# NeRD: Neuro-Symbolic Rule Distillation for Efficient Ontology-Grounded Chain-of-Thought in Medical Image Diagnosis

Hongxi Yang[1,2], Yiwen Jiang[2,3(✉)], Siyuan Yan[1,2], Jamie Chow[4], Eunis Li[4], Charlotte Poon[4], Stephanie Fong[1,2], Xiangyu Zhao[1,2], Deval Mehta[1,5], Yasmeen George[1,2], and Zongyuan Ge[1,2(✉)]

[1] Department of Data Science & AI, Faculty of Information Technology, Monash University, Melbourne, Australia
[2] AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia
[3] Faculty of Engineering, Monash University, Melbourne, Australia
[4] Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China
[5] School of Computing Technologies, RMIT University, Melbourne, Australia
yiwen.jiang@monash.edu, zongyuan.ge@monash.edu

**Abstract.** Interpretability is essential for trustworthy medical image diagnosis. However, existing concept-driven interpretable methods have key limitations: Concept Bottleneck Models (CBMs) require scoring all predefined concepts during both inference and manual intervention, imposing a substantial burden on clinicians, while rationale-based generative approaches often select concepts by class discriminability, which can drift from diagnostic ontologies. To address these issues, we propose **Neuro-Symbolic Rule Distillation (NeRD)**, a framework that produces efficient, ontology-grounded reasoning chains that are sufficient yet non-redundant, without manually crafting diagnostic rules. Experiments on three datasets demonstrate strong diagnostic performance and interpretability, and blinded expert evaluation confirms the clinical plausibility of NeRD rationales. Our method further enables a first intervention study for Multimodal Chain-of-Thought-based diagnosis, achieving efficient and effective concept-level intervention. Our code is available at: https://github.com/HongxiY/NeRD.

**Keywords:** Explainable AI · Chain-of-Thought · Concept · Skin Disease Diagnosis.

## 1 Introduction

Real-world clinical decision-making relies on inherently interpretable and structured reasoning, where visual evidence is evaluated and explicitly mapped to established diagnostic criteria [3, 12, 25]. Recently, deep learning (DL) models have shown strong potential in medical image diagnosis [21, 23, 24], yet their opaque nature limits clinical trust and safe adoption in practice [14].

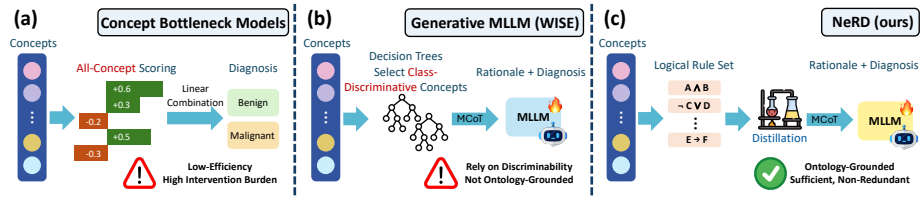


Fig. 1: Comparison of concept-driven interpretable methods. (a) CBM (discriminative) predicts the diagnosis via a linear combination of all concept scores. (b) WISE (generative) uses decision-tree guidance to select class-discriminative concepts and reformulate them into MCoTs. (c) Ours distills ontology-grounded diagnostic MCoTs from an induced logical rule set.

Existing interpretability methods for DL largely fall into two paradigms [11]: discriminative concept-based explanations such as Concept Bottleneck Models (CBMs) [7, 10], and generative rationale-based stepwise reasoning such as Multimodal Chain-of-Thought (MCoT) [27, 28] for Multimodal Large Language Models (MLLMs) [26]. Despite their potential, CBMs require evaluating all predefined concepts at inference time, whereas MCoT depends on substantial expert-authored rationales, both introducing computational and annotation burden. Bridging these paradigms while addressing limitations, WISE [8] is the first to automate MCoT generation from the bottleneck layer of CBMs. It leverages an ensemble of decision trees to selectively activate class-discriminative concepts, yielding concept-grounded rationales while avoiding the exhaustive, all-concept scoring pipeline of CBMs (Fig. 1(a, b)).

However, concept selection guided purely by class discriminability can diverge from established diagnostic ontologies, which leads to several limitations of WISE for medical image diagnosis. First, it may favor statistically predictive but clinically irrelevant shortcuts, selecting concepts that fall outside the ontology's criterion-based decision pathway and thereby bypassing the guideline-encoded diagnostic logic. Second, the generated MCoTs lack systematic validation by human experts, raising uncertainty about their alignment with clinical reasoning and limiting confidence in deployment. Third, CBMs natively support concept-level intervention, allowing clinicians to correct intermediate concept predictions and propagate these corrections to the final diagnosis [22]. However, comparable intervention in generative MLLM reasoning remains largely unexplored, leaving a critical gap for human-in-the-loop clinical workflows.

To address these limitations, we propose **Neuro-symbolic Rule Distillation (NeRD)**, a framework that produces efficient, ontology-grounded MCoT rationales for interpretable medical image diagnosis (Fig. 1(c)). Rather than prioritizing class discriminability, NeRD first induces explicit logical diagnostic rules over clinical concepts in a data-driven manner, without manual rule engineering. It then distills the activated rules into compact, case-specific reasoning chains through rule selection, logic simplification, and concept grounding, thereby removing redundant concepts and yielding concise, ontology-aligned ev-

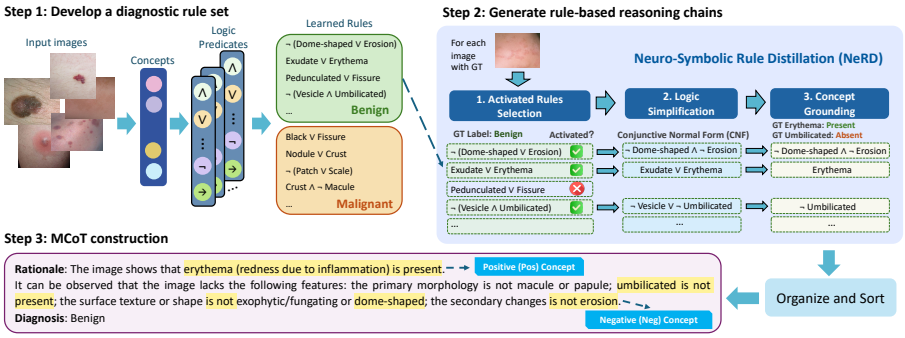


Fig. 2: Framework of our method. Step 1: Induce a diagnostic rule set from concepts. Step 2: Execute Neuro-Symbolic Rule Distillation (NeRD) to select, simplify, and ground rules. Step 3: Organize the reasoning chains into MCoT.

idence for each prediction. **Our contributions are threefold**: (1) We propose NeRD, a novel framework that bridges rule-based concept interpretability with MCoT reasoning for medical image diagnosis. (2) We design a blinded human expert study to assess the clinical plausibility, showing that NeRD rationales best support expert diagnostic reasoning. (3) We present the first study of concept intervention in MCoT-based diagnosis, demonstrating that NeRD enables both efficient and effective concept-level intervention.

## 2 Method

**Task Formulation.** Let $\mathcal{S} = \{(x_i, \mathbf{c}_i, y_i)\}_{i=1}^{N}$ denote a dataset, where $x_i$ is an input image, $y_i \in \mathcal{Y}$ is the diagnostic label, and $\mathbf{c}_i = [c_i^1, c_i^2, \ldots, c_i^K]$ represents annotations for $K$ concepts. Our goal is to derive, for each $x_i$, a compact textual MCoT that selects a small subset of concepts and verbalizes them into a reasoning chain to fine-tune MLLMs. As illustrated in Fig. 2, our framework operates in three steps. First, we begin by inducing a diagnostic rule set to capture the underlying logic between concept pairs. Next, NeRD distills these rules into compact, case-specific reasoning chains through rule activation, logic simplification, and concept grounding. Finally, the refined chains are structured into MCoTs.

### 2.1 Diagnostic Rule Set Construction

We induce a diagnostic rule set $\mathcal{R}$ using LogicCBM [18] in Step 1 of Fig. 2. LogicCBM learns differentiable logical gates (e.g., $\wedge$, $\vee$, $\neg$) to compose pairs of concept predicates into human-interpretable rules such as $r_j = c_i^a \wedge \neg c_i^b$, defined over $\mathbf{c}_i$. Each derived rule $r_j \in \mathcal{R}$ is associated with a class-wise weight vector. We assign $r_j$ to a diagnosis $y \in \mathcal{Y}$ based on these weights: if signs differ across classes, $r_j$ is assigned to the class with a positive weight; otherwise, it is assigned to the class with the largest weight. This yields a class-conditioned partition $\{\mathcal{R}^y\}_{y \in \mathcal{Y}}$, where each subset $\mathcal{R}^y$ collects rules attributed to diagnosis $y$.

### 2.2 Neuro-Symbolic Rule Distillation

**Activated Rule Selection.** In Step 2 of Fig. 2, for each training sample $(x_i, \mathbf{c}_i, y_i)$, we restrict candidate rules to the class-specific subset $\mathcal{R}^{y_i}$. We then perform rule-level selection by applying each $r_j \in \mathcal{R}^{y_i}$ to the concept annotation $\mathbf{c}_i$, and retain only the activated rules (i.e., those whose logical predicates are satisfied). This activation-based screening prunes rules that do not explain the current case, yielding a diagnostically aligned rule set for subsequent refinement.

**Logic Simplification.** Since activated rules in $\mathcal{R}^{y_i}$ are selected independently without cross-rule interaction, they may contain overlapping or redundant conditions. To obtain a compact representation, we combine all selected rules via conjunction and convert the resulting expression into Conjunctive Normal Form (CNF). Specifically, an selected rule $r_j$ is rewritten as a conjunction of disjunctive clauses, formulated as $r_j^{\text{CNF}} = \bigwedge_{l=1}^{L_j} D_{j,l}$, where $D_{j,l} = \bigvee_k \ell_k$. Here, each literal $\ell_k$ corresponds to either a positive concept assertion ($c^k$) or its negation ($\neg c^k$). This conversion utilizes standard logical equivalences, including De Morgan's laws [4]. After applying this transformation across all activated rules, we pool the resulting disjunctive clauses and eliminate duplicates, yielding a simplified, unique clause set $\mathcal{C}_i = \{D_1, D_2, \ldots, D_P\}$.

**Concept Grounding.** $\mathcal{C}_i$ may still contain residual disjunctions (uncertainties) that need to be resolved for the specific sample. To ground this logic, we evaluate every literal $\ell_k$ against the concept annotations $\mathbf{c}_i$. For each clause $D_l$, we extract only the literals that evaluate to True, forming a set of verified facts: $D'_l = \{\ell \in D_l \mid \ell(\mathbf{c}_i) = \text{True}\}$. For instance, given $D_l = \text{Exudate} \vee \text{Erythema}$, if Erythema is present and Exudate is absent in $\mathbf{c}_i$, the clause reduces to {Erythema}. We then aggregate these verified literals into a unified grounded set $\mathcal{G}_i = \bigcup_{l=1}^{P} D'_l$. At this stage, $\mathcal{G}_i$ consists purely of conjunctive assertions.

Additionally, some clinical concepts form mutually exclusive groups (e.g., pigment network $\in$ {absent, atypical, typical}). We leverage this structural prior to prune redundant negations in $\mathcal{G}_i$. Specifically, if $\mathcal{G}_i$ contains a positive assertion $c^v$ (e.g., typical) and a negation $\neg c^t$ (e.g., $\neg$atypical) belonging to the same group, $\neg c^t$ is safely discarded; the presence of $c^v$ implicitly guarantees the absence of all other states in that concept group. Conversely, if all concepts in a group except one are negated in $\mathcal{G}_i$ (e.g., both $\neg$typical and $\neg$atypical are present), we logically infer the remaining concept (i.e., absent) as a positive assertion and update the set accordingly.

### 2.3 MCoT Construction.

In Step 3 of Fig. 2, we construct MCoT rationale from the grounded set $\mathcal{G}_i$. We first partition $\mathcal{G}_i$ into supportive (positive) concepts $\mathcal{G}_i^+$ and refutational (negative) concepts $\mathcal{G}_i^-$. To produce a coherent and clinically intuitive narrative, concepts within each partition are prioritized based on their class-conditional

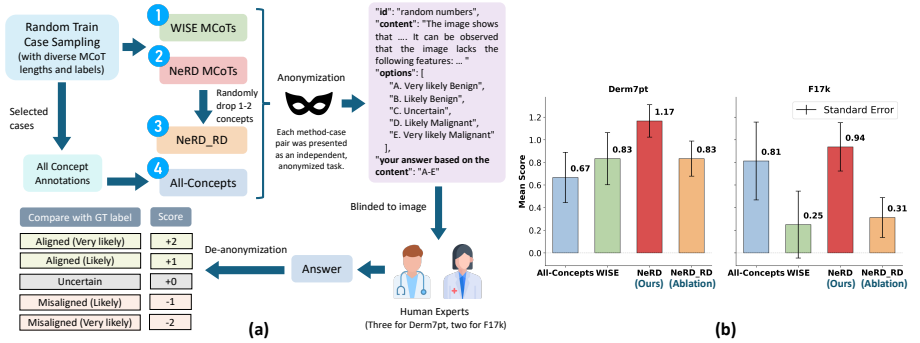


Fig. 3: Human expert evaluation overview. (a) Workflow of the human expert evaluation. We sample training cases and rationales from four methods. The rationales are anonymized and presented to experts, who are asked to select a diagnosis based solely on the rationale content. (b) Evaluation results. Mean scores (± standard error) across evaluators on Derm7pt and F17k. Higher scores indicate stronger alignment between expert judgments and diagnostic labels.

statistics estimated on the training set. Concretely, concepts in $\mathcal{G}_i^+$ are ranked by decreasing prevalence under the predicted label, whereas concepts in $\mathcal{G}_i^-$ are ranked by the label-conditional probability of being absent. To preserve structural cohesion, we further group concepts by clinical attribute category (e.g., primary morphology, texture) and order concepts within each group accordingly. Finally, we render the rationale with a fixed template: (i) report the salient positive findings, (ii) enumerate clinically relevant absent features, and (iii) conclude with the diagnostic prediction.

## 3 Experiments

### 3.1 Experimental Datasets

We evaluate NeRD on three publicly available datasets. Derm7pt [9] and F17k [5] are dermatology datasets sourced from MAKE [19, 20]. Derm7pt: seven categorical clinical concept groups with mutually exclusive values, yielding a 28-dim one-hot concept vector; we follow the official 413/203/395 train/val/test split for melanoma vs. non-melanoma classification. F17k: 32 independently annotated binary concepts; following [13] we cast the task as benign vs. malignant with a stratified 2253/322/644 split. PBC [1, 17]: white blood cell classification with 5 classes and 24 morphological attributes; to simulate a low-data scenario we train on 30% of the official training set (1030 images).

### 3.2 Experiment 1: Human Expert Evaluation of Clinical Plausibility

**Evaluation Method.** To assess the clinical plausibility of the reasoning chains, we conduct a blinded evaluation with dermatology-trained experts (three for

Derm7pt, two for F17k) experienced in concept-based skin lesion diagnosis. As illustrated in Fig. 3 (a), for each evaluator we randomly sample 8 training cases, spanning diverse diagnostic labels and chain lengths; each case yields 4 distinct text-only rationales from All-Concepts (verbalizing all annotated concepts), WISE [8], NeRD (ours), and NeRD-RD (an ablation that randomly removes 1-2 concepts). To prevent cueing, each rationale is detached from its originating case and method and presented as an independent item; all items are globally shuffled across cases and methods. Evaluators assign a 5-point ordinal likelihood from A (very likely benign/non-melanoma) to E (very likely malignant/melanoma). We score responses by agreement with the ground-truth label: correct predictions receive $+2$ (A/E) or $+1$ (B/D), incorrect predictions receive $-2$ or $-1$ with the same confidence mapping, and uncertain responses (C) receive 0. Scores are averaged over all items to obtain a mean score per method.

**Main Results.** Fig. 3(b) summarizes the results of the human expert study. NeRD attains the highest mean score on both datasets (Derm7pt: 1.17; F17k: 0.94), surpassing WISE (0.83; 0.25) and All-Concepts (0.67; 0.81). These results suggest that NeRD mitigates the redundancy in All-Concepts, where redundant or irrelevant details can dilute salient evidence, potentially increasing cognitive load and reducing diagnostic confidence. NeRD also selects concepts that are more clinically aligned than WISE. To further assess the contribution of NeRD's selection, we perform an ablation by randomly removing one to two concepts (NeRD-RD). This drives a marked performance drop (Derm7pt: 1.17 $\rightarrow$ 0.83; F17k: 0.94 $\rightarrow$ 0.31), confirming the importance and non-redundancy of NeRD's selected concepts for reaching correct conclusions without image access.

### 3.3 Experiment 2: Fine-Tuned MLLMs for Interpretable Diagnosis

**Baselines and Implementation Details.** We compare NeRD against three categories of baselines: (1) Black-box models. An InceptionV3 [16] model trained end-to-end from images to diagnostic labels, without concept supervision. (2) Discriminative concept-based models. Sequential CBM [10], Joint CBM [10], as well as LogicCBM [18] in both joint and sequential training variants, all using InceptionV3 as the backbone. (3) Generative reasoning models. Zero-shot prompting with Qwen3-VL-8B-Instruct [2] using the predefined concept set and task description to elicit concept-based rationales without any fine-tuning; and WISE [8], which generates MCoTs via decision-tree–guided reformulation. For MLLM fine-tuning, we adopt Qwen3-VL-8B-Instruct as the backbone and perform parameter-efficient adaptation with LoRA [6] via LLaMA-Factory [29]. Fine-tuning is conducted for 8 epochs on a single NVIDIA RTX 4090. All models are trained on the training split, selected by validation diagnostic accuracy, and evaluated on the test split.

**Evaluation Method.** We evaluate diagnostic performance using accuracy and F1-score. To measure interpretability, we compute concept accuracy over the

Table 1: Comparison of diagnostic and concept prediction performance. ACC (%): diagnostic accuracy; F1 (%): diagnostic F1-score; Intp. (%): interpretability, quantified by concept accuracy (at the bottleneck layer or along the MCoT).

| Method | Derm7pt | | | F17k | | | PBC | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | F1 | Intp. | ACC | F1 | Intp. | ACC | F1 | Intp. |
| *Black-box* | | | | | | | | | |
| InceptionV3 | **81.77** | 55.00 | - | 88.04 | 54.97 | - | **99.29** | **99.02** | - |
| *Discriminative concept-based* | | | | | | | | | |
| Joint CBM | 79.49 | 50.91 | 75.01 | 88.35 | 52.83 | 79.71 | 96.10 | 95.06 | 86.33 |
| Sequential CBM | 78.99 | 34.65 | **78.80** | 76.86 | 13.87 | 91.59 | 95.39 | 94.28 | 86.26 |
| Joint LogicCBM | 79.75 | <u>58.33</u> | 62.75 | 87.73 | 55.87 | 69.01 | 94.13 | 91.66 | <u>87.42</u> |
| Sequential LogicCBM | 77.97 | 54.45 | 64.76 | <u>88.51</u> | <u>57.95</u> | 71.77 | 96.35 | 94.81 | 86.98 |
| *Generative reasoning* | | | | | | | | | |
| Zero-shot | 77.22 | 32.84 | 56.78 | 84.63 | 34.44 | 88.62 | 35.43 | 30.45 | 36.67 |
| WISE | 77.72 | 43.59 | 57.06 | 86.65 | 56.12 | 85.90 | 96.13 | 94.60 | 85.87 |
| NeRD (Ours) | <u>80.76</u> | **60.00** | <u>76.76</u> | **88.66** | **61.70** | **94.10** | <u>96.64</u> | <u>95.40</u> | **92.81** |

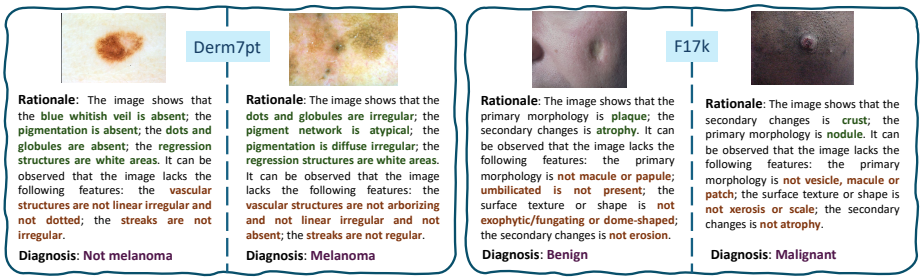


Fig. 4: NeRD generated MCoTs.

concepts explicitly mentioned in the generated MCoT. Since Zero-shot MCoT outputs are not emitted in a fixed schema, we use GPT-5 Mini [15] to parse the predicted concepts and diagnosis from free-form text for evaluation.

**Main Results.** Table 1 reports diagnostic performance and concept prediction accuracy on Derm7pt, F17k, and PBC datasets. Among generative methods, NeRD achieves the strongest overall results: 80.76% ACC / 60.00% F1 on Derm7pt, 88.66% ACC / 61.70% F1 on F17k, and 96.64% ACC / 95.40% F1 on PBC, outperforming WISE and the zero-shot baseline. On Derm7pt, NeRD attains the best ACC among interpretable methods and ranks second overall, trailing only the black-box InceptionV3; on F17k, it yields the highest ACC across all compared methods. For concept prediction, NeRD substantially improves concept accuracy over WISE on all three datasets (76.76% vs. 57.06% on Derm7pt; 94.10% vs. 85.90% on F17k; 92.81% vs. 85.87% on PBC). Fig. 4

Table 2: Average concept count. Pos / Neg: present / absent per case (CBM); supportive / refutational per MCoT (WISE and NeRD).

| Method | Derm7pt | | | F17k | | |
|---|---|---|---|---|---|---|
| | Pos | Neg | Total | Pos | Neg | Total |
| CBM | 7 | 21 | 28 | 3 | 29 | 32 |
| WISE | 3 | 1 | 4 | 2 | 11 | 13 |
| NeRD | 3 | 3 | 6 | 1 | 5 | 6 |

Table 3: Intervention results on Derm7pt. N: intervened samples; Concepts: average corrected concepts per sample; Before / After: diagnostic accuracy (%) pre / post intervention; Change: accuracy gain.

| Method | N | Concept | Before | After | Change |
|---|---|---|---|---|---|
| CBM | 82 | 6.56 | 78.99 | 83.04 | +4.05 |
| NeRD | 73 | 2.26 | 80.76 | 85.82 | +5.06 |

provides qualitative examples of the generated MCoTs. Table 2 further shows that NeRD uses 6 concepts per MCoT for both datasets, whereas CBM-style baselines evaluate all predefined concepts (28 for Derm7pt; 32 for F17k). This indicates that NeRD can preserve competitive interpretability while producing compact, non-exhaustive reasoning chains. Notably, on F17k, WISE yields more negative concepts per chain (11 vs. 5), consistent with its tendency to continue exploring additional discriminative cues when the current concept set fails to separate classes.

### 3.4 Experiment 3: Clinician Intervention on Reasoning Concepts

**Intervention Setup and Evaluation.** We evaluate the efficiency and efficacy of concept intervention on Derm7pt by comparing NeRD with a Sequential CBM [10]. For both methods, we intervene only on test cases where the model (i) predicts an incorrect diagnosis and (ii) makes at least one concept prediction error. This protocol mirrors a realistic workflow in which clinicians review and correct concept-level mistakes only for flagged, high-risk cases. For NeRD, we locate the incorrect concept statements in the generated MCoTs, replace them with the ground-truth concept annotations, and prepend the corrected rationale to prompt the model to regenerate the diagnosis. For the Sequential CBM, we overwrite the incorrectly predicted concept variables with their annotations and forward-propagate the corrected concept vector through the concept-to-label predictor. No parameters are updated during intervention. We report post-intervention diagnostic performance, along with the intervention cost measured by the number of corrected concepts.

**Main Results.** Table 3 summarizes intervention results on Derm7pt. Across 73 intervened cases, NeRD requires only 2.26 concept corrections per sample on average while improving accuracy by 5.06% (80.76%→85.82%). By comparison, Sequential CBM requires 6.56 corrections per sample, yet yields a smaller gain of 4.05% (78.99%→83.04%) across 82 intervened cases. Overall, NeRD demonstrates higher intervention-efficient refinement by enabling larger diagnostic improvements with substantially fewer concept edits.

## 4 Conclusion

We propose NeRD, a neuro-symbolic framework that bridges rule-based concept interpretability with MCoT reasoning for medical image diagnosis. By selecting concepts based on diagnostic necessity, NeRD distills induced logical rules into compact, criterion-aligned reasoning chains. Experiments on three image datasets demonstrate strong diagnostic performance and concept fidelity. Blinded expert evaluation further confirms the clinical plausibility of NeRD rationales, while intervention experiments demonstrate efficient concept-level correction with fewer edits and larger diagnostic gains.

**Acknowledgments.** This work was supported by the Australian Government under Grant CRCPXIII000022.

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

## References

1. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in Brief **30**, 105474 (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Croskerry, P.: Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education **14**(Suppl 1), 27–35 (2009)
4. De Morgan, A.: Formal logic: or, the calculus of inference, necessary and probable. Taylor and Walton (1847)
5. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1820–1828 (2021)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
7. Jiang, Y., Mehta, D., Feng, W., Ge, Z.: Enhancing interpretable image classification through LLM agents and conditional concept bottleneck models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12285–12297 (Jul 2025)
8. Jiang, Y., Mehta, D., Yan, S., Shen, Y., Wang, Z., Ge, Z.: WISE: Weak-supervision-guided step-by-step explanations for multimodal LLMs in image classification. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 14674–14685 (2025)
9. 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 (2018)

10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
11. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
12. Mehta, D., Jiang, Y., Jan, C., He, M., Jadhav, K., Ge, Z.: Interpretable few-shot retinal disease diagnosis with concept-guided prompting of vision-language models. In: Information Processing in Medical Imaging. pp. 263–277. Springer Nature Switzerland (2026)
13. Pang, W., Ke, X., Tsutsui, S., Wen, B.: Integrating clinical knowledge into concept bottleneck models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 243–253. Springer (2024)
14. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
15. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
17. Tsutsui, S., Pang, W., Wen, B.: WBCAtt: A white blood cell dataset annotated with detailed morphological attributes. In: Advances in Neural Information Processing Systems. vol. 36, pp. 50796–50824. Curran Associates, Inc. (2023)
18. Vemuri, D.S., Bellamkonda, G., Pola, A., Balasubramanian, V.N.: LogicCBMs: Logic-enhanced concept-based learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6039–6048 (2026)
19. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H., Tschandl, P., Kittler, H., Ge, Z.: Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12681–12690 (2025)
20. Yan, S., Li, X., Hu, M., Jiang, Y., Yu, Z., Ge, Z.: Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–379. Springer (2025)
21. Yan, S., Li, X., Mo, D., Tschandl, P., Jiang, Y., Wang, Z., Hu, M., Ju, L., Vico-Alonso, C., Zheng, Y., et al.: A vision-language foundation model for zero-shot clinical collaboration and automated concept discovery in dermatology. arXiv preprint arXiv:2602.10624 (2026)
22. Yan, S., Yu, Z., Zhang, X., Mahapatra, D., Chandra, S.S., Janda, M., Soyer, P., Ge, Z.: Towards trustable skin cancer diagnosis via rewriting model's decision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11568–11577 (2023)
23. Yang, H., George, Y., Mehta, D., Lin, L., Chen, C., Yang, D., Sun, J., Lau, K.F., Bain, C., Yang, Q., et al.: Multimodal CT perfusion–based deep learning for predicting stroke lesion outcomes in complete and no recanalization scenarios. American Journal of Neuroradiology **47**(4), 920–928 (2026)
24. Yang, H., Yu, Z., Mehta, D., George, Y., Ge, Z.: Semantic-guided 3D CT generation for stroke lesion segmentation. In: 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2026)

25. Yang, H., Yuwen, C., Cheng, X., Fan, H., Wang, X., Ge, Z.: Deep learning: A primer for neurosurgeons. Computational Neurosurgery pp. 39–70 (2024)
26. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
27. Zhang, Y., Unell, A., Wang, X., Ghosh, D., Su, Y., Schmidt, L., Yeung-Levy, S.: Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems **37**, 51727–51753 (2024)
28. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
29. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z.: LlamaFactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). pp. 400–410. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)