Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncovering neural decisions: How textual interpretation bridges AI and clinical reasoning

Hui Lu

huilu@sjtu.edu.cn

Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University https://orcid.org/0000-0001-8347-0830

Article

Keywords: Explainable artificial intelligence, Medical imaging, Clinical concept alignment, Neuron interpretation, Breast ultrasound, BI-RADS

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10368446/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Uncovering neural decisions: How textual interpretation bridges AI and clinical reasoning

Siqiong Yao[1,2†], Yueting Wu[1†], Peng Sun[1], Wentao Tao[1], Yun Wang[3], Wenqiang Xu[4*], Caifeng Wan[5*], Hui Lu[1,2,6*]

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China.
[2]SJTU-Yale Joint Center for Biostatistics and Data Science, Technical Center for Digital Medicine, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, 200240, China.
[3]Hepatobiliary Pancreatic Center, Xuzhou Central Hospital, Xuzhou, Jiangsu Province, 221009, China.
[4]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.
[5]Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China.
[6]Institute of Bioinformatics, Shanghai Academy of Experimental Medicine, Shanghai, 200240, China.

*Corresponding author(s). E-mail(s): vinjohn@sjtu.edu.cn; China.wancaifeng@renji.com; huilu@sjtu.edu.cn;
Contributing authors: yaosiqiong@sjtu.edu.cn; wyt522@sjtu.edu.cn;
†These authors contributed equally to this work.

Abstract

Medical imaging artificial intelligence (AI) can achieve strong diagnostic performance, yet its decision evidence often remains difficult to verify, raising concerns about reliability and safety in clinical use. Conventional heatmaps mainly localize model-attended lesion regions and have limited capacity to distinguish multiple overlapping clinical features, whereas image–text explanations may describe the image content without being coupled to the actual decision pathway of the target model. Here, we present CRysTal, a neuron-level post-hoc explanation framework that maps decision-related neuronal activations to standardized clinical

terminology through two-stage non-negative sparse regression and cross-category fusion. In breast ultrasound diagnosis, CRysTal linked neurons from EfficientNet-B0, ResNet-50, VGG-19, and ViT-B models to BI-RADS concepts, generating explanations that were more consistent with radiology reports than large-model reasoning, CLIP-Dissect, and linear explanation baselines while preserving the original classification performance. Across the four diagnostic models, CRysTal improved feature-level agreement by 3.4–12.2% and tendency-level agreement by 3.5–19.2% over the second-best baseline. CRysTal further tracked a shift in model explanations during training, from non-specific visual cues toward clinical concepts commonly used by physicians, with concept alignment positively associated with diagnostic performance in ResNet-50 and EfficientNet-B0 (Pearson r = 0.89 and 0.84, respectively). Its reasoning conflict (RCF) score identified erroneous AI predictions with an AUC of 0.745, outperforming saliency-map-based risk signals (AUC range, 0.542–0.625) and a concept bottleneck model (AUC = 0.643). In a reader study involving 20 ultrasound physicians, CRysTal explanations received mean scores close to 97 for explanation accuracy, readability and guideline concordance, and reduced mean reading time by 10.7% for junior physicians and 4.7% for senior experts. Generalization experiments on external breast ultrasound, thyroid ultrasound and dermoscopy datasets provide initial evidence that CRysTal can support auditable AI explanations across standardized medical imaging terminology systems.

Keywords: Explainable artificial intelligence, Medical imaging, Clinical concept alignment, Neuron interpretation, Breast ultrasound, BI-RADS

1 Introduction

Artificial intelligence (AI) has improved the efficiency and accuracy of medical image analysis, but its clinical translation remains limited, with safety concerns being one of the major barriers [1–11]. Small prediction errors may lead to missed malignant lesions or diagnostic misclassification, affect downstream clinical decisions, compromise patient safety and raise medico-legal and public-trust concerns [12]. A study in *JAMA* reported that clinicians' diagnostic accuracy decreased by 11.3% when they referred to biased AI diagnostic results [13]. These findings indicate that AI models with systematic bias may shift from clinical decision-support tools to potential sources of clinical risk [14, 15]. Medical imaging AI systems therefore need to provide not only diagnostic outputs but also decision rationales that clinicians can understand. This requirement is consistent with evidence-based medicine, which emphasizes traceable evidence and reproducible reasoning, and is particularly important when human interpretation and AI predictions disagree [16–21].

Early work on explainable AI sought to make model decision processes more transparent through visual attention or attribution [22, 23]. Classical attribution methods, including Grad-CAM, SHAP, LRP, CRP and class activation mapping, link model outputs, neuronal activation responses and image pixels to visualize decision-relevant regions [24–28]. However, these methods have important limitations in medical imaging. A local region in a medical image often contains multiple lesion attributes,

including texture, morphology, size and location. Highlighting a region can only provide a coarse indication of where the model attends, without specifying the corresponding clinical or pathological meaning. As a result, such explanations often lack clinically actionable semantics (Fig. 1a) [18–21, 29]. Subsequent image–text explanation methods, including approaches based on UniMed-CLIP and biomedical vision–language models, can associate image features with multiple clinical terms [30–33]. Yet these methods largely describe the image rather than the internal decision process of the target diagnostic model (Fig. 1a). More recently, CLIP-Dissect introduced an automated framework for describing hidden neurons using natural-language concepts, enabling associations among decision-relevant neurons, activation regions and text concepts [34]. Directly transferring this strategy to medical image analysis remains challenging. In our experiments, CLIP-Dissect often assigned a single local concept to an individual neuron in medical images, failing to cover the composite clinical semantics enriched in the same neuron, such as morphology, margin, echo pattern, calcification and posterior acoustic features. This limited both coverage and diagnostic-tendency consistency. These observations motivate algorithmic designs that are adapted to the structured and multi-attribute nature of medical imaging.

Here we propose Clinical Reasoning Analysis by Text-Neuron Feature Elucidation (CRysTal), a neuron-level explanation framework that maps decision-related neuronal responses to structured clinical concepts (Fig. 1b). CRysTal does not alter the parameters or prediction pathway of the original diagnostic model. Instead, it learns a mapping from decision-related neuronal responses in a fixed model to a structured clinical concept system, thereby generating case-level, auditable clinical evidence chains. In a breast ultrasound benign–malignant classification task, we constructed neuron–concept mappings based on the Breast Imaging Reporting and Data System (BI-RADS) lexicon [35, 36], and evaluated CRysTal across EfficientNet-B0, ResNet-50, VGG-19 and ViT-B diagnostic models [37–40]. We then assessed CRysTal from five perspectives: case-level explanation consistency, single-neuron semantic representation, training dynamics of clinical concept alignment, human–AI collaborative diagnosis and cross-dataset generalization. In ResNet-50 and EfficientNet-B0, model explanations became increasingly aligned with standardized imaging signs used by clinicians, such as morphology, margin, calcification and echo pattern, as training progressed and diagnostic performance improved. Overall, CRysTal recovered structured explanations of decision-related neurons that were highly consistent with clinical reports while preserving the diagnostic performance of the original model. Reader studies further suggest that CRysTal may help clinicians identify erroneous AI predictions and improve diagnostic efficiency.

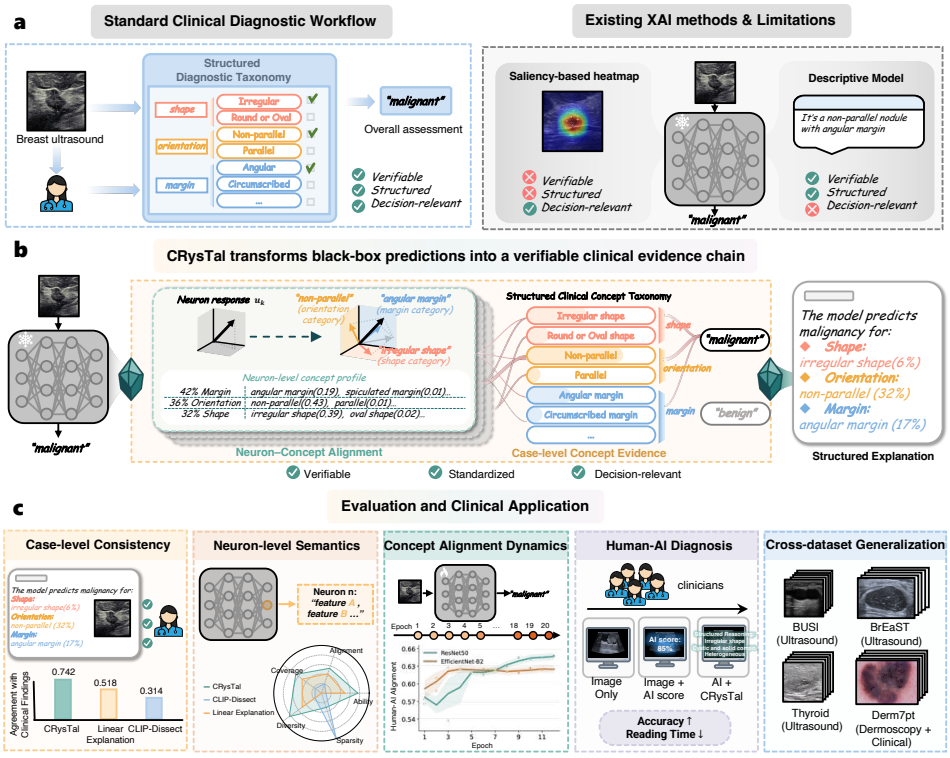


Fig. 1 Clinical evidence requirements for medical imaging AI explanations and the CRysTal study framework. **a**, Standard clinical diagnosis relies on verifiable, standardized and decision-relevant imaging evidence, such as BI-RADS descriptors for lesion shape, margin and orientation. By contrast, saliency heatmaps mainly localize attended regions, whereas image–text descriptive models may generate explanations that are not coupled to the internal decision pathway of the target diagnostic model, limiting verifiability, standardization and decision relevance. **b**, CRysTal maps neuronal responses from a black-box diagnostic model to a structured clinical concept taxonomy and aggregates them into case-level concept evidence, enabling model predictions to be translated into auditable clinical semantic explanations. **c**, CRysTal was evaluated across five aspects: case-level explanation consistency, single-neuron semantic representation, training dynamics of clinical concept alignment, human–AI collaborative diagnosis and cross-dataset generalization.

2 Results

2.1 Overall CRysTal framework

We developed CRysTal (Fig. 2) to characterize the internal reasoning process of medical imaging diagnostic models. The method uses neuronal feature representations and standardized clinical terminology to provide structured explanations of model decision evidence. In this study, we focused on breast ultrasound models for benign–malignant lesion classification. Based on neuronal responses extracted from a fixed diagnostic model, CRysTal links neuronal representations, visual features and structured clinical

concepts, thereby converting abstract visual representations into textual diagnostic evidence that follows clinical conventions.

Fig. 2a shows the paired inputs used to develop CRysTal: structured BI-RADS clinical concept annotations and neuronal response vectors extracted by feeding the same images into a fixed diagnostic model. Fig. 2b illustrates the training workflow for interpreting individual neurons using clinical text. The workflow consists of two stages. The first stage performs within-category estimation. For each independent BI-RADS category, such as shape, margin, echo pattern and calcification, non-negative sparse regression is used to estimate the association weights between each neuron and specific clinical descriptors within that category. The second stage performs cross-category integration. Association weights learned across all clinical feature categories are jointly fused to form a neuron-specific clinical semantic representation. This design allows each neuron to be associated with multiple clinical descriptors with different weights, rather than being constrained to a single category label, and is intended to better approximate the multidimensional clinical semantics carried by individual neurons.

Fig. 2c shows how CRysTal explains the activation of multiple neurons. After a breast ultrasound image is processed by the diagnostic model, the model first produces a classification output. CRysTal then uses the learned neuron–concept mappings to generate clinically structured evidence for that decision. During training, neuronal activation vectors extracted from the fixed diagnostic model are used as input, and structured BI-RADS concept annotations from the same cases are used as supervision. Through within-category non-negative sparse regression and cross-category concept fusion, CRysTal learns association weights between each neuron and different clinical signs. During inference, it reads the neuronal responses of the original diagnostic model and computes output-specific neuron relevance scores to aggregate the learned neuron–concept profiles into case-level structured explanations. Details of the model decomposition are provided in Section 4.4.

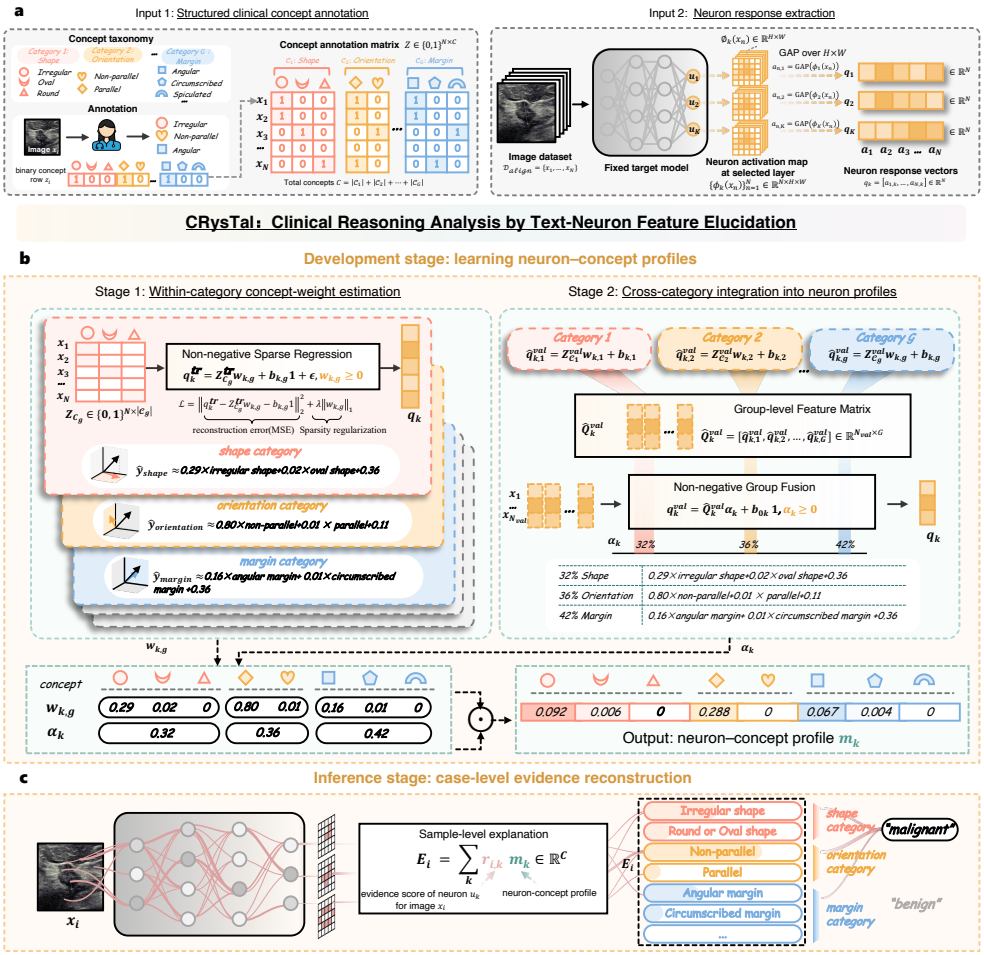


Fig. 2 Overall framework and workflow of CRysTal. **a**, Data preparation. Structured BI-RADS clinical concept annotations are organized into a category-aware concept taxonomy and encoded as a binary concept annotation matrix. In parallel, the same images are passed through a fixed breast ultrasound diagnostic model, and neuronal activation vectors are extracted from a selected layer. **b**, Neuron–concept alignment. For each neuron, non-negative sparse regression first estimates concept weights within each BI-RADS category. Cross-category fusion then integrates category-specific fits into a neuron-specific concept profile, allowing individual neurons to be associated with multiple clinical concepts across diagnostic categories. **c**, Case-level evidence reconstruction. For a test image, output-specific neuron relevance scores are used to aggregate the learned neuron–concept profiles into a structured evidence vector. The evidence vector is summarized within diagnostic categories to generate clinically interpretable decision evidence.

2.2 Evaluation of CRysTal for explaining model decisions

We systematically evaluated the ability of CRysTal to explain diagnostic decisions across four common classification models: EfficientNet-B0, ResNet-50, VGG-19 and

ViT-B. CRysTal was compared with representative explanation baselines, including large-model-based chain-of-thought (CoT), CLIP-Dissect, medically adapted CLIP-Dissect* and Linear Explanation (LE). We also included a Concept Bottleneck Model (CBM) as an explicitly concept-supervised reference model to assess the trade-off between explanation consistency and classification performance under a concept bottleneck constraint. All explanations were compared against the corresponding clinical reports.

Fig. 3a shows that the four breast ultrasound diagnostic models achieved strong classification performance, with AUCs of 0.937, 0.950, 0.981 and 0.976, respectively. Fig. 3b,c further shows that CRysTal explanations were more consistent with clinical diagnostic reports than those generated by the baselines. Across the four classifiers, CRysTal achieved the highest feature-level accuracy, reaching 0.717, 0.680, 0.673 and 0.697, corresponding to improvements of 3.4–12.2% over the second-best method. CRysTal also achieved the highest tendency-level accuracy, reaching 0.727, 0.751, 0.769 and 0.778, with improvements of 3.5–19.2% over the second-best method. These results indicate that CRysTal more accurately recovered lesion features described in clinical reports and more correctly identified their benign or malignant diagnostic tendency. CRysTal also achieved the highest tendency-level F1 and MCC, suggesting that its concept explanations more reliably distinguished clinical evidence supporting benign versus malignant decisions. F1 reflects the overall ability to identify diagnosis-related features, whereas MCC provides a more robust correlation measure under class imbalance. Additional quantitative comparisons across breast ultrasound backbones and external clinical datasets are provided in Extended Data Table 1.

Fig. 3d shows that CBM achieved a benign–malignant classification AUC of 0.786, with feature-level agreement, tendency-level agreement, F1 and MCC of 0.798, 0.762, 0.656 and 0.474, respectively. Because CBM follows a "clinical concept prediction followed by diagnosis" training paradigm, its explanations are directly derived from an explicit concept bottleneck, which explains its advantage in feature-level agreement. However, this explicit concept constraint was accompanied by a marked decrease in classification performance relative to the original black-box models. By contrast, CRysTal did not alter the prediction pathway of the original classifier and therefore preserved its high diagnostic AUC. At the same time, its tendency-level agreement, F1 and MCC were comparable to, or in some cases slightly higher than, those of CBM. These results suggest that although CRysTal does not necessarily exceed an explicit concept bottleneck model in recovering every individual feature, it can recover diagnostically directional clinical evidence without sacrificing classification performance.

Fig. 3e compares the explanations generated by different methods in representative cases. CRysTal produced fine-grained lesion features consistent with clinical reports in both malignant and benign cases, including angular margin, irregular shape, shadowing, microcalcifications, taller-than-wide orientation, hypoechoic echo pattern, circumscribed margin, oval shape, no posterior features, no calcification, wider-than-tall orientation and homogeneous echo pattern. CoT generated report-like descriptions, but some features were inconsistent with the reference clinical report; for example, it described posterior features or calcification in cases where these findings were

absent. CLIP-Dissect, CLIP-Dissect* and LE identified some correct concepts, but they still missed key features, confused concepts or produced evidence inconsistent with the diagnostic tendency, such as omissions of microcalcifications, shadowing or taller-than-wide orientation.

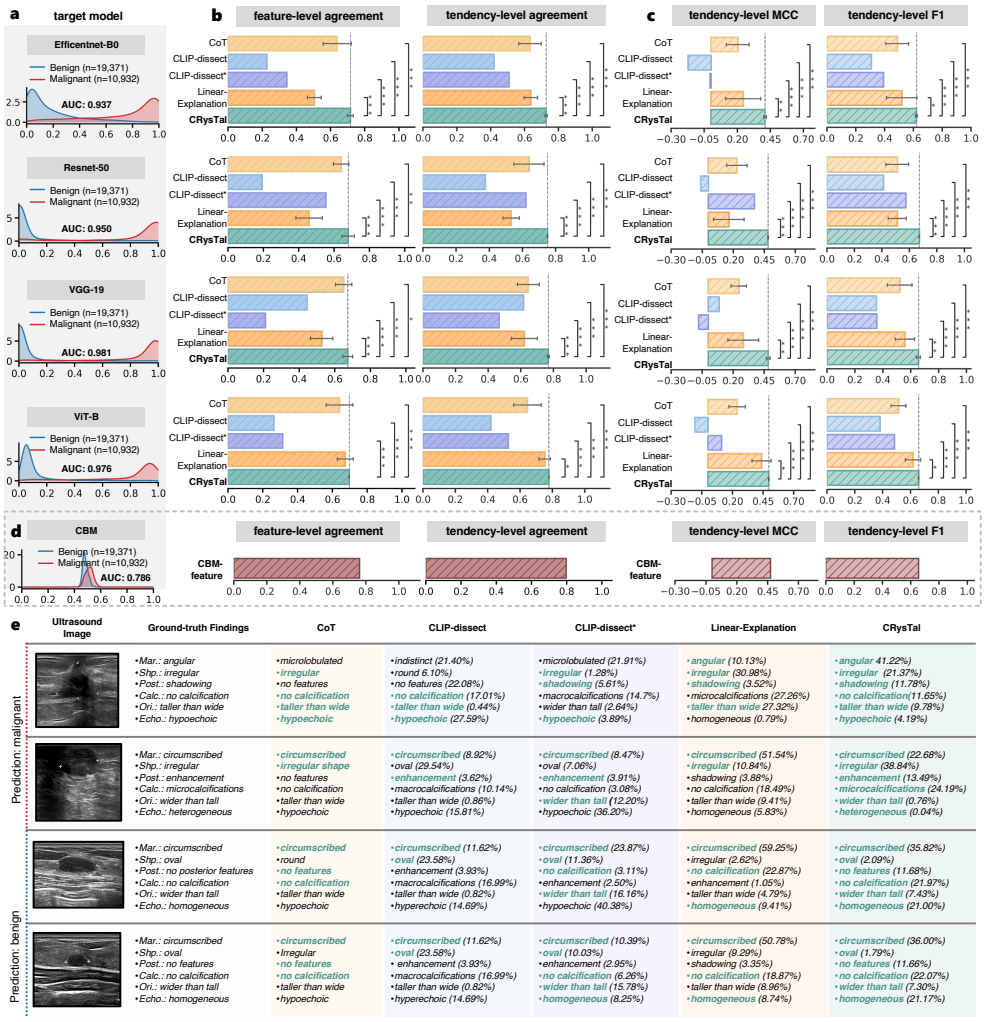


Fig. 3 Performance comparison of different explainability methods for breast ultrasound diagnostic decision explanations. **a**, Area under the receiver operating characteristic curve (AUC) of four common classification models, EfficientNet-B0, ResNet-50, VGG-19 and ViT-B, for benign–malignant breast ultrasound classification. **b,c**, Consistency between diagnostic decision explanations generated by different explainability methods and clinical reports. CRysTal overall achieved higher explanation consistency than existing methods and more accurately captured imaging evidence relevant to clinical diagnosis. **d**, Decision explanation performance of the Concept Bottleneck Model (CBM). CRysTal achieved consistency scores close to those of CBM, suggesting that it can generate decision explanations aligned with clinical diagnostic logic without requiring an explicitly predefined clinical concept bottleneck. **e**, Representative clinical cases. CRysTal generated benign and malignant decision evidence that was correct and covered key clinical assessment features. By contrast, other methods typically provided only local saliency regions or discrete concept cues, making it difficult to form complete, verifiable and report-consistent diagnostic explanations.

2.3 Evaluation of CRysTal for interpreting individual neuron features

We evaluated the ability of CRysTal to provide clinical interpretations of individual neurons across multiple classification models (Fig. 4a–d). The results showed that decision-related neuronal activations associated with imaging features such as calcification, margin, shape, echo pattern, posterior acoustic features and orientation had mean AUCs of 0.637, 0.649, 0.544 and 0.628 in representative neurons from VGG-19, ResNet-50, EfficientNet-B0 and ViT-B, respectively. AUCs for most BI-RADS feature dimensions were above the random level. These findings indicate that single-neuron explanations generated by CRysTal can distinguish benign- and malignant-oriented tendencies embedded in corresponding imaging features and contain clinically meaningful discriminative semantics.

We further evaluated individual-neuron clinical semantic interpretation using five metrics: Ability, Alignment, Coverage, Diversity and Sparsity (Fig. 4e–h and Extended Data Table 2). Ability assesses whether an explanation preserves the core diagnostic discriminative information carried by a neuron. Alignment measures the consistency between explanation concepts and neuronal activation patterns. Coverage evaluates whether the explanation covers key clinical signs associated with the neuron. Diversity measures whether the explanation contains complementary concepts from different BI-RADS categories. Sparsity evaluates whether neuronal semantics can be expressed using a small number of concepts.

Averaged across four backbone networks, CRysTal achieved Ability, Coverage and Diversity scores of 0.611, 0.803 and 0.767, respectively, outperforming LE (0.242, 0.235 and 0.316) and CLIP-Dissect (0.041, 0.137 and 0.137). The mean Sparsity of CRysTal was 0.321, similar to that of LE (0.313), indicating that the construction of multi-category concept combinations did not substantially increase explanation complexity. For Alignment, LE achieved a mean value of 0.393, slightly higher than CRysTal (0.332), suggesting a local advantage in matching candidate concepts to activation patterns. However, when considered together with Coverage and Diversity, this local advantage did not translate into more comprehensive clinical semantic coverage.

Fig. 4e shows that CLIP-Dissect interpreted a representative neuron only as "indistinct margin"; in Fig. 4g, it output only "hypoechoic appearance". However, the neuron-level discriminative analyses in Fig. 4a and Fig. 4c show that these representative neurons had above-random AUCs across multiple BI-RADS feature dimensions. This suggests that their activations were not associated with a single imaging feature, but with multidimensional clinical semantics, including shape, margin, calcification, echo pattern and posterior acoustic features. The single-concept explanations produced by CLIP-Dissect therefore had limited ability to cover the composite feature information carried by these neurons, resulting in lower Coverage and Diversity. Fig. 4g further shows that the high-weight concepts output by LE were semantically similar and concentrated on boundary-related features, limiting coverage of other BI-RADS categories. In the neuron shown in Fig. 4f, posterior acoustic features had the highest discriminative AUC, indicating that the dominant clinical semantics of this neuron were more related to posterior acoustic features. By contrast, the high-weight concepts

Table 1 Reliability of neuron explanations across repeated runs and backbones. Aggregate neuron-level reliability metrics for CLIP-Dissect, Linear Explanation (LE), and CRysTal on ResNet, VGG, EfficientNet, and ViT. Val Corr measures agreement with validation signals, Consistency@K measures top-K agreement across 10 repeated explanations, and Redundancy@K and Conflict@K quantify repetition and contradiction within the recovered concepts. Because CLIP-Dissect is deterministic, repeated-run consistency, redundancy, and conflict statistics are not applicable and are therefore omitted. Best and second-best results within each backbone are highlighted in bold and underline, respectively.

Model	Method	Val Corr↑	Consistency@K↑			Redundancy@K↓		Conflict@K↓	
			K=1	K=3	K=5	K=3	K=5	K=3	K=5
ResNet	CLIP-Dissect	0.189±0.000	–	–	–	–	–	–	–
	LE	0.511±0.131	**0.682±0.298**	0.435±0.151	0.432±0.150	0.071±0.063	0.099±0.082	0.050±0.079	0.079±0.094
	CRysTal	**0.527±0.136**	0.573±0.300	**0.462±0.134**	**0.480±0.135**	**0.032±0.034**	**0.027±0.030**	**0.002±0.017**	**0.002±0.018**
VGG	CLIP-Dissect	0.235±0.000	–	–	–	–	–	–	–
	LE	0.510±0.161	**0.922±0.196**	0.491±0.134	0.488±0.136	0.011±0.028	0.018±0.036	0.024±0.025	0.052±0.025
	CRysTal	**0.524±0.135**	0.771±0.286	**0.496±0.139**	**0.495±0.139**	**0.008±0.020**	**0.008±0.028**	**0.006±0.023**	**0.008±0.028**
EfficientNet	CLIP-Dissect	0.027±0.000	–	–	–	–	–	–	–
	LE	**0.236±0.101**	0.286±0.227	0.240±0.164	0.238±0.162	0.059±0.063	0.066±0.065	0.073±0.104	0.100±0.120
	CRysTal	0.225±0.098	**0.337±0.215**	**0.279±0.157**	**0.279±0.157**	**0.032±0.046**	**0.032±0.045**	**0.030±0.071**	**0.031±0.072**
ViT	CLIP-Dissect	0.227±0.000	–	–	–	–	–	–	–
	LE	**0.592±0.184**	0.723±0.295	0.550±0.203	0.530±0.206	0.022±0.054	0.035±0.058	0.015±0.057	0.057±0.107
	CRysTal	0.534±0.176	**0.803±0.301**	**0.580±0.193**	**0.580±0.191**	**0.012±0.031**	**0.012±0.031**	**0.012±0.049**	**0.012±0.049**

generated by LE were mainly concentrated in shape, calcification and orientation categories, suggesting a mismatch between the explanation and the neuron's dominant clinical representation.

To assess the stability of neuron-level explanations, we performed ten repeated experiments on the same feature layer of each backbone network and quantified the results using four metrics: Val Corr, Consistency@K, Redundancy@K and Conflict@K (Table 1). Val Corr measures the agreement between explanation contents and clinical diagnostic reports. Consistency@K reflects the stability of the top-K clinical features across repeated experiments. Redundancy@K measures the proportion of semantically repeated features among the top-K explanations, whereas Conflict@K measures the proportion of mutually contradictory or clinically inconsistent features among them.

The results showed that CRysTal generated stable neuron-level explanations with low redundancy and low semantic conflict across the four major backbone networks. Specifically, the mean Consistency@3 and Consistency@5 values of CRysTal across the four models were 0.454 and 0.459, respectively, higher than those of LE (0.429 and 0.422). Meanwhile, the mean Redundancy@3 and Redundancy@5 values of CRysTal were 0.021 and 0.020, respectively, lower than those of LE (0.041 and 0.055). The mean Conflict@3 and Conflict@5 values of CRysTal were 0.013 and 0.013, respectively, also lower than those of LE (0.041 and 0.072).

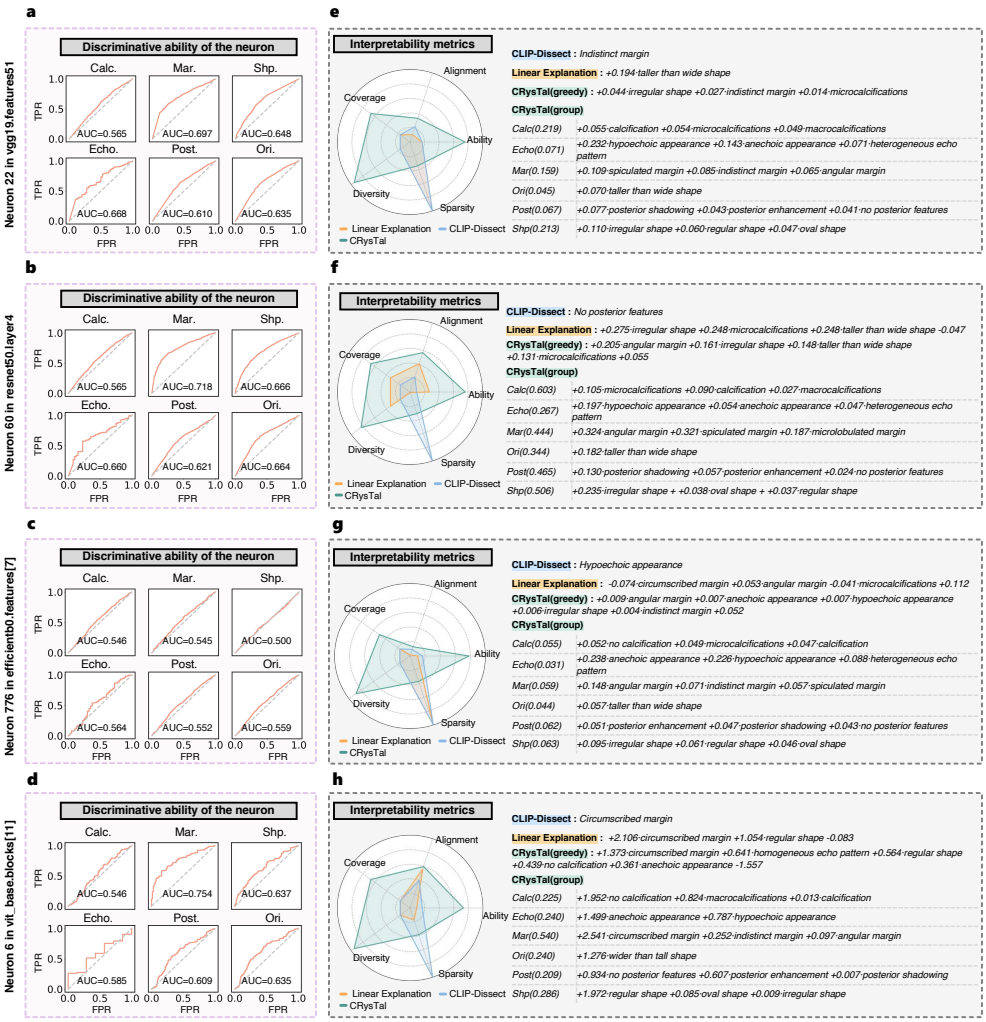


Fig. 4 Individual-neuron interpretation and quality evaluation. **a**–**d**, Evaluation of the clinical semantic representation ability of representative individual neurons across multiple classification models, including VGG-19 (features51, neuron 22), ResNet-50 (layer4, neuron 60), EfficientNet-B0 (features7, neuron 776) and ViT-Base (blocks11, neuron 6). The evaluated BI-RADS-related imaging features included calcification, margin, shape, echo pattern, posterior acoustic features and orientation. **e**–**h**, Quality evaluation across five dimensions: Ability, Alignment, Coverage, Diversity and Sparsity. Compared with CLIP-Dissect and Linear Explanation, CRysTal generated more complete and diagnostically relevant concept combinations, with stronger preservation of diagnostic signals, better coverage of key features and greater semantic diversity across BI-RADS categories, while maintaining comparable activation-pattern matching and explanation sparsity. In the representative neuron examples, CRysTal (group) denotes the full category-aware neuron–concept profile, whereas CRysTal (greedy) denotes the compressed ranked concept readout derived from the same profile for top-K-based comparison with baseline methods.

2.4 Training dynamics of clinically aligned model reasoning

We used clinical text explanations of decision-related neurons to examine whether the decision process of medical imaging diagnostic models became increasingly aligned with expert clinical interpretation during training. Specifically, we selected model checkpoints from different training epochs, used CRysTal to map internal decision-related neurons to BI-RADS clinical concepts and analyzed dynamic changes in model explanations across training. As shown in Fig. 5a, model checkpoints were first saved at different training stages. CRysTal explanations were then generated for each checkpoint to identify the clinical concepts used by the model during breast ultrasound diagnosis. The predicted concepts were matched with expert-annotated BI-RADS concepts to calculate concept alignment. BI-RADS is a standardized evidence system developed by clinical experts based on long-term imaging experience, associations with pathological risk and the need for diagnostic consistency. It structures breast lesion description across dimensions such as shape, margin, orientation, echo pattern, posterior acoustic features and calcification, thereby providing reusable and communicable evidence for benign–malignant risk assessment. Changes in the agreement between model explanations and the BI-RADS concepts can therefore be used to assess whether internal model representations increasingly rely on clinically meaningful imaging evidence.

As shown in Fig. 5b, the mean model–clinical concept alignment of both ResNet-50 and EfficientNet-B0 increased with training epoch and stabilized in the middle-to-late training stages. In early epochs, concept alignment was relatively low; it then increased as training proceeded, suggesting that, while optimizing the classification objective, the models' decision evidence shifted from non-specific visual statistical cues toward clinical semantic concepts described by BI-RADS. Fig. 5c shows a positive association between mean concept alignment and diagnostic performance. In ResNet-50 and EfficientNet-B0, increases in model AUC were accompanied by increases in mean concept alignment, with Pearson correlation coefficients of 0.89 and 0.84, respectively. These results suggest that improved model performance was accompanied by increasing alignment between internal representations and clinically interpretable diagnostic semantics; they should not be interpreted as proof of a causal mechanism.

Fig. 5d shows that concept alignment across different BI-RADS categories followed category-specific trajectories during training. In early training, alignment was generally low for shape, orientation, margin, posterior acoustic features and calcification. As training progressed, alignment increased across several categories, with more pronounced increases for orientation and posterior acoustic features. This indicates that model explanations became more focused on professional medical signs defined by BI-RADS rather than remaining limited to low-level textures or simple grayscale changes. Fig. 5e shows an overall positive association between the alignment of individual BI-RADS categories and model AUC. As the diagnostic performance of ResNet-50 and EfficientNet-B0 improved, concept alignment for most feature categories also increased, suggesting that the models' internal discriminative evidence became more consistent with the standardized clinical evidence system used by human experts.

Together, Fig. 5 indicates, from the perspectives of training dynamics and performance association, that the decision logic of medical imaging diagnostic models did

not emerge as a static property but underwent progressive alignment with BI-RADS clinical semantic evidence. CRysTal captured this dynamic process and provided neuron-level evidence for assessing whether model decision evidence is consistent with clinical interpretation standards.

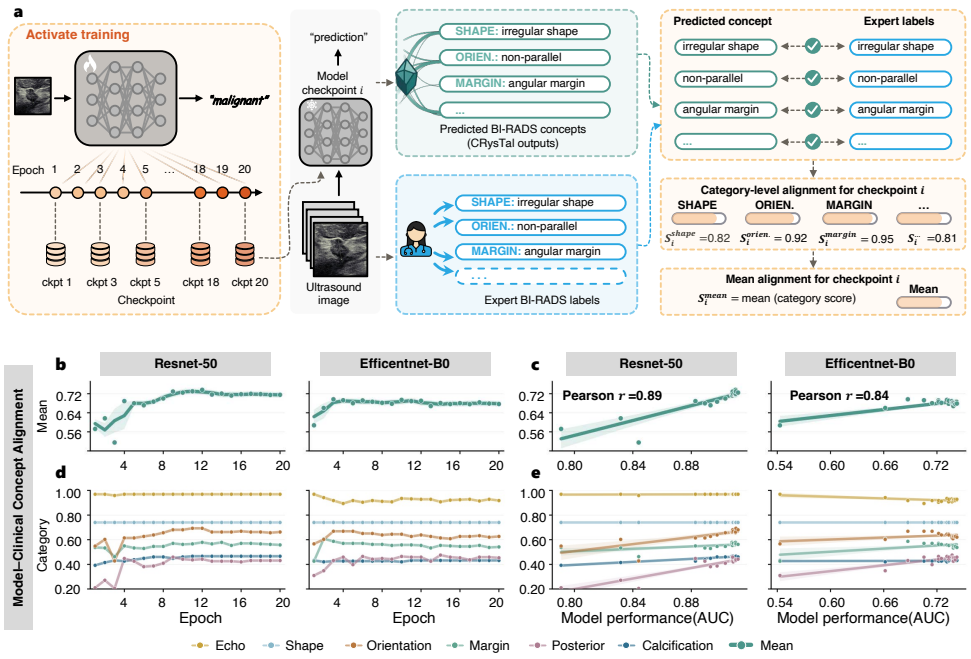


Fig. 5 Semantic evolution of model explanations during training. **a**, Using model checkpoints from different training epochs, CRysTal tracked changes in the alignment between internal diagnostic model representations and the BI-RADS clinical concept system. For each checkpoint, CRysTal first generated case-level predicted concepts and matched them with expert-annotated BI-RADS concepts, then calculated concept consistency across categories including shape, margin, orientation, echo pattern, posterior acoustic features and calcification. **b,c**, Mean-level model–clinical concept consistency analysis. **b**, Mean model–clinical concept consistency across training epochs for ResNet-50 and EfficientNet-B0. **c**, Association between model classification AUC and mean model–clinical concept consistency. Each point corresponds to one checkpoint, and the fitted line indicates the overall trend. **d,e**, Category-level model–clinical concept consistency analysis. **d**, Consistency trajectories across BI-RADS feature categories, including echo, shape, orientation, margin, posterior feature and calcification. **e**, Association between model AUC and category-level concept consistency. The results indicate that, as training progressed and model performance improved, model explanations shifted from non-specific image cues toward standardized clinical signs, suggesting increasing alignment between internal model representations and the BI-RADS clinical evidence system.

2.5 CRysTal decision explanations support clinical diagnosis

Explainability can improve the trustworthiness of individualized AI-assisted diagnosis when it provides information that clinicians can interpret and verify. Conventional

heatmaps mainly show model-attended regions, whereas CRysTal links decision-related neuronal outputs to standardized clinical semantics (Fig. 6a, left). When model predictions are biased or incorrect, clinicians may use the lesion features recovered by CRysTal as supporting evidence to reassess the prediction (Fig. 6a, right). We defined Reasoning Conflict (RCF) to quantify the proportion of recovered clinical imaging evidence whose benign–malignant tendency conflicts with the model's final prediction. CRysTal and CBM used RCF as the prediction-risk assessment signal, whereas conventional visual attribution methods, including Grad-CAM, IG, SHAP, LRP and CRP, used the intersection over union (IoU) between attribution heatmaps and ground-truth lesion regions as the risk-assessment signal. CRysTal-based RCF more clearly separated correctly and incorrectly predicted samples (Fig. 6b). In the task of identifying erroneous model predictions, CRysTal achieved an AUC of 0.745, exceeding visual attribution methods (AUC range, 0.542–0.625) and CBM (AUC = 0.643; Fig. 6c). These results indicate that structured decision explanations may better capture case-level diagnostic evidence than visual region localization alone.

To further evaluate whether CRysTal explanations could be understood and used by clinicians, we conducted two clinician validation experiments using independent image sets. In the first experiment, 20 ultrasound physicians, including 10 senior experts and 10 junior residents, rated CRysTal-generated structured explanations from 100 breast ultrasound images. Explanations were evaluated on a 0–100 scale across three dimensions: explanation accuracy, readability and guideline concordance. The mean scores for all three dimensions were close to 97, with no significant differences between senior experts and junior physicians (explanation accuracy, $P = 0.612$; readability, $P = 0.613$; guideline concordance, $P = 0.645$; Fig. 6d and Extended Data Table 3). These results suggest that CRysTal outputs were readable and consistent with clinical terminology across clinicians with different levels of experience.

The second experiment used a separate set of 80 breast ultrasound images and assessed CRysTal in a three-phase multi-reader, multi-case diagnostic workflow. In Phase 1, physicians interpreted the original ultrasound images alone. In Phase 2, the AI-predicted malignancy probability was provided. In Phase 3, explainability assistance was further provided. The 80 cases were divided into two balanced 40-case subsets: one subset was interpreted with Grad-CAM heatmap assistance, and the other was interpreted with CRysTal-generated structured feature explanations. The two subsets were balanced by diagnosis and AI prediction correctness. Diagnostic accuracy and reading time were recorded at each phase and analyzed after stratification by reader experience level.

Fig. 6g,h shows that CRysTal assistance improved physicians' use of AI predictions. For cases in which AI predictions were correct, junior physicians' accuracy increased from 0.82 during independent reading to 0.91 with black-box AI assistance, and further to 0.93 with CRysTal assistance, exceeding the 0.88 achieved with Grad-CAM assistance. Senior experts achieved an accuracy of 0.98 with CRysTal assistance, also higher than the 0.96 achieved with Grad-CAM. For cases in which AI predictions were incorrect, junior physicians' accuracy decreased from 0.35 to 0.27 with black-box AI assistance, indicating susceptibility to incorrect AI predictions. With CRysTal assistance, accuracy increased to 0.48, substantially higher than the 0.19 observed with

Grad-CAM. Among senior experts, accuracy in AI-error cases was 0.68 with CRysTal and 0.64 with Grad-CAM. These results suggest that CRysTal may help clinicians accept correct AI predictions while also identifying and correcting erroneous predictions, with a greater benefit for junior physicians. Mean reading time decreased by 10.7% for junior physicians and 4.7% for senior experts after CRysTal assistance, suggesting that structured explanations improved decision support without increasing the reading burden.

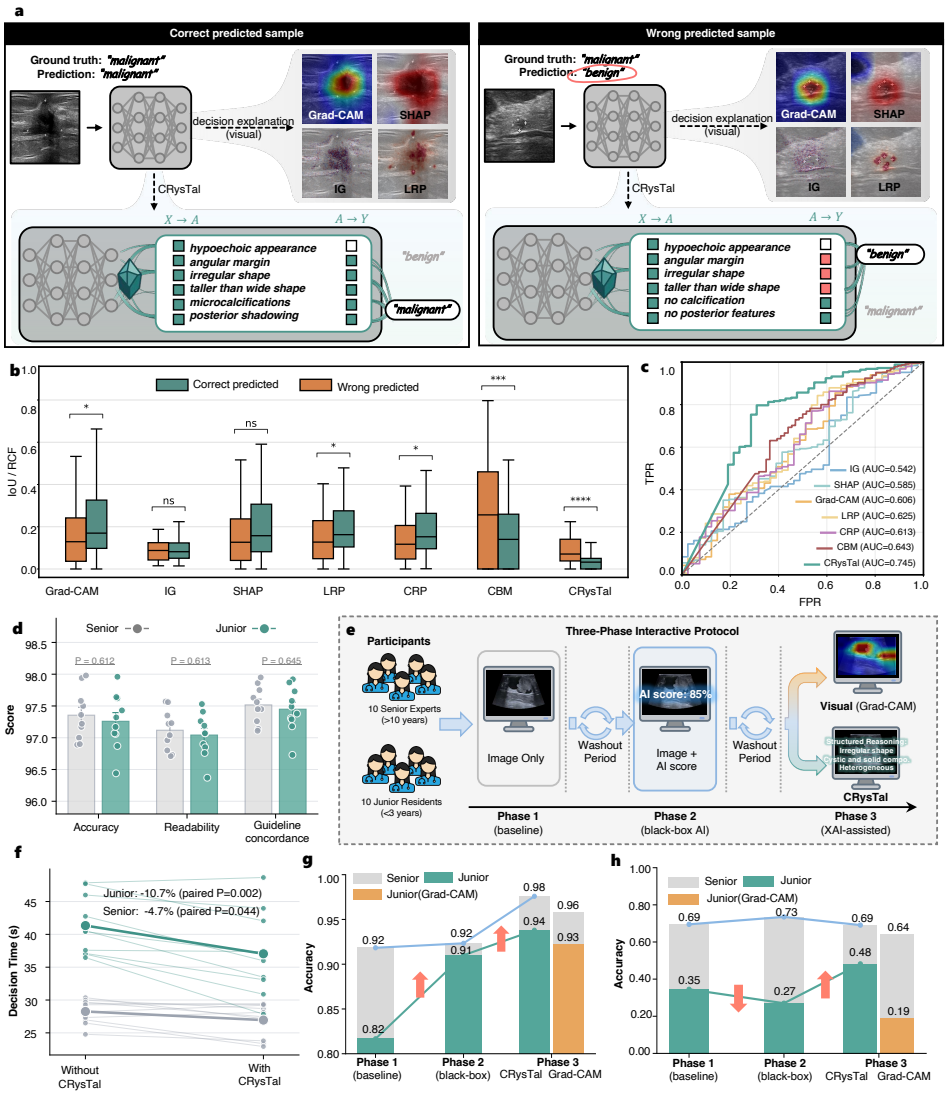


Fig. 6 Evaluation of CRysTal in clinical application. **a**, Examples of case-level reliability analysis. The left case is a correctly predicted malignant case, and the right case is a malignant case misclassified as benign. Grad-CAM, SHAP, IG and LRP show visual attribution results. CRysTal outputs case-level structured clinical attributes and indicates whether the benign–malignant tendency of each attribute is consistent or conflicting with the final model prediction. **b**, Distribution of risk signals in correctly and incorrectly predicted cases. CRysTal and CBM use Reasoning Conflict (RCF) as the risk signal, whereas visual attribution methods use IoU between attribution maps and lesion regions as a surrogate signal. Statistical significance is denoted as ns, $P < 0.05$, $P < 0.01$, $P < 0.001$ and $P < 0.0001$. **c**, ROC analysis for distinguishing correct from incorrect AI predictions. CRysTal uses RCF as the risk score and achieves an AUC of 0.745, higher than visual attribution methods and CBM. **d**, Clinician ratings of CRysTal structured explanations, including explanation accuracy, readability and guideline concordance. Differences between senior experts and junior physicians were not statistically significant across the three dimensions. **e**, Three-phase interactive reading workflow. Twenty clinicians, including 10 senior experts and 10 junior residents, evaluated breast ultrasound cases under image-only, black-box AI score-assisted and explainability-assisted conditions, with a washout interval of at least 3 days between phases. In Phase 3, Grad-CAM and CRysTal were evaluated on two balanced 40-case subsets. **f**, Change in reading time before and after CRysTal use. Thin lines indicate individual physicians, and the thick line indicates the group mean. Reading time decreased by 10.7% for junior physicians and 4.7% for senior experts. **g**, Change in overall diagnostic accuracy across the three-stage workflow, compared with the performance of junior physicians using Grad-CAM assistance. **h**, Diagnostic accuracy in more challenging case subsets, showing that structured concept explanations provided additional decision support for junior physicians compared with black-box scores or visual heatmaps alone.

2.6 Cross-dataset generalization

To further evaluate the generalization of CRysTal to external data distributions and different medical imaging tasks, we tested it on two independent public breast ultrasound datasets, BrEaST and BUSI, one internal thyroid ultrasound dataset and the public dermoscopy dataset Derm7pt. The BrEaST dataset contains 256 breast ultrasound scans covering benign lesions, malignant lesions and normal tissue. Lesion regions and BI-RADS signs/lexicon annotations were provided by radiologists with breast ultrasound experience, enabling assessment of feature-level consistency between CRysTal explanations and breast ultrasound clinical signs. The BUSI dataset contains 780 breast ultrasound images from approximately 600 female patients, including normal, benign and malignant images, with lesion segmentation annotations. Because BUSI does not contain BI-RADS feature labels, we mainly used it to assess case-level explanation reliability and diagnostic-tendency consistency on external breast ultrasound data. The internal thyroid ultrasound dataset contains 3,741 benign nodule images and 7,079 malignant nodule images, with TI-RADS feature annotations and benign–malignant labels, enabling evaluation of CRysTal generalization to a standardized clinical sign system in thyroid ultrasound. The Derm7pt dataset contains 1,011 skin lesion cases, each with dermoscopic images, clinical images, patient metadata, 7-point checklist feature labels and diagnostic labels, and was used to evaluate cross-task mapping to clinical concepts in dermoscopy.

Extended Data Table 4 shows that, on the external breast ultrasound datasets, the fixed ResNet-50 black-box predictor achieved diagnostic AUCs of 0.849 on BrEaST and 0.813 on BUSI, indicating that the target diagnostic model retained some classification generalization on independent datasets. On this basis, CRysTal achieved a feature-level consistency of 0.522, a diagnostic-tendency consistency of 0.728 and an RCF-AUC of 0.674 on BrEaST. Its diagnostic-tendency consistency was higher than those of CLIP-Dissect, CLIP-Dissect*, LE and CBM, and its reliability matched CBM while exceeding other fixed-predictor explanation methods. By comparison, CLIP-Dissect, CLIP-Dissect* and LE, which also explained the fixed predictor, had feature-level consistencies of 0.379, 0.344 and 0.320 and diagnostic-tendency consistencies of 0.469, 0.448 and 0.493, respectively, all lower than CRysTal. On BUSI, CRysTal achieved an RCF-AUC of 0.867, exceeding CLIP-Dissect, CLIP-Dissect*, LE and CBM by 0.172, 0.053, 0.015 and 0.049, respectively. These results suggest that CRysTal maintained relatively stable case-level explanation quality on independent breast ultrasound datasets.

Derm7pt uses the dermoscopic 7-point checklist as its structured concept system, whereas the thyroid ultrasound dataset uses features from the Thyroid Imaging Reporting and Data System (TI-RADS), including composition, echogenicity, shape, margin and echogenic foci. Fig. 7b,c shows that, on the thyroid ultrasound dataset, CRysTal achieved feature-level agreement, tendency-level agreement, MCC and F1 scores of 0.591, 0.656, 0.189 and 0.753, respectively. On Derm7pt, the corresponding scores were 0.538, 0.703, 0.300 and 0.520. Compared with CoT, CLIP-Dissect, CLIP-Dissect* and LE, CRysTal maintained an advantage across these metrics. In the thyroid task, in particular, some baseline methods had MCC values close to or below

0, suggesting weak agreement between recovered concept tendencies and diagnostic direction. By contrast, CRysTal maintained a positive and higher correlation.

Fig. 7d,e further illustrates explanation differences among methods using thyroid ultrasound and dermoscopy cases. In the thyroid ultrasound task, CRysTal recovered TI-RADS-related findings consistent with the reference clinical findings, including shape, margin, echo pattern, composition and echogenic foci, while maintaining a relatively correct diagnostic tendency. Other methods more frequently produced errors in key signs or deviations in diagnostic tendency. In the Derm7pt task, CRysTal also captured 7-point checklist concepts such as pigment network, color distribution, structural asymmetry and blue-whitish structures, whereas CoT, CLIP-Dissect, CLIP-Dissect* and LE more often produced repeated concepts, overemphasized non-key concepts or disagreed with the reference clinical findings. These qualitative observations were consistent with the quantitative metrics and suggest that CRysTal can be adapted to different disease tasks and medical terminology systems.

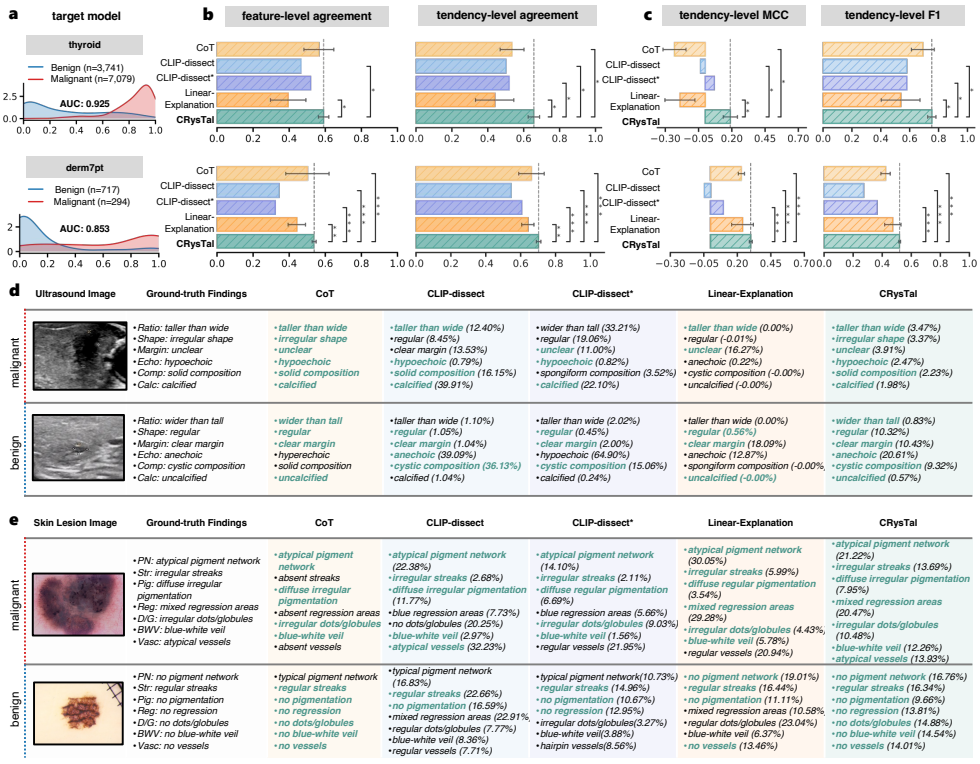


Fig. 7 Generalization validation on public datasets. **a**, Prediction distribution and AUC of the fixed target diagnostic model in the thyroid ultrasound task. The results show that the model had sufficient baseline discriminative ability for explanation generalization analysis. **b**, Comparison of feature-level agreement and tendency-level agreement across different explanation methods in thyroid ultrasound and Derm7pt dermoscopy tasks. Bars indicate mean results over repeated experiments, and error bars indicate standard deviation. CRysTal achieved higher agreement in both tasks, suggesting adaptability to different medical concept systems. **c**, Comparison of tendency-level MCC and tendency-level F1 across different explanation methods. Bars indicate mean results over repeated experiments, and error bars indicate standard deviation. These metrics evaluate diagnostic-direction consistency and overall recognition ability of explanation concepts. CRysTal maintained an advantage in both aspects. **d,e**, Representative thyroid ultrasound and dermoscopy cases. Clinical concepts recovered by CRysTal were more consistent with the reference clinical findings.

3 Discussion

In this study, we introduced CRysTal, a decision-explanation framework that maps neuronal responses from predictive models to standardized clinical terminology. CRysTal addresses several limitations of existing medical imaging XAI approaches, including explanations that are semantically detached from clinical consensus, insufficiently coupled to the target model's decision pathway or poorly adapted to medical images with multiple overlapping features. Systematic evaluations across EfficientNet-B0, ResNet-50, VGG-19 and ViT-Base showed that CRysTal generated verifiable, standardized and decision-relevant structured explanations that were more consistent

with radiology reports than the evaluated baselines. In case-level decision explanation, CRysTal achieved the highest explanation consistency across the four diagnostic models; feature-level accuracy improved by up to 12.2% over the second-best method, and tendency-level accuracy improved by up to 19.2%. CRysTal recovered key malignant breast ultrasound signs, such as irregular shape, angular margin and microcalcifications, together with their diagnostic tendencies. Methodologically, the two-stage design based on non-negative sparse regression and cross-category fusion allowed individual neurons to encode multidimensional clinical semantics, addressing the limitations of single-label neuron explanations. Across Ability, Alignment, Coverage, Diversity and Sparsity, CRysTal provided stronger preservation of discriminative signals, broader coverage of key signs and better integration of complementary concepts across categories, while maintaining activation-pattern matching and sparsity comparable to Linear Explanation. Repeated experiments further showed stable neuron-level explanations with low redundancy and low semantic conflict across different backbones. Together, these results establish a quantifiable mapping from internal neuronal representations to standardized clinical concepts and suggest a route for converting black-box predictions into structured evidence that clinicians can inspect.

Analysis of explanation dynamics during training showed that, as classification performance improved, model explanations became increasingly aligned with expert-defined clinical concepts. CRysTal captured a transition from low-level or non-specific texture cues toward higher-level clinical concepts, and model AUC was strongly correlated with concept alignment (Pearson $r = 0.84$–0.89). These findings support the interpretation that the evaluated models acquired representations increasingly consistent with radiological evidence. CRysTal also supported clinical risk assessment and reader decision-making. The Reasoning Conflict (RCF) metric derived from structured explanations identified erroneous model predictions with an AUC of 0.745, outperforming visual attribution methods (0.542–0.625) and CBM (0.643). In the multi-reader study involving 20 ultrasound physicians, CRysTal explanations received mean scores close to 97 for explanation accuracy, readability and guideline concordance, with no significant differences between senior experts and junior physicians. In the interactive reading workflow, CRysTal reduced diagnostic time by 4.7–10.7% and improved accuracy, with a larger benefit among junior physicians. These findings suggest that structured concept explanations can provide practical decision support, especially when clinicians need to evaluate whether an AI prediction is reliable.

CRysTal also showed cross-dataset and cross-task adaptability across breast ultrasound, thyroid ultrasound and dermoscopy. On independent public datasets, including BrEaST, BUSI and Derm7pt, CRysTal outperformed the evaluated baselines on metrics including feature agreement, tendency agreement, F1 and MCC, and could be adapted to different terminology systems, including BI-RADS, TI-RADS and the 7-point checklist. Variation in performance across tasks suggests that explainability depends on data quality, annotation granularity and the structure of the clinical terminology system, which should guide future domain-specific optimization.

This study has several limitations. First, CRysTal is most naturally suited to medical imaging domains with complex and fine-grained clinical sign systems, such as BI-RADS, TI-RADS and the 7-point checklist. These systems provide explicit

references for neuron–concept mapping, but public medical imaging datasets with fine-grained feature annotations remain limited. The generalization of CRysTal to additional disease types, imaging modalities and multicenter datasets therefore requires further validation. Second, although CRysTal may help identify candidate diagnostic signs associated with decision-related neurons, clinical guidelines are built from long-term clinical experience, pathological associations and multicenter validation. Model explanations should not be treated as direct evidence for modifying guidelines. Future work should combine expert review, pathological outcomes and clinical studies to validate whether candidate signs are truly associated with disease risk, diagnostic decisions or patient prognosis.

Overall, CRysTal helps address the limited human interpretability of internal decision evidence in medical imaging AI by establishing an auditable mapping between neuronal responses and standardized clinical concepts. The framework generates stable and traceable structured explanations and may support human–AI collaborative diagnosis, risk stratification and diagnostic consistency. From ethical and regulatory perspectives, CRysTal provides a transparent decision-evidence chain for AI diagnosis, which may help reduce risks associated with black-box models and improve the auditability of medical imaging AI. Further validation across diseases, institutions and prospective clinical workflows will be needed before broad clinical deployment.

4 Methods

4.1 Ethics approval and consent

This study was designed as a retrospective multi-centre clinical study and was conducted in accordance with the Declaration of Helsinki and relevant institutional guidelines. Ethical approvals for the breast ultrasound cohort were obtained from the participating centres under the following reference numbers: prylz2020-082 (Shanghai Pudong New Area People's Hospital), XZXY-LK-20240710-0110 (Xuzhou Central Hospital), 2024-KY-189 (The Second People's Hospital of Wuhu), 050432-4-2307E (Fudan University Shanghai Cancer Center), 2024-762-02 (Nanjing Drum Tower Hospital), KY2023-054-B (Renji Hospital, Shanghai Jiao Tong University School of Medicine) and 2023GLLH0461 (Xing'an League People's Hospital). Ethical approval for the thyroid ultrasound cohort was obtained from the Biomedical Research Ethics Review Committee of Xuzhou Central Hospital (approval no. XZXY-LJ-20201208-0194). The requirement for informed consent was waived for the retrospective component. To protect patient privacy, all ultrasound images and clinical data were anonymised before inclusion in the analysis.

The clinician reader studies were conducted using anonymised breast ultrasound images. Participation was voluntary, and all participating ultrasound physicians agreed to participate before the study.

4.2 Datasets and clinical concept annotations

The primary experiments were conducted on an internal multi-centre breast ultrasound cohort collected from the same participating centres covered by the ethics

approvals. The cohort contained 30,303 breast ultrasound images from 12,184 subjects, including 19,371 benign and 10,932 malignant images, together with benign–malignant diagnostic labels and structured clinical concept annotations provided by physicians. The breast ultrasound concept vocabulary was organized according to six BI-RADS-related categories: shape, margin, orientation, echo pattern, posterior acoustic features and calcification. Each clinical concept was encoded as a binary variable indicating whether the corresponding feature was present in the image or clinical annotation.

The internal breast ultrasound cohort was split at the patient level into training, validation and test sets with a ratio of 60%/20%/20%. Patient-level splitting was used to avoid image-level leakage between model development and evaluation. The diagnostic labels were used to train the target benign–malignant classifiers, whereas the structured clinical concept annotations were used to fit and evaluate CRysTal after the target models had been trained and frozen.

External breast ultrasound evaluation was performed on BrEaST and BUSI. BrEaST contains 256 breast ultrasound scans from 256 patients with lesion annotations and BI-RADS feature/lexicon labels, and was used to assess feature-level and tendency-level explanation consistency [41]. BUSI contains 780 breast ultrasound images categorized as normal, benign and malignant, with lesion masks, and was used for external case-level reliability evaluation because it does not provide complete BI-RADS feature annotations [42]. All available cases from BrEaST and BUSI were used in the corresponding external evaluations.

For cross-task evaluation, we further used an internal thyroid ultrasound dataset containing 3,741 benign nodule images and 7,079 malignant nodule images with TI-RADS-related feature annotations and benign–malignant labels. We also used the public Derm7pt dermoscopy dataset, which provides dermoscopic 7-point checklist annotations and diagnostic labels [43]. All available Derm7pt clinical concepts were used. For each task, the original clinical terminology system was mapped to a task-specific structured concept vocabulary and used to evaluate whether CRysTal could generalize across standardized medical imaging concept systems.

All private clinical data were anonymised before analysis. Public datasets were used according to their respective data-use terms.

4.3 Target model training

Target diagnostic models were trained to perform benign–malignant classification from medical images. In the primary breast ultrasound task, we trained four representative backbone architectures: EfficientNet-B0, ResNet-50, VGG-19 and ViT-B. All models were initialized with ImageNet-pretrained weights and fine-tuned on the training split of the internal breast ultrasound cohort. Input images were resized to 224×224 pixels and normalized using the mean and standard deviation calculated from the corresponding training dataset. Model selection was performed on the validation split using AUC, and the selected models were evaluated once on the held-out test split.

For CRysTal analysis, the trained target models were fixed. CRysTal did not modify the target-model parameters, retrain the classifier or change the diagnostic prediction pathway. Instead, activations were extracted from selected internal layers

of each fixed target model and paired with structured clinical concept annotations to learn neuron–concept profiles.

For external breast ultrasound evaluation, the fixed ResNet-50 breast ultrasound target model was applied to BrEaST and BUSI without additional fine-tuning. For thyroid ultrasound and Derm7pt dermoscopy, task-specific target models were trained using the corresponding task data and were then frozen before fitting the corresponding CRysTal explainer. For all target models, diagnostic performance was evaluated using AUC calculated from continuous prediction scores.

Target models were implemented in Python 3.9.17 using PyTorch 1.13.0 and torchvision 0.14.0. Model evaluation and statistical analyses were performed using scikit-learn 1.3.0, scipy 1.11.1 and statsmodels 0.14.4. Visual attribution baselines were implemented using Captum 0.7.0, grad-cam 1.5.0, SHAP 0.45.1, zennit 0.4.6 and zennit-crp 0.6.0. Vision–language baselines used open-clip-torch 2.32.0 and transformers 4.36.0 where applicable. The complete software environment is provided in the accompanying code repository.

4.4 Decomposition of the CRysTal model architecture

CRysTal is a neuron-level post-hoc explanation framework designed to reconstruct structured clinical evidence from the internal activations of a fixed diagnostic model. The framework consists of three main components that correspond to the workflow shown in Fig. 2: data preparation, neuron–concept alignment and case-level evidence reconstruction. First, structured clinical descriptors are organized into a category-aware concept taxonomy and paired with neuronal activation vectors extracted from the fixed target model. Second, CRysTal learns a neuron–concept profile for each analysed neuron through within-category concept-weight estimation and cross-category integration. Third, for a test image, output-specific neuron relevance scores are used to aggregate the learned neuron–concept profiles into a case-level evidence vector.

4.4.1 Task formulation and data organization

We consider a setting in which a diagnostic model is first trained to perform a clinical prediction task, and a separate post-hoc explainer, CRysTal, is then trained to reconstruct the structured clinical evidence associated with the model's individual predictions. We refer to the diagnostic classifier as the target model. Given an input image x_i, the target model f_θ outputs a predictive probability

$$\hat{p}_i = f_\theta(x_i) \in [0, 1],$$

which is converted into a binary diagnostic decision

$$\hat{y}_i = \mathbb{I}(\hat{p}_i \geq \tau_y), \qquad \hat{y}_i \in \{0, 1\},$$

where $\hat{y}_i = 0$ denotes benign and $\hat{y}_i = 1$ denotes malignant.

CRysTal is trained after the target model has been obtained and is applied with the target model fixed. It does not modify the target-model parameters, re-estimate the diagnostic probability or serve as a replacement classifier. Instead, CRysTal operates

on the internal activations of a selected layer of the fixed target model and reconstructs a structured case-level evidence vector that describes the clinical concepts associated with the current prediction.

In the primary breast ultrasound experiments, the target model and CRysTal were developed using the same clinical cohort, but with different supervision signals and objectives. Each sample contained an ultrasound image x_i, an image-level diagnostic label y_i and a set of structured clinical descriptors derived from dataset-provided annotations. The diagnostic label y_i was used to train the target model for benign–malignant classification, whereas the structured descriptors were paired with frozen target-model activations to train CRysTal after the target model had been obtained.

The structured descriptors were reorganized into a binary concept annotation matrix

$$\mathbf{Z} \in \{0,1\}^{N\times C},$$

where N denotes the number of samples in the CRysTal alignment split and C denotes the number of clinical concepts in the predefined vocabulary. The i-th row,

$$\mathbf{z}_i = \mathbf{Z}_{i,:} \in \{0,1\}^{C},$$

records the reference clinical findings for image x_i, with $z_{i,c} = 1$ indicating that concept c is present in the annotation and $z_{i,c} = 0$ otherwise. The concept vocabulary consisted of BI-RADS-related descriptors, including shape, margin, orientation, echo pattern, posterior acoustic features and calcification. These concepts were grouped into G diagnostic categories, and $\mathcal{C}_g \subseteq \{1,\dots,C\}$ denotes the subset of concept indices belonging to category g.

Once the target model had been trained, its parameters were frozen. The images in the CRysTal alignment split were then passed through the fixed target model, and activations were extracted from a selected layer ℓ. After applying the layer-specific aggregation operation, we obtained an activation vector

$$h_\ell(x_i) \in \mathbb{R}^K,$$

where K is the number of analysed neurons. In this work, the term neuron is used operationally to denote a scalar component of this aggregated activation vector. For convolutional backbones, this corresponds to a channel-level response obtained after spatial aggregation. For transformer-based backbones, it corresponds to a scalar feature dimension after the selected token or feature aggregation. The k-th component of $h_\ell(x_i)$ is denoted by $a_{i,k}$ and represents the response of neuron u_k to image x_i.

Across the N alignment samples, these responses form an activation matrix

$$\mathbf{A} = [a_{i,k}] \in \mathbb{R}^{N\times K}.$$

CRysTal uses the paired matrices $(\mathbf{A}, \mathbf{Z})$ to learn neuron–concept alignments. Thus, although the target model and CRysTal are developed from the same cohort, they serve distinct roles: the target model learns the image-to-label mapping, whereas CRysTal learns how the frozen model's internal neuron responses are associated with structured clinical concepts.

In the main implementation, the target model was instantiated as a ResNet-50 initialized with ImageNet pre-training and fine-tuned on an internal breast ultrasound cohort. The CRysTal alignment procedure was fitted on the training split using activations extracted from the fixed target model and the corresponding structured concept annotations. To examine architectural robustness, we additionally instantiated target models using VGG, EfficientNet and Vision Transformer backbones and trained a corresponding CRysTal explainer for each fixed target model. To assess robustness under domain shift, the trained target model and its associated CRysTal explainer were evaluated without adaptation on two external public breast ultrasound datasets, BrEaST and BUSI. To examine generalization beyond breast ultrasound, we further trained separate target models and CRysTal explainers on thyroid ultrasound and Derm7pt dermoscopy datasets, which provide distinct imaging modalities and clinical concept taxonomies.

4.4.2 Neuron–concept alignment

The neuron–concept alignment component learns how each analysed neuron in the selected layer of the fixed target model is associated with the predefined clinical concept vocabulary. For neuron u_k, let

$$\mathbf{q}_k = [a_{1,k}, a_{2,k}, \ldots, a_{N,k}]^\top \in \mathbb{R}^N$$

denote its response vector across the N CRysTal alignment samples. CRysTal estimates a neuron–concept profile

$$\mathbf{m}_k \in \mathbb{R}^C$$

for this neuron, where each entry of $\mathbf{m}_k$ quantifies the association between the response pattern of u_k and a clinical concept.

A direct fit from all concepts to a neuron response may entangle clinically distinct descriptor groups. To preserve the structure of the diagnostic taxonomy, CRysTal estimates the neuron–concept profile in two stages. The first stage performs within-category concept-weight estimation. The second stage performs cross-category integration to obtain a unified neuron–concept profile.

Within-category concept-weight estimation.

For diagnostic category g, we use the concept submatrix

$$\mathbf{Z}_{\mathcal{C}_g} \in \{0,1\}^{N \times |\mathcal{C}_g|}$$

that contains only the concepts in $\mathcal{C}_g$. The CRysTal alignment split is further divided into an alignment-training subset and an alignment-validation subset. On the alignment-training subset, the response of neuron u_k is modelled as a sparse non-negative combination of concepts within category g:

$$\mathbf{q}_k^{\mathrm{tr}} = \mathbf{Z}_{\mathcal{C}_g}^{\mathrm{tr}} \mathbf{w}_{k,g} + b_{k,g}\mathbf{1} + \boldsymbol{\epsilon},$$

where $\mathbf{w}_{k,g} \in \mathbb{R}^{|\mathcal{C}_g|}$ is the category-specific concept-weight vector and $b_{k,g}$ is an intercept. The weights are estimated by solving

$$\min_{\mathbf{w}_{k,g} \geq \mathbf{0},\, b_{k,g}} \left\| \mathbf{q}_k^{\mathrm{tr}} - \mathbf{Z}_{\mathcal{C}_g}^{\mathrm{tr}} \mathbf{w}_{k,g} - b_{k,g}\mathbf{1} \right\|_2^2 + \lambda \|\mathbf{w}_{k,g}\|_1.$$

The sparsity coefficient λ was fixed to 1×10^{-3} in the main experiments and was used for both within-category concept-weight estimation. The non-negativity constraint makes the resulting associations additive and avoids cancellation between positive and negative concept weights. The ℓ_1 penalty encourages a concise profile in which each neuron is associated with a limited number of clinical concepts within each category. Estimating weights within each diagnostic category preserves the clinical structure of the concept space and prevents unrelated descriptor types, such as shape and margin, from being mixed in the same local fit.

Cross-category integration.

A neuron may be informative for multiple diagnostic categories to different degrees. CRysTal therefore learns non-negative category-fusion coefficients to determine how much each category-specific fit contributes to the final neuron–concept profile. For each category g, the within-category regression fitted above is used to predict the neuron response on the alignment-validation subset:

$$\hat{\mathbf{q}}_{k,g}^{\mathrm{val}} = \mathbf{Z}_{\mathcal{C}_g}^{\mathrm{val}} \mathbf{w}_{k,g} + b_{k,g}\mathbf{1}.$$

Stacking the predictions across all G categories yields

$$\hat{\mathbf{Q}}_k^{\mathrm{val}} = [\hat{\mathbf{q}}_{k,1}^{\mathrm{val}}, \hat{\mathbf{q}}_{k,2}^{\mathrm{val}}, \ldots, \hat{\mathbf{q}}_{k,G}^{\mathrm{val}}] \in \mathbb{R}^{N_{\mathrm{val}} \times G}.$$

Let

$$\boldsymbol{\alpha}_k = [\alpha_{k,1}, \ldots, \alpha_{k,G}]^\top$$

denote the category-fusion vector for neuron u_k. We estimate it by fitting the validation responses:

$$\min_{\boldsymbol{\alpha}_k \geq \mathbf{0},\, b_{0k}} \left\| \mathbf{q}_k^{\mathrm{val}} - \left(\hat{\mathbf{Q}}_k^{\mathrm{val}} \boldsymbol{\alpha}_k + b_{0k}\mathbf{1} \right) \right\|_2^2.$$

The final neuron–concept profile of neuron u_k is obtained by scaling each category-specific concept-weight vector by its fusion coefficient and concatenating the results according to the global concept order:

$$\mathbf{m}_k = [\alpha_{k,1}\mathbf{w}_{k,1}^\top, \alpha_{k,2}\mathbf{w}_{k,2}^\top, \ldots, \alpha_{k,G}\mathbf{w}_{k,G}^\top]^\top \in \mathbb{R}^C.$$

The collection $\{\mathbf{m}_k\}_{k=1}^K$ constitutes the learned neuron–concept alignment of the selected target-model layer.

4.4.3 Case-level evidence reconstruction

The neuron–concept profiles learned above describe the clinical concepts with which each neuron is associated across the CRysTal alignment samples. To explain a specific prediction, CRysTal assigns an output-specific relevance score to every analysed neuron in the selected target layer. For a test image x_i, we compute a neuron-level relevance score

$$r_{i,k}$$

for each neuron u_k. This score quantifies the contribution of neuron u_k to the target model's predicted output for that particular image. In the main implementation, $r_{i,k}$ is computed using Layer-wise Relevance Propagation (LRP) from the model output to layer ℓ. The same aggregation framework can accommodate other neuron-level relevance estimators; we evaluated alternatives, including Grad×Act, in the ablation analysis.

CRysTal reconstructs the case-level evidence vector by relevance-weighted aggregation of neuron–concept profiles:

$$\mathbf{E}_i = \sum_{k=1}^{K} r_{i,k}\mathbf{m}_k \in \mathbb{R}^C.$$

The entry $\mathbf{E}_i(c)$ is the evidence score assigned to clinical concept c for image x_i. Because the concept vocabulary is organized by diagnostic category, the evidence vector can be partitioned as

$$\mathbf{E}_i = [\mathbf{E}_i^{(1)\top}, \ldots, \mathbf{E}_i^{(G)\top}]^\top, \qquad \mathbf{E}_i^{(g)} \in \mathbb{R}^{|\mathcal{C}_g|}.$$

This yields a structured case-level explanation in which the model's current prediction is supported by explicit clinical concepts rather than by an unstructured heatmap or free-text rationale.

For presentation and evaluation, the evidence vector is summarized within each diagnostic category. Specifically, the highest-scoring concept, or the top-ranked concepts, within each category are retained to form a compact clinical evidence chain. This chain can be directly compared with reference clinical findings and reviewed by clinicians.

For neuron-level evaluation requiring a ranked list of concepts, we further derived a compact greedy readout from the full CRysTal neuron–concept profile. The original CRysTal profile is category-aware and may assign weights to multiple concepts across several BI-RADS categories. To make it comparable with baselines that output ranked concept lists, we greedily selected the highest-weight concepts from the fused neuron profile to obtain a compact top-K readout. This compressed readout was used only for top-K-based evaluation, including Val Corr, Consistency@K, Redundancy@K and Conflict@K. The full category-aware CRysTal profile was retained for group-level interpretation and case-level evidence reconstruction.

4.5 Reasoning-conflict score for risk-aware verification

The CRysTal evidence vector can also be used to derive a case-level warning signal. This analysis is separate from the evidence reconstruction procedure itself: CRysTal first reconstructs the case-level evidence vector $\mathbf{E}_i$, and the reasoning-conflict score then evaluates whether this evidence is directionally consistent with the target model's predicted class.

A clinically useful evidence reconstruction should not only summarize the model's apparent decision evidence, but also indicate when the reconstructed evidence conflicts with the final prediction. We therefore define Reasoning Conflict (RCF) to quantify the extent to which the extracted evidence supports the opposite diagnostic direction.

For each diagnostic category g, we identify the dominant concept in the case-level evidence vector:

$$j_g(x_i) = \arg \max_{j \in \{1,\dots,|\mathcal{C}_g|\}} E_{i,j}^{(g)}, \qquad s_{i,g} = E_{i,j_g(x_i)}^{(g)}.$$

The magnitude of the dominant evidence is

$$v_{i,g} = |s_{i,g}|.$$

Each clinical concept is assigned a predefined diagnostic tendency

$$\pi(g, j) \in \{-1, 0, +1\},$$

where -1 denotes benign-associated evidence, 0 denotes neutral or unspecified evidence and $+1$ denotes malignant-associated evidence. The diagnostic direction of the target model's prediction is

$$t(\hat{y}_i) = 2\hat{y}_i - 1 \in \{-1, +1\}.$$

The reasoning-conflict score is defined as the magnitude-weighted proportion of dominant category-level evidence whose diagnostic tendency opposes the predicted class:

$$\mathrm{RCF}(x_i) = \frac{\sum_{g=1}^{G} v_{i,g} \cdot \mathbb{I}\left[\pi(g, j_g(x_i)) = -t(\hat{y}_i)\right]}{\sum_{g=1}^{G} v_{i,g} + \varepsilon},$$

where $\varepsilon > 0$ is a small constant for numerical stability. A high $\mathrm{RCF}(x_i)$ indicates that the concepts reconstructed from the model's internal activations predominantly support the opposite diagnostic direction from the final prediction. We use this score as a case-level warning signal for potentially unreliable or failure-prone decisions.

4.6 Baseline explanation methods

CRysTal was compared with representative explanation baselines spanning language-based reasoning, vision–language neuron description, linear neuron–concept association, explicit concept bottleneck modelling and visual attribution.

For case-level language-based explanation, we used a Chain-of-Thought (CoT) baseline generated by Qwen2.5-VL-7B-Instruct (Qwen/Qwen2.5-VL-7B-Instruct), a locally deployed instruction-tuned vision–language model [44]. For each case, the model was given the input image, the black-box model prediction label and a restricted list of candidate clinical concepts. The prompt instructed the model not to re-diagnose the image, but to describe visible image features that supported or contradicted the given black-box prediction. The model was not provided with the reference clinical report, ground-truth diagnostic label or reference concept annotations. For the main case-level comparison, each case was processed once using constrained decoding. For repeated-sampling analysis, the CoT baseline was run five times with sampling enabled and temperature set to 0.7. The model output was constrained to a structured JSON-like format. The generated outputs were parsed and converted into binary concept profiles using rule-based exact matching after normalization against the same task-specific clinical concept vocabulary used by CRysTal.

CLIP-Dissect was used as a neuron-description baseline that assigns natural-language concepts to hidden neurons using image–text similarity [34]. The original CLIP-Dissect baseline used the RN50 CLIP backbone. We also evaluated a medically adapted variant, denoted CLIP-Dissect*, in which the CLIP image–text encoder was replaced by BiomedCLIP [32]. For both CLIP-Dissect variants, the candidate concept vocabulary was restricted to the same task-specific structured clinical concept vocabulary used by CRysTal, including the BI-RADS-related vocabulary for breast ultrasound, the TI-RADS-related vocabulary for thyroid ultrasound and the Derm7pt vocabulary for dermoscopy. The resulting concept-level neuron descriptions were used to generate neuron-level explanations and case-level concept scores.

Linear Explanation (LE) was implemented following Linear Explanations for Individual Neurons [45]. LE explains each neuron as a sparse linear combination of interpretable concepts. In our experiments, we used the label-based LE setting, in which the concept activation matrix was constructed from the same structured clinical concept annotations used by CRysTal. For each analysed neuron, the neuron activation vector was treated as the regression target. Candidate concepts were first screened using a sparse linear model, and a greedy concept-selection procedure was then used to obtain a compact linear explanation that predicted neuron activations on held-out samples. The resulting concept weights were used to generate neuron-level explanations and case-level concept scores. LE was evaluated using the same analysed neurons, concept vocabulary and reference annotations as CRysTal.

A sequential Concept Bottleneck Model (CBM) was trained as an explicitly concept-supervised reference model [46]. The CBM first predicted structured clinical concepts from the input image and then used the predicted concepts to make the final benign–malignant diagnosis. In the breast ultrasound task, the CBM used all six BI-RADS-related concept categories: shape, margin, orientation, echo pattern, posterior acoustic features and calcification. This baseline was used to compare CRysTal with an interpretable model whose diagnostic pathway was explicitly constrained by clinical concepts.

For visual attribution baselines, we evaluated Grad-CAM, integrated gradients, SHAP, layer-wise relevance propagation and concept relevance propagation [24–27, 47].

These methods generated spatial attribution maps indicating image regions associated with model predictions. For ResNet-based models, Grad-CAM was applied to the last convolutional block. Integrated gradients used a zero-image baseline with 50 integration steps. SHAP was implemented using a gradient-based explainer on the last convolutional feature activations. Layer-wise relevance propagation was implemented using an epsilon-plus-flat composite with batch-normalization merging. Concept relevance propagation was applied to the last convolutional block, and the top 10 relevance-ranked channels were used to generate concept-conditioned attribution maps.

For localization-based evaluation, attribution maps were converted to non-negative maps by retaining positive attribution values; if no positive attribution was present, absolute attribution values were used. Attribution maps were compared with lesion masks using pointing-game top-1 hit, in-mask relevance ratio, in/out relevance ratio, top-20% in-mask relevance ratio and top-20% in/out relevance ratio. For risk-aware verification, visual attribution metrics were used as localization-based risk signals, whereas CRysTal and CBM were evaluated using Reasoning Conflict.

All baseline methods were evaluated on the same test cases and, where applicable, using the same target models, concept vocabulary and reference annotations as CRysTal.

4.7 Clinician reader studies

We conducted two clinician validation experiments using two independent image sets. The first experiment included 100 breast ultrasound images and was designed to assess physicians' subjective evaluation of CRysTal-generated explanations. The second experiment used a separate set of 80 breast ultrasound images and was designed to assess the effect of CRysTal on diagnostic decision-making. The two image sets were non-overlapping, so that prior exposure to CRysTal explanations in the first experiment did not influence reader decisions in the second experiment.

In the first experiment, 20 ultrasound physicians, including 10 junior residents and 10 senior experts, evaluated CRysTal-generated explanations from 100 breast ultrasound images. Each explanation was scored on a 0–100 scale in terms of explanation accuracy, readability and guideline concordance. Higher scores indicated better perceived clinical quality.

In the second experiment, the same 20 physicians participated in a three-phase multi-reader, multi-case interactive reading workflow. They classified 80 breast ultrasound images as benign or malignant under three sequential conditions. In Phase 1, readers interpreted the original ultrasound images alone, establishing independent physician performance. In Phase 2, the AI-predicted malignancy probability was provided, and readers made a new diagnosis. In Phase 3, AI explainability information was further provided, and readers made a final diagnosis. A washout interval of at least 3 days was used between phases to reduce recall bias.

For Phase 3, the 80 cases were divided into two balanced 40-case subsets. One subset was interpreted with Grad-CAM heatmap assistance, and the other subset was interpreted with CRysTal-generated structured feature explanations. The two subsets were balanced by diagnosis and AI prediction correctness to compare visual

heatmap explanations with structured concept explanations under comparable case composition. Diagnostic accuracy and reading time were recorded at each phase and analysed overall and after stratification by reader experience level.

4.8 Statistical analysis

Continuous variables are reported as mean ± standard deviation unless otherwise specified. For repeated experiments, results were summarized across independent runs or repeated explanation generations. For case-level explanation evaluation, micro-averaged metrics were computed across all evaluable case–category pairs, whereas macro-averaged metrics were computed by first averaging within each clinical concept category and then averaging across categories.

For neuron-level repeated-run reliability analysis, explanation generation was repeated ten times for each selected feature layer and backbone. Method-level summary statistics were calculated across analysed neurons. For each metric, we reported the mean, standard deviation, standard error and bootstrap 95% confidence interval where applicable. Bootstrap confidence intervals were estimated by resampling neuron-level metric values with replacement using 2,000 bootstrap iterations.

For ROC analyses, AUC values were computed from continuous prediction scores or risk scores. In risk-aware verification, the ability of Reasoning Conflict or localization-based risk signals derived from visual attribution maps to distinguish correctly and incorrectly predicted cases was evaluated using ROC-AUC. Differences in risk-score distributions between correctly and incorrectly predicted cases were assessed using Welch's two-sample t-test. Statistical significance was defined as $P < 0.05$. Significance annotations were defined as follows: ns, $0.05 < P \leq 1.00$; *, $0.01 < P \leq 0.05$; **, $0.001 < P \leq 0.01$; ***, $0.0001 < P \leq 0.001$; and ****, $P \leq 0.0001$.

For the clinician reader studies, diagnostic accuracy and reading time were analysed overall and after stratification by reader experience level. In the explanation-rating experiment, differences between junior residents and senior experts were assessed using Welch's two-sample t-test. Changes in reading time before and after CRysTal assistance were assessed using paired t-tests within readers. Between-group comparisons of percentage time reduction between junior residents and senior experts were assessed using Welch's two-sample t-test. In the multi-reader, multi-case diagnostic study, results were further stratified by whether the AI prediction was correct or incorrect. A baseline-adjusted generalized estimating equation (GEE) sensitivity analysis was performed for Phase 3 binary diagnostic accuracy, with correctness as the binary outcome, reader ID as the clustering variable, a binomial family with logit link and an exchangeable working correlation structure. All statistical analyses were performed using Python 3.9.17 with scikit-learn 1.3.0, scipy 1.11.1 and statsmodels 0.14.4.

4.9 Evaluation metrics

Diagnostic classification performance was evaluated using the area under the receiver operating characteristic curve (AUC), calculated from continuous prediction scores.

Case-level explanation performance was evaluated by comparing the clinical concepts recovered by each explanation method with reference annotations or clinical

reports. For each case, recovered concepts and reference concepts were organized according to the same clinical concept taxonomy. Feature-level agreement was evaluated as a category-wise top-1 hit rate. For each case and each clinical concept category, candidate concepts were ranked according to the explanation score assigned by the method, and the highest-scoring concept was selected. A category was counted as correctly recovered if the selected concept matched any reference concept annotated as present in that category. Case–category pairs without any positive reference concept were excluded from this calculation. Micro feature-level agreement was calculated by averaging the hit indicator over all evaluable case–category pairs, whereas macro feature-level agreement was calculated by first averaging within each category and then averaging across categories.

Tendency-level agreement evaluated whether the recovered concepts were directionally consistent with the reference clinical evidence. Each clinical concept was assigned a predefined diagnostic tendency: benign-associated, malignant-associated or uncertain. For each case and each clinical concept category, the reference tendency was determined from the annotated positive concepts in that category, and the predicted tendency was determined from the top-1 recovered concept. Case–category pairs with uncertain reference tendency were excluded from tendency-level accuracy calculation. Micro tendency-level agreement was calculated by averaging the correctness of predicted tendency over all valid case–category pairs, whereas macro tendency-level agreement was calculated by first averaging within each category and then averaging across categories. Tendency-level F1 score and Matthews correlation coefficient (MCC) were further reported to evaluate diagnostic-direction consistency under class imbalance.

Neuron-level explanation quality was evaluated using five complementary metrics: Ability, Alignment, Coverage, Diversity and Sparsity. Ability measured whether the concepts recovered for a neuron preserved malignant-oriented discriminative information. For each diagnostic category, concept scores were aggregated from the concept weights produced by the explanation method and evaluated against malignant-associated concept labels using AUC; the score was transformed as $2|\mathrm{AUC} - 0.5|$ and averaged over valid categories. Alignment measured the absolute Pearson correlation between the observed neuron activation vector and the concept-based activation vector reconstructed from the recovered concept weights. Coverage measured the fraction of diagnostic categories whose normalized positive concept weight exceeded a predefined threshold. Diversity combined the fraction of diagnostic categories represented among the top-ranked concepts and the fraction represented among lower-weight concepts. Sparsity was measured using Hoyer sparsity after removing near-zero concept weights.

To assess repeated-run reliability, neuron-level explanation generation was repeated ten times for the same selected feature layer of each backbone model. For each method and run, the top-K concepts of each neuron were extracted by ranking concepts according to the absolute value of their explanation weights. Val Corr measured the validation correlation reported by each explanation run. Consistency@K measured the mean pairwise Jaccard similarity of top-K concept sets across repeated runs. Redundancy@K measured the fraction of repeated clinical concept groups among the top-K concepts. Conflict@K measured the fraction of explanations containing at

least one predefined mutually exclusive or clinically contradictory concept pair. These repeated-run metrics were averaged across neurons shared by all repeated runs.

For risk-aware verification, CRysTal and the concept bottleneck model used Reasoning Conflict (RCF) as the case-level risk score. Visual attribution methods, including Grad-CAM, integrated gradients, SHAP, layer-wise relevance propagation and concept relevance propagation, were evaluated using the intersection over union (IoU) between attribution heatmaps and ground-truth lesion masks. The ability of each risk score to distinguish correct from incorrect AI predictions was evaluated using ROC-AUC. Detailed mathematical definitions of the case-level agreement metrics, neuron-level quality metrics and repeated-run reliability metrics are provided in Supplementary Methods.

Code availability. The code used for CRysTal implementation, model evaluation and figure reproduction is available from the corresponding author upon reasonable request. The code will be released in a public repository upon publication. Code involving institutional data access or protected clinical infrastructure is not publicly available owing to institutional data-governance restrictions.

Data availability. The private breast and thyroid ultrasound datasets used in this study are not publicly available because they contain patient-level medical images and clinical reports and are subject to institutional ethics and data-governance restrictions. De-identified data may be made available from the corresponding author upon reasonable request, subject to approval by the relevant institutional review board and completion of a data-use agreement. Public datasets used in this study, including BrEaST, BUSI and Derm7pt, are available from their original sources.

Acknowledgements. We thank the ultrasound physicians and clinical experts who participated in image annotation, clinical review and reader studies. This work was supported by the National Natural Science Foundation of China (grant no. 62406191), the Fundamental Research Funds for the Central Universities (grant no. YG2025LC03), the Innovation Ecosystem Construction Plan of Shanghai Municipal Science and Technology Commission (grant no. 25010701400), the Shanghai Special Fund for Promoting High-Quality Industrial Development–Pilot Industry Innovation Development (AI Special Topic) Project (grant no. 2025-GZL-RGZN-02078), the National Key Research and Development Program of China (grant no. 2025YFC2511603), and Neil Shen's SJTU Medical Research Fund. The funders had no role in study design, data collection, data analysis, data interpretation or manuscript preparation.

Competing interests. The authors declare no competing interests.

References

[1] Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**, 44–56 (2019) https://doi.org/10.1038/s41591-018-0300-7

[2] Esteva, A., *et al.*: A guide to deep learning in healthcare. Nature Medicine **25**, 24–29 (2019) https://doi.org/10.1038/s41591-018-0316-z

[3] McKinney, S.M., *et al.*: International evaluation of an ai system for breast cancer screening. Nature **577**, 89–94 (2020) https://doi.org/10.1038/s41586-019-1799-6

[4] Ardila, D., *et al.*: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine **25**, 954–961 (2019) https://doi.org/10.1038/s41591-019-0447-x

[5] Gulshan, V., *et al.*: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**, 2402–2410 (2016) https://doi.org/10.1001/jama.2016.17216

[6] Shen, P., *et al.*: Explainable multimodal deep learning for predicting thyroid cancer lateral lymph node metastasis using ultrasound imaging. Nature Communications **16**, 7052 (2025) https://doi.org/10.1038/s41467-025-62042-z

[7] Moor, M., *et al.*: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023) https://doi.org/10.1038/s41586-023-05881-4

[8] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**, 195 (2019) https://doi.org/10.1186/s12916-019-1426-2

[9] Wiens, J., *et al.*: Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine **25**, 1337–1340 (2019) https://doi.org/10.1038/s41591-019-0548-6

[10] Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**, 745–750 (2021) https://doi.org/10.1016/S2589-7500(21)00208-9

[11] McNamara, S.L., Yi, P.H., Lotter, W.: The clinician-ai interface: intended use and explainability in fda-cleared ai devices for medical image interpretation. npj Digital Medicine **7**, 80 (2024) https://doi.org/10.1038/s41746-024-01080-1

[12] Chen, I.Y., *et al.*: Ethical machine learning in healthcare. Annual Review of Biomedical Data Science **4**, 123–144 (2021) https://doi.org/10.1146/annurev-biodatasci-092820-114757

[13] Jabbour, S., *et al.*: Measuring the impact of ai in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA **330**, 2277–2285 (2023) https://doi.org/10.1001/jama.2023.19582

[14] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**, 447–453

(2019) https://doi.org/10.1126/science.aax2342

[15] Seyyed-Kalantari, L., *et al.*: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**, 2176–2182 (2021) https://doi.org/10.1038/s41591-021-01595-0

[16] Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn't. BMJ **312**, 71–72 (1996) https://doi.org/10.1136/bmj.312.7023.71

[17] Guyatt, G.H., *et al.*: Grade: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ **336**, 924–926 (2008) https://doi.org/10.1136/bmj.39489.470347.AD

[18] Chen, H., Gomez, C., Huang, C.-M., Unberath, M.: Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. npj Digital Medicine **5**, 156 (2022) https://doi.org/10.1038/s41746-022-00699-2

[19] Jin, W., Li, X., Fatehi, M., Hamarneh, G.: Guidelines and evaluation of clinical explainable ai in medical image analysis. Medical Image Analysis **84**, 102684 (2023) https://doi.org/10.1016/j.media.2022.102684

[20] Borys, K., *et al.*: Explainable ai in medical imaging: an overview for clinical practitioners—beyond saliency-based xai approaches. European Journal of Radiology **162**, 110786 (2023) https://doi.org/10.1016/j.ejrad.2023.110786

[21] Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., Viergever, M.A.: Explainable artificial intelligence in medical imaging: a survey. Medical Image Analysis **79**, 102470 (2022) https://doi.org/10.1016/j.media.2022.102470

[22] Samek, W., Wiegand, T., Müller, K.-R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. IEEE Signal Processing Magazine **34**, 80–105 (2017) https://doi.org/10.1109/MSP.2017.2746515

[23] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Transactions on Neural Networks and Learning Systems **32**, 4793–4813 (2021) https://doi.org/10.1109/TNNLS.2020.3027314

[24] Selvaraju, R.R., *et al.*: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74

[25] Lundberg, S.M., *et al.*: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**, 56–67 (2020) https://doi.org/10.1038/s42256-019-0138-9

[26] Bach, S., *et al.*: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**, 0130140 (2015) https://doi.org/10.1371/journal.pone.0130140

[27] Achtibat, R., *et al.*: From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence **5**, 1006–1019 (2023) https://doi.org/10.1038/s42256-023-00711-8

[28] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

[29] Adebayo, J., *et al.*: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, vol. 31, pp. 9505–9515 (2018). https://doi.org/10.48550/arXiv.1810.03292

[30] Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: towards a unified image-text pretraining paradigm for diverse medical imaging modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 153–163. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_15

[31] Zhang, K., *et al.*: Biomedgpt: a generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine **30**, 3129–3141 (2024) https://doi.org/10.1038/s41591-024-03185-2

[32] Zhang, S., *et al.*: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17513–17524 (2023). https://doi.org/10.1109/CVPR52729.2023.01751

[33] Lu, M.Y., *et al.*: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024) https://doi.org/10.1038/s41591-024-02856-4

[34] Oikarinen, T., Weng, T.-W.: Clip-dissect: automatic description of neuron representations in deep vision networks. In: Proceedings of the International Conference on Learning Representations (2023). https://doi.org/10.48550/arXiv.2204.10965

[35] Stavros, A.T., *et al.*: Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology **196**, 123–134 (1995) https://doi.org/10.1148/radiology.196.1.7784555

[36] Costantini, M., *et al.*: Characterization of solid breast masses: use of the sonographic breast imaging reporting and data system lexicon. Journal of Ultrasound

in Medicine **25**, 649–659 (2006) https://doi.org/10.7863/jum.2006.25.5.649

[37] Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, vol. 97, pp. 6105–6114 (2019). https://doi.org/10.48550/arXiv.1905.11946

[38] 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). https://doi.org/10.1109/CVPR.2016.90

[39] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (2015). https://doi.org/10.48550/arXiv.1409.1556

[40] Dosovitskiy, A., *et al.*: An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (2021). https://doi.org/10.48550/arXiv.2010.11929

[41] 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) https://doi.org/10.1038/s41597-024-02984-z

[42] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020) https://doi.org/10.1016/j.dib.2019.104863

[43] Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics **23**(2), 538–546 (2019) https://doi.org/10.1109/JBHI.2018.2824327

[44] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) https://doi.org/10.48550/arXiv.2502.13923 arXiv:2502.13923 [cs.CV]

[45] Oikarinen, T., Iyer, R.A., Bai, N., Weng, T.-W.: Linear explanations for individual neurons. arXiv preprint arXiv:2405.06855 (2024) https://doi.org/10.48550/arXiv.2405.06855 arXiv:2405.06855 [cs.LG]

[46] Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept Bottleneck Models (2020). https://arxiv.org/abs/2007.04612

[47] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks.

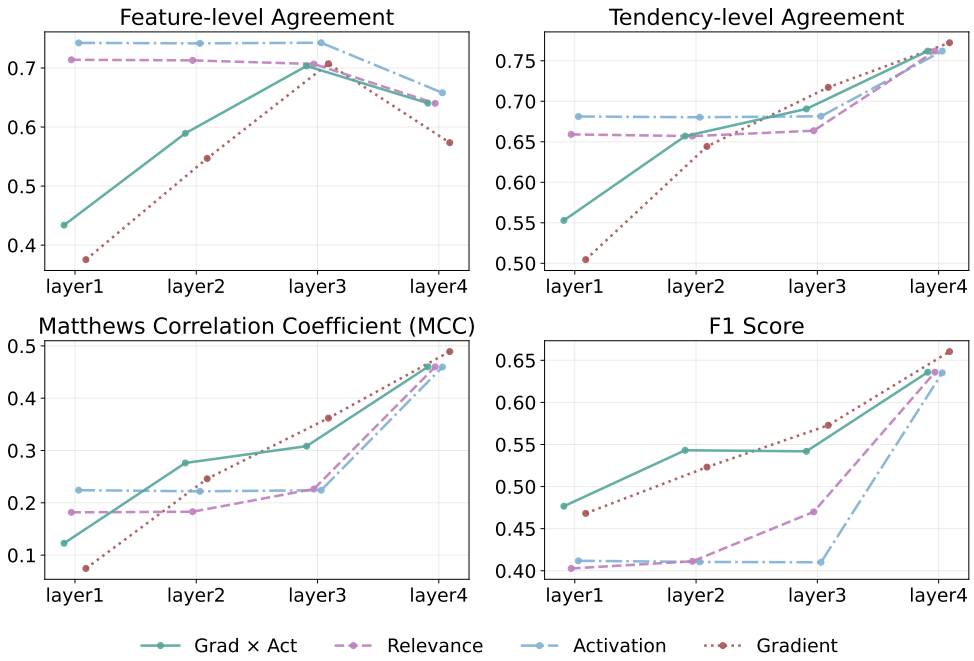


Fig. Extended Data 1 Effects of relevance signal and target layer on case-level evidence reconstruction. CRysTal aggregates neuron-level concept profiles into case-level evidence using an output-specific relevance signal. To assess the sensitivity of this aggregation step, we varied both the neuron relevance signal and the target layer from which neuron responses were extracted. Performance is shown across four ResNet-50 layer blocks (*layer1*–*layer4*) and four relevance signals: Activation, Gradient, Relevance, and Grad×Act. The four panels report feature-level agreement, diagnostic-tendency agreement, Matthews correlation coefficient (MCC) and F1 score, respectively. Earlier and intermediate layers tend to yield stronger feature-level recovery, suggesting that lower-level representations preserve more localized visual attributes. In contrast, the penultimate block (*layer4*) provides the most balanced downstream diagnostic performance across tendency-level metrics, indicating that deeper representations are better organized around decision-relevant semantic evidence. Pure Gradient is consistently weaker than the other relevance signals, whereas Activation, Relevance, and Grad×Act remain competitive at *layer4*. These results support the use of the penultimate target layer in the main experiments and show that CRysTal is relatively robust to the choice of relevance signal among the stronger alternatives.

In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR, ??? (2017)

Extended Data

Table Extended Data Table 1 Comparison of explanation methods for X2A evaluation across model architectures and external clinical datasets. We report feature-level semantic recovery, diagnostic-tendency agreement, and diagnostic-tendency prediction performance for each explanation method across breast ultrasound model architectures, Derm7pt, and thyroid ultrasound datasets. Values are shown as mean ± standard deviation when available; otherwise, point estimates are reported. Higher values indicate better performance across all metrics.

Dataset / Backbone	Method	Feature Acc ↑	Tendency Acc ↑	F1 ↑	MCC ↑
Breast(EfficientNet-B0)	CoT	0.639 ± 0.083	0.633 ± 0.066	0.488 ± 0.079	0.203 ± 0.086
	CLIP-Dissect	0.226	0.420	0.308	-0.170
	CLIP-Dissect*	0.344	0.508	0.391	-0.005
	LE	0.505 ± 0.040	0.637 ± 0.040	0.519 ± 0.106	0.245 ± 0.133
	CRysTal	**0.717 ± 0.017**	**0.727 ± 0.007**	**0.621 ± 0.006**	**0.410 ± 0.011**
Breast(ResNet-50)	CoT	0.640 ± 0.043	0.640 ± 0.091	0.511 ± 0.083	0.230 ± 0.082
	CLIP-Dissect	0.191	0.376	0.410	-0.062
	CLIP-Dissect*	0.554	0.622	0.574	0.370
	LE	0.459 ± 0.075	0.532 ± 0.050	0.512 ± 0.066	0.164 ± 0.123
	CRysTal	**0.680 ± 0.035**	**0.751 ± 0.002**	**0.672 ± 0.002**	**0.481 ± 0.003**
Breast(VGG-19)	CoT	0.650 ± 0.046	0.645 ± 0.067	0.521 ± 0.089	0.244 ± 0.063
	CLIP-Dissect	0.445	0.616	0.353	0.085
	CLIP-Dissect*	0.210	0.467	0.355	-0.079
	LE	0.527 ± 0.062	0.621 ± 0.080	0.559 ± 0.069	0.282 ± 0.124
	CRysTal	**0.673 ± 0.027**	**0.769 ± 0.006**	**0.655 ± 0.014**	**0.482 ± 0.016**
Breast(ViT-Base)	CoT	0.640 ± 0.077	0.645 ± 0.085	0.514 ± 0.053	0.238 ± 0.067
	CLIP-Dissect	0.264	0.419	0.381	-0.107
	CLIP-Dissect*	0.313	0.527	0.485	0.111
	LE	0.674 ± 0.046	0.752 ± 0.035	0.619 ± 0.055	0.437 ± 0.078
	CRysTal	**0.697 ± 0.001**	**0.778 ± 0.001**	**0.659 ± 0.002**	**0.496 ± 0.003**
Thyroid	CoT	0.564 ± 0.083	0.535 ± 0.066	0.691 ± 0.079	-0.231 ± 0.086
	CLIP-Dissect	0.465	0.500	0.578	-0.035
	CLIP-Dissect*	0.519	0.519	0.580	0.068
	LE	0.394 ± 0.098	0.438 ± 0.107	0.537 ± 0.134	-0.190 ± 0.113
	CRysTal	**0.591 ± 0.029**	**0.656 ± 0.032**	**0.753 ± 0.029**	**0.189 ± 0.053**
Derm7pt	CoT	0.501 ± 0.120	0.662 ± 0.073	0.423 ± 0.032	0.230 ± 0.023
	CLIP-Dissect	0.343	0.545	0.271	-0.045
	CLIP-Dissect*	0.322	0.606	0.365	0.096
	LE	0.442 ± 0.048	0.641 ± 0.035	0.473 ± 0.056	0.239 ± 0.081
	CRysTal	**0.538 ± 0.011**	**0.703 ± 0.013**	**0.520 ± 0.006**	**0.300 ± 0.011**

Table Extended Data Table 2 Aggregate neuron-level interpretability metrics across backbones. We report backbone-level averages of Ability, Alignment, Coverage, Diversity and Sparsity for CLIP-Dissect, Linear Explanation (LE) and CRysTal.

Backbone	Method	Ability	Alignment	Coverage	Diversity	Sparsity
Breast(ResNet-50)	CLIP-Dissect	0.038	0.126	0.137	0.137	**1.000**
	LE	0.276	**0.430**	0.276	0.370	0.199
	CRysTal	**0.665**	0.363	**0.836**	**0.804**	0.350
Breast(VGG-19)	CLIP-Dissect	0.037	0.206	0.139	0.139	**0.998**
	LE	0.230	**0.461**	0.216	0.280	0.281
	CRysTal	**0.536**	0.355	**0.819**	**0.769**	0.247
Breast(EfficientNet-B0)	CLIP-Dissect	0.046	0.072	0.133	0.134	**0.998**
	LE	0.166	**0.156**	0.180	0.240	0.536
	CRysTal	**0.679**	0.141	**0.806**	**0.742**	0.267
Breast(ViT-B)	CLIP-Dissect	0.042	0.190	0.138	0.138	**1.000**
	LE	0.294	**0.526**	0.268	0.372	0.235
	CRysTal	**0.563**	0.467	**0.750**	**0.753**	0.419

Table Extended Data Table 3 Physician ratings of CRysTal-generated explanations in the prospective explanation assessment study. Twenty ultrasound physicians, including 10 junior residents and 10 senior experts, evaluated CRysTal-generated explanations from 100 ultrasound images. Explanations were scored from 0 to 100 in terms of explanation accuracy, readability and guideline concordance. Higher scores indicate better perceived clinical quality.

Reader group	Explanation accuracy ↑	Readability ↑	Guideline concordance ↑	Overall ↑
Junior	97.26	97.04	97.45	97.25
Senior	97.35	97.12	97.52	97.30
Overall	97.31	97.08	97.48	97.29

Table Extended Data Table 4 External generalization of diagnostic predictors and explanation methods on breast ultrasound datasets. Diagnostic AUC is reported for diagnostic models, whereas feature-level agreement, tendency-level agreement and RCF-AUC evaluate the quality and reliability of generated explanations. Higher values indicate better performance across all metrics.

Dataset	Method	Diagnostic AUC ↑	Feature-level agreement ↑	Tendency-level agreement ↑	RCF-AUC ↑
BrEaST	ResNet-50 backbone	0.849	–	–	–
	CLIP-Dissect	–	0.379	0.469	0.344
	CLIP-Dissect*	–	0.344	0.448	0.463
	Linear Explanation	–	0.320	0.493	0.605
	CBM	0.689	**0.565**	0.623	**0.674**
	CRysTal	–	0.522	**0.728**	**0.674**
BUSI	ResNet-50 backbone	0.813	–	–	–
	CLIP-Dissect	–	–	–	0.695
	CLIP-Dissect*	–	–	–	0.814
	Linear Explanation	–	–	–	0.852
	CBM	0.608	–	–	0.818
	CRysTal	–	–	–	**0.867**

Supplementary Methods

Case-level explanation evaluation

Case-level explanation performance was evaluated by comparing the clinical concepts recovered by each explanation method with the corresponding reference annotations or clinical reports. For each case, both recovered and reference concepts were organized according to the same clinical concept taxonomy. In the breast ultrasound experiments, this taxonomy was based on BI-RADS-related categories, including shape, margin, orientation, echo pattern, posterior acoustic features and calcification. For thyroid ultrasound and dermoscopy, the same evaluation framework was applied using task-specific terminology systems, including TI-RADS features and the dermoscopic 7-point checklist.

Let $\mathcal{G}$ denote the set of clinical concept categories and let $\mathcal{C}_g$ denote the set of candidate concepts belonging to category $g \in \mathcal{G}$. For case i and concept c, the reference binary annotation is denoted by $z_{i,c} \in \{0, 1\}$, where $z_{i,c} = 1$ indicates that concept c is present in the reference annotation or clinical report. The explanation method assigns a score $s_{i,c}$ to each concept c.

Feature-level agreement

Feature-level agreement was evaluated as a category-wise top-1 hit rate. For case i and category g, the reference positive concept set was defined as

$$\mathcal{R}_{i,g} = \{c \in \mathcal{C}_g : z_{i,c} = 1\}.$$

Concepts within category g were ranked according to the explanation scores $s_{i,c}$, and the highest-scoring concept was selected:

$$\hat{c}_{i,g} = \arg\max_{c \in \mathcal{C}_g} s_{i,c}.$$

For case–category pairs with at least one reference positive concept, feature recovery was defined as

$$h_{i,g} = \mathbb{I}\left[\hat{c}_{i,g} \in \mathcal{R}_{i,g}\right].$$

Case–category pairs without any positive reference concept were excluded from feature-level agreement calculation.

Micro feature-level agreement was calculated by averaging the hit indicator over all evaluable case–category pairs:

$$\text{FeatureAgreement}_{\text{micro}} = \frac{1}{|\mathcal{D}_{\text{eval}}|} \sum_{(i,g) \in \mathcal{D}_{\text{eval}}} h_{i,g},$$

where $\mathcal{D}_{\text{eval}}$ denotes the set of evaluable case–category pairs.

Macro feature-level agreement was calculated by first averaging within each category and then averaging across categories:

$$\text{FeatureAgreement}_{\text{macro}} = \frac{1}{|\mathcal{G}_{\text{eval}}|} \sum_{g \in \mathcal{G}_{\text{eval}}} \frac{1}{|\mathcal{D}_g|} \sum_{i:(i,g)\in\mathcal{D}_g} h_{i,g},$$

where $\mathcal{D}_g$ denotes the evaluable cases for category g, and $\mathcal{G}_{\text{eval}}$ denotes the set of categories with at least one evaluable case.

Tendency-level agreement

Tendency-level agreement evaluated whether the recovered concepts were directionally consistent with the reference clinical evidence. Each clinical concept was assigned to one of three tendency labels: benign-associated, malignant-associated or uncertain. Let $\mathcal{C}^{\text{ben}}$ and $\mathcal{C}^{\text{mal}}$ denote the sets of benign-associated and malignant-associated concepts, respectively.

For case i and category g, the reference positive concept set $\mathcal{R}_{i,g}$ was mapped to a reference tendency:

$$T_{i,g}^{\text{ref}} = \begin{cases} \text{benign}, & \mathcal{R}_{i,g} \cap \mathcal{C}^{\text{ben}} \neq \emptyset \text{ and } \mathcal{R}_{i,g} \cap \mathcal{C}^{\text{mal}} = \emptyset, \\ \text{malignant}, & \mathcal{R}_{i,g} \cap \mathcal{C}^{\text{mal}} \neq \emptyset \text{ and } \mathcal{R}_{i,g} \cap \mathcal{C}^{\text{ben}} = \emptyset, \\ \text{uncertain}, & \text{otherwise.} \end{cases}$$

The predicted tendency was derived from the top-1 recovered concept $\hat{c}_{i,g}$:

$$T_{i,g}^{\text{pred}} = \begin{cases} \text{benign}, & \hat{c}_{i,g} \in \mathcal{C}^{\text{ben}}, \\ \text{malignant}, & \hat{c}_{i,g} \in \mathcal{C}^{\text{mal}}, \\ \text{uncertain}, & \text{otherwise.} \end{cases}$$

Only case–category pairs with benign or malignant reference tendency were included in tendency-level accuracy calculation. A binary reference label and predicted label were then defined as

$$y_{i,g} = \mathbb{I}\left[T_{i,g}^{\text{ref}} = \text{malignant}\right], \qquad \hat{y}_{i,g} = \mathbb{I}\left[T_{i,g}^{\text{pred}} = \text{malignant}\right].$$

Micro tendency-level agreement was calculated by averaging $\mathbb{I}[y_{i,g} = \hat{y}_{i,g}]$ over all valid case–category pairs. Macro tendency-level agreement was calculated by first averaging tendency agreement within each category and then averaging across categories.

For AUC-based tendency evaluation, a continuous tendency score was computed as

$$r_{i,g} = \max_{c \in \mathcal{C}_g \cap \mathcal{C}^{\text{mal}}} s_{i,c} - \max_{c \in \mathcal{C}_g \cap \mathcal{C}^{\text{ben}}} s_{i,c}.$$

If a category had no malignant-associated or benign-associated concepts, the corresponding maximum term was set to zero. Higher values of $r_{i,g}$ indicate stronger

malignant-oriented evidence. Tendency-level AUC was computed using $r_{i,g}$ and the binary reference tendency label $y_{i,g}$. F1 score, balanced accuracy and average precision were also computed for repeatability analyses where applicable.

Neuron-level explanation quality metrics

Neuron-level explanation quality was evaluated using five complementary metrics: Ability, Alignment, Coverage, Diversity and Sparsity. These metrics were designed to assess whether an explanation preserved diagnostic information, matched neuronal activation patterns, covered multiple clinical categories, avoided overly narrow concept assignments and remained compact.

For neuron u, an explanation method produced a concept-weight vector

$$\mathbf{w}_u = \{w_{u,c}\}_{c \in \mathcal{C}},$$

where $\mathcal{C}$ denotes the full clinical concept vocabulary. Let

$$\mathbf{P} \in \mathbb{R}^{N \times C}$$

denote the concept annotation matrix across N validation samples, and let

$$\mathbf{a}_u \in \mathbb{R}^N$$

denote the activation vector of neuron u across the same samples.

Ability

Ability measured whether the recovered concepts preserved malignant-oriented discriminative information carried by the neuron. For each diagnostic category g, the concept scores of neuron u were aggregated as

$$s_{u,g} = \sum_{c \in \mathcal{C}_g} w_{u,c} \mathbf{P}_{:,c}.$$

For each category, we constructed a binary malignant-concept label

$$\mathbf{y}_g^{\text{mal}} \in \{0,1\}^N,$$

where a sample was assigned 1 if any malignant-associated concept in category g was present and 0 otherwise. Category-level ability was calculated using the area under the receiver operating characteristic curve:

$$\text{Ability}_{u,g} = 2 \left| \text{AUC}\left(s_{u,g}, \mathbf{y}_g^{\text{mal}}\right) - 0.5 \right|.$$

Categories without both positive and negative samples were excluded. The final Ability score for neuron u was calculated as

$$\text{Ability}_u = \frac{1}{|\mathcal{G}_u^{\text{valid}}|} \sum_{g \in \mathcal{G}_u^{\text{valid}}} \text{Ability}_{u,g},$$

where $\mathcal{G}_u^{\text{valid}}$ denotes the valid categories for which AUC could be computed. The transformation $2|\text{AUC} - 0.5|$ maps random discrimination to 0 and perfect or perfectly reversed discrimination to 1.

Alignment

Alignment measured whether the recovered concepts could reconstruct the neuron's activation pattern. A concept-based predicted activation vector was computed as

$$\hat{\mathbf{a}}_u = \sum_{c \in \mathcal{C}} w_{u,c} \mathbf{P}_{:,c}.$$

Alignment was defined as the absolute Pearson correlation between the predicted and observed neuron activation vectors:

$$\text{Alignment}_u = \left|\text{corr}\left(\hat{\mathbf{a}}_u, \mathbf{a}_u\right)\right|.$$

If either vector had zero variance, the Alignment score was treated as not evaluable for that neuron.

Coverage

Coverage quantified how many diagnostic categories were represented by the recovered concept profile. For each category g, positive concept weights were summed as

$$q_{u,g} = \sum_{c \in \mathcal{C}_g} \max(w_{u,c}, 0).$$

The category weights were normalized by the largest category weight for the neuron. A category was considered covered if its normalized weight exceeded the threshold $\tau_{\text{cov}} = 0.05$:

$$\text{Coverage}_u = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{I}\left[\frac{q_{u,g}}{\max_{g' \in \mathcal{G}} q_{u,g'} + \varepsilon} \geq \tau_{\text{cov}}\right],$$

where ε is a small constant for numerical stability. A higher Coverage score indicates that the explanation spans more clinical feature categories.

Diversity

Diversity measured whether the explanation contained concepts from multiple diagnostic categories among both dominant and non-dominant concept groups. First, the

top-K concepts were selected by ranking concepts according to the absolute value of their weights, with $K = 10$. The top-concept diversity term was calculated as

$$D_u^{\text{top}} = \frac{|\{g(c) : c \in \text{TopK}(|w_{u,c}|), K = 10\}|}{|\mathcal{G}|},$$

where $g(c)$ denotes the diagnostic category of concept c.

We also computed a tail category-coverage term using the normalized positive category weights:

$$D_u^{\text{tail}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{I}\left[\frac{q_{u,g}}{\max_{g' \in \mathcal{G}} q_{u,g'} + \varepsilon} \geq \tau_{\text{tail}}\right],$$

where $\tau_{\text{tail}} = 0.05$. The final Diversity score was defined as

$$\text{Diversity}_u = \alpha D_u^{\text{top}} + (1 - \alpha) D_u^{\text{tail}},$$

with $\alpha = 0.7$. For CRysTal, Diversity was computed using unscaled greedy concept weights to reflect the breadth of concepts identified before category-level scaling. For other methods, the method-specific concept weights were used directly.

Sparsity

Sparsity measured whether neuronal semantics could be represented using a compact set of concepts. We used Hoyer sparsity calculated from the absolute concept-weight vector after removing near-zero weights:

$$\mathbf{v}_u = \{|w_{u,c}| : |w_{u,c}| \geq \tau_{\text{sp}}\},$$

where $\tau_{\text{sp}} = 0.01$, and $n_u = |\mathbf{v}_u|$. Sparsity was defined as

$$\text{Sparsity}_u = \frac{\sqrt{n_u} - \frac{\|\mathbf{v}_u\|_1}{\|\mathbf{v}_u\|_2}}{\sqrt{n_u} - 1}.$$

If only one concept remained after thresholding, Sparsity was set to 1; if no concept remained, Sparsity was set to 0. All five neuron-level quality metrics were calculated for each analysed neuron and then averaged within each backbone.

Repeated-run reliability metrics

To assess the robustness of neuron-level explanations to repeated runs, explanation generation was repeated ten times for the same selected feature layer of each backbone model. For each method and each run, concepts were ranked according to the absolute value of their explanation weights, and the top-K concepts were extracted for each neuron.

Let

$$\mathcal{E}_u^{(r,K)}$$

denote the set of top-K concepts recovered for neuron u in repeated run r, where $r = 1, \ldots, R$ and $R = 10$.

Val Corr

Val Corr measured the validation correlation reported by each explanation run. All available validation-correlation values were collected across neurons and repeated runs for each method. The mean, standard deviation, standard error and bootstrap 95% confidence interval were then computed across the collected values.

Consistency@K

Consistency@K measured the stability of top-K concept explanations across repeated runs. For each neuron, Consistency@K was calculated as the mean pairwise Jaccard similarity across all pairs of repeated runs:

$$\text{Consistency@}K(u) = \frac{2}{R(R-1)} \sum_{1 \le r < s \le R} \frac{\left| \mathcal{E}_u^{(r,K)} \cap \mathcal{E}_u^{(s,K)} \right|}{\left| \mathcal{E}_u^{(r,K)} \cup \mathcal{E}_u^{(s,K)} \right|}.$$

The final Consistency@K score was obtained by averaging over neurons shared by all repeated runs.

Redundancy@K

Redundancy@K measured semantic repetition within the top-K concepts. Each concept was mapped to its corresponding clinical concept group. For neuron u in run r, let

$$n_{u,r,g}^{(K)}$$

denote the number of top-K concepts assigned to group g. Group-level redundancy was calculated as

$$\text{Redundancy@}K(u, r) = \frac{\sum_{g \in \mathcal{G}} \max\left(0, n_{u,r,g}^{(K)} - 1\right)}{K}.$$

The final Redundancy@K score was averaged over repeated runs and neurons. When a concept co-occurrence matrix was available, an optional co-occurrence-based redundancy score was also calculated as the mean pairwise co-occurrence value among the top-K concepts; the group-level redundancy score was used for the main analysis.

Conflict@K

Conflict@K measured whether the top-K concepts contained mutually exclusive or clinically contradictory concept pairs. A predefined set of mutually exclusive concept pairs was denoted by

$$\mathcal{P}_{\text{excl}}.$$

For each neuron and run, Conflict@K was assigned 1 if at least one exclusive pair appeared in the top-K concepts and 0 otherwise:

$$\text{Conflict@}K(u,r) = \mathbb{I}\left[\exists (c_i, c_j) \in \mathcal{P}_{\text{excl}},\ c_i \in \mathcal{E}_u^{(r,K)},\ c_j \in \mathcal{E}_u^{(r,K)}\right].$$

The final Conflict@K score was averaged over repeated runs and neurons. The mean number of conflicting pairs per neuron was also recorded as an auxiliary statistic, but the binary conflict rate was used for the main comparison.

Summary statistics

For each method and each value of K, neuron-level reliability scores were summarized as mean, standard deviation, standard error and bootstrap 95% confidence interval. Bootstrap confidence intervals were estimated by resampling neuron-level metric values with replacement.

Risk-aware verification

For risk-aware verification, CRysTal and the concept bottleneck model used Reasoning Conflict (RCF) as the case-level risk score. RCF quantified the proportion of reconstructed clinical evidence whose diagnostic tendency opposed the final model prediction.

For each diagnostic category g, the dominant concept in the case-level evidence vector was identified as

$$j_g(x_i) = \arg \max_{j \in \{1,\dots,|\mathcal{C}_g|\}} E_{i,j}^{(g)},$$

with the corresponding evidence magnitude

$$v_{i,g} = \left| E_{i,j_g(x_i)}^{(g)} \right|.$$

Each concept was assigned a predefined diagnostic tendency

$$\pi(g,j) \in \{-1, 0, +1\},$$

where -1 denotes benign-associated evidence, 0 denotes neutral or unspecified evidence and $+1$ denotes malignant-associated evidence. The diagnostic direction of the model prediction was defined as

$$t(\hat{y}_i) = 2\hat{y}_i - 1.$$

RCF was calculated as

$$\text{RCF}(x_i) = \frac{\sum_{g=1}^{G} v_{i,g} \cdot \mathbb{I}\left[\pi(g, j_g(x_i)) = -t(\hat{y}_i)\right]}{\sum_{g=1}^{G} v_{i,g} + \varepsilon}.$$

A higher RCF indicates stronger conflict between the reconstructed clinical evidence and the final model prediction.

For visual attribution methods, including Grad-CAM, integrated gradients, SHAP, layer-wise relevance propagation and concept relevance propagation, attribution maps were compared with ground-truth lesion masks using intersection over union (IoU). The ability of RCF or IoU-based risk signals to distinguish correct from incorrect AI predictions was evaluated using ROC-AUC.