global Training Hyperparameters |                       |              |                    |
|-----------------------------------------|-----------------------|--------------|--------------------|
| Parameter                               | Value                 | Parameter    | Value              |
| Hardware                                | GPU (80GB)            | Base LR      | $1 \times 10^{-5}$ |
| Base Model                              | Qwen3-VL-Embed-8B     | LR Scheduler | Cosine             |
| LoRA Config                             | $r = 16, \alpha = 32$ | Seeds        | 42, 666, 2026      |

| Part II: Dual-ASL Configuration Matrix |                               |            |            |
|----------------------------------------|-------------------------------|------------|------------|
| Reasoning Step ( $t$ )                 | Ground-Truth ( $c_t$ )        | Weight $w$ | Margin $m$ |
| $t = 1$ : Margin                       | Circumscribed (Benign)        | 2.5        | 0.3        |
|                                        | Not Circumscribed (Malignant) | 1.0        | 0.2        |
| $t = 2$ : calcification                | No suspicious (Majority)      | 1.0        | 0.2        |
|                                        | Suspicious (Malignant)        | 2.0        | 0.3        |
| $t = 3$ : orientation                  | Parallel (Majority)           | 1.0        | 0.2        |
|                                        | Non-parallel (Rare)           | 4.0        | 0.4        |
| $t = 4$ : Final                        | Benign (Majority)             | 1.0        | 0.2        |
|                                        | Malignant (Target)            | 6.0        | 0.5        |

defaulting to the prediction of benign outcomes and causing significant specificity degradation that ranges from 21.28% to 77.04%. Latent-CURE circumvents these generative bottlenecks by executing the reasoning trajectory within a protected latent space guided by Dual-ASL. Explicitly aligning this trajectory with clinical morphological descriptors effectively mitigates false positive susceptibility, which restores diagnostic precision to 97.16% specificity while achieving an approximately sixfold inference acceleration over explicit text-based adaptation strategies.

### 3.3 Ablation Studies

To evaluate individual module contributions, we conducted a systematic ablation study (Table 3). While the Direct SFT baseline provides a strong performance

**Table 2.** Quantitative diagnostic performance comparison across different model architectures. Metrics are reported with standard deviations.

| Model Architecture        | Accuracy (%)                       | Precision (%)                      | Specificity (%)                    | MCC (%)                            |
|---------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| ResNet18 [8]              | 89.80 $\pm$ 2.70                   | 97.01 $\pm$ 0.60                   | 95.04 $\pm$ 1.10                   | 79.50 $\pm$ 4.64                   |
| ViT-B/16 [5]              | 92.29 $\pm$ 1.02                   | 94.69 $\pm$ 2.88                   | 90.07 $\pm$ 5.81                   | 83.27 $\pm$ 2.61                   |
| Gemini-2.5-Pro [21]       | 73.13 $\pm$ 6.46                   | 73.88 $\pm$ 3.45                   | 40.43 $\pm$ 11.26                  | 38.52 $\pm$ 17.57                  |
| Qwen3-VL-Plus [2]         | 77.36 $\pm$ 1.55                   | 76.40 $\pm$ 0.63                   | 46.10 $\pm$ 2.46                   | 48.47 $\pm$ 4.57                   |
| GPT-5 [17]                | 78.11 $\pm$ 1.88                   | 75.88 $\pm$ 2.23                   | 42.55 $\pm$ 7.67                   | 51.16 $\pm$ 4.29                   |
| Baichuan-M2-Plus [22]     | 88.12 $\pm$ 10.51                  | 89.44 $\pm$ 11.36                  | 77.04 $\pm$ 23.34                  | 73.33 $\pm$ 23.27                  |
| Qwen3-VL-Embed-8B [27]    | 58.71 $\pm$ 4.56                   | 63.82 $\pm$ 2.42                   | 12.06 $\pm$ 5.35                   | -5.13 $\pm$ 11.98                  |
| Qwen3-VL-Embed-COT        | 64.93 $\pm$ 0.00                   | 64.93 $\pm$ 0.00                   | 0.00 $\pm$ 0.00                    | 0.00 $\pm$ 0.00                    |
| Qwen3-VL-8B-Instruct      | 64.43 $\pm$ 0.77                   | 71.91 $\pm$ 0.98                   | 46.10 $\pm$ 5.49                   | 20.83 $\pm$ 0.67                   |
| Qwen3-VL-8B-lora [10]     | 63.43 $\pm$ 2.00                   | 73.22 $\pm$ 2.13                   | 53.19 $\pm$ 5.03                   | 21.70 $\pm$ 4.80                   |
| Qwen3-VL-8B-COT           | 66.92 $\pm$ 0.77                   | 68.28 $\pm$ 0.28                   | 21.28 $\pm$ 0.00                   | 18.33 $\pm$ 2.20                   |
| <b>Latent-CURE (Ours)</b> | <b>93.78 <math>\pm</math> 0.77</b> | <b>98.37 <math>\pm</math> 1.26</b> | <b>97.16 <math>\pm</math> 2.20</b> | <b>87.07 <math>\pm</math> 1.85</b> |

**Table 3.** Ablation studies on the core components of Latent-CURE. ✓ indicates the inclusion of a module. Metrics are reported as percentages (%).

| Model Variant      | Implicit CoT | Dual-ASL | Accuracy                           | Specificity                        | MCC                                |
|--------------------|--------------|----------|------------------------------------|------------------------------------|------------------------------------|
| Direct SFT         | ✗            | ✗        | 91.79 $\pm$ 0.75                   | 96.45 $\pm$ 1.23                   | 83.26 $\pm$ 1.40                   |
| Standard CoT       | ✓            | ✗        | 62.69 $\pm$ 1.77                   | 4.26 $\pm$ 3.30                    | -3.95 $\pm$ 6.12                   |
| <b>Latent-CURE</b> | <b>✓</b>     | <b>✓</b> | <b>93.78 <math>\pm</math> 0.77</b> | <b>97.16 <math>\pm</math> 2.20</b> | <b>87.07 <math>\pm</math> 1.85</b> |

foundation (91.79% accuracy, 96.45% specificity), it remains an opaque classifier. Transitioning to a Standard CoT variant without targeted optimization triggers a catastrophic diagnostic collapse, where accuracy drops to 62.69% and specificity plummets to 4.26% with a negative -3.95% MCC. This reveals the inherent instability of unconstrained reasoning in large multimodal models, where generating morphological anchors without gradient protection exacerbates majority class bias and shortcut learning. Integrating both Implicit CoT and Dual-ASL, the Latent-CURE framework effectively mitigates this collapse and surpasses the baseline across all metrics. By shielding scarce malignant anchors from gradient starvation, Dual-ASL elevates accuracy to 93.78% and specificity to 97.16%, with MCC improving from 83.26% to 87.07%. These gains validate that explicit gradient protection is mandatory to transform clinical reasoning into a deterministic, highly discriminative diagnostic trajectory.

## 4 Discussion and Conclusion

We present Latent-CURE, a framework bridging multimodal models and clinical logic by shifting breast ultrasound diagnosis from open-ended generation to an implicit reasoning trajectory within a constrained latent space. Rather than relying on free-form CoT decoding, Latent-CURE performs determinis-

tic retrieval over fixed BI-RADS descriptor anchors, thereby providing bounded intermediate evidence before the final diagnosis. While our dual-asymmetric optimization helps protect rare malignant indicators from dominant benign priors, the current reasoning trajectory follows a fixed descriptor order from margin, calcification, and orientation to the final diagnosis, and is not assumed to be permutation-invariant. Future work should further study how alternative descriptor sequences and feature-specific contributions affect latent representations and diagnostic outcomes. Our conclusions are limited to the submitted breast ultrasound setting with histopathology-derived labels. Overall, Latent-CURE demonstrates that standardized latent-anchor reasoning combined with asymmetric optimization can improve both interpretability and robustness in imbalanced clinical diagnosis.

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

## References

1. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Consortium, P.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making* **20**(1), 310 (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Brem, R.F., Lenihan, M.J., Lieberman, J., Torrente, J.: Screening breast ultrasound: past, present, and future. *American Journal of Roentgenology* **204**(2), 234–240 (2015)
4. Choi, J., Kim, J.W., Lee, Y.S., Tae, J.H., Choi, S.Y., Chang, I.H., Kim, J.H.: Availability of chatgpt to provide medical information for patients with kidney cancer. *Scientific Reports* **14**(1), 1542 (2024)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)
7. Hao, Y., Qiu, Z., Holmes, J., Löckenhoff, C.E., Liu, W., Ghassemi, M., Kalantari, S.: Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *npj Digital Medicine* **8**(1), 450 (2025)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
9. Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T.T., McGee, L.A., Ashman, J.B., Li, X., Liu, T., Shen, J., et al.: Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology* **13**, 1219326 (2023)

10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *Iclr* **1**(2), 3 (2022)
11. Kolla, L., Parikh, R.B.: Uses and limitations of artificial intelligence for oncology. *Cancer* **130**(12), 2101–2107 (2024)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**, 28541–28564 (2023)
13. 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)
14. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 82–91 (2021)
15. Saleem, A., Lewis, J.R., Gilani, S.Z.: A hybrid contrastive ordinal regression method for advancing disease severity assessment in imbalanced medical datasets. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 14–23. Springer (2025)
16. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074* (2025)
17. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025)
18. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (2023)
19. Su, D., Zhu, H., Xu, Y., Jiao, J., Tian, Y., Zheng, Q.: Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275* (2025)
20. Tan, W., Li, J., Ju, J., Luo, Z., Luan, J., Song, R.: Think silently, think fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint arXiv:2505.16552* (2025)
21. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024)
22. Team, M., Dou, C., Liu, C., Yang, F., Li, F., Jia, J., Chen, M., Ju, Q., Wang, S., Dang, S., Li, T., Zeng, X., Zhou, Y., Zhu, C., Pan, D., Deng, F., Ai, G., Dong, G., Zhang, H., Tai, J., Hong, J., Lu, K., Sun, L., Guo, P., Ma, Q., Xin, R., Yang, S., Zhang, S., Mo, Y., Liang, Z., Zhang, Z., Cui, H., Zhu, Z., Wang, X.: Baichuan-m2: Scaling medical capability with large verifier system (2025), <https://arxiv.org/abs/2509.02208>
23. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature medicine* **29**(8), 1930–1940 (2023)
24. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
25. Xian, M., Zhang, Y., Cheng, H.D., Xu, F., Zhang, B., Ding, J.: Automatic breast ultrasound image segmentation: A survey. *Pattern Recognition* **79**, 340–355 (2018)

26. Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., Carson, P.L.: Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* **91**, 1–9 (2019)
27. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (2025)
28. Zhang, Z., He, X., Yan, W., Shen, A., Zhao, C., Wang, S., Shen, Y., Wang, X.E.: Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778* (2025)
29. Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., Chen, J.: A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys* **57**(8), 1–35 (2025)
30. Zhu, R.J., Peng, T., Cheng, T., Qu, X., Huang, J., Zhu, D., Wang, H., Xue, K., Zhang, X., Shan, Y., Cai, T., Kergan, T., Kembay, A., Smith, A., Lin, C., Nguyen, B., Pan, Y., Chou, Y., Cai, Z., Wu, Z., Zhao, Y., Liu, T., Yang, J., Zhou, W., Zheng, C., Li, C., Zhou, Y., Li, Z., Zhang, Z., Liu, J., Zhang, G., Huang, W., Eshraghian, J.: A survey on latent reasoning (2025), <https://arxiv.org/abs/2507.06203>