DIAGNOSE WHAT YOU DESCRIBE: ALIGNING VISUAL EVIDENCE WITH MAMMOGRAPHY DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable predictions require agreement between outputs and supporting evidence, which is essential in high-stakes settings. However, recent Vision–Language Models often produce predictions that are not supported by the evidence they present. This issue becomes critical in medical imaging, where such inconsistencies can undermine trust in automated systems. In mammography diagnosis, for example, a model may predict malignancy while describing features that are typically associated with benign findings. To address this problem, we propose a Knowlwdge-Driven Reasoning (KDR) framework. The model generates multiple candidate predictions for each image, each paired with its supporting evidence. We evaluate each candidate by checking whether the evidence is consistent with the predicted diagnosis, which allows us to identify and discard unsupported outputs. The final prediction is selected based on this consistency criterion. A metric to measure agreement between evidence and diagnosis is introduces, which complements standard accuracy. Experiments on multiple public datasets show that the proposed framework improves classification performance and consistency.

Keywords: Vision–Language Models, Knowledge-Driven, Reasoning, Consistency-Aware

1 INTRODUCTION

Reliable decision-making requires that predictions are supported by corresponding evidence. However, recent Vision–Language Models often violate this requirement. They can produce predictions that are not consistent with the evidence they present, which leads to unreliable outputs even when predictions are correct. This issue becomes critical in high-stakes domains such as medical imaging, where inconsistencies between evidence and diagnosis can undermine trust in automated systems. In mammography, diagnostic decisions depend on subtle visual patterns and their alignment with established clinical criteria. A model may predict malignancy while describing features that are typically associated with benign findings, which creates a mismatch between evidence and diagnosis. Such inconsistencies are not addressed by standard training objectives, which focus on prediction accuracy rather than the agreement between evidence and conclusions. As a result, models may achieve strong performance while still producing unreliable reasoning. Radiological diagnosis is a complex cognitive task, which requires the integration of bottom-up perceptual signals with top-down semantic schemas Li et al. (2023a); Tu et al. (2024); Thirunavukarasu et al. (2023). Breast cancer is one of the most widely occurred cancer. Mammographic interpretation is a complex cognitive task that involves integrating perceptual information with clinical knowledge and experience Alamudun et al. (2016); Amirkhani et al. (2018); Mirzaei-Alavijeh et al. (2018). Radiologists must discriminate subtle abnormalities from normal anatomical variations and imaging artefacts, often under high workload and time constraints. Research has shown that diagnostic performance in mammography is influenced by a range of perceptual and cognitive factors, including the ability to recognize clinically relevant features, visual search strategies across multiple views, and decision-making under uncertainty Qiao et al. (2018); Azevedo & Lajoie (2022). Studies of expert and novice radiologists indicate that extensive experience and training markedly improve the interpretive consistency, highlighting the critical role of cognitive skill development in mammography Purtzer (2012); Tolma et al. (2003). Despite expert experience, breast imaging is prone to diagnostic errors and variability, with a considerable proportion of lesions missed or misinterpreted, highlighting the fundamental limitations of human perception and decision-making Branch et al. (2022); Tan et al. (2007). Awareness

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

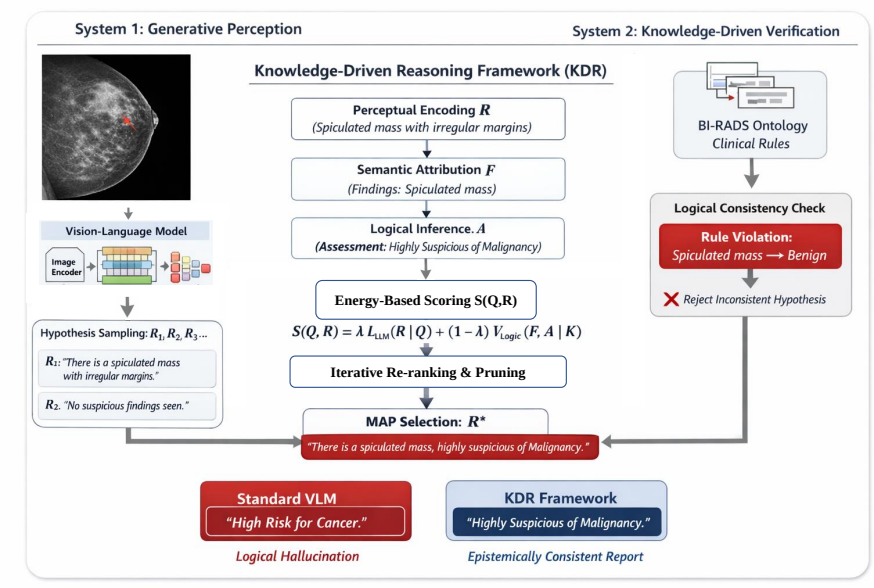

Figure 1: Overview of the Knowledge-Driven Reasoning (KDR) framework.

of these human factors has spurred work in medical image perception to characterize how radiologists gather and integrate visual information during mammogram interpretation, including efforts to relate eye-tracking metrics to diagnostic choices and to analyze search patterns across views Petersen et al. (2011); Warnecke et al. (1997). Concurrently, developments in machine learning and interpretable AI reveal how models can enhance radiologists' interpretations by encoding both fine-grained image features and broader structural context, reinforcing the necessity of linking perceptual processing with sophisticated clinical reasoning Reani et al. (2018); Liu et al. (2021). As imaging technology evolves, algorithmic development must be informed by the cognitive mechanisms of expert mammographic interpretation, guaranteeing that AI tools support clinicians' decision-making rather than interfere with it. Mammogram interpretation is a decision-focused process, which aims to judge whether the image contains signs of disease rather than produce a complete description of what is visible. Mammographic patterns are recognized by experienced radiologists through established clinical categories defined in BI-RADS Balleyguier et al. (2007). Expert perception is shaped by prior knowledge, which allows visual evidence to be interpreted in light of diagnostic goals.

Despite the rapid advancement of large-scale Vision-Language Models (VLMs) in automated report generation Liu et al. (2023); Li et al. (2023b); Alayrac et al. (2022), these systems continue to exhibit a notable lack of cognitive fidelity, particularly in high-stakes clinical scenarios. A critical failure mode arises when a model's final clinical assessment contradicts the visual evidence it describes Zhai et al. (2022). For example, a model may label a lesion as highly malignant while simultaneously describing imaging characteristics that are commonly associated with benign findings. Such inconsistencies undermine the reliability of automated reports and pose significant risks for clinical decision-making. Human experts actively monitor their reasoning process, detect inconsistencies between observed evidence and diagnostic conclusions, and iteratively resolve them through top-down control and domain knowledge. In contrast, current vision–language models typically generate outputs in a feed-forward manner, without explicit mechanisms for self-consistency checking or conflict resolution He et al. (2022); Wang et al. (2023). As a result, they tend to prioritize local statistical correlations over the global coherence required for clinically grounded reasoning.

In this work, we propose Knowledge-Driven Reasoning (KDR), a cognitively inspired architecture that formulates radiological report generation as constrained inference under clinical knowledge, instead of an end-to-end black-box image-to-text mapping. Rather than committing to a single decoding path, KDR maintains multiple candidate reasoning trajectories and evaluates their consistency from visual evidence to clinical assessment. Our framework is grounded in Dual-Process Theory,

which is shown in Figure 1. It first generates multiple candidate diagnostic hypotheses from visual evidence through stochastic sampling, capturing the uncertainty inherent in mammographic interpretation. These candidates are not treated as final outputs, but as provisional hypotheses that are subject to further evaluation. To regulate hypothesis generation, we employ a knowledge-driven verification stage that performs structured, rule-based evaluation rather than token-level decoding. Candidate trajectories are mapped to BI-RADS–aligned representations and checked for consistency under a formal clinical ontology. This process filters out hypotheses that violate diagnostic constraints, such as malignancy claims without high-risk visual support. The interaction between hypothesis generation and verification is implemented via a winner-take-all selection over the candidate trajectories. Each trajectory is assigned an energy score that combines its generative plausibility with its consistency under domain knowledge. The final report is selected as the trajectory that maximizes this joint criterion, balancing generative plausibility with consistency under clinical knowledge. This architecture introduces an explicit control mechanism that constrains neural generation using structured domain knowledge.

2 Related Work

2.1 Vision-Language Models (VLMs)

VLMs have shown strong capability in multimodal understanding by jointly modeling visual and textual information Liu et al. (2023); Li et al. (2023b); Alayrac et al. (2022); He et al. (2022); Wang et al. (2023). Built on large-scale pretraining, these models can connect image regions with textual descriptions and generate coherent outputs across a range of tasks. In medical imaging, they have been used for report generation and disease classification, where the model produces both a prediction and a corresponding description. Despite these advances, existing methods are still largely optimized for prediction accuracy or text quality. The generated descriptions are often treated as auxiliary outputs, rather than as evidence that must support the final decision. A model may assign a malignant label while describing features that are more consistent with benign findings. This limitation becomes more pronounced in high-stakes scenarios such as medical diagnosis, where the validity of a prediction depends not only on whether it is correct, but also on whether it is supported by observable evidence. VLMs may rely on spurious correlations or partial cues, which produce outputs that appear reasonable but are not well grounded in the image.

2.2 Clinic Reasoning

Improving reasoning ability in large models has attracted increasing attention Wei et al. (2022); Kojima et al. (2022). Rather than producing predictions in a single pass, recent approaches encourage models to generate intermediate reasoning steps that make the decision process more transparent Liu et al. (2021). Techniques such as Chain-of-Thought prompting guide the model to articulate step-by-step reasoning, while self-consistency decoding samples multiple reasoning paths and aggregates their outputs to improve robustness Wei et al. (2022); Yao et al. (2023). These methods have shown clear benefits in complex reasoning tasks, where a single forward pass is often insufficient. In practice, however, these reasoning processes are still largely driven by the model's internal representations, without explicit constraints on how evidence should support the final decision. Despite these advances, most existing methods focus on consistency across reasoning trajectories, such as agreement among multiple sampled outputs, rather than on the alignment between evidence and conclusions Kojima et al. (2022); Reani et al. (2018). As a result, a model may generate internally coherent reasoning steps while still relying on weak, incomplete, or even irrelevant evidence. This limitation becomes particularly important in medical imaging, where diagnostic decisions must be grounded in observable features.

2.3 Reliable Diagnosis

Reliability is a central concern in medical AI, where incorrect or unsupported predictions can have serious consequences Branch et al. (2022); Moor et al. (2023); Warnecke et al. (1997). In clinical settings, decisions are rarely based on predictions alone. They are typically supported by observable evidence, clinical context, and established diagnostic criteria, which together form the basis for reliable judgment. When a model produces an output that is not well supported by the underlying

image, it becomes difficult to assess whether the prediction can be trusted in practice. Recent studies have drawn attention to inconsistencies between model predictions and the evidence they provide, often described as right answers for the wrong reasons D'Orsi et al. (2018). A model may arrive at the correct diagnosis while relying on irrelevant cues or incomplete observations, which are not aligned with established clinical criteria Tan et al. (2007). Although such cases do not affect accuracy metrics directly, they raise concerns about robustness and generalization, especially when the data distribution shifts.

3 METHODOLOGY

The Knowledge-Driven Reasoning (KDR) framework is designed as a computational analogue of the executive control mechanisms employed by expert radiologists. Rather than framing diagnosis as a direct image-to-text mapping, KDR formulates radiological interpretation as a constrained search problem over an epistemic space structured by clinical knowledge. As illustrated in Figure 1, the framework explicitly separates hypothesis generation from hypothesis verification, reflecting the interaction between intuitive perception and deliberate reasoning in human cognition. Given a mammographic image Q, a Vision–Language Model (System 1) generates multiple candidate reasoning trajectories via stochastic sampling, where each trajectory consists of a predicted diagnosis and its textual explanation. Each trajectory is then converted into a structured representation. Specifically, the textual explanation is parsed into BI-RADS–aligned features F (e.g., shape, margin, density) and a provisional assessment A. This step makes the implicit evidence explicit and allows comparison across candidates. A knowledge-driven verification module (System 2) then evaluates each candidate by checking whether the extracted features F support the predicted assessment A under clinical rules derived from the BI-RADS ontology K. Candidates that violate these rules, such as assigning high malignancy without corresponding high-risk features, are rejected. The remaining candidates are re-ranked using an energy-based scoring function that combines the model likelihood of the trajectory with its consistency under the verification module. The final report R^* is selected as the candidate with the highest score.

Unlike end-to-end vision–language models that directly generate free-form reports from images, KDR interprets them through a structured, multi-stage reasoning process. Given a mammographic image Q, a generative perception module (System 1) produces a set of candidate reasoning trajectories $R = \{R^{(1)}, \ldots, R^{(k)}\}$ via stochastic sampling. Each trajectory represents a complete provisional interpretation, including visual descriptions, inferred clinical features, and an associated diagnostic assessment. Each candidate trajectory is subsequently grounded into structured clinical representations. Specifically, free-form textual descriptions are parsed into BI-RADS-aligned feature sets F and a provisional assessment A. A knowledge-driven verification module (System 2) evaluates each candidate against symbolic constraints derived from the BI-RADS ontology K. Trajectories that violate established clinical logic are explicitly rejected, corresponding to the elimination of logically inconsistent or hallucinatory interpretations. The remaining candidates are re-ranked using an energy-based scoring function that jointly accounts for neural generative likelihood and epistemic consistency. The final output R^* is selected via maximum a posteriori inference, yielding a report that is clinically consistent and logically justified.

3.1 GENERATIVE PERCEPTION

Inspired by the analysis-by-synthesis paradigm, we view radiological interpretation as generating possible explanatory hypotheses and systematically assessing them. Given an image Q, the system generates a reasoning trajectory R that explains the observed visual evidence. To ensure interpretability and logical grounding, each trajectory is decomposed into three representational layers, $i.e.$, perceptual description, structured feature attribution, and diagnostic assessment.

Formally, the diagnostic process is modeled as:

$$P(A, F, R \mid Q) = \underbrace{P(A \mid F, K)}_{\text{Inference } \iota} \cdot \underbrace{P(F \mid R)}_{\text{Attribution } \sigma} \cdot \underbrace{P(R \mid Q; M_\theta)}_{\text{Perception } \omega}, \tag{1}$$

where $P(R \mid Q; M_\theta)$ captures the likelihood that System 1 generates a narrative interpretation given the visual input. The attribution term $P(F \mid R)$ maps unstructured language into a discrete BI-RADS feature space $F \in \mathcal{F}_{\text{BI-RADS}}$ via deterministic parsing. $P(A \mid F, K)$ represents the normative

diagnostic likelihood governed by the clinical knowledge base K. This factorization explicitly separates perceptual uncertainty from clinical inference, enabling top-down regulation of the diagnostic process.

3.2 KNOWLEDGE-DRIVEN VERIFICATION

To operationalize top-down control, we define a verification score $S(Q, R)$ that evaluates each reasoning trajectory according to both neural plausibility and clinical validity. This score formalizes how System 2 regulates the outputs of System 1 by enforcing goal-directed constraints.

$$S(Q, R) = \lambda \cdot \mathcal{L}_{\text{LLM}}(R \mid Q) + (1 - \lambda) \cdot \mathcal{V}_{\text{Logic}}(F, A \mid K), \tag{2}$$

where $\mathcal{L}_{\text{LLM}} = \frac{1}{|R|} \log P(R \mid Q)$ is the normalized generative log-likelihood. The epistemic consistency term is defined as:

$$\mathcal{V}_{\text{Logic}}(F, A \mid K) = \sum_{i=1}^{M} w_i \cdot \Psi(\phi_i, A, K), \tag{3}$$

where w_i denotes the diagnostic saliency of feature ϕ_i. The consistency kernel Ψ is defined as:

$$\Psi(\phi_i, A, K) = \begin{cases} \log \sigma(\text{dist}(\phi_i, A)) & \text{if Logic}(\phi_i \to A) \vdash K, \\ -\infty & \text{otherwise.} \end{cases} \tag{4}$$

Assigning an infinite penalty to logical violations ensures that highly fluent but clinically impossible trajectories are decisively eliminated.

3.3 LOGICAL HALLUCINATION

To characterize failure modes of unconstrained vision-language models, we define a logical hallucination as a divergence between the model's inferred diagnosis and the normative clinical inference implied by extracted evidence.

$$\mathcal{H} \iff \mathbb{E}_{P(F,A|M_\theta)} \left[\log \frac{P(A \mid F, M_\theta)}{P(A \mid F, K)} \right] > \epsilon. \tag{5}$$

Minimizing this divergence forces the model's posterior beliefs to align with expert clinical priors, transforming unconstrained stochastic generation into knowledge-consistent reasoning.

3.4 ITERATIVE HYPOTHESIS REFINEMENT

The KDR algorithm performs a stochastic search over the hypothesis space, acknowledging that initial generative outputs may include implausible or inconsistent interpretations. Our evaluation protocol is designed to probe fundamentally the degree of cognitive alignment between model reasoning and expert radiological schemas. We hypothesize that imposing explicit knowledge-driven constraints will improve accuracy and reshape the internal reasoning distribution toward epistemic coherence and clinical plausibility.

4 EXPERIMENTS

4.1 DATASETS

We evaluate KDR across five heterogeneous mammography datasets, selected to reflect variation in imaging modality, population demographics, annotation granularity, and diagnostic difficulty.

CBIS-DDSM Sawyer-Lee et al. (2016) is a curated subset of DDSM containing digitized film mammograms with lesion-level annotations. While widely used, it exhibits moderate noise and relatively coarse BI-RADS descriptors.

Table 1: Diagnostic Performance and Reasoning Consistency across datasets.

Dataset	Qwen2.5-VL		Human-CoT		KDR (Ours)	
	ACC	RCS	ACC	RCS	ACC	RCS
CBIS-DDSM	0.681	0.452	0.714	0.588	**0.792**	**0.844**
INbreast	0.754	0.510	0.772	0.631	**0.841**	**0.892**
VinDr-Mammo	0.712	0.495	0.748	0.602	**0.825**	**0.887**
CMMD	0.705	0.477	0.729	0.594	**0.803**	**0.856**
MIAS	0.648	0.421	0.672	0.553	**0.754**	**0.821**

INbreast Moreira et al. (2012) provides full-field digital mammograms with high-quality expert annotations, often serving as an upper bound for perceptual performance.

VinDr-Mammo Nguyen et al. (2023) is a large-scale, multi-institutional dataset characterized by scanner variability and population shift, making it a strong testbed for robustness.

CMMD Cui et al. (2021) focuses on an Asian cohort and introduces systematic shifts in breast density distribution.

MIAS Suckling (1994) is an earlier dataset with lower spatial resolution and increased perceptual ambiguity, serving as a stress test for reasoning stability.

Across all datasets, we adopt standardized train-validation-test splits and report results averaged over five runs.

4.2 Metrics for Cognitive Alignment

Traditional NLP metrics (e.g., BLEU, ROUGE, CIDEr) prioritize lexical overlap and are insufficient for evaluating medical reasoning, where logical validity and feature-label consistency are paramount. We therefore adopt three complementary metrics:

- **Accuracy (ACC):** Macro-averaged F1-score for binary malignancy classification.
- **Reasoning Consistency Score (RCS):** Measures whether generated evidence F logically entails the assessment A under the BI-RADS ontology K:

$$RCS = \frac{1}{N}\sum_{i=1}^{N} \nVdash[\text{Logic}(F_i, A_i) \vdash K]. \tag{6}$$

- **Clinical Information Gain (CIG):** Importance-weighted recall of diagnostically critical descriptors (e.g., margin, shape, density) relative to expert annotations.

Specifically, RCS decouples *process validity* from outcome correctness, allowing us to detect cases where predictions are correct for logically incorrect reasons.

4.3 Diagnostic Performance and Reasoning Consistency

To evaluate the logical robustness of the proposed KDR framework, we conduct extensive benchmarking against Qwen2.5-VL and a Human-Chain-of-Thought (Human-CoT) prompting strategy. The evaluation spans five heterogeneous mammography environments: CBIS-DDSM, INbreast, VinDr-Mammo, CMMD, and MIAS. As summarized in Table 1, KDR consistently outperforms other methods across both classification accuracy (ACC) and epistemic alignment (RCS). A salient empirical pattern emerges: while Qwen2.5-VL can achieve competitive surface-level accuracy, they suffer from a profound reasoning-gap. For instance, on the INbreast dataset, Qwen2.5-VL attains a 75.4% ACC but an RCS of merely 0.510. This suggests that a significant portion of the model's correct predictions are effectively right for the wrong reasons, stemming from logically inconsistent evidentiary paths. Although Human-CoT improves consistency by providing structured prompts, its gains remain limited by the model's tendency to drift into stochastic hallucinations during long-form generation.In contrast, KDR simultaneously elevates ACC and RCS, indicating that its performance gains are rooted in epistemically valid reasoning rather than the exploitation of spurious correlations.

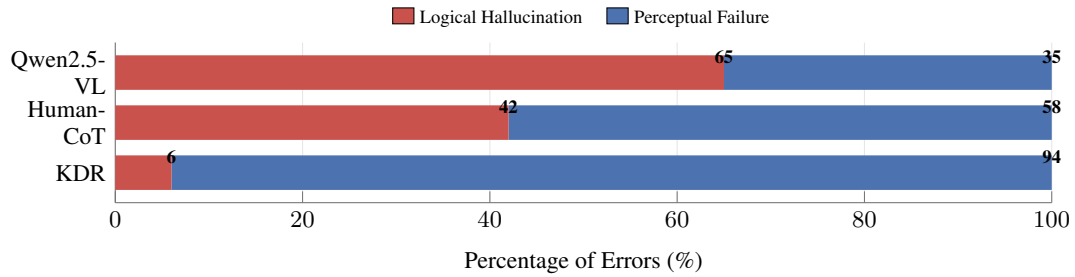

Figure 2: **Error-type distribution across models.** Qwen2.5-VL produces more logical hallucinations, while the remaining errors of KDR are primarily perceptual.

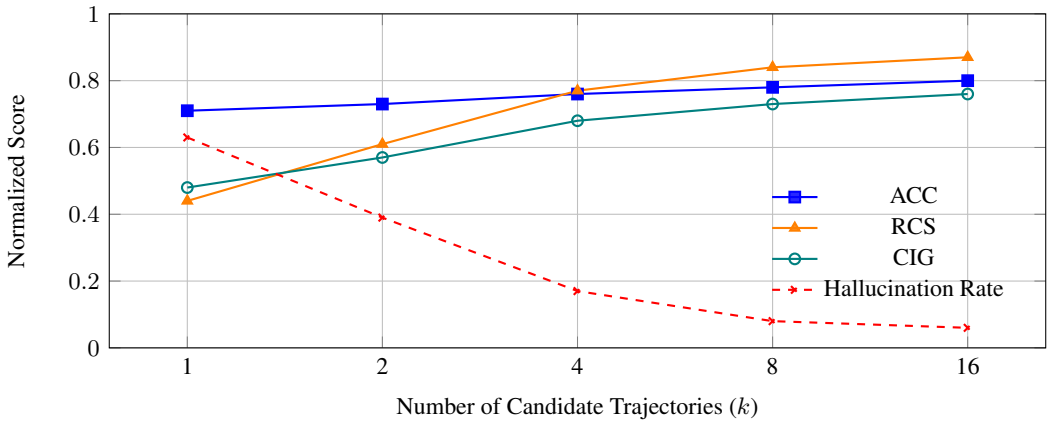

Figure 3: **Effect of deliberation depth k.** The KDR framework exhibits a robust *inference-time scaling* effect. Increasing k facilitates the transition from stochastic decoding to knowledge-aligned reasoning, significantly boosting RCS while suppressing logical hallucinations.

This stabilizing effect is most pronounced in datasets characterized by significant population shifts and varying acquisition protocols (e.g., VinDr-Mammo and CMMD). By imposing the BI-RADS axiomatic manifold as a top-down constraint, KDR effectively filters out noise-induced trajectories, ensuring that the model remains within the bounds of clinical plausibility. These results underscore the efficacy of KDR as a domain-agnostic reasoning layer that bridges the gap between neural perception and symbolic medical logic.

4.4 Logical vs. Perceptual Failure

To better understand model failures, we manually audited 100 error cases and categorized them into: (1) Logical Hallucinations, where evidence contradicts the diagnosis; and (2) Perceptual Failures, where visual cues are misidentified.

Figure 2 shows that Qwen2.5-VL is dominated by logical hallucinations, reflecting internal epistemic dissonance. KDR nearly eliminates this failure mode, shifting remaining errors toward perceptual limitations. From a clinical trust perspective, perceptual misses reflect sensor uncertainty, whereas logical hallucinations undermine diagnostic integrity.

4.5 Scaling with Deliberation

The efficacy of the KDR framework is intrinsically governed by the depth of the deliberation process, parameterized by the number of candidate trajectories k. As illustrated in Figure 3, we observe a distinct **Inference-time Scaling** effect: while classification accuracy (ACC) shows a steady but

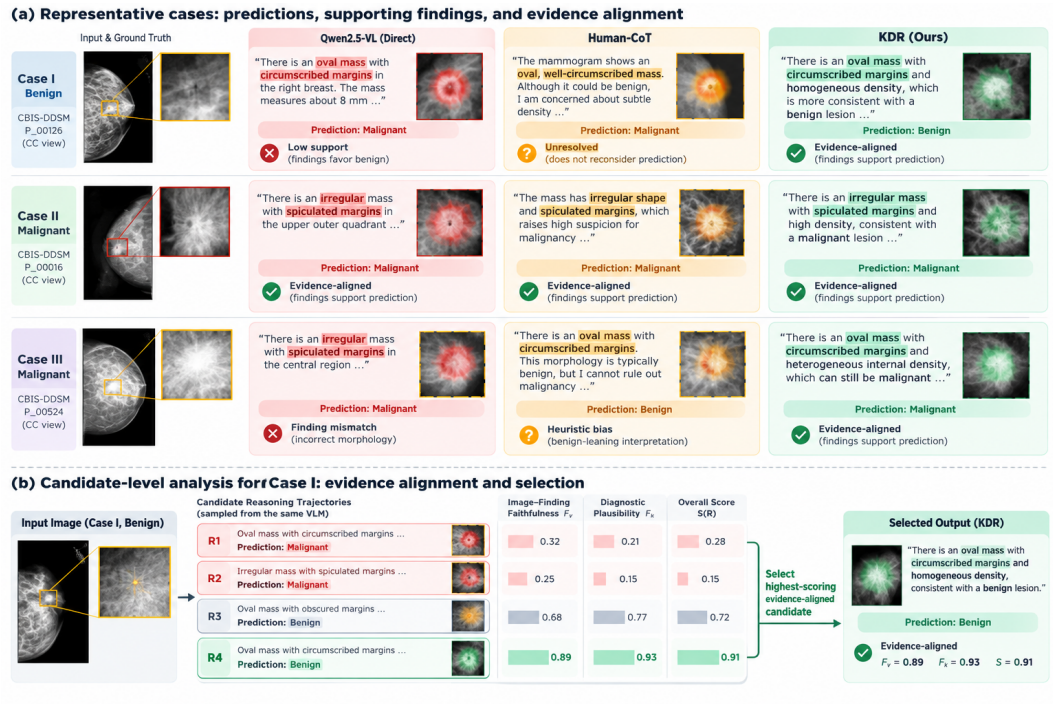

Figure 4: **Qualitative evidence–diagnosis alignment analysis.** **(a)** Representative CBIS-DDSM cases comparing direct Qwen2.5-VL decoding, Human-CoT prompting, and KDR. Each output is decomposed into its predicted imaging findings and diagnostic prediction. KDR retains the observed morphology while selecting the correct malignant prediction. **(b)** Candidate-level analysis of Case I. KDR jointly evaluates diagnostic plausibility and image–finding faithfulness rather than treating individual BI-RADS descriptors as deterministic rules. The final report is selected from the candidate with the highest evidence-alignment score. Green, red, and orange indicate evidence-aligned, evidence-mismatched, and unresolved outputs, respectively.

marginal ascent, the Reasoning Consistency Score (RCS) exhibits a sharp monotonic leap before reaching an asymptote.

This logarithmic scaling behavior reveals that the Qwen2.5-VL's latent distribution inherently encapsulates high-fidelity diagnostic rationales, which are nonetheless obscured by the stochastic noise of greedy decoding or local probability traps at $k = 1$. By augmenting the inference budget k, our System 2 operates as a knowledge-driven filter, effectively re-aligning the model's trajectory with the BI-RADS axiomatic manifold. Crucially, the precipitous decline in the Hallucination Rate, coupled with the surge in RCS, suggests that deliberation optimizes **epistemic alignment** rather than mere linguistic fluency. These results reinforce the central thesis: the path toward reliable radiological AI lies not merely in parameter expansion, but in the strategic allocation of computational-logical budget during inference to prune spurious correlations.

4.6 Ablation analysis of the KDR framework components

The ablation results presented in Table 2 empirically validate the structural necessity of each component within the KDR framework. Specifically, we observe that reducing the deliberation depth to a single-pass inference ($k = 1$) leads to a catastrophic decay in the Reasoning Consistency Score (RCS) from 0.887 to 0.441. While Qwen2.5-VL possesses latent diagnostic knowledge, they lack the internal regulatory mechanisms to ensure logical coherence without explicit symbolic intervention.Furthermore, the performance gap between the full KDR and the uniform weighting configuration ($w_j = 1$) underscores the importance of clinically-informed priors. In expert radiological practice, diagnostic features are not stochastically equivalent; certain visual primitives (e.g., spiculated

Table 2: Ablation analysis of the KDR framework components on VinDr-Mammo.

Configuration	ACC ↑	RCS ↑	CIG ↑
Full KDR (**Ours**)	**0.825**	**0.887**	**0.784**
Uniform Weights ($w_j = 1$)	0.784	0.763	0.691
No Trade-off ($\lambda = 1$)	0.712	0.495	0.512
No Deliberation ($k = 1$)	0.692	0.441	0.485

Note: ACC = Accuracy; RCS = Reasoning Consistency Score; CIG = Clinical Information Gain.

margins) carry significantly higher pathognomonic weight than others. By relaxing these weights, the model loses its ability to prioritize high-stakes evidence, resulting in sub-optimal alignment with the BI-RADS ontology. The consistent superiority of the "Full KDR" configuration demonstrates that the synergy between probabilistic neural generation and axiomatic symbolic weighting is essential for achieving expert-level cognitive fidelity.

4.7 Qualitative Evidence–Diagnosis Audit

Figure 4 examines whether KDR improves the alignment between the predicted diagnosis and its supporting imaging findings, rather than merely changing the final prediction. For each case, we separately report the findings identified by the model, its diagnostic prediction, and their agreement with the image and ground truth. In Case I, direct decoding and Human-CoT both predict malignancy while describing an oval mass with circumscribed margins. These descriptors do not exclude malignancy, but provide weaker support for it than for the selected benign hypothesis. KDR jointly considers diagnostic plausibility and finding faithfulness, selecting the highest-scoring evidence-aligned candidate.

Case II demonstrates that KDR preserves a correct prediction when the reported irregular shape and spiculated margins faithfully support the malignant diagnosis. More importantly, Case III distinguishes diagnostic correctness from rationale faithfulness. Although direct decoding predicts malignancy correctly, it incorrectly describes the lesion as irregular and spiculated. Human-CoT faithfully recognizes the oval, circumscribed morphology but incorrectly treats these benign-leaning descriptors as determinative. KDR instead preserves the observed findings while selecting the correct malignant diagnosis. These cases show that KDR does not enforce a brittle one-to-one mapping between individual BI-RADS descriptors and pathology; it ranks candidates according to their combined diagnostic plausibility and evidence faithfulness.

5 Conclusion

This paper studies the problem that model predictions are not always supported by the evidence described in medical diagnosis. Such mismatches can undermine the reliability of automated systems, especially in high-stakes settings. A Knowledge-Driven Reasoning framework is introduced to select predictions based on their consistency with supporting evidence. Instead of relying on a single output, the model considers multiple candidates and filters out those not well supported by their own descriptions. Experimental results show that enforcing evidence–diagnosis agreement improves both accuracy and reasoning reliability. The approach reduces cases where correct predictions are made for inconsistent reasons, which are difficult to detect using standard metrics. These findings suggest that consistency provides a useful signal beyond accuracy and can be used to guide model decisions in a practical way. This work highlights the value of checking whether predictions are justified by the evidence. Incorporating such checks at inference time offers a lightweight way to improve the reliability of medical AI systems without modifying the underlying models.

AI use statement

In this work, we used generative AI tools for writing assistance. We did not use generative AI tools to produce research ideas, experimental designs, proofs or any reported measurement.

Ethics statement

This work does not collect new patient data or involve prospective clinical intervention. The proposed framework is intended for research evaluation and should not be interpreted as a medical device or a substitute for radiologist judgment.

Reproducibility statement

The code is available at https://anonymous.4open.science/r/ICLR-KDR-DF33.

References

Folami Alamudun, Hong-Jun Yoon, Tracy Hammond, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi. Shapelet analysis of pupil dilation for modeling visuo-cognitive behavior in screening mammography. In *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment*, volume 9787, pp. 160–172. SPIE, 2016.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Abdollah Amirkhani, Mojtaba Kolahdoozi, Elpiniki I Papageorgiou, and Mohammad R Mosavi. Classifying mammography images by using fuzzy cognitive maps and a new segmentation algorithm. In *Advanced data analytics in health*, pp. 99–116. Springer, 2018.

Roger Azevedo and Susanne P Lajoie. The cognitive basis for the design of a mammography interpretation tutor. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 78–83. Routledge, 2022.

Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal. Birads™ classification in mammography. *European journal of radiology*, 61(2):192–194, 2007.

Fallon Branch, Isabella Santana, and Jay Hegdé. Biasing influence of 'mental shortcuts' on diagnostic decision-making: radiologists can overlook breast cancer in mammograms when prior diagnostic information is available. *Diagnostics*, 12(1):105, 2022.

Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, and Jinghua Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast. *The Cancer Imaging Archive*, 1, 2021.

Carl D'Orsi, L Bassett, and S Feig. Breast imaging reporting and data system (bi-rads). *Oxford University Press, New York*, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu. Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5947–5961, 2021.

Mehdi Mirzaei-Alavijeh, Parvaneh Ghorbani, and Farzad Jalilian. Socio-cognitive determinants of the mammography screening uptake among iranian women. *Asian Pacific journal of cancer prevention: APJCP*, 19(5):1351, 2018.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

Suni Petersen, Jennifer Hydeman, and Kristin Flowers. The decisional processing model: How cognitive processing affects adherence to mammography among african american women. *Journal of Black Psychology*, 37(3):357–380, 2011.

Mary A Purtzer. A transformative decision-making process for mammography screening among rural, low-income women. *Public health nursing*, 29(3):247–255, 2012.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Manuele Reani, Alan Davies, Niels Peek, and Caroline Jay. How do people use information presentation to make decisions in bayesian reasoning tasks? *International Journal of Human-Computer Studies*, 111:62–77, 2018.

Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). *(No Title)*, 2016.

John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, volume 1069, pp. 375–378, 1994.

Tuan Zea Tan, Chai Quek, Geok See Ng, and Eddie Y-K Ng. A novel cognitive interpretation of breast cancer thermography with complementary learning fuzzy neural memory structure. *Expert Systems with Applications*, 33(3):652–666, 2007.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Eleni L Tolma, Belinda M Reininger, John Ureda, and Alexandra Evans. Cognitive motivations associated with screening mammography in cyprus. *Preventive medicine*, 36(3):363–373, 2003.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14408–14419, 2023.

Richard B Warnecke, Seymour Sudman, Timothy P Johnson, Diane O'Rourke, Andrew M Davis, and Jared B Jobe. Cognitive aspects of recalling and reporting health-related events: Papanicolaou smears, clinical breast examinations, and mammograms. *American Journal of Epidemiology*, 146 (11):982–992, 1997.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

A Appendix

A.1 Knowledge-Driven Methods

Incorporating structured knowledge into machine learning systems has been widely studied as a way to improve interpretability and robustness. Rather than relying purely on data-driven correlations, these approaches aim to guide model behavior using prior knowledge about the domain. In medical applications, such knowledge is often expressed through clinical ontologies, diagnostic rules, or expert-defined criteria, which provide a more transparent link between observed features and final decisions. In the context of mammography, BI-RADS offers a standardized schema that connects visual findings, such as mass shape or margin, to diagnostic categories. This structured representation plays an important role in clinical practice, as it allows radiologists to reason from observable evidence to diagnostic assessment in a consistent manner. Most existing approaches integrate knowledge during training, which often requires careful design and limits flexibility when adapting to new models or datasets. Post-hoc verification provides an alternative perspective, where model outputs are evaluated against external knowledge without modifying the underlying model. This allows constraints to be applied in a modular way, making it easier to extend across tasks and domains.

A.2 Qualitative Reasoning Trajectory Analysis

Figure 5 illustrates the decisional geometry of the KDR framework. In Qwen2.5-VL, the inference process is often biased toward the High-Probability/Low-Consistency regime (red dot), where neural fluency overshadows symbolic rigor, leading to logical hallucinations. Conversely, our KDR framework performs a constrained search in the latent reasoning space, navigating toward the High-Consistency manifold (blue dot). By optimizing for the joint objective of neural likelihood and clinical axiom adherence, KDR identifies trajectories that are not merely statistically frequent, but epistemically sound.

A.3 Additional Discussion about the Results

Across datasets, metrics, and ablations, the Knowledge-Driven Reasoning (KDR) framework consistently reshapes performance and failure modes, which suggests that the limitations of current vision–language models stem less from insufficient capacity than from a lack of executive control aligned with expert schemas. From a cognitive science perspective, these findings provide a concrete instantiation of dual-process theory in a high-stakes perceptual domain. KDR introduces an explicit System 2 analogue, in which multiple candidate reasoning trajectories are held in parallel and evaluated against the symbolic constraints of the BI-RADS ontology. The resulting gains in RCS indicate a qualitative shift toward cognitively interpretable inference. As shown in the error distribution analysis, unconstrained VLMs predominantly fail through logical hallucinations, which produces diagnoses that contradict their own visual descriptions. Perceptual errors reflect limitations

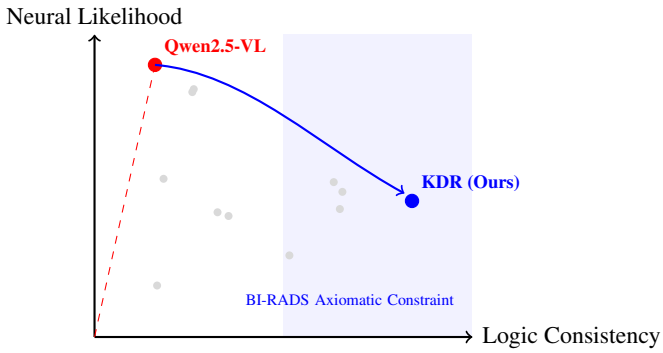

Figure 5: Visualization of reasoning trajectories. KDR selects hypotheses that maximize joint neural likelihood and symbolic validity.

of sensory evidence and are readily understood by human experts, whereas logical inconsistencies undermine the internal integrity of the diagnostic process. By enforcing symbolic coherence, KDR preserves the rational structure of the model's internal representations even under uncertainty.

These findings also bear on ongoing debates around explainability and safety in medical AI. Standard Chain-of-Thought approaches often rely on post-hoc rationalizations that may themselves drift from clinical logic. KDR instead operationalizes ante-hoc constraint, which ensures that the selected reasoning trajectory is logically valid by construction. This distinction is particularly important in radiology, where explanations must be defensible with respect to established reporting standards rather than surface-level plausibility. KDR also inherits the perceptual limitations of its underlying vision–language backbone. When System 1 fails to extract diagnostically relevant features, System 2 has no valid hypotheses to evaluate. This limitation parallels human expertise, where executive reasoning cannot compensate for absent or degraded sensory input. Future extensions can explore bidirectional architectures, in which logical inconsistencies detected by the verification layer trigger targeted re-inspection of the image.