CLAIR-MED: SELECTIVE NEURAL–SYMBOLIC REASONING WITH WEAKLY SUPERVISED CONCEPTS FOR MEDICAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rule-based concept models expose an explicit decision process but are often less accurate than neural predictors; however, always falling back to the neural model renders the symbolic explanation irrelevant to the final decision. We investigate a narrower question: *how often can an explicit symbolic branch make the prediction while retaining competitive task utility?* CLAIR-Med constructs medical concept vocabularies using a language model, derives pseudo concept targets from vision–language similarity, and learns an image-to-concept interface without manual concept annotations. The predicted concept state is consumed by two structurally different pathways: a concept-augmented neural predictor and a Deep Concept Reasoner (DCR) that executes fuzzy rules on concept truth values. We transfer class-level predictive information to DCR through asymmetric distillation and use DCR uncertainty to select between symbolic and neural predictions at inference. Experiments on five medical image datasets and CUB evaluate branch utility, routing, concept quality, prediction sensitivity, and controlled edits. At selected operating points, symbolic coverage is 78.30–99.31%, with displayed-rule agreement of 91.05–98.52% on selected cases. On Chest X-Ray and Fitzpatrick17k, routed accuracy is 78.21% and 81.88%, respectively. In a blinded evaluation of 100 cases, three medical-domain evaluators consistently rate its concepts above LF-CBM, with Fleiss' κ of 0.62–0.76; deleting the three most activated concepts also causes substantially larger probability drops than random or bottom-ranked deletion. The results support selective use of an explicit symbolic pathway while delimiting the evidence from clinical deployment validation.

1 INTRODUCTION

Concept bottleneck models (CBMs) make a prediction inspectable by associating intermediate variables with human-readable concepts (Koh et al., 2020). Label-free variants reduce the cost of concept annotation (Oikarinen et al., 2023), and neural–symbolic approaches such as Deep Concept Reasoner (DCR) can execute explicit rules over concept states (Barbiero et al., 2023). These developments create a practical tension. A rule-based predictor can expose its computation, but it may trail a higher-capacity neural predictor. Conversely, a system that always uses the neural output and displays the rule only afterward provides an explanation that did not determine the reported prediction.

We formulate this tension as *selective symbolic prediction*. For each input, the system either returns the output of an explicit concept-rule pathway or falls back to a concept-augmented neural predictor. The objective is therefore not to claim that the symbolic model is universally superior; it is to characterize how symbolic coverage changes with routed predictive risk. This framing separates three questions that are often conflated: whether the learned concepts are meaningful, whether the symbolic pathway is competitive enough to use, and whether uncertainty identifies cases on which that pathway should decide.

CLAIR-Med instantiates this setting without human concept annotations. A language model proposes textual medical concepts, CLIP (Radford et al., 2021) converts image–text similarity into pseudo concept targets, and a task model predicts the resulting concept state from image features.

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.

Two predictors consume this shared state. The first is a concept-augmented neural predictor that also retains an unrestricted visual feature path; we do not call it a strict concept bottleneck. The second is DCR (Barbiero et al., 2023), which generates relevance and polarity gates from concept embeddings and executes the resulting fuzzy rule on predicted concept truth values. Class-distribution distillation transfers utility from the neural pathway to DCR, after which an entropy threshold controls the symbolic coverage–risk operating point. Figure 1 summarizes the full concept-supervision, dual-pathway, distillation, and routing pipeline. Optional concept and rule edits are evaluated as controllability probes, not as evidence of a validated clinical workflow.

We evaluate the method on five medical datasets—BloodMNIST (Yang et al., 2023), COVID-CT (Zhao et al., 2020), OCT2017 (Kermany et al., 2018), Chest X-Ray (Govi, 2020), and Fitzpatrick17k (Groh et al., 2021)—and on CUB-200-2011 (Wah et al., 2011). Beyond predictive and routing metrics, we conduct a blinded concept-quality study with three medical-domain evaluators and a dataset-wide deletion analysis. CLAIR-Med receives higher mean ratings than LF-CBM on every dataset and criterion, with moderate-to-substantial inter-rater agreement ($\kappa = 0.62$–0.76), while top-3 concept deletion produces approximately four times the ground-truth probability drop of random deletion. The two additional challenging medical benchmarks further test whether the gains persist beyond the original controlled datasets.

Our contributions are:

- We formalize **selective symbolic prediction** over a shared, pseudo-supervised concept interface and evaluate it through symbolic coverage and routed risk rather than treating the presence of concepts as sufficient evidence of interpretability.
- We introduce **asymmetric predictive-to-symbolic distillation** followed by uncertainty-based branch selection, allowing the explicit DCR pathway to participate in the final prediction on a validation-controlled subset of inputs.
- We provide a **mechanism-oriented evaluation** spanning five medical datasets and CUB, including blinded medical-domain concept ratings, top/random/bottom concept deletion, branch utility, routing behavior, and controlled intervention.

2 Related Work

Weakly supervised concept interfaces. CBMs predict through semantically named intermediate variables (Koh et al., 2020). Post-hoc CBMs, concept embeddings, concept whitening, and label-free CBMs relax the annotation or capacity constraints of the original formulation (Yuksekgonul et al., 2022; Zarlenga et al., 2022; Chen et al., 2020; Oikarinen et al., 2023). These relaxations also weaken what can be inferred from the interface: a concept name does not guarantee concept correctness, and an unrestricted feature bypass means that the final neural prediction is not fully explained by concept scores (Margeloiu et al., 2021). Medical concept-model studies consequently evaluate clinically meaningful vocabularies, faithfulness, and stability in addition to task performance (Lai et al., 2024; Hu et al., 2024a; Daneshjou et al., 2022; Groh et al., 2021; Hu et al., 2025). CLAIR-Med uses language- and vision–language-model supervision as an enabling component, not as its primary algorithmic novelty.

Neural–symbolic concept reasoning. Differentiable logic systems combine neural representation learning with structured operators (Xu et al., 2018; Manhaeve et al., 2018; Badreddine et al., 2022; Garcez et al., 2022). DCR specifically generates fuzzy rule structure from concept embeddings and executes the rule on concept truth values (Barbiero et al., 2023). CLAIR-Med retains this separation between neural rule generation and symbolic rule execution. Its technical focus is instead the interface between a higher-utility neural predictor and DCR: output-level distillation is used to improve DCR utility, but is not treated as proof that the learned concepts or rules are correct.

Selective prediction and intervention. Selective prediction uses confidence or uncertainty to trade coverage for risk (Geifman & El-Yaniv, 2019; Guo et al., 2017; Kull et al., 2019). Our router applies this idea to branch selection: low-uncertainty examples are decided by DCR, while the remaining examples use the neural pathway. This is not learning to defer to a clinician and does not establish clinical safety. Concept interventions provide a complementary diagnostic by changing

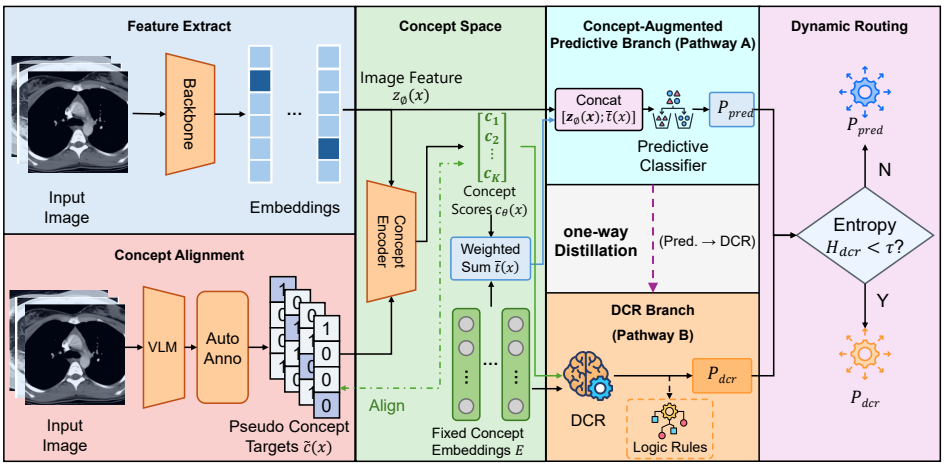

Figure 1: Overview of CLAIR-Med. The lower-left pathway provides pseudo concept targets for learning the shared concept state. Pathway A combines image features with concept-weighted embeddings for predictive classification. Pathway B uses fixed concept embeddings to parameterize DCR rules and executes them on predicted concept scores. One-way distillation transfers class-level information from Pathway A to Pathway B, and DCR entropy selects the final branch at inference.

semantically indexed intermediate states (Chauhan et al., 2023; Hu et al., 2024b). We combine controlled edits with blinded medical-domain ratings and dataset-wide deletion analysis: the former evaluates perceived relevance and usefulness, whereas the latter measures activation-based prediction sensitivity.

3 Method

3.1 Problem Setup and Information Flow

Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ contain images and class labels but no human concept annotations. CLAIR-Med introduces a K-dimensional concept interface $c_\theta(x) \in [0,1]^K$ and two class predictors: a higher-capacity concept-augmented neural predictor, $a_{\text{pred}}(x) \in \mathbb{R}^C$, and an explicit rule-based predictor, $a_{\text{dcr}}(x) \in [0,1]^C$. Figure 1 summarizes the architecture. The vision–language model supplies pseudo concept targets and precomputed textual concept embeddings, while the learned concept head supplies the input-dependent concept state at inference. Appendix C summarizes the component interfaces.

This terminology is deliberate. Because the neural pathway directly consumes the unrestricted feature $z_\phi(x)$, it is *concept augmented* rather than a strict CBM. Only DCR produces its class score by executing an explicit rule on the concept truth values. The two pathways share a concept interface but make predictions through different computation graphs.

3.2 Pseudo-Supervised Concept Interface

For each medical dataset, GPT-4o proposes candidate visual attributes from class names and modality-specific prompts. Candidates from repeated prompts are pooled and filtered for excessive length, class-name leakage, semantic redundancy, poor image alignment, and low projectability. Exact prompts and filtering thresholds are given in Appendix H. CUB is a distinct vocabulary setting: it uses 98 attributes selected from the original CUB annotations (Appendix G) rather than a GPT-generated medical vocabulary. The complete pseudo-label generation pipeline is illustrated in Appendix D.

For a concept κ_k, the vision–language text encoder returns $t_k \in \mathbb{R}^{d_t}$; its image encoder returns $v(x) \in \mathbb{R}^{d_t}$. Their cosine similarity defines

$$s_k(x) = \frac{v(x)^\top t_k}{\|v(x)\|_2 \|t_k\|_2}, \qquad \tilde{c}_k(x) = \mathbb{I}[s_k(x) > \delta]. \tag{1}$$

The main experiments use the global threshold δ documented in Appendix K.

The task backbone and concept head are separate from the pseudo-label generator:

$$z_\phi(x) = h_\phi(x) \in \mathbb{R}^{d_z}, \qquad c_\theta(x) = \mathrm{sigmoid}(q_\theta(z_\phi(x))) \in [0,1]^K. \tag{2}$$

They are trained with

$$\mathcal{L}_{\mathrm{concept}} = \mathrm{BCE}(c_\theta(x), \tilde{c}(x)). \tag{3}$$

Thus CLIP similarity supplies supervision, whereas $c_\theta(x)$ is the concept state used by the two task pathways. Learning this separate head permits the task backbone to produce the concept interface without querying CLIP during final prediction; the resulting concepts are evaluated through blinded medical-domain ratings and prediction-sensitivity analysis.

3.3 Two Predictors over the Shared Concept State

Concept-augmented neural predictor. Let $E = [t_1; \ldots; t_K] \in \mathbb{R}^{K \times d_t}$ collect the textual concept embeddings. The predicted truth scores form an image-specific semantic summary

$$\bar{t}(x) = g(c_\theta(x), E) = \sum_{k=1}^{K} c_{\theta,k}(x) t_k, \qquad a_{\mathrm{pred}}(x) = W_{\mathrm{pred}}[z_\phi(x) \| \bar{t}(x)] + b_{\mathrm{pred}}. \tag{4}$$

The semantic summary makes the concept state available to the high-capacity predictor, while the direct $z_\phi(x)$ path preserves information not captured by the pseudo concepts. This pathway is therefore a utility reference and fallback, not an intrinsically complete explanation of its own output.

Deep Concept Reasoner. DCR (Barbiero et al., 2023) uses a concept embedding to generate rule structure and a concept truth value to execute that rule. For class j and concept k, a role network $\pi_j(t_k) \in [0,1]$ determines whether the literal is positive or negated, and a relevance network $\rho_j(t_k) \in [0,1]$ determines whether it participates in the rule. With differentiable fuzzy equivalence $\Leftrightarrow$, implication $\Rightarrow$, and conjunction $\wedge$, the literal and relevance-filtered literal are

$$\ell_{jk}(x) = \pi_j(t_k) \Leftrightarrow c_{\theta,k}(x), \tag{5}$$

$$\ell^r_{jk}(x) = \rho_j(t_k) \Rightarrow \ell_{jk}(x) = \neg \rho_j(t_k) \vee \ell_{jk}(x). \tag{6}$$

The class score is obtained by executing the conjunction

$$a_{\mathrm{dcr},j}(x) = \bigwedge_{k=1}^{K} \ell^r_{jk}(x). \tag{7}$$

The specific fuzzy operators used in Eqs. (5)–(7) are given in Appendix F. Equation 7 depends on the input through $c_\theta(x)$. The role and relevance networks are optimized through the DCR task loss; after training, their gates and polarities expose the literals used by the rule. DCR provides an explicit rule execution path, but the correctness of that rule still depends on the concept interface and must be evaluated separately.

3.4 Asymmetric Predictive-to-Symbolic Distillation

The DCR branch produces fuzzy truth scores $a_{\mathrm{dcr}}(x) \in [0,1]^C$. We convert them into class logits by applying an element-wise log-odds transformation:

$$s_{\mathrm{dcr}}(x) = \mathrm{logit}(\mathrm{clip}(a_{\mathrm{dcr}}(x), \varepsilon_{\mathrm{clip}}, 1 - \varepsilon_{\mathrm{clip}})), \qquad \varepsilon_{\mathrm{clip}} = 10^{-8}, \tag{8}$$

where $\mathrm{logit}(u) = \log \frac{u}{1-u}$ and clip denotes element-wise clipping.

Let

$$\mathcal{L}_{\mathrm{pred}} = \mathrm{CE}(a_{\mathrm{pred}}, y), \qquad \mathcal{L}_{\mathrm{dcr}} = \mathrm{CE}(s_{\mathrm{dcr}}, y) \tag{9}$$

denote the two supervised task losses. From epoch t_0 onward, we form temperature-scaled class distributions

$$p^T_{\text{pred}} = \text{softmax}(a_{\text{pred}}/T), \qquad p^T_{\text{dcr}} = \text{softmax}(s_{\text{dcr}}/T), \tag{10}$$

and minimize the one-way distillation loss

$$\mathcal{L}_{\text{KD}} = T^2 \,\text{KL}\big(\text{sg}[p^T_{\text{pred}}] \,\|\, p^T_{\text{dcr}}\big)\,, \tag{11}$$

where sg stops gradients through the teacher distribution. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{dcr}}\mathcal{L}_{\text{dcr}} + \lambda_{\text{c}}\mathcal{L}_{\text{concept}} + \lambda_{\text{kd}}\mathcal{L}_{\text{KD}}. \tag{12}$$

Equation 11 aligns class distributions only. It can reduce the predictive–symbolic utility gap, but it neither changes the pseudo concept targets nor certifies the semantic correctness or faithfulness of the extracted rules.

3.5 Selective Symbolic Inference

At inference, we obtain the DCR class distribution from the logits defined above, $p_{\text{dcr}}(x) = \text{softmax}(s_{\text{dcr}}(x))$, and compute

$$H_{\text{dcr}}(x) = -\sum_{j=1}^{C} p_{\text{dcr},j}(x) \log p_{\text{dcr},j}(x). \tag{13}$$

Let $p_{\text{pred}}(x) = \text{softmax}(a_{\text{pred}}(x))$ denote the predictive-branch class distribution. The router uses the explicit rule-based output only for low-entropy inputs:

$$p_\tau(x) = \begin{cases} p_{\text{dcr}}(x), & H_{\text{dcr}}(x) < \tau, \\ p_{\text{pred}}(x), & \text{otherwise.} \end{cases} \tag{14}$$

The final predicted class is $\hat{y}_\tau(x) = \arg\max_j p_{\tau,j}(x)$. The corresponding symbolic coverage is

$$\alpha(\tau) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[H_{\text{dcr}}(x_n) < \tau]. \tag{15}$$

We select τ on held-out validation data and report performance across the threshold sweep, because a single accuracy-maximizing threshold hides the coverage–risk trade-off. This is branch-selective prediction, not a guarantee that low entropy implies clinical reliability.

3.6 Optional Post-hoc Intervention Interface

The trained computation graph also permits controlled edits. A concept intervention replaces selected coordinates of the per-sample state $c_\theta(x)$ with an edited state $c'(x)$ and recomputes the chosen predictor. A rule intervention either activates an additional antecedent or flips the polarity of a selected literal before re-executing Eq. 7. Neither operation updates the backbone, concept head, or predictor parameters.

These edits establish functional controllability: changing an exposed intermediate state can mechanically change the recomputed output. The current experiments treat them as what-if probes. They do not establish how an editor obtains the correct value, whether a persistent rule update is safe on future cases, or whether the interface improves clinical decisions.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on five medical image datasets: BloodMNIST (Yang et al., 2023), COVID-CT (Zhao et al., 2020), OCT2017 (Kermany et al., 2018), Chest X-Ray (Govi, 2020), and Fitzpatrick17k (Groh et al., 2021), spanning relatively controlled benchmarks and more challenging medical image recognition settings. CUB-200-2011 (Wah et al., 2011) is additionally used to evaluate transfer beyond medicine to fine-grained recognition.

Table 1: Main benchmark performance averaged over five runs with different random seeds. Best results are boldfaced and the CLAIR-Med row is shaded.

Method	BloodMNIST	OCT2017	COVID-CT	CUB	Chest X-Ray	Fitzpatrick17k
	Acc. (%)	Acc. (%)	Acc. (%)	Acc. (%)	Acc. (%)	Acc. (%)
Standard	97.16	99.71	86.51	77.58	76.29	80.26
LF-CBM	95.18	96.80	77.63	74.51	72.68	74.39
P-CBM	94.89	97.23	74.56	65.60	70.32	72.27
SVCT	97.01	99.49	84.67	76.86	75.77	77.56
Static fusion	97.58	99.59	90.41	78.88	76.85	80.12
Confidence fallback	97.05	99.56	91.06	78.89	77.32	80.27
Random routing	96.99	99.28	90.26	75.68	75.23	76.84
CLAIR-Med (Ours)	**97.87**	**99.79**	**92.06**	**79.41**	**78.21**	**81.88**

Training and model-selection protocol. All methods use matched backbones, concept dimensionality, and training budget. Hyperparameters and routing thresholds are selected on held-out validation splits. We use ViT-B/16 (Dosovitskiy et al., 2021) for BloodMNIST and OCT2017, ResNet50 (He et al., 2016) for COVID-CT and Fitzpatrick17k, and DenseNet169 (Huang et al., 2017) for Chest X-Ray and CUB; All competing methods use the same dataset-specific backbone and matched training protocol. Backbone robustness across all three encoder families is reported below.

Baselines. Baselines are a matched standard backbone, LF-CBM (Oikarinen et al., 2023), Post-hoc CBM (Yuksekgonul et al., 2022), SVCT (Hu et al., 2025), static fusion of the predictive and symbolic branches, confidence-based fallback, and random routing.

Metrics. We report task metrics (accuracy, macro-F1, AUROC, precision, recall) and operational metrics: branch accuracy, utility gap, gap reduction, routed accuracy, and symbolic usage. Concept quality is evaluated on four 1–5 rating criteria with Fleiss' κ for inter-rater agreement. Prediction sensitivity is measured as the decrease in ground-truth class probability after concept deletion.

Appendix G gives dataset, concept-count and baseline details. Appendix J tests prompt and concept-size robustness. Appendix K records the CLIP pseudo-label configuration. Appendix M reports the full metric tables, and Appendix N reports backbone robustness. Appendix X formalizes the operational metrics and routing/intervention decompositions used by the analysis. Calibration analysis of the DCR branch is provided in Appendix S, and robustness experiments under input perturbations are reported in Appendix U.

4.2 Main Results: Competitive Utility With a Usable Symbolic Pathway

Table 1 reports the main predictive results across all six benchmarks, with five-seed variability for CLAIR-Med provided in Appendix L. CLAIR-Med achieves the highest accuracy among the compared methods on all datasets, while exposing concept states, DCR rules, and branch-selection decisions. Relative to the matched standard backbone, the gains are particularly pronounced on COVID-CT (92.06% vs. 86.51%), with improvements also observed on the more challenging Chest X-Ray (78.21% vs. 76.29%) and Fitzpatrick17k (81.88% vs. 80.26%) benchmarks. CLAIR-Med also outperforms the concept-based baselines (LF-CBM, P-CBM, and SVCT) and the routing/fusion alternatives across all six datasets. Compared with confidence fallback, the strongest routing baseline, CLAIR-Med improves accuracy across all six datasets, with gains ranging from 0.23 to 1.61 percentage points. On Chest X-Ray and Fitzpatrick17k benchmarks, the improvements are 0.89 and 1.61 points, respectively. These results indicate that the selective symbolic pathway maintains strong predictive utility across medical benchmarks of varying difficulty, while also transferring to fine-grained recognition on CUB.

Table 2: Utility and symbolic participation at the selected operating points. Pred. and DCR are whole-dataset branch accuracies; DCR (no KD) gives the undistilled reference. Coverage is the fraction assigned to DCR. Replay is displayed-rule agreement with DCR on that selected subset. All values are percentages. Full paired branch results and replay definitions appear in Appendices O and W.

Dataset	Pred.	DCR (no KD)	DCR	Routed	Coverage	Replay
BloodMNIST	97.90	97.26	97.78	97.87	96.17	97.81
OCT2017	99.79	98.87	99.66	99.79	99.31	98.52
COVID-CT	92.19	89.23	91.66	92.06	89.10	96.36
CUB	80.24	76.96	78.81	79.41	78.30	92.03
Chest X-Ray	78.24	73.22	75.57	78.21	80.35	93.37
Fitzpatrick17k	81.81	77.37	79.92	81.88	83.75	91.05

4.3 Mechanism Analysis: Distillation Narrows the Utility Gap

The central mechanistic claim of CLAIR-Med is that asymmetric distillation improves the symbolic branch without degrading the predictive one. Table 2 summarizes branch utility, symbolic coverage, and displayed-rule fidelity. Appendix O provides the full paired branch accuracies and utility-gap reductions. The key diagnostic is whether the gap shrinks through symbolic improvement rather than through degradation of the predictive branch.

As shown in Appendix O, distillation reduces the reported utility gap across all six datasets, with a stable predictive branch and a stronger symbolic branch. The reduction is particularly pronounced on CUB, Chest X-Ray, and Fitzpatrick17k, where the gap decreases from 3.25 to 1.43, 4.69 to 2.67, and 4.30 to 1.89 percentage points, respectively. The effect is smaller on the near-saturated BloodMNIST and OCT2017 benchmarks, where the two branches are already closely matched without distillation. Appendix O provides the formal utility definitions and a five-seed analysis of the run-to-run stability of the branch accuracies and paired gap reductions.

The routed accuracy is close to, but usually slightly below, the predictive branch. The supported claim is therefore that DCR can be used at non-zero coverage while retaining near-predictive utility, not that routing is an oracle that inherits the better branch on every input. Routing behavior and the resulting coverage–utility trade-off are analyzed separately in the following subsection.

4.4 Routing Analysis: The Utility–Interpretability Frontier

We analyze the routing threshold τ as a symbolic coverage–risk frontier rather than as a single tuned hyperparameter. Across the completed sweeps, increasing τ generally raises symbolic coverage, while routed accuracy remains stable near the selected operating point and degrades when the symbolic branch is overused. Full τ tables, together with additional λ_{kd} and t_0 sensitivity plots, are reported in Appendix P.

A coverage-matched comparison of DCR entropy against margin, MSP, random selection, and a retrospective oracle is provided in Appendix V. The results show that DCR entropy is generally competitive with the alternative confidence selectors, although its advantage is not uniform across all datasets and coverage levels.

At the selected operating points, symbolic coverage ranges from 78.30% to 99.31% (Table 2). Displayed-rule agreement on those samples ranges from 91.05% to 98.52%, making explicit the distinction between the soft rule that supplies the class score and its displayed discrete approximation. These results support a measurable utility–coverage trade-off rather than uniformly superior routing. Appendix X.3 relates coverage and replay agreement to a bound on the risk change under fixed-router rule replacement.

4.5 Blinded Concept-Quality Evaluation

We randomly sample 100 test cases from the five medical datasets. Three medical-domain evaluators—one radiologist and two PhD researchers with medical-imaging backgrounds—

Table 3: Blinded medical-domain evaluation of concept explanations on a 1–5 scale. Fleiss' κ measures inter-rater agreement within each dataset.

Dataset	Method	Clinical rel.	Consistency	Usefulness	Completeness	κ
COVID-CT	LF-CBM	3.62	3.42	3.58	3.12	0.76
	CLAIR-Med	**4.42**	**4.52**	**4.35**	**4.10**	
OCT2017	LF-CBM	3.65	3.89	3.96	3.39	0.72
	CLAIR-Med	**4.31**	**4.27**	**4.29**	**4.08**	
BloodMNIST	LF-CBM	3.97	3.84	3.78	3.56	0.65
	CLAIR-Med	**4.18**	**4.12**	**4.05**	**4.02**	
Chest X-Ray	LF-CBM	3.52	3.57	3.68	3.43	0.62
	CLAIR-Med	**4.29**	**4.38**	**4.27**	**4.06**	
Fitzpatrick17k	LF-CBM	3.61	3.86	3.79	3.57	0.68
	CLAIR-Med	**4.21**	**4.42**	**4.12**	**4.02**	

Table 4: Ground-truth class-probability drop after concept deletion (mean $\pm$ standard deviation). The LF-CBM column uses its three highest-activation concepts.

	CLAIR-Med			LF-CBM
Dataset	Top-3	Random-3	Bottom-3	Top-3
COVID-CT	0.167 ± 0.026	0.041 ± 0.022	0.021 ± 0.008	0.102 ± 0.018
OCT2017	0.152 ± 0.018	0.036 ± 0.015	0.015 ± 0.007	0.081 ± 0.015
BloodMNIST	0.128 ± 0.011	0.032 ± 0.006	0.009 ± 0.005	0.084 ± 0.008
Chest X-Ray	0.151 ± 0.025	0.038 ± 0.012	0.011 ± 0.005	0.076 ± 0.017
Fitzpatrick17k	0.137 ± 0.017	0.034 ± 0.013	0.019 ± 0.008	0.069 ± 0.009

independently assess CLAIR-Med and LF-CBM explanations while blinded to method identity. Each case contains the image, prediction, and top-ranked concepts. Evaluators score clinical relevance, concept consistency, explanation usefulness, and completeness from 1 to 5.

Table 3 shows that CLAIR-Med receives a higher mean score than LF-CBM for every dataset and every criterion. The largest improvements occur on COVID-CT and Chest X-Ray, while BloodMNIST shows smaller but consistent gains. Fleiss' κ ranges from 0.62 to 0.76 across datasets, indicating moderate-to-substantial agreement among the three evaluators. Complementary rule-replay analysis in Appendix W evaluates whether the displayed discrete rules preserve the decision behavior of the underlying soft DCR reasoner.

4.6 Concept-Deletion Analysis

We next test whether highly activated concepts are more strongly associated with the prediction than arbitrary concepts. For each sample, concepts are ranked by their predicted activation $c_{\theta,k}(x)$. Given a deleted set S, we measure the drop in the probability assigned to the ground-truth class,

$$\Delta p_y(S) = p_y(x; c_\theta(x)) - p_y(x; c_\theta^{\setminus S}(x)). \tag{16}$$

This is an activation-based sensitivity measure rather than a claim of causal attribution.

Table 4 compares deletion of the top three, three random, and bottom three concepts. Top-3 deletion yields a probability drop of 0.128–0.167, approximately four times the 0.032–0.041 drop from random deletion, while bottom-3 deletion has the smallest effect on every dataset. Under the same top-3 protocol, CLAIR-Med also produces larger drops than LF-CBM across all five datasets. Together with the blinded ratings, this indicates that the exposed concepts are both judged more meaningful and more closely associated with the prediction pathway.

4.7 Post-hoc Intervention Examples

CLAIR-Med exposes three complementary internal objects: predicted concept scores, DCR rules, and the branch-selection decision. In a representative COVID-CT case (Figure 2), the routed system initially selects the predictive branch and predicts the wrong class. Editing two concept scores changes the recomputed prediction from non-COVID to COVID without retraining, illustrating model-internal controllability.

The same interface extends to DCR-routed cases, where an evaluator can edit a rule by adding an antecedent or flipping a negated condition. These examples demonstrate inspect–edit–recompute behavior; they remain distinct from the blinded concept-quality study, which evaluates the unedited explanations across a sampled cohort.

Appendix Q provides additional routed and rule-level intervention examples. Appendix R includes dataset-specific qualitative examples for COVID-CT, CUB, and BloodMNIST.

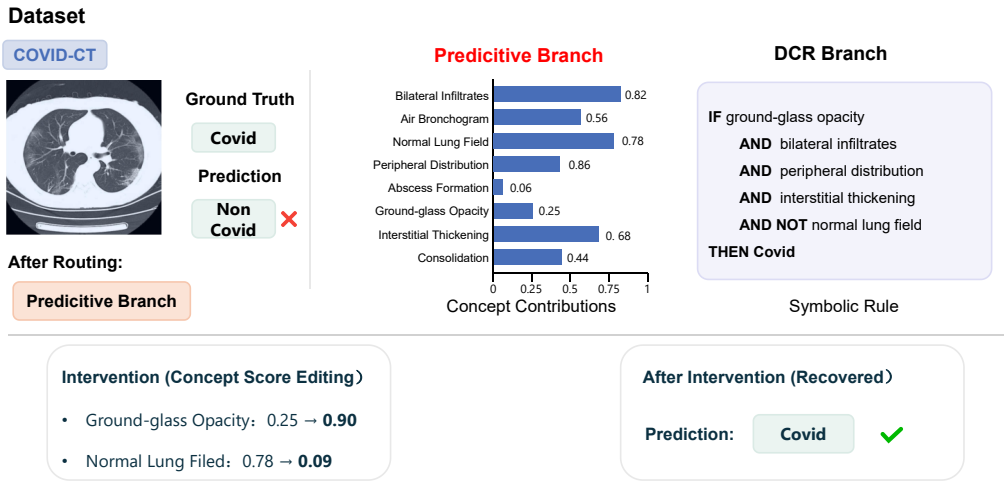

Figure 2: Illustrative per-sample concept-state edit. The model initially routes to the predictive branch and predicts the wrong class; editing selected scores changes the recomputed output.

4.8 Ablation and Robustness

Appendix B separates gains attributable to distillation and routing from gains attributable merely to extra capacity or favorable parameter choices. Table 18 further evaluates CLAIR-Med across ViT-B/16, ResNet50, and DenseNet169.

5 Conclusion

We presented CLAIR-Med for selective neural–symbolic prediction over weakly supervised concepts. A concept-augmented neural predictor transfers class-level utility to a concept-constrained DCR pathway, narrowing the utility gap between the two branches, while DCR entropy controls how often the explicit rule pathway supplies the final class score. Across five medical datasets and CUB, the method combines competitive predictive performance with an explicit symbolic pathway. Blinded medical-domain ratings and concept deletion support the semantic quality and prediction relevance of the exposed concepts, while displayed-rule replay shows that the user-facing discrete rules preserve most of the decision behavior of the underlying soft DCR reasoner. Controlled edits further demonstrate inspect–edit–recompute behavior. The study is best viewed as a benchmark framework for retrospective selective neural–symbolic medical image classification rather than evidence of clinical readiness.

AI USE STATEMENT

In this work, we used generative AI tools to assist with concept generation, manuscript drafting and editing, limited code checking, and verification of mathematical derivations and proofs. We did not use generative AI tools to originate the core research idea or make final scientific conclusions. All AI-assisted content was reviewed and verified by the authors before inclusion. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies retrospective image classification on public benchmark datasets and does not evaluate autonomous diagnosis or clinical decision support. The reported concept and rule interfaces are intended for model inspection and methodological evaluation under the studied benchmark settings. Any use in clinical workflows would require additional external validation, monitoring under distribution shift, and appropriate human oversight.

REPRODUCIBILITY STATEMENT

We provide dataset and concept-count details in Appendix G, with dataset-specific backbones, training hyperparameters, distillation settings, and routing thresholds reported in Appendix E and Table 7. Concept construction and filtering are documented in Appendix H, and the complete CLIP pseudo-label configuration is given in Appendix K and Table 10.

REFERENCES

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lió, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, pp. 1801–1825. PMLR, 2023.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pp. 5948–5955, 2023.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto A. Novoa, and James Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Luis C Lamb, Leo de Penning, BV Illuminoo, and Hoifung Poon. Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342(1):327, 2022.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2151–2159. PMLR, 2019.

Praveen Govi. Coronahack -chest x-ray-dataset. Kaggle dataset, 2020. URL `https://www.kaggle.com/datasets/praveengovi/coronahack-chestxraydataset`.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Lijie Hu, Songning Lai, Wenshuo Chen, Hongru Xiao, Hongbin Lin, Lu Yu, Jingfeng Zhang, and Di Wang. Towards multi-dimensional explanation alignment for medical classification. *Advances in Neural Information Processing Systems*, 37:129640–129671, 2024a.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Hui Xiong, Jingfeng Zhang, and Di Wang. Editable concept bottleneck models. *arXiv preprint arXiv:2405.15476*, 2024b.

Lijie Hu, Songning Lai, Yongkang Hua, Shun Yang, Jingfeng Zhang, and Di Wang. Stable vision concept transformers for medical diagnosis. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2025)*, volume 16015 of *Lecture Notes in Computer Science*. Springer, 2025.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

Daniel S Kermany, Michael H Goldbaum, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.e9, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5338–5348. PMLR, 2020.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Equille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *The Twelfth International Conference on Learning Representations*, 2024.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. *Advances in Neural Information Processing Systems*, 31, 2018.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5502–5511. PMLR, 2018.

Jiancheng Yang, Rui Shi, Donglai Wei, and et al. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):1–15, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Liò, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Han Xu, Jaspreet Bagga, Ruijia Zhang, Song Wang, Hoifung Poon, Jianwei Wei, Tristan Naumann, et al. Biomedclip: a multimodal biomedical foundation model. *arXiv preprint arXiv:2303.00915*, 2023.

Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020.

A Discussion, Limitations, and Broader Impact

CLAIR-Med is designed for selective symbolic prediction rather than replacing the neural predictor. Asymmetric distillation narrows the utility gap between the predictive and DCR branches, allowing the explicit rule pathway to supply the final prediction on a validation-controlled subset of inputs. The concept ratings, deletion analysis, routing experiments, and rule-replay evaluation examine complementary aspects of this interface, including semantic quality, prediction sensitivity, selective utility, and agreement between the displayed and executed rules.

The study has several scope limitations. The medical concepts are learned from automatically constructed vocabularies and weak vision–language supervision rather than exhaustive concept-level annotations. In addition, the predictive branch retains an unrestricted visual-feature path, so our interpretability claims concern the concept-constrained DCR pathway and the cases for which it supplies the final prediction. The reported routing operating points are selected on held-out validation data and are evaluated retrospectively on benchmark datasets; their behavior under external distribution shifts remains to be established.

The broader potential benefit of CLAIR-Med is to make a substantial fraction of model predictions inspectable through an explicit concept-rule pathway while retaining a higher-capacity predictive fallback. At the same time, interpretable interfaces should not be taken as evidence of clinical validity when their concepts or rules are imperfect. Accordingly, the present results support retrospective methodological evaluation rather than autonomous diagnosis or clinical decision support; deployment would require additional external validation and appropriate human oversight.

B Main-Text Ablation and Robustness Details

Table 5 separates gains attributable to distillation and routing from gains attributable merely to extra capacity or favorable parameter choices. Each ablation configuration uses its own validation-selected checkpoint. For the "w/o routing" configuration, both checkpoint selection and test evaluation use the predictive branch directly, whereas the full model uses the routing mechanism during validation-based checkpoint selection and evaluation. Across all datasets, the full model outperforms both ablated variants, confirming that distillation and routing contribute complementary benefits. The ablation shows complementary numerical effects: distillation improves DCR utility, while routing exposes a tunable symbolic-coverage operating point. Under "w/o routing," predictive-branch accuracy is reported as Final Acc because it typically exceeds DCR accuracy. This is a strong utility baseline, but it does not by itself test concept correctness, rule fidelity, or coverage-matched selector quality.

C Component Interfaces and Information Flow

The shared concept state connects training supervision to two distinct prediction pathways. Table 6 separates the offline vocabulary and pseudo-target construction from the components used for final prediction.

D Pseudo-Label Generation Pipeline

Figure 3 illustrates the complete pipeline used to construct the concept vocabulary and pseudo concept labels for the medical datasets.

GPT-based concept set generation. As shown in Figure 3(a), class names together with representative medical images are provided to GPT-4o to generate candidate visual concepts for each class. The resulting raw concept lists are pooled, cleaned, and formatted to remove undesirable or redundant candidates, yielding the final concept set $\mathcal{C} = \{\kappa_1, \ldots, \kappa_K\}$.

CLIP-based pseudo concept labeling. Figure 3(b) shows the pseudo-label generation stage. The concepts in $\mathcal{C}$ are encoded by the CLIP text encoder to obtain concept text embeddings $\{t_1, \ldots, t_K\}$, while each medical image is encoded by the CLIP image encoder. Cosine similarity between the

Table 5: Ablation study across datasets. The full model combines predictive-to-symbolic distillation with uncertainty-aware routing. Each configuration uses its own validation-selected checkpoint. All values are percentages except categorical entries. The DCR Usage metric is only defined for configurations with entropy-based routing enabled, entries are marked N/A when routing is disabled.

Dataset	KD	Routing	Final Acc.	DCR Ratio	DCR Acc.
OCT2017	✓	×	99.65	N/A	99.58
	×	✓	99.02	97.83	98.87
	✓	✓	**99.79**	99.31	99.66
CUB	✓	×	79.27	N/A	77.82
	×	✓	78.88	65.82	76.96
	✓	✓	**79.41**	78.30	78.81
BloodMNIST	✓	×	97.61	N/A	97.58
	×	✓	97.46	94.96	97.26
	✓	✓	**97.87**	96.17	97.78
COVID-CT	✓	×	90.87	N/A	90.75
	×	✓	90.07	84.92	89.23
	✓	✓	**92.06**	89.10	91.66
Chest X-Ray	✓	×	77.25	N/A	74.99
	×	✓	76.87	69.94	73.22
	✓	✓	**78.21**	80.35	75.57
Fitzpatrick17k	✓	×	80.64	N/A	78.96
	×	✓	79.43	77.92	77.37
	✓	✓	**81.88**	83.75	79.92

Table 6: Component interfaces in CLAIR-Med. The vision–language model supplies pseudo-target supervision and fixed textual concept embeddings, while final predictions use the learned concept state. The neural branch also has direct access to visual features.

Component	Output	Use in the pipeline
LLM vocabulary	$\mathcal{C} = \{\kappa_k\}_{k=1}^{K}$	Constructs the dataset-level vocabulary before task training.
Vision–language model	$v(x), t_k, \tilde{c}(x)$	Supplies pseudo concept targets during training and provides the precomputed textual concept embeddings t_k used by the task pathways.
Task backbone/head	$z_\phi(x),\ c_\theta(x)$	Produces visual features and concept scores at training and inference.
Neural predictor	$a_{\mathrm{pred}}(x)$	Predicts from both visual features and the concept summary.
DCR predictor	$a_{\mathrm{dcr}}(x)$	Executes learned fuzzy rules on the predicted concept scores.
Router	$r_\tau(x)$	Chooses the final prediction branch at inference.

image embedding and the concept embeddings produces a K-dimensional similarity vector for each image. Each similarity score is then compared with the fixed threshold δ to obtain a binary pseudo concept label. Stacking these labels across the training and validation images produces the pseudo concept label matrix used to supervise the learned concept head $c_\theta(x)$.

The vision–language model is not queried on test images during final prediction. Its textual concept embeddings $\{t_1, \ldots, t_K\}$ are precomputed and retained as fixed concept representations, while the learned concept head supplies the input-dependent concept scores $c_\theta(x)$.

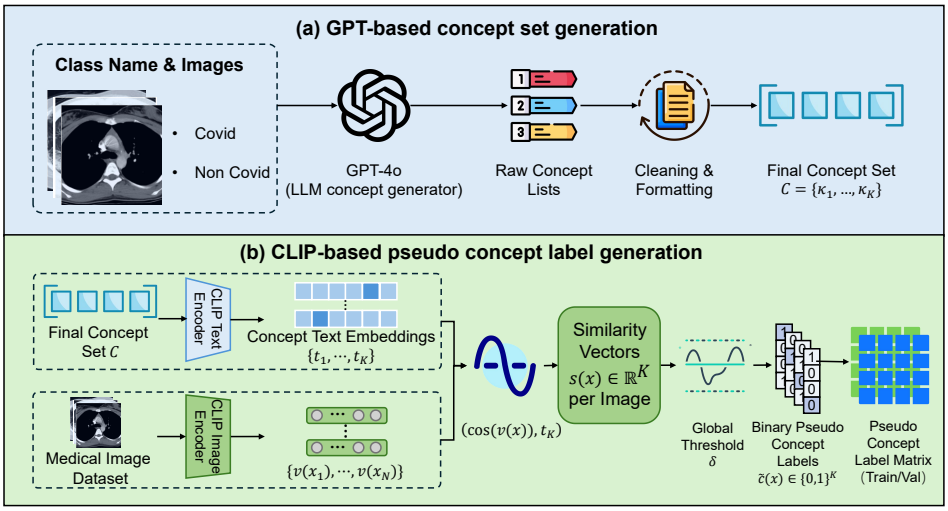

Figure 3: Pseudo-label generation pipeline for the medical datasets. (a) GPT-4o generates candidate visual concepts, which are cleaned and formatted to form the final concept set $\mathcal{C}$. (b) CLIP encodes the concepts and images, and the resulting image–text similarities are thresholded to produce binary pseudo concept targets $\tilde{c}(x)$ for supervising the learned concept head. The concept text embeddings are retained as the fixed embeddings E used by the prediction pathways.

E Training and Reproducibility Details

Table 7 summarizes the dataset-specific architecture, concept, optimization, and routing configurations used throughout the experiments. When a setting differs across datasets, we report the corresponding dataset-specific value. Routing thresholds are selected on held-out validation data and then fixed for test evaluation. Unless otherwise stated, the reported benchmark results are averaged over five runs with different random seeds.

Table 7: Dataset-specific training configurations for CLAIR-Med. K denotes the number of concepts, t_0 is the epoch at which predictive-to-symbolic distillation is activated, and τ is the DCR-entropy routing threshold.

Dataset	Backbone	K	Epochs	Batch	LR	λ_{kd}	t_0	τ
BloodMNIST	ViT-B/16	36	100	64	5×10^{-5}	0.1	50	0.5
OCT2017	ViT-B/16	68	20	8	5×10^{-5}	0.1	6	0.6
COVID-CT	ResNet50	16	40	8	5×10^{-5}	0.5	3	0.4
CUB	DenseNet169	98	50	64	1×10^{-4}	0.8	20	2.5
Chest X-Ray	DenseNet169	32	50	16	5×10^{-5}	0.5	15	0.7
Fitzpatrick17k	ResNet50	82	60	32	5×10^{-5}	0.8	20	2.2

The distillation temperature is fixed to $T = 4$ across all datasets. The predictive, concept, and DCR objectives are optimized jointly according to Eq. (12), with distillation activated from epoch t_0. When routing is enabled, checkpoint selection is based on validation routed accuracy. The routing threshold τ is selected using the held-out validation split and is kept fixed during test evaluation.

Pseudo concept targets are generated using the fixed CLIP configuration reported in Table 10; we therefore do not repeat the CLIP model, text template, similarity function, or pseudo-label threshold here.

F FUZZY OPERATORS FOR DCR

For completeness, we specify the fuzzy operators used in Eqs. (5)–(7). For $u, v \in [0, 1]$, negation, disjunction, and conjunction are defined as

$$\neg u = 1 - u, \qquad u \vee v = \max(u, v), \qquad u \wedge v = \min(u, v). \tag{17}$$

The fuzzy implication and equivalence are defined as

$$u \Rightarrow v = \max(1 - u, v), \tag{18}$$

and

$$u \Leftrightarrow v = \min\{\max(1 - u, v), \max(u, 1 - v)\}. \tag{19}$$

Accordingly, with $\pi_j(t_k)$, $\rho_j(t_k)$, and $c_{\theta,k}(x)$ in $[0, 1]$, Eqs. (5)–(7) produce DCR class truth scores $a_{\mathrm{dcr}}(x) \in [0, 1]^C$.

G MORE EXPERIMENTAL SETTINGS

G.1 DATASETS

BloodMNIST. BloodMNIST is a subset of the MedMNIST benchmark dataset collection (Yang et al., 2023). It contains images of individual normal blood cells from individuals without any infections, blood disorders, or tumor diseases. The dataset is categorized into eight types of cells: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, red blood cells, and platelets.

OCT2017. Optical coherence tomography (OCT) is an imaging modality capable of viewing the morphology of the retina layer. The OCT2017 dataset (Kermany et al., 2018) comprises 84,484 OCT images with four retinal disease classes: Normal, CNV (choroidal neovascularization), DME (diabetic macular edema), and Drusen. The dataset is divided into training, validation, and test folders.

COVID-CT. The COVID-CT dataset (Zhao et al., 2020) consists of 746 lung CT images, with 349 COVID-19 cases and 397 non-COVID cases. It provides default divisions for training, validation, and test data. For COVID-CT, we use a randomized cropping strategy for data augmentation.

Chest X-Ray. The CoronaHack Chest X-Ray dataset (Govi, 2020) contains chest radiographs from healthy subjects and patients with pneumonia-related infections. We formulate the dataset as a three-class classification task consisting of Normal, Bacterial Pneumonia, and Viral Pneumonia.

Fitzpatrick17k. Fitzpatrick17k (Groh et al., 2021) is a dermatology dataset containing 16,577 clinical images covering 114 skin conditions and a wide range of skin tones. We use the fine-grained disease labels and formulate the dataset as a 114-class classification task.

CUB-200-2011. The CUB-200-2011 dataset (Wah et al., 2011) is a fine-grained bird species classification benchmark containing 11,788 images of 200 bird species. It is split into 5,994 training images and 5,794 test images. We use CUB to evaluate the generalization capability of our method beyond the medical domain.

G.2 NUMBER OF CONCEPTS

In our framework, the number of concepts used in each dataset is determined empirically based on the complexity of the classification task. For COVID-CT (binary classification), we use 16 clinically relevant concepts (e.g., ground-glass opacity, consolidation, bilateral involvement, crazy-paving pattern). For BloodMNIST (8 cell types), we use 36 biologically meaningful concepts covering nucleus morphology, granules, cytoplasm characteristics, and cell size. For OCT2017 (4 retinal diseases), we use 68 concepts describing fluid accumulation, drusen deposits, retinal layer integrity, and hyperreflective foci. For Chest X-Ray (3 classes), we use 32 radiographic visual concepts describing patterns relevant to pneumonia-related chest radiograph classification. For Fitzpatrick17k (114 skin conditions), we use 82 dermatological visual concepts describing the morphology and appearance of

skin lesions. For CUB (200 bird species), we use 98 fine-grained visual attributes (e.g., wing color, beak shape, body pattern) selected from the original CUB annotations based on their discriminative power.

G.3 Baseline Models

Standard backbone. The standard model is a plain classifier (ResNet50, ViT-B/16, or DenseNet169 depending on the dataset) that extracts image features and connects to a fully connected layer for classification. It serves as a strong non-interpretable reference. Within each dataset, the same backbone is used for all methods to ensure fair comparison.

LF-CBM. Label-free CBM (Oikarinen et al., 2023) is a fully automated method that transforms any neural network into an interpretable concept bottleneck model without requiring labeled concept data. It uses CLIP to generate concept pseudo-labels and trains a sparse linear classifier on the concept scores. In our experiments, we use the same backbone as our method for each dataset.

Post-hoc CBM (PCBM). Post-hoc CBM (Yuksekgonul et al., 2022) converts any pre-trained model into a concept bottleneck model by learning a concept subspace via CAVs (Concept Activation Vectors) or CLIP text embeddings. It trains a sparse linear predictor on the projected concept representations. We evaluate PCBM with the same backbone configuration as our method.

SVCT (Stable Vision Concept Transformer). SVCT (Hu et al., 2025) is a ViT-based interpretable model that combines a concept bottleneck layer with feature fusion and denoised diffusion smoothing to provide stable explanations under input perturbations. We include SVCT as a strong interpretable baseline. For datasets where the primary backbone is not ViT (e.g., COVID-CT with ResNet50, CUB with DenseNet169), we use the corresponding backbone for SVCT to ensure fair comparison.

Static fusion. Static fusion combines the predictive and symbolic branches with fixed equal weights (0.5 and 0.5) without routing. The final score is $a = 0.5a_{\text{pred}} + 0.5a_{\text{dcr}}$. This baseline isolates the effect of dynamic branch selection.

Confidence fallback. Confidence fallback is a validation-tuned branch-selection baseline that uses predictive confidence as the fallback signal. It tests whether CLAIR-Med's gains can be reproduced by a simpler confidence rule that lacks the explicit entropy-based symbolic-coverage frontier.

Random routing. Random routing selects between the predictive and DCR branches without using uncertainty. It controls for the possibility that gains come merely from occasionally using the symbolic branch rather than from the selection score.

H Concept Set Generation and Filtering Rules

H.1 Concept Set Generation

In the original CBM paradigm, the concept set is typically curated by domain experts, a process that requires significant expertise and effort. To automate this process, we leverage a large multimodal model (GPT-4o) via the OpenAI API to generate candidate concepts for each dataset. Specifically, we prompt GPT-4o with class names and, when applicable, sample images from the dataset to ground the generation in actual visual patterns. Note that the image here refers to a randomly sampled image from the dataset, and the type represents the modality of the medical image (e.g., CT, OCT, microscopy).

For each dataset, we initiate the procedure with the following conversation:

- **important.** List the most important features for recognizing something as a {dataset image-class} of {class}:
- **around.** List the things most commonly seen around a {class}:
- **superclass.** Give superclasses for the word {class}:

Here, {dataset image-class} refers to the medical image modality (e.g., CT, OCT, etc.,) and {class} corresponds to the target class name in the classification task (e.g., Neutrophil, CNV). We provide

two example outputs in the prompt for few-shot adaptation; these examples are shared across all datasets and require no additional user input. To reduce variance, we run each prompt three times and merge the results.

Combining the concepts received from different classes and prompts gives us a large, somewhat noisy set of initial concepts, which we further improve by filtering. Specific examples of the generated concept sets for each dataset are provided in Appendix I.

H.2 Concept Set Filtering

After obtaining the initial set of concepts, a noisy concept set with some redundancy can be obtained. The following filters are applied to enhance the quality and interpretability of the concept set. The filtering process consists of the following main aspects:

1. Length filtering. Remove concepts longer than 40 characters to keep concepts simple and avoid unnecessary complexity.

2. Remove concepts too similar to class names. A two-step strategy is adopted: first, remove concepts that are exact matches, plural forms, or prefixed with "a/an/the" of any class name; then, compute cosine similarity using CLIP + MPNet weighted embeddings and remove concepts with similarity > 0.85 to any class name.

3. Remove concepts too similar to each other. Compute cosine similarity using CLIP + MPNet weighted embeddings and remove concepts with pairwise similarity > 0.9 to avoid semantic redundancy.

4. Remove concepts poorly aligned with images. Perform one-way ANOVA on CLIP similarities between training images and each concept across classes. Apply FDR multiple testing correction ($\alpha = 0.05$) and require the difference between the maximum and minimum class means to exceed 0.1. Concepts must satisfy both criteria to be retained.

5. Remove concepts that cannot be projected accurately. Compute an automatic projectability score for each concept using image-text alignment consistency on high-activation samples. Concepts with consistently low projectability are removed.

I Examples of Generated Concept Sets

Figure 4 provides examples of our full prompts and the corresponding GPT-4o outputs. For all experiments, we use the GPT-4o model available through the OpenAI API.

J Concept set ablations

To demonstrate that CLAIR-Med is robust to different concept set configurations, we conducted two additional experiments. First, we varied the prompt templates used to generate the initial concept sets while keeping the final concept set size fixed (prompt ablation). Second, we varied the number of concepts after filtering while keeping the prompt fixed (size ablation). Table 8 reports the results under alternative prompts, and Table 9 reports the results under different concept set sizes.

J.1 Prompt Ablation

To evaluate the sensitivity of CLAIR-Med to different prompt templates, we generated concept sets using alternative prompts while keeping the final concept set size fixed. Table 8 reports the classification accuracy on all datasets. The results show that CLAIR-Med maintains stable performance across different prompt choices.

List the most important features for recognizing something as a lymphocyte:
- high nuclear-cytoplasmic ratio
- dense round nucleus
- coarse clumped chromatin
- scant pale-blue cytoplasmic rim
- non-lobulated nucleus
- absence of prominent specific granules

Give superclasses for the word lymphocyte:
- red blood cells
- platelets
- other white blood cells
- nucleus
- cytoplasmic rim

List the things most commonly seen around a lymphocyte:
- mononuclear leukocyte
- small lymphoid cell
- mature lymphocyte morphology
- lymphoid lineage cell
- non-granulocytic white blood cell

List the most important features for recognizing something as a DRUSEN:
- sub-RPE drusen deposits
- RPE elevation
- small nodular drusen
- irregular RPE contour
- drusen-related shadowing
- hyperreflective sub-RPE material

Give superclasses for the word lymphocyte:
- retinal pigment epithelium
- Bruch's membrane
- drusen morphology
- age-related macular degeneration
- outer retinal layers
- foveal contour

List the things most commonly seen around a lymphocyte:
- drusen lesion
- sub-RPE deposit
- AMD-related feature
- RPE-Bruch complex abnormality
- macular degenerative sign

Figure 4: Examples of generated concept sets.

Table 8: Results under different prompt variants.

Method	BloodMNIST	OCT2017	COVID-CT	CUB	Chest X-Ray	Fitzpatrick17k
Standard (No interpretability)	97.16	99.71	86.51	77.58	76.29	80.26
Label-Free CBM (LF-CBM)	94.76	96.72	77.51	74.24	72.13	74.05
Post-hoc CBM (P-CBM)	92.61	95.78	73.16	65.76	70.41	71.53
Stable Vision Concept Transformer (SVCT)	97.02	99.41	84.56	76.25	75.36	76.87
Static fusion	97.32	99.53	90.12	76.17	76.10	80.24
CLAIR-Med (Ours)	**97.82**	**99.69**	**91.79**	**79.16**	**78.08**	**81.65**

J.2 Concept Set Size Ablation

To evaluate the sensitivity of CLAIR-Med to different concept set sizes, we varied the number of concepts after filtering while keeping the prompt fixed. Table 9 reports the classification accuracy on all datasets. The results demonstrate that CLAIR-Med is robust to the choice of concept set size.

Table 9: Results under different concept set sizes.

Method	BloodMNIST	OCT2017	COVID-CT	CUB	Chest X-Ray	Fitzpatrick17k
Standard (No interpretability)	97.16	99.71	86.51	77.58	76.29	80.26
Label-Free CBM (LF-CBM)	94.81	96.07	75.76	74.62	71.99	73.84
Post-hoc CBM (P-CBM)	94.56	96.57	74.89	74.12	70.24	72.03
Stable Vision Concept Transformer (SVCT)	97.03	99.36	84.13	76.35	75.23	77.68
Static fusion	97.18	99.06	91.01	77.57	76.34	79.79
CLAIR-Med (Ours)	**97.79**	**99.62**	**91.88**	**78.68**	**78.02**	**81.36**

K Pseudo concept label construction

This section records the CLIP configuration used to convert textual concepts into weak supervision. We use the OpenAI CLIP (Radford et al., 2021) implementation from the official repository. The detailed configuration is summarized in Table 10.

Table 10: CLIP pseudo-label construction details.

Parameter	**Value**
CLIP model	ViT-B/32
Text wrapper	"a photo of {concept}"
Image preprocessing	Resize(224), CenterCrop(224), ToTensor(), Normalize(mean=[0.48145466, 0.4578275, 0.40821073], std=[0.26862954, 0.26130258, 0.27577711])
Similarity score	Cosine similarity between L2-normalized image/text embeddings
Threshold rule	Fixed global threshold $\delta = 0.25$

L Main-Benchmark Multi-Seed Stability

The main benchmark results in Table 1 are averaged over five independent runs. For each dataset, we use the same train/validation/test split across all methods and runs, and repeat training with five different seeds. To make the run-to-run variability of the final CLAIR-Med model explicit, we report the mean and sample standard deviation of the routed test accuracy across the five runs.

For dataset d, let $A_{d,s}$ denote the routed test accuracy obtained in run s. We compute

$$\bar{A}_d = \frac{1}{5}\sum_{s=1}^{5} A_{d,s}, \tag{20}$$

and the sample standard deviation

$$s_d = \sqrt{\frac{1}{4}\sum_{s=1}^{5}\left(A_{d,s} - \bar{A}_d\right)^2}. \tag{21}$$

Table 11 shows that the final routed performance is stable across the five runs. BloodMNIST and OCT2017 exhibit particularly small run-to-run variation, while COVID-CT, CUB, Chest X-Ray, and Fitzpatrick17k show moderately larger variability. The standard deviation remains below one percentage point on all six benchmarks. These results make explicit that the main reported CLAIR-Med accuracies summarize multiple independently trained runs rather than a single favorable run.

Table 11: Multi-seed stability of the final routed CLAIR-Med accuracy. Results are reported as mean ± standard deviation over five independent runs with five different seeds.

Dataset	Accuracy (%)
BloodMNIST	97.87 ± 0.07
OCT2017	99.79 ± 0.03
COVID-CT	92.06 ± 0.37
CUB	79.41 ± 0.32
Chest X-Ray	78.21 ± 0.35
Fitzpatrick17k	81.88 ± 0.30

M Full Benchmark Tables

This section presents the complete multi-metric benchmark results on all six datasets, including Accuracy, Precision, Recall, F1, and AUROC. Table 12 presents the results on BloodMNIST, Table 13 on OCT2017, Table 14 on COVID-CT, Table 15 on CUB-200-2011, Table 16 on Chest X-Ray, and Table 17 on Fitzpatrick17k.

Table 12: Results for BloodMNIST, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	97.16	95.87	95.27	95.57	96.67
LF-CBM	95.18	93.57	92.79	93.18	88.19
P-CBM	94.89	93.95	95.17	94.56	96.92
SVCT	97.01	95.56	95.12	95.34	98.50
Static Fusion	97.58	96.27	96.25	96.26	98.91
CLAIR-Med (Ours)	**97.87**	**97.94**	**97.47**	**97.70**	**99.28**

Table 13: Results for OCT2017, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	99.71	99.11	99.61	99.36	99.76
LF-CBM	96.80	94.98	94.87	94.92	97.22
P-CBM	97.23	96.73	95.21	95.96	97.63
SVCT	99.49	98.49	96.32	97.39	99.21
Static Fusion	99.59	99.46	99.59	99.52	99.56
CLAIR-Med (Ours)	**99.79**	**99.72**	**99.78**	**99.75**	**99.92**

N Backbone Robustness

Table 18 reports results across ViT-B/16, ResNet50, and DenseNet169. For each dataset, we compare the standard classifier and CLAIR-Med under the same backbone architecture and training setting.

O Branch-Wise Utility Definitions and Multi-Seed Stability

Table 2 in the main paper summarizes utility, coverage, and replay fidelity. Table 19 provides the complete paired branch accuracies, utility gaps, and gap reductions.

We define the predictive–DCR utility gap as

$$G = \mathrm{Acc}(\hat{y}_{\mathrm{pred}}) - \mathrm{Acc}(\hat{y}_{\mathrm{dcr}}). \tag{22}$$

Table 14: Results for COVID-CT, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	86.51	86.63	86.42	86.52	90.48
LF-CBM	77.63	77.81	77.65	77.73	78.09
P-CBM	74.56	72.96	75.79	74.35	78.05
SVCT	84.67	86.47	88.06	87.26	89.39
Static Fusion	90.41	90.48	90.42	90.45	96.17
CLAIR-Med (Ours)	**92.06**	**92.01**	**92.03**	**92.02**	**98.35**

Table 15: Results for CUB-200-2011, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	77.58	77.62	77.25	77.43	80.87
LF-CBM	74.51	75.24	74.33	74.78	78.72
P-CBM	65.60	63.75	64.83	64.29	69.01
SVCT	76.86	77.03	77.26	77.14	79.17
Static Fusion	78.88	81.13	79.68	80.40	84.18
CLAIR-Med (Ours)	**79.41**	**82.52**	**80.18**	**81.33**	**85.29**

Table 16: Results for Chest X-Ray, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	76.29	75.85	76.37	76.12	80.84
LF-CBM	72.68	72.34	72.71	72.25	74.29
P-CBM	70.32	70.01	73.95	70.67	67.83
SVCT	75.77	75.67	76.48	75.67	78.54
Static Fusion	76.85	76.10	77.16	76.84	79.96
CLAIR-Med (Ours)	**78.21**	**77.79**	**78.38**	**78.23**	**85.68**

Table 17: Results for Fitzpatrick17k, averaged over five runs. Best results are boldfaced. CLAIR-Med achieves the highest accuracy among the compared methods while retaining an explicit concept-rule prediction pathway.

Method	Acc. (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Standard (No interpretability)	80.26	80.58	81.52	80.15	83.17
LF-CBM	74.39	77.49	75.88	74.22	79.63
P-CBM	72.27	71.72	72.29	71.89	73.46
SVCT	77.56	77.23	77.76	77.51	80.78
Static Fusion	80.12	79.94	80.37	80.08	84.29
CLAIR-Med (Ours)	**81.88**	**81.26**	**82.05**	**81.82**	**86.32**

Table 18: Backbone robustness. Each CLAIR-Med configuration is compared against a matched standard classifier using the same visual encoder.

Method	Backbone	BloodMNIST	OCT2017	COVID-CT	CUB	Chest X-Ray	Fitzpatrick17k
Standard	DenseNet169	97.08	99.62	87.65	77.58	76.29	80.31
Standard	ResNet50	97.36	99.70	86.51	73.98	75.87	80.26
Standard	ViT-B/16	97.16	99.71	83.06	76.28	75.96	79.82
CLAIR-Med	DenseNet169	97.43	99.69	90.48	**79.41**	**78.21**	81.53
CLAIR-Med	ResNet50	97.49	99.78	**92.06**	74.48	77.83	**81.88**
CLAIR-Med	ViT-B/16	**97.87**	**99.79**	88.51	78.87	78.02	81.16

Table 19: Branch-wise utility analysis. Gap is defined as $\mathrm{Acc}(\hat{y}_{\mathrm{pred}}) - \mathrm{Acc}(\hat{y}_{\mathrm{dcr}})$, and ΔGap is the reduction relative to the no-distillation configuration.

Dataset	Configuration	Pred. Acc.	DCR Acc.	Gap	ΔGap	Routed Acc.
BloodMNIST	w/o distill.	97.52	97.26	0.26	+0.14	97.46
	CLAIR-Med	**97.90**	**97.78**	**0.12**		**97.87**
CUB	w/o distill.	80.21	76.96	3.25	+1.82	78.88
	CLAIR-Med	**80.24**	**78.81**	**1.43**		**79.41**
COVID-CT	w/o distill.	91.03	89.23	1.80	+1.27	90.07
	CLAIR-Med	**92.19**	**91.66**	**0.53**		**92.06**
OCT2017	w/o distill.	99.08	98.87	0.21	+0.08	99.02
	CLAIR-Med	**99.79**	**99.66**	**0.13**		**99.79**
Chest X-Ray	w/o distill.	77.91	73.22	4.69	+2.02	76.87
	CLAIR-Med	**78.24**	**75.57**	**2.67**		**78.21**
Fitzpatrick17k	w/o distill.	81.67	77.37	4.30	+2.41	79.43
	CLAIR-Med	**81.81**	**79.92**	**1.89**		**81.88**

The reduction in this gap from the configuration without distillation to the full CLAIR-Med configuration is defined as

$$\Delta G = G_{\mathrm{w/o\,distill}} - G_{\mathrm{CLAIR\text{-}Med}}. \tag{23}$$

A positive ΔG therefore indicates that distillation narrows the reported predictive–DCR utility gap.

Multi-seed stability. We further evaluate whether the observed gap reduction is stable across training randomness using five independent runs with different random seeds. For each run s, we compute

$$G_s = A_{\mathrm{pred},s} - A_{\mathrm{dcr},s}, \tag{24}$$

where $A_{\mathrm{pred},s}$ and $A_{\mathrm{dcr},s}$ denote the predictive- and DCR-branch accuracies, respectively. We then compute the paired gap reduction for the same run as

$$\Delta G_s = G_s^{\mathrm{w/o\,distill}} - G_s^{\mathrm{CLAIR\text{-}Med}}. \tag{25}$$

We report mean $\pm$ standard deviation across the five runs. Importantly, the variability of ΔG is computed from the per-run paired reductions $\{\Delta G_s\}_{s=1}^{5}$ rather than derived from the separately aggregated branch statistics.

Results. Table 20 shows that distillation reduces the mean predictive–DCR utility gap on all six benchmarks. The reduction is most pronounced on the datasets with larger initial branch discrepancies. On CUB, Chest X-Ray, and Fitzpatrick17k, the mean gap decreases from 3.25 to 1.43, 4.69 to 2.67, and 4.30 to 1.89 percentage points, corresponding to paired mean reductions of 1.82, 2.02, and 2.41 points, respectively. COVID-CT exhibits the same pattern, with the gap decreasing from 1.80 to 0.53 points and a paired mean reduction of 1.27 points.

Table 20: Multi-seed stability of the distillation mechanism. Pred. and DCR denote predictive- and symbolic-branch accuracy, respectively, and Gap is their accuracy difference. Gap Reduction is the paired reduction in the predictive–DCR gap after distillation. Values are mean ± standard deviation over five independent runs with different random seeds.

	w/o Distillation			With Distillation			
Dataset	Pred.	DCR	Gap	Pred.	DCR	Gap	Gap Reduction
BloodMNIST	97.52±0.09	97.26±0.11	0.26±0.07	97.90±0.07	97.78±0.08	0.12±0.06	**0.14±0.09**
OCT2017	99.08±0.06	98.87±0.08	0.21±0.05	99.79±0.03	99.66±0.05	0.13±0.04	**0.08±0.06**
COVID-CT	91.03±0.41	89.23±0.52	1.80±0.27	92.19±0.35	91.66±0.38	0.53±0.22	**1.27±0.31**
CUB	80.21±0.31	76.96±0.42	3.25±0.28	80.24±0.29	78.81±0.36	1.43±0.24	**1.82±0.34**
Chest X-Ray	77.91±0.38	73.22±0.49	4.69±0.35	78.24±0.34	75.57±0.43	2.67±0.31	**2.02±0.42**
Fitzpatrick17k	81.67±0.35	77.37±0.46	4.30±0.31	81.81±0.32	79.92±0.41	1.89±0.28	**2.41±0.37**

The reduction is smaller on BloodMNIST and OCT2017, where the initial gaps are only 0.26 and 0.21 points, respectively. This is consistent with a ceiling effect: when the predictive and symbolic branches are already closely matched, there is less discrepancy for distillation to reduce. Across the more challenging benchmarks, the larger paired gap reductions relative to their run-to-run variability indicate that the observed mechanism is not driven solely by a single training run.

Importantly, the narrowing of the utility gap is not explained by degradation of the predictive branch. Mean predictive accuracy is maintained or improved after distillation on every benchmark, while mean DCR accuracy increases throughout. For example, on Fitzpatrick17k, DCR accuracy increases from 77.37 ± 0.46 to 79.92 ± 0.41, whereas predictive accuracy changes only from 81.67 ± 0.37 to 81.81 ± 0.32. On CUB, predictive accuracy is nearly unchanged (80.21 ± 0.31 versus 80.24 ± 0.29), while DCR accuracy increases from 76.96 ± 0.42 to 78.81 ± 0.36.

Together, these results support the intended role of asymmetric output-level distillation: it narrows the utility discrepancy between the predictive and concept-constrained symbolic pathways primarily by improving DCR utility rather than by weakening the predictive branch. The remaining non-zero gaps also show that distillation does not make the two pathways equivalent.

P Sensitivity Analysis

We analyze the sensitivity of CLAIR-Med to three key hyperparameters: the routing threshold τ, the distillation weight λ_{kd}, and the distillation start epoch t_0 (total training epochs fixed at 100). The expanded routing-threshold sweep covers all six datasets, while the λ_{kd} and t_0 sweeps are reported only on BloodMNIST. For each analysis, we fix the other hyperparameters to their default values and vary the target hyperparameter.

P.1 Sensitivity to Routing Threshold τ

We vary τ while keeping λ_{kd} and t_0 fixed. For each value of τ, checkpoint selection is performed independently on the validation set under the corresponding routing configuration, and the selected checkpoint is subsequently evaluated on the test set. Consequently, different values of τ may correspond to different checkpoints, so the predictive- and DCR-branch accuracies are not required to remain constant across operating points.

Figure 5 visualizes the BloodMNIST sweep over the routing threshold τ, and Table 21 reports the numeric sweeps for all six datasets: BloodMNIST, CUB, COVID-CT, OCT2017, Chest X-Ray, and Fitzpatrick17k. As τ increases, the DCR usage ratio rises, while routed accuracy remains stable near the best operating point before degrading when the symbolic branch is overused. Routed accuracy peaks at $\tau = 0.5$ on BloodMNIST (95.91% DCR usage), $\tau = 2.5$ on CUB (78.30% DCR usage), $\tau = 0.4$ on COVID-CT (89.10% DCR usage), $\tau = 0.6$ on OCT2017 (99.31% DCR usage), $\tau = 0.7$ on Chest X-Ray (80.35% DCR usage), and $\tau = 2.2$ on Fitzpatrick17k (83.75% DCR usage).

These results characterize the coverage–utility trade-off across validation-selected routing operating points: lower values of τ are more conservative, whereas higher values expose more symbolic decisions at a measurable utility cost.

Table 21: Expanded routing-threshold sensitivity. All values except τ are percentages. Bold τ values indicate the default operating points used in the main benchmark.

Dataset	τ	DCR Acc.	Pred. Acc.	DCR Usage	Routed Acc.
BloodMNIST	0.1	97.72	97.81	82.94	97.75
	0.3	97.75	97.82	93.98	97.79
	0.5	**97.82**	**97.89**	95.91	**97.87**
	0.7	97.78	97.83	97.78	97.81
	0.9	97.66	97.84	**98.77**	97.76
CUB	1.5	75.59	78.51	68.01	77.52
	2.0	77.02	79.83	75.45	78.82
	2.5	**78.81**	**80.24**	78.30	**79.41**
	3.0	77.92	79.92	80.26	78.67
	3.5	77.58	80.11	**82.39**	78.05
COVID-CT	0.1	88.96	90.45	72.38	90.41
	0.2	91.01	91.70	88.06	91.68
	0.3	90.84	91.32	88.86	91.27
	0.4	**91.66**	**92.15**	89.10	**92.06**
	0.5	91.42	91.92	**91.28**	91.82
OCT2017	0.2	99.42	99.49	95.87	99.48
	0.4	99.56	99.69	97.12	99.67
	0.6	**99.66**	**99.79**	99.31	**99.79**
	0.8	99.31	99.56	99.68	99.55
	1.0	99.12	99.18	**99.85**	99.17
Chest X-Ray	0.3	72.84	77.57	64.72	77.34
	0.5	74.67	78.02	73.46	77.86
	0.7	**75.57**	**78.24**	80.35	**78.21**
	0.9	75.21	78.08	86.73	77.22
	1.0	74.63	77.86	**91.28**	76.82
Fitzpatrick17k	1.6	77.84	81.05	68.42	80.73
	1.9	79.01	81.32	77.26	81.42
	2.2	**79.92**	**81.81**	83.75	**81.88**
	2.5	79.47	81.68	88.61	81.15
	2.8	78.97	80.99	**92.37**	80.58

P.2 Sensitivity to Distillation Weight λ_{KD}

On BloodMNIST, we vary λ_{kd} while keeping τ and t_0 fixed. Figure 6 reports the results. The routed accuracy remains consistently high across all tested values, with a slight increase in DCR usage ratio as λ_{kd} grows. This indicates that CLAIR-Med performs stably over a broad range of distillation strengths.

P.3 Sensitivity to Distillation Start Epoch t_0

On BloodMNIST, we vary t_0 while keeping τ and λ_{kd} fixed. Figure 7 shows the results. The routed accuracy remains stable across the reported start epochs once the predictive branch has been trained sufficiently.

Q Routed and Intervention Examples

This appendix provides additional qualitative examples for routed explanations and intervention behavior. Figure 8 shows a routed OCT2017 explanation. Figure 2 shows concept-level correction, and Figures 9–10 show rule-level intervention operators. For easy distinction, two different forms of rule editing are denoted as "logic editing" and "rule editing" in the figures, respectively.

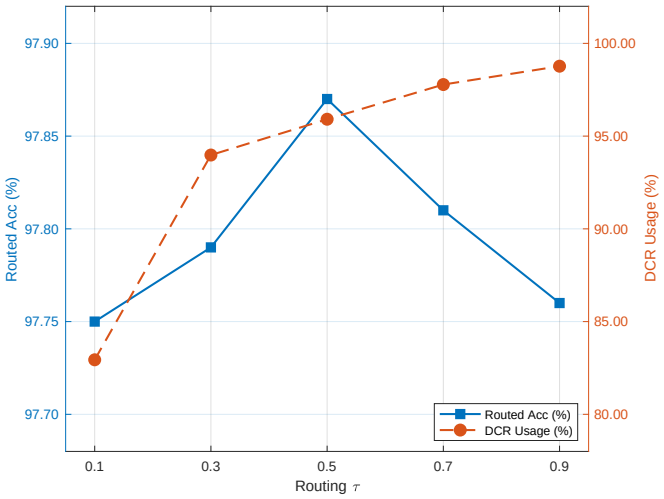

Figure 5: τ sensitivity analysis.

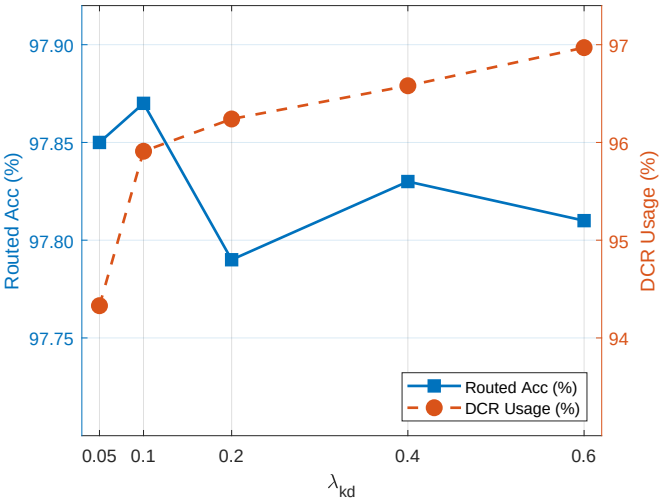

Figure 6: λ_{kd} sensitivity analysis.

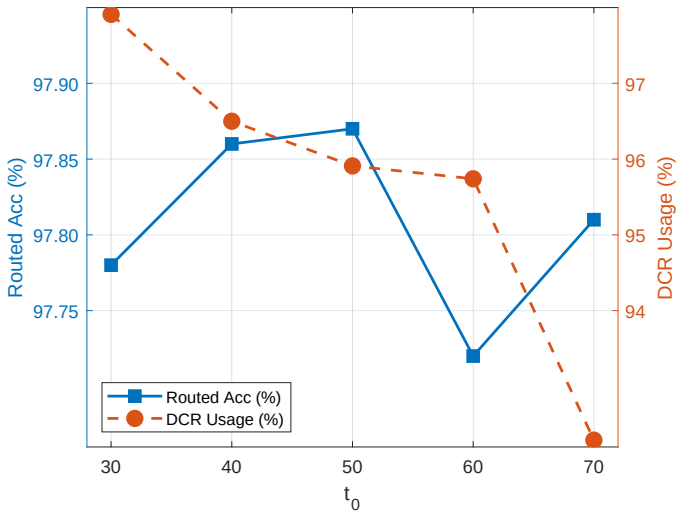

Figure 7: t_0 sensitivity analysis.

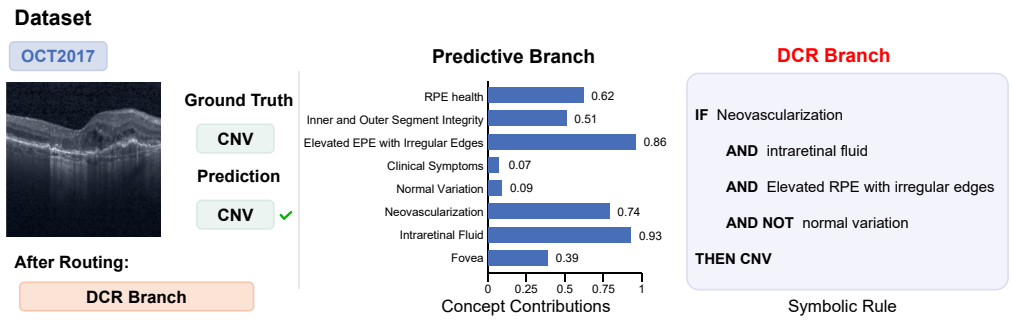

Figure 8: Correctly predicted OCT2017 sample with concept scores, symbolic rule, and routing decision.

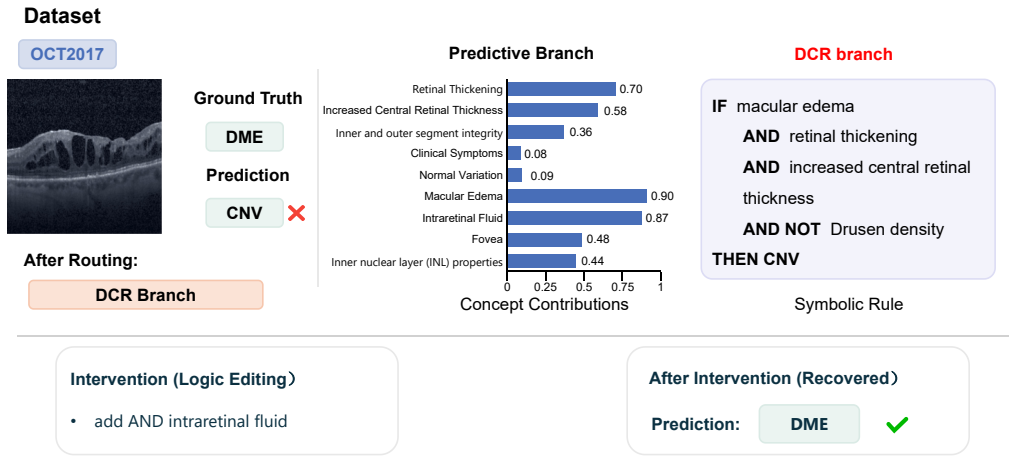

Figure 9: Rule-level intervention example.

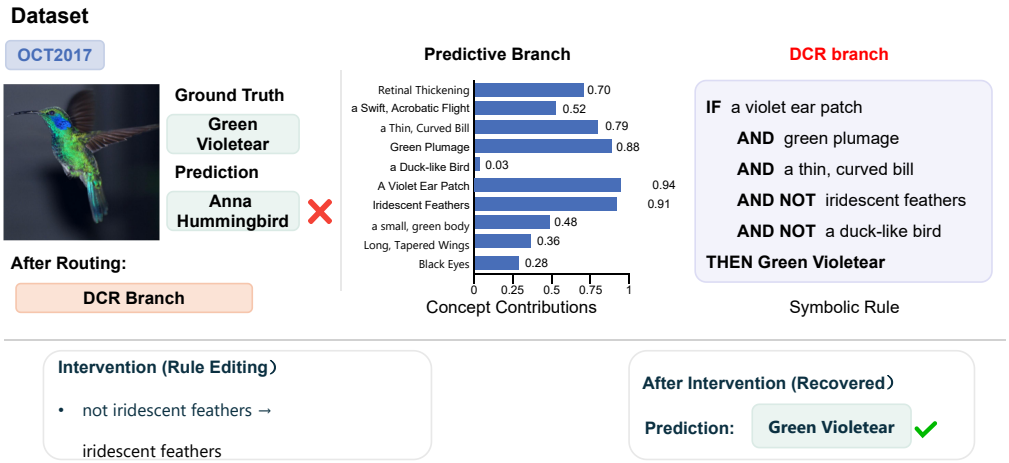

Figure 10: Rule-level intervention example.

R Additional Dataset Visualizations

This appendix provides additional visualization examples across the remaining three datasets: COVID-CT, CUB, and BloodMNIST. Each example is presented in the same format as the main paper, illustrating how the model routes to the appropriate branch based on predictive uncertainty, together with the corresponding concept scores and symbolic reasoning.

Figure 11 shows a correctly predicted COVID-CT sample. Figure 12 shows a correctly predicted CUB sample. Figures 13 and 14 show two correctly predicted BloodMNIST samples.

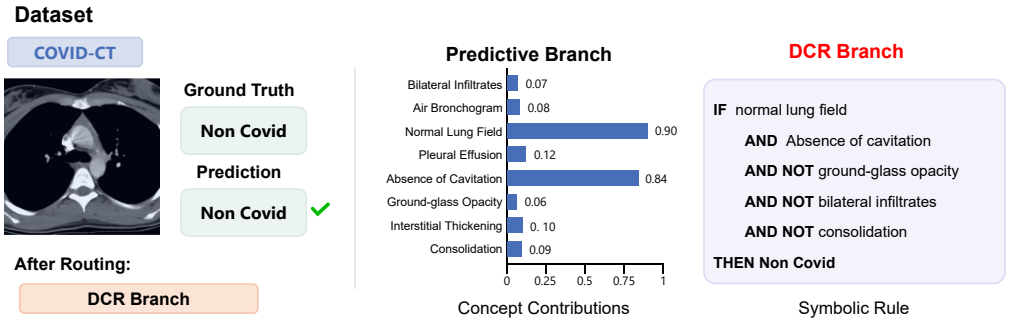

Figure 11: Correctly predicted COVID-CT sample.

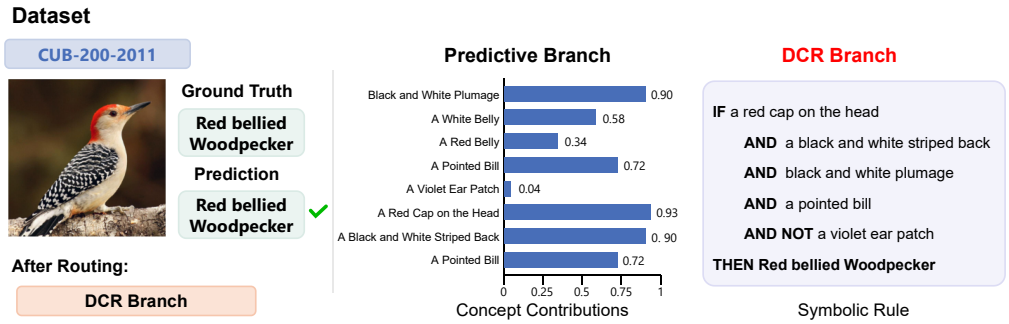

Figure 12: Correctly predicted CUB sample.

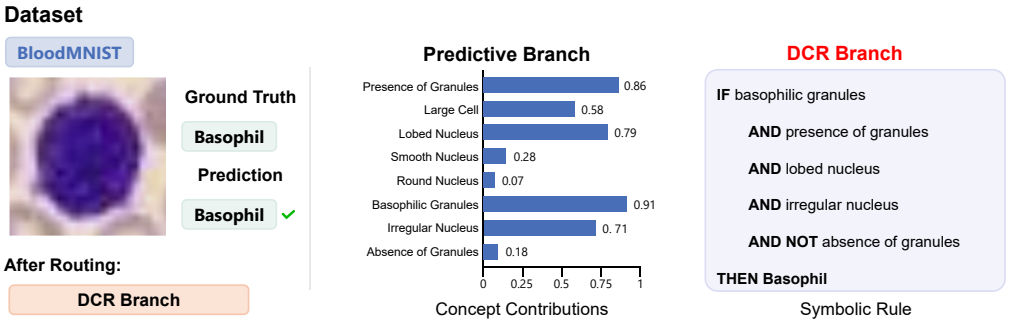

Figure 13: Correctly predicted BloodMNIST sample (example 1).

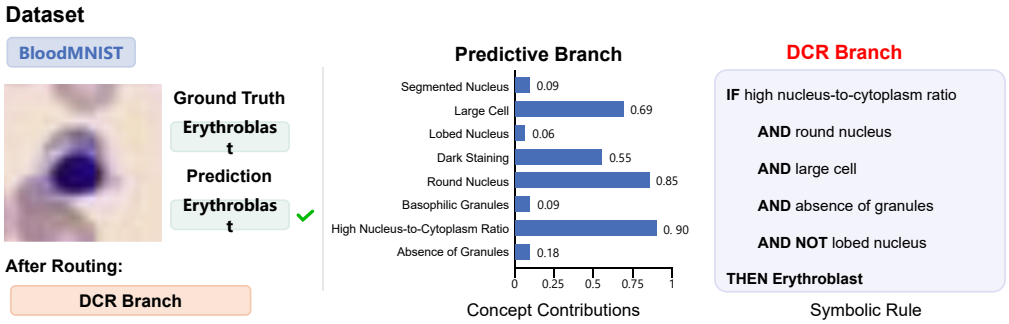

Figure 14: Correctly predicted BloodMNIST sample (example 2).

S Calibration of the DCR Branch

To assess whether the confidence estimates underlying entropy-based routing are reasonably calibrated, we evaluate the DCR branch using top-label Expected Calibration Error (ECE) (Guo et al., 2017). ECE is computed from the same DCR class probabilities used for routing:

$$p_{\mathrm{dcr}}(x) = \mathrm{softmax}(s_{\mathrm{dcr}}(x)), \qquad \mathrm{conf}(x) = \max_j p_{\mathrm{dcr},j}(x),$$

where $s_{\mathrm{dcr}}(x)$ is defined in Eq. (8). These probabilities are evaluated at temperature 1, without applying the distillation temperature. We use $M = 15$ equal-width bins on $[0, 1]$, weighted by their sample proportions, with empty bins contributing zero. ECE is computed on all examples in the reported split, irrespective of routing selection.

Table 22 reports the ECE across all datasets. The reported ECE values characterize marginal DCR calibration in the listed validation/test settings. They do not by themselves establish that entropy identifies the subset on which DCR outperforms the predictive branch; the coverage-matched routing analysis in Appendix V tests that comparative question.

Table 22: Expected Calibration Error (ECE) of the DCR branch across datasets.

Dataset	Split	ECE (%)
BloodMNIST	test	1.26
OCT2017	test	1.10
COVID-CT	test	4.98
CUB	val	2.46
Chest X-Ray	test	4.21
Fitzpatrick17k	test	3.75

T Additional Analysis of CLIP-Based Concept Supervision

T.1 CLIP Threshold Sensitivity

We perform a sensitivity analysis on the COVID-CT validation set to determine the optimal threshold δ for CLIP-based concept pseudo-labels. Table 23 shows that $\delta = 0.25$ yields the best downstream accuracy (92.06%), which is used throughout the paper.

Table 23: CLIP pseudo-label threshold δ sensitivity analysis on COVID-CT validation set. $\delta = 0.25$ yields the best downstream accuracy and is used throughout the paper.

δ	Concept Activation (%)	Avg. Concepts per Image	Task Acc. (%)
0.10	68.3	10.9	90.77
0.15	51.2	8.2	91.01
0.20	42.2	6.8	91.52
0.25	**35.8**	**5.7**	**92.06**
0.30	22.5	3.6	91.74
0.35	18.1	2.9	91.02

T.2 Task Accuracy: General CLIP vs. BiomedCLIP

We compare downstream task accuracy when concept pseudo-labels are produced by General CLIP (Radford et al., 2021) versus BiomedCLIP (Zhang et al., 2023), a medical-domain vision–language model. For each dataset, both settings use the same five random seeds, data splits, model architecture, and training protocol, only the pseudo-labeling model differs. We report mean test accuracy over the five runs.

The effect of the pseudo-labeling model varies across datasets. BiomedCLIP yields higher accuracy on COVID-CT and Chest X-Ray, with gains of 0.21 and 0.98 percentage points, respectively. General CLIP yields higher accuracy on OCT2017, BloodMNIST, and Fitzpatrick17k, with differences of 0.07, 0.24, and 2.08 percentage points, respectively. On CUB-200-2011, General CLIP exceeds BiomedCLIP by 6.44 percentage points under the evaluated setting. Overall, BiomedCLIP does not consistently improve downstream task accuracy across the evaluated medical datasets, and the largest observed difference occurs on CUB-200-2011. Table 24 presents the detailed results.

Table 24: Task accuracy comparison between General CLIP and BiomedCLIP on the test sets, averaged over five runs. Δ denotes BiomedCLIP accuracy minus General CLIP accuracy in percentage points.

Dataset	General CLIP Acc. (%)	BiomedCLIP Acc. (%)	Δ (%)
COVID-CT	92.06	92.27	+0.21
OCT2017	99.79	99.72	-0.07
BloodMNIST	97.87	97.63	-0.24
CUB	79.41	72.97	-6.44
Chest X-Ray	78.21	79.19	+0.98
Fitzpatrick17k	81.88	79.80	-2.08

T.3 Word Cloud Visualization: General CLIP vs. BiomedCLIP

To further validate the quality of automatically induced concepts, we compare word clouds generated by General CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2023), a medical-domain vision-language model, on the same concept set. The results show that the concept rankings and activation frequencies are broadly similar between the two models, indicating that General CLIP already provides sufficiently reliable signals for medical concept induction. Note that this analysis is conducted on COVID-CT, OCT2017, and BloodMNIST, with the word clouds shown in Figure 15, Figure 16, and Figure 17. CUB is excluded as it is a natural-image benchmark where medical-domain CLIP offers no expected advantage.

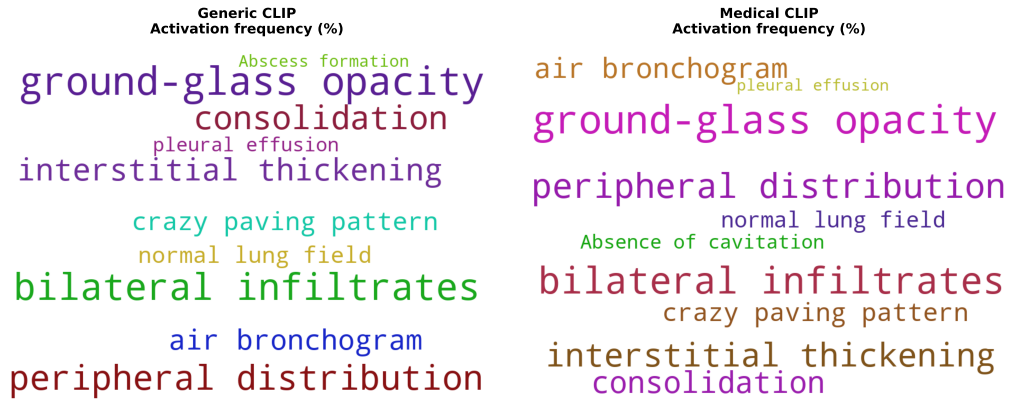

Figure 15: COVID-CT top-10 concept word clouds ($\delta = 0.25$). Left: General CLIP. Right: BiomedCLIP. BiomedCLIP shows higher activation for clinically critical features such as bilateral infiltrates and peripheral distribution.

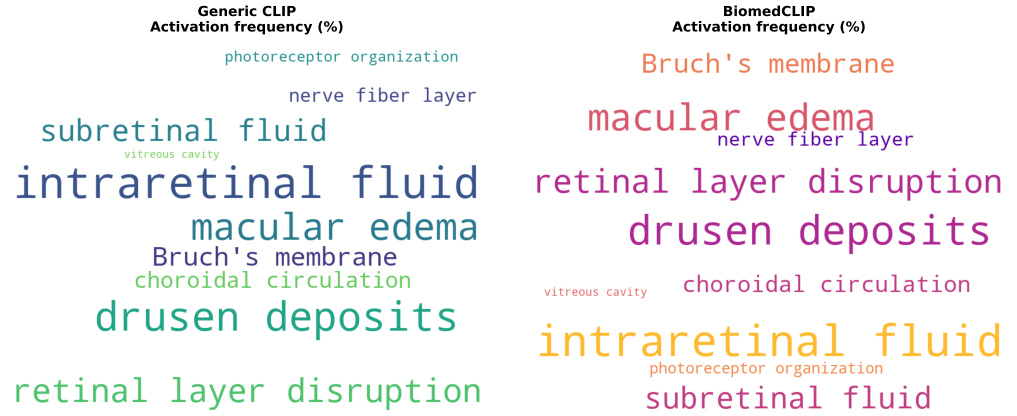

Figure 16: OCT2017 top-10 concept word clouds ($\delta = 0.25$). Left: General CLIP. Right: BiomedCLIP. BiomedCLIP assigns higher relative weight to disease-specific features such as intraretinal fluid and macular edema.

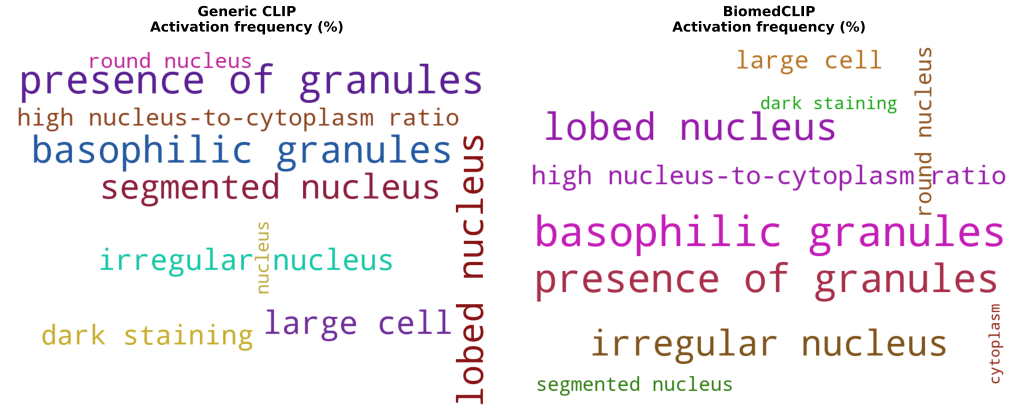

Figure 17: BloodMNIST top-10 concept word clouds ($\delta = 0.25$). Left: General CLIP. Right: BiomedCLIP. BiomedCLIP better captures cell-type specific morphological features such as basophilic granules and irregular nucleus.

U Robustness to Input Perturbations

To evaluate entropy-based routing under a controlled input corruption, we inject additive Gaussian noise with standard deviation σ into *preprocessed* test inputs for COVID-CT and BloodMNIST. We report predictive-branch accuracy, DCR accuracy, the fraction routed to DCR, and final routed accuracy. OCT2017 is omitted because accuracy is near-saturated within the evaluated σ range, and CUB is excluded as a natural-image benchmark. This synthetic corruption test is not a substitute for site or modality shift evaluation.

Across both datasets, DCR usage decreases as σ increases, while routed accuracy remains above DCR and tracks the predictive baseline under the tested corruptions. This suggests that the entropy score reduces symbolic usage as these perturbations grow; it does not establish robustness under clinically realistic distribution shifts. Tables 25 and 26 report the details.

Table 25: COVID-CT results under Gaussian noise. DCR usage decreases from 89.1% to 62.6%; routed accuracy exceeds DCR and stays within 1.5% of the predictive branch at $\sigma = 0.10$.

σ	Pred. (%)	DCR (%)	DCR Usage (%)	Routed (%)
0.00	92.19	91.66	89.10	92.06
0.02	91.50	90.88	86.53	91.42
0.04	91.27	90.48	82.79	90.98
0.06	90.45	88.67	76.65	89.95
0.08	88.50	84.50	71.12	87.29
0.10	80.78	74.52	62.56	79.27

Table 26: BloodMNIST results under Gaussian noise. DCR usage decreases from 96.2% to 64.0%; routed accuracy remains above DCR and stays within 2.4% of the predictive branch at $\sigma = 0.10$.

σ	Pred. (%)	DCR (%)	DCR Usage (%)	Routed (%)
0.00	97.90	97.78	96.17	97.87
0.02	96.22	96.01	92.38	96.10
0.04	94.89	93.92	87.26	94.23
0.06	92.66	90.48	81.78	91.05
0.08	89.52	86.11	73.23	87.73
0.10	84.57	80.53	64.02	82.21

V Coverage-Matched Routing Analysis

The main routing analysis varies the entropy threshold τ, thereby changing the fraction of examples assigned to DCR. We therefore perform an additional coverage-matched analysis to isolate the quality of the routing criterion itself. Specifically, we ask whether DCR entropy identifies a more useful subset for symbolic prediction than alternative confidence measures when all selectors are constrained to use DCR on the same fraction of test examples.

Protocol. For each dataset and random seed, we fix the validation-selected checkpoint and evaluate both prediction branches on the same test set. The resulting branch predictions are held fixed across all selectors and target coverage levels. We report the mean routed accuracy over five random seeds, using the same fixed checkpoint across selectors and coverage levels within each seed.

For a target symbolic coverage α, let $k_\alpha = \lfloor \alpha N \rfloor$. Each selector ranks the test examples according to a score $s(x)$, and the top k_α examples are assigned to DCR. The remaining examples use the predictive branch:

$$\hat{y}_{\alpha,s}(x_i) = \begin{cases} \hat{y}_{\mathrm{dcr}}(x_i), & x_i \in S_{\alpha,s}, \\ \hat{y}_{\mathrm{pred}}(x_i), & x_i \notin S_{\alpha,s}, \end{cases} \tag{26}$$

where $S_{\alpha,s}$ contains the k_α highest-ranked examples under selector s. This construction ensures identical symbolic coverage across selectors within each run.

Table 27: Coverage-matched routing accuracy at 50% and 80% symbolic coverage. Within each run, all selectors and coverage levels use the same fixed checkpoint and branch predictions. Bold indicates the best confidence-based selector; Random is included as a coverage-matched reference, and the retrospective label-dependent Oracle provides an upper bound on routed accuracy under the same coverage constraint.

Dataset	Selector	Acc.@50 (%) ↑	Acc.@80 (%) ↑
BloodMNIST	Random	97.84	97.80
	MSP	97.84	97.81
	Margin	97.86	**97.83**
	Entropy	**97.87**	97.74
	Oracle	98.07	98.07
OCT2017	Random	99.73	99.69
	MSP	99.74	99.71
	Margin	**99.77**	99.73
	Entropy	**99.77**	**99.75**
	Oracle	99.89	99.89
COVID-CT	Random	91.93	91.77
	MSP	**92.13**	**91.82**
	Margin	**92.13**	**91.82**
	Entropy	**92.13**	**91.82**
	Oracle	93.31	93.31
CUB	Random	79.53	79.10
	MSP	79.70	78.43
	Margin	79.82	78.58
	Entropy	**79.91**	**78.68**
	Oracle	82.18	82.18
Chest X-Ray	Random	76.91	76.10
	MSP	77.31	77.54
	Margin	**77.64**	77.78
	Entropy	77.60	**78.19**
	Oracle	80.26	80.26
Fitzpatrick17k	Random	80.87	80.30
	MSP	81.18	81.25
	Margin	81.32	81.43
	Entropy	**81.46**	**81.61**
	Oracle	83.52	83.52

We compare DCR entropy with two alternative confidence measures computed from the DCR class distribution: maximum softmax probability (MSP) and the top-two probability margin. For entropy, we use $s_{\mathrm{ent}}(x) = -H_{\mathrm{dcr}}(x)$ so that larger values consistently indicate greater confidence for all three selectors. Random routing provides a coverage-matched reference independent of DCR confidence.

We additionally report a retrospective label-dependent oracle. At each target coverage, the oracle maximizes routed accuracy using the fixed branch predictions and ground-truth labels. It first assigns DCR to examples where only DCR is correct, then fills the remaining coverage with examples on which the two branches have the same correctness. Examples where only the predictive branch is correct are assigned to DCR only when required by the coverage constraint. The oracle is not deployable and is included only to quantify the headroom available to an ideal branch selector.

Results. Table 27 shows that the advantage of entropy-based routing is not explained solely by assigning a particular fraction of examples to the symbolic branch. Among the confidence-based selectors, DCR entropy achieves the highest or tied accuracy at most of the reported operating points, while margin performs slightly better on BloodMNIST at 80% coverage and Chest X-Ray at 50% coverage. On the near-saturated BloodMNIST and OCT2017 benchmarks, differences among confidence-based selectors are small.

For the binary COVID-CT task, entropy, MSP, and top-two margin induce the same ordering of examples up to ties. With consistent tie handling, they therefore yield identical coverage-matched results. On the multiclass benchmarks, the differences among selectors are more visible. In particular, entropy obtains 78.19% accuracy at 80% symbolic coverage on Chest X-Ray and 81.61% on Fitzpatrick17k, compared with 77.78% and 81.43% for margin, respectively. On CUB, entropy-based routed accuracy decreases from 79.91% at 50% symbolic coverage to 78.68% at 80% coverage.

Random routing outperforms entropy-based routing at some high-coverage operating points, including BloodMNIST and CUB at 80% coverage. This indicates that DCR confidence does not perfectly identify the examples on which the symbolic branch is preferable.

The retrospective oracle achieves the same accuracy at 50% and 80% symbolic coverage across all six datasets. Under the fixed-checkpoint evaluation, the additional symbolic assignments can be drawn from examples on which both branches have the same correctness, without introducing harmful switches. At both coverage levels, the oracle remains above all deployable selectors, indicating remaining headroom for improved branch selection.

Together, these results support DCR entropy as a useful, but imperfect, criterion for selective symbolic prediction.

W Displayed-Rule Replay Fidelity

Although the DCR branch executes an explicit fuzzy rule over the predicted concept state, the explanation presented to users is a discrete conjunction of positive and negated concept literals. We therefore evaluate whether this displayed rule preserves the decision behavior of the underlying soft DCR reasoner.

Rule replay. For an input x, let $a_{\mathrm{dcr}}(x)$ denote the original DCR class scores and $\hat{y}_{\mathrm{dcr}}(x)$ the corresponding prediction. For each class j, we extract the same discrete rule representation used for visualization,

$$R_j(x) = \{(k, s_{jk}) : k \in S_j(x)\}, \tag{27}$$

where $S_j(x)$ contains the displayed antecedents and $s_{jk} \in \{+1, -1\}$ denotes whether concept k appears as a positive or negated literal.

Replay uses only the information exposed by the displayed rule and the predicted concept truth values. The truth value of each displayed literal is

$$r_{jk}(x) = \begin{cases} c_{\theta,k}(x), & s_{jk} = +1, \\ 1 - c_{\theta,k}(x), & s_{jk} = -1. \end{cases} \tag{28}$$

The displayed antecedents are then recombined using the same fuzzy conjunction operator as the DCR branch:

$$a_j^{\mathrm{replay}}(x) = \bigwedge_{k \in S_j(x)} r_{jk}(x), \tag{29}$$

with the replay prediction defined as

$$\hat{y}_{\mathrm{replay}}(x) = \arg\max_j a_j^{\mathrm{replay}}(x). \tag{30}$$

The replay intentionally omits continuous internal quantities that are not visible in the displayed discrete rule. It therefore measures the fidelity of the presented symbolic explanation rather than exact numerical reproduction of the internal soft reasoner.

Metrics. We first report overall prediction agreement,

$$\mathrm{Agreement} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_{\mathrm{replay}}(x_i) = \hat{y}_{\mathrm{dcr}}(x_i)\right]. \tag{31}$$

Because CLAIR-Med uses DCR selectively, we additionally report agreement on the subset assigned to the symbolic branch by the validation-selected router:

$$\mathrm{RoutedAgreement} = \frac{\sum_i \mathbb{I}[H_{\mathrm{dcr}}(x_i) < \tau]\, \mathbb{I}[\hat{y}_{\mathrm{replay}}(x_i) = \hat{y}_{\mathrm{dcr}}(x_i)]}{\sum_i \mathbb{I}[H_{\mathrm{dcr}}(x_i) < \tau]}. \tag{32}$$

Table 28: Displayed-rule replay fidelity. Agreement measures whether replaying the displayed discrete rule reproduces the prediction of the original soft DCR reasoner. Routed Agreement is computed only on examples assigned to DCR by the validation-selected router. Score Rank Corr. reports the class-averaged Spearman correlation between original and replayed DCR scores across test examples.

Dataset	Agreement (%) ↑	Routed Agreement (%) ↑	Score Rank Corr. ↑
BloodMNIST	96.73	97.81	0.952
OCT2017	98.42	98.52	0.968
COVID-CT	95.67	96.36	0.941
CUB	90.13	92.03	0.881
Chest X-Ray	91.87	93.37	0.906
Fitzpatrick17k	89.43	91.05	0.868

The routing decision is always computed from the original soft DCR output rather than from the replayed rule.

Finally, we measure whether replay preserves the ordering of DCR scores rather than only the final argmax decision. For each class j, we compute the Spearman rank correlation between the original and replayed class scores across the test set and average across classes:

$$\rho_{\text{rank}} = \frac{1}{C}\sum_{j=1}^{C} \text{Spearman}\left(\{a_{\text{dcr},ij}\}_{i=1}^{N}, \{a_{ij}^{\text{replay}}\}_{i=1}^{N}\right). \tag{33}$$

Results. Table 28 shows that the displayed discrete rules preserve most of the decision behavior of the underlying soft DCR reasoner. Overall prediction agreement ranges from 89.43% on Fitzpatrick17k to 98.42% on OCT2017. Agreement is consistently higher on the subset assigned to DCR by the validation-selected router, ranging from 91.05% to 98.52%.

The difference between overall and routed agreement is small on OCT2017, where the selected operating point assigns nearly all examples to the symbolic branch, and becomes more pronounced on CUB, Chest X-Ray, and Fitzpatrick17k. The score-level analysis shows the same overall pattern: replayed rules retain strong rank agreement with the original DCR scores, with class-averaged Spearman correlations ranging from 0.868 to 0.968 across the six benchmarks.

These results indicate that the displayed discrete rules retain a substantial fraction of the decision structure of the underlying soft DCR reasoner, particularly on examples for which the symbolic branch is selected. At the same time, the non-zero replay discrepancy makes the effect of discretization explicit. Rule replay therefore supports the displayed rule as a faithful approximation of the DCR decision process, rather than as an exact reconstruction of its continuous internal computation. This computational agreement does not establish that the underlying concepts or rule semantics are clinically correct.

X Operational Metrics and Supporting Analysis

We condition on a fitted model and any validation-selected threshold. All expectations below are taken under a specified evaluation distribution of (X, Y); all score maps and selectors are measurable, class scores are finite, $C \geq 2$, and ties use a fixed deterministic rule. Logarithms are natural. These are distribution-level statements, not generalization guarantees inferred from a finite training loss. For completeness, we make the main evaluation quantities explicit. Let the two branch predictions be

$$\hat{y}_{\text{pred}}(x) = \arg\max_j a_{\text{pred},j}(x), \qquad \hat{y}_{\text{dcr}}(x) = \arg\max_j s_{\text{dcr},j}(x). \tag{34}$$

where $s_{\text{dcr}}(x)$ denotes the element-wise log-odds of the clipped DCR truth scores, as defined in Eq. (8), and ties are resolved by a fixed deterministic rule. For $b \in \{\text{pred}, \text{dcr}\}$, define $R_b = \mathbb{E}[\ell(\hat{y}_b(X), Y)]$. Let $\text{Acc}(\cdot)$ denote accuracy. The branch-wise utility gap is

$$\text{Gap} = \text{Acc}(\hat{y}_{\text{pred}}) - \text{Acc}(\hat{y}_{\text{dcr}}). \tag{35}$$

For routing threshold τ, define the symbolic coverage using $p_{\mathrm{dcr}}(x) = \mathrm{softmax}(s_{\mathrm{dcr}}(x))$ and $H_{\mathrm{dcr}}(x) = -\sum_j p_{\mathrm{dcr},j}(x) \log p_{\mathrm{dcr},j}(x)$ from Section 3:

$$\alpha(\tau) = \mathbb{P}(H_{\mathrm{dcr}}(X) < \tau). \tag{36}$$

The routed predictor is

$$\hat{y}_\tau(x) = \begin{cases} \hat{y}_{\mathrm{dcr}}(x), & H_{\mathrm{dcr}}(x) < \tau, \\ \hat{y}_{\mathrm{pred}}(x), & \text{otherwise}, \end{cases} \tag{37}$$

and the routed risk is $R_{\mathrm{route}}(\tau) = \mathbb{E}[\ell(\hat{y}_\tau(X), Y)]$.

The following results are supporting analyses of the training objective and inference rule. They delimit what can be concluded from distillation, routing, and concept editing; they are not presented as optimization, calibration, or clinical-safety guarantees.

X.1 Theoretical perspective on one-way alignment

Let

$$p(x) = \mathrm{softmax}(a_{\mathrm{pred}}(x)/T), \qquad q(x) = \mathrm{softmax}(s_{\mathrm{dcr}}(x)/T) \tag{38}$$

denote the softened predictive and symbolic class distributions for $T > 0$. Temperature scaling preserves score order, so $\arg\max_j p_j(x) = \hat{y}_{\mathrm{pred}}(x)$ and $\arg\max_j q_j(x) = \hat{y}_{\mathrm{dcr}}(x)$. The empirical distillation loss encourages alignment on training samples. The following bounds instead assume control of the expected KL under the specified evaluation distribution; that assumption is not implied by minimizing the empirical objective alone.

Proposition X.1 (Distillation controls branch discrepancy)**.** *For a finite $\varepsilon \geq 0$, if*

$$\mathbb{E}_X \, \mathrm{KL}(p(X) \| q(X)) \leq \varepsilon \tag{39}$$

then

$$\mathbb{E}_X \, \|p(X) - q(X)\|_1 \leq \sqrt{2\varepsilon}. \tag{40}$$

Proof. Pinsker's inequality gives

$$\|p(x) - q(x)\|_1 \leq \sqrt{2 \, \mathrm{KL}(p(x) \| q(x))} \tag{41}$$

for every x. Taking expectations and using Jensen's inequality for the concave square-root function yields

$$\begin{aligned} \mathbb{E}_X \, \|p(X) - q(X)\|_1 &\leq \mathbb{E}_X \sqrt{2 \, \mathrm{KL}(p(X) \| q(X))} && (42) \\ &\leq \sqrt{2 \, \mathbb{E}_X \, \mathrm{KL}(p(X) \| q(X))} && (43) \\ &\leq \sqrt{2\varepsilon}. && (44) \end{aligned}$$

□

Corollary X.1 (Margin preservation under small branch discrepancy)**.** *Let $y^\star(x) = \arg\max_j p_j(x)$ and define the predictive top-two margin*

$$m(x) = p_{y^\star(x)}(x) - \max_{j \neq y^\star(x)} p_j(x). \tag{45}$$

If $\|p(x) - q(x)\|_1 < m(x)$, then $\arg\max_j q_j(x) = y^\star(x)$.

Proof. Because $p(x)$ and $q(x)$ are probability vectors, the positive and negative parts of $p(x) - q(x)$ have the same total mass. Hence

$$\delta = \|p(x) - q(x)\|_\infty \leq \tfrac{1}{2} \|p(x) - q(x)\|_1 < m(x)/2. \tag{46}$$

Moreover,

$$q_{y^\star}(x) \geq p_{y^\star}(x) - \delta, \qquad q_j(x) \leq p_j(x) + \delta \quad \forall j \neq y^\star. \tag{47}$$

Therefore, for every $j \neq y^\star$,

$$q_{y^\star}(x) - q_j(x) \geq p_{y^\star}(x) - p_j(x) - 2\delta > 0, \tag{48}$$

which preserves the argmax. □

Corollary X.2 (Expected KL controls branch disagreement)**.** *Under the hypothesis of Proposition X.1, for every $\gamma > 0$,*

$$\mathbb{P}\left(\arg\max_j p_j(X) \neq \arg\max_j q_j(X)\right) \leq \mathbb{P}(m(X) \leq \gamma) + \frac{2\varepsilon}{\gamma^2}. \tag{49}$$

Proof. On the event $m(X) > \gamma$, disagreement implies, by contraposition of the preceding corollary, that $\|p(X) - q(X)\|_1 \geq m(X) > \gamma$. Pinsker's inequality then implies $\mathrm{KL}(p(X)\|q(X)) > \gamma^2/2$. A union bound over the low-margin event and Markov's inequality give

$$\mathbb{P}(\arg\max p \neq \arg\max q) \leq \mathbb{P}(m \leq \gamma) + \mathbb{P}\big(\mathrm{KL}(p\|q) > \gamma^2/2\big) \tag{50}$$

$$\leq \mathbb{P}(m \leq \gamma) + 2\varepsilon/\gamma^2. \tag{51}$$

□

Remark X.1. *These results do not claim that symbolic accuracy must match predictive accuracy. They formalize a weaker statement: a small expected one-way KL bounds distributional discrepancy, while decision disagreement is additionally controlled by the teacher's low-margin mass. Neither result implies concept correctness or rule faithfulness.*

X.2 Theoretical perspective on uncertainty-aware routing

Define the routing indicator

$$r_\tau(x) = \mathbb{I}[H_{\mathrm{dcr}}(x) < \tau], \tag{52}$$

and let $\ell(\hat{y}, y)$ denote the task loss for a prediction-label pair. Both branch losses are assumed integrable throughout this subsection; bounded loss is imposed explicitly where needed.

Proposition X.2 (Risk decomposition of routed inference)**.** *Suppose both branch losses are integrable and $0 < \alpha(\tau) < 1$, and let*

$$\alpha(\tau) = \mathbb{P}(r_\tau(X) = 1), \tag{53}$$

$$R^{<}_{\mathrm{dcr}}(\tau) = \mathbb{E}\big[\ell(\hat{y}_{\mathrm{dcr}}(X), Y) \mid r_\tau(X) = 1\big], \tag{54}$$

$$R^{\geq}_{\mathrm{pred}}(\tau) = \mathbb{E}\big[\ell(\hat{y}_{\mathrm{pred}}(X), Y) \mid r_\tau(X) = 0\big]. \tag{55}$$

Then the routed risk satisfies

$$R_{\mathrm{route}}(\tau) = \alpha(\tau) R^{<}_{\mathrm{dcr}}(\tau) + \big(1 - \alpha(\tau)\big) R^{\geq}_{\mathrm{pred}}(\tau). \tag{56}$$

Proof. By definition,

$$\ell(\hat{y}_\tau(X), Y) = r_\tau(X)\, \ell(\hat{y}_{\mathrm{dcr}}(X), Y) + \big(1 - r_\tau(X)\big)\, \ell(\hat{y}_{\mathrm{pred}}(X), Y). \tag{57}$$

Taking expectations and conditioning on whether $r_\tau(X) = 1$ or 0 gives

$$R_{\mathrm{route}}(\tau) = \mathbb{E}\big[r_\tau(X)\, \ell(\hat{y}_{\mathrm{dcr}}(X), Y)\big] + \mathbb{E}\big[(1 - r_\tau(X))\, \ell(\hat{y}_{\mathrm{pred}}(X), Y)\big] \tag{58}$$

$$= \alpha(\tau) R^{<}_{\mathrm{dcr}}(\tau) + \big(1 - \alpha(\tau)\big) R^{\geq}_{\mathrm{pred}}(\tau). \tag{59}$$

□

At the boundary values $\alpha(\tau) \in \{0, 1\}$, the conditional risk on the zero-probability branch is not defined; the unconditional identity in the first line of the proof remains valid and reduces to the risk of the single selected branch.

Corollary X.3 (When routing improves over always-predictive inference)**.** *For every τ,*

$$R_{\mathrm{route}}(\tau) - R_{\mathrm{pred}} = \mathbb{E}\big[r_\tau(X)\big(\ell(\hat{y}_{\mathrm{dcr}}(X), Y) - \ell(\hat{y}_{\mathrm{pred}}(X), Y)\big)\big]. \tag{60}$$

If $\alpha(\tau) > 0$, define

$$R^{<}_b(\tau) = \mathbb{E}\big[\ell(\hat{y}_b(X), Y) \mid r_\tau(X) = 1\big], \qquad b \in \{\mathrm{pred}, \mathrm{dcr}\}. \tag{61}$$

Then the difference is also

$$R_{\mathrm{route}}(\tau) - R_{\mathrm{pred}} = \alpha(\tau)\Big(R^{<}_{\mathrm{dcr}}(\tau) - R^{<}_{\mathrm{pred}}(\tau)\Big). \tag{62}$$

Whenever $\alpha(\tau) > 0$, routing improves over always using the predictive branch exactly when the average loss difference on accepted examples is negative; when $\alpha(\tau) = 0$, the risk difference is zero.

Proof. Subtract the always-predictive loss from the pointwise routed-loss identity in the proof above and take expectations. Conditioning the resulting expectation on $r_\tau(X) = 1$ gives the second equality whenever $\alpha(\tau) > 0$. □

Proposition X.3 (Regret of a branch selector relative to the oracle selector)**.** *Assume* $0 \le \ell \le B$ *and define the conditional branch risks*

$$L_b(x) = \mathbb{E}[\ell(\hat{y}_b(X), Y) \mid X = x], \qquad b \in \{\mathrm{pred}, \mathrm{dcr}\}. \tag{63}$$

Let $r^\star(x) = \mathbb{I}[L_{\mathrm{dcr}}(x) \le L_{\mathrm{pred}}(x)]$ *be the oracle branch selector. For any measurable selector* $r : \mathcal{X} \to \{0, 1\}$ *on the input space* $\mathcal{X}$*, with* $r = 1$ *denoting DCR, define*

$$R(r) = \mathbb{E}[r(X)\ell(\hat{y}_{\mathrm{dcr}}(X), Y) + (1 - r(X))\ell(\hat{y}_{\mathrm{pred}}(X), Y)]. \tag{64}$$

Its excess risk over the oracle is

$$R(r) - R(r^\star) = \mathbb{E}[|L_{\mathrm{dcr}}(X) - L_{\mathrm{pred}}(X)|\, \mathbb{I}[r(X) \ne r^\star(X)]] \le B\, \mathbb{P}(r(X) \ne r^\star(X)). \tag{65}$$

Proof. Condition on $X = x$. If $L_{\mathrm{dcr}}(x) \le L_{\mathrm{pred}}(x)$, the oracle selects DCR and a mismatch incurs excess conditional risk $L_{\mathrm{pred}}(x) - L_{\mathrm{dcr}}(x)$. Otherwise it selects the predictive branch and a mismatch incurs $L_{\mathrm{dcr}}(x) - L_{\mathrm{pred}}(x)$. This yields the equality after averaging over X. The inequality follows because both conditional risks lie in $[0, B]$. □

Remark X.2. *These identities separate coverage from selection quality. The threshold* τ *controls coverage* $\alpha(\tau)$*, but entropy is a useful selector only to the extent that* r_τ *agrees with the lower-conditional-risk branch. The oracle* $r^\star$ *depends on unknown conditional risks and is not a deployable router. It is distinct from the empirical, label-dependent hindsight oracle in Appendix V, which additionally imposes a fixed coverage. Consequently, Eq. 65 does not directly characterize that coverage-constrained empirical oracle. The propositions therefore motivate empirical coverage–risk and oracle-regret analyses; they do not prove that low symbolic entropy is intrinsically reliable.*

X.3 Linking displayed-rule fidelity to routed risk

The replay metric can be connected to the risk of replacing the soft DCR prediction by the displayed discrete rule, while keeping the original router fixed. This replacement is an analysis device, not an additional experiment. Let $r : \mathcal{X} \to \{0, 1\}$ be the original selector and let $\hat{y}_{\mathrm{replay}}$ be defined as in Appendix W. For this general selector, $\hat{y}_r(x)$ returns $\hat{y}_{\mathrm{dcr}}(x)$ when $r(x) = 1$ and $\hat{y}_{\mathrm{pred}}(x)$ otherwise. Define

$$\hat{y}_{r,\mathrm{replay}}(x) = \begin{cases} \hat{y}_{\mathrm{replay}}(x), & r(x) = 1, \\ \hat{y}_{\mathrm{pred}}(x), & r(x) = 0, \end{cases} \tag{66}$$

and let $R_{\mathrm{replay}}(r)$ denote its expected task loss.

Proposition X.4 (Coverage-weighted replay discrepancy)**.** *Suppose* $0 \le \ell \le B < \infty$*. Write* $\alpha = \mathbb{P}(r(X) = 1)$*. If* $\alpha > 0$*, define the selected-subset fidelity*

$$\mathcal{F}_r = \mathbb{P}\big(\hat{y}_{\mathrm{replay}}(X) = \hat{y}_{\mathrm{dcr}}(X) \mid r(X) = 1\big). \tag{67}$$

Then

$$|R_{\mathrm{replay}}(r) - R(r)| \le B\alpha(1 - \mathcal{F}_r). \tag{68}$$

If $\alpha = 0$*, the risks are equal and no conditional fidelity is needed. The same inequality holds for empirical risks and empirical frequencies computed on the same evaluation sample with a nonempty selected subset.*

Proof. The two routed predictors coincide whenever $r(x) = 0$ or the replayed and original DCR predictions agree. For every (x, y), boundedness gives

$$|\ell(\hat{y}_{r,\mathrm{replay}}(x), y) - \ell(\hat{y}_r(x), y)| \le B\, r(x)\, \mathbb{I}[\hat{y}_{\mathrm{replay}}(x) \ne \hat{y}_{\mathrm{dcr}}(x)]. \tag{69}$$

Taking expectations and applying $|\mathbb{E}Z| \le \mathbb{E}|Z|$ yields the bound. When $\alpha = 0$, its right-hand event has probability zero. Replacing the expectation by the sample average proves the empirical statement without an independence assumption. □

Remark X.3. *For zero–one loss, $B = 1$ and the bound also limits the absolute change in accuracy. It motivates reporting coverage and selected-subset replay fidelity together, as in Table 2. Agreement does not imply correctness: both predictions may make the same error. The bound also neither validates concept semantics nor guarantees generalization from empirical fidelity. When averaging runs, the empirical bound must be computed per run before averaging; a product of separately averaged coverage and infidelity is not generally the corresponding bound.*

X.4 Theoretical perspective on concept-level intervention

Let $F(c, z)$ denote the score map of a fixed branch for image features z and concept state c. Parameters, textual embeddings, visual features, and the chosen branch are held fixed; the analysis does not include a router switch after editing. Assume that, in a local neighborhood of interest, F is L-Lipschitz with respect to the concept state:

$$\|F(c, z) - F(c', z)\|_2 \leq L\|c - c'\|_2. \tag{70}$$

Let $c^\star$ denote the corrected concept state for a sample.

Proposition X.5 (Bound on post-edit output deviation)**.** *Suppose c, c', and $c^\star$ belong to the neighborhood on which the Lipschitz condition holds. If an edit changes the concept state from c to c' and reduces the concept error relative to the corrected state, i.e.,*

$$\|c' - c^\star\|_2 \leq \rho\|c - c^\star\|_2 \tag{71}$$

for some $\rho \in [0, 1)$, then

$$\|F(c', z) - F(c^\star, z)\|_2 \leq \rho L\|c - c^\star\|_2. \tag{72}$$

Proof. Applying the Lipschitz condition with $(c', c^\star)$ gives

$$\|F(c', z) - F(c^\star, z)\|_2 \leq L\|c' - c^\star\|_2. \tag{73}$$

Substituting the assumed contraction of concept error yields the claim. □

Corollary X.4 (Explicit sensitivity of the predictive branch)**.** *Partition $W_{\text{pred}} = [W_z\ W_t]$ according to Eq. 4. For fixed z, the predictive score map is*

$$F_{\text{pred}}(c, z) = W_z z + W_t E^\top c + b_{\text{pred}}, \tag{74}$$

and is globally Lipschitz in c with

$$L_{\text{pred}} = \|W_t E^\top\|_2. \tag{75}$$

Consequently, Proposition X.5 applies to the predictive branch with this explicit constant.

Proof. For any c, c', the feature-only term cancels and

$$\|F_{\text{pred}}(c, z) - F_{\text{pred}}(c', z)\|_2 = \|W_t E^\top (c - c')\|_2 \leq \|W_t E^\top\|_2\, \|c - c'\|_2. \tag{76}$$

□

Remark X.4. *These results are sensitivity bounds on score vectors. They do not prove that an edit improves classification, contracts output error relative to the pre-edit state, identifies a clinically correct concept value, or produces a clinically valid intervention. For DCR, an analogous constant additionally depends on the chosen fuzzy operators and their parameterization.*