A. Appendix

A.1. Motivation of Using Agentic Framework: LLMs as Reasoning - Planning and Executing Engines

LLMs have emerged as powerful agents capable of solving multi step tasks across domains, including mathematical reasoning [31], tool usage [21, 25], robotic navigation and planning [2, 28], and interactive code generation [34]. Most contemporary LLM-based agents rely on *chain of thought* (CoT) prompting [31] to decompose problems into intermediate reasoning steps, interleaved with environment specific actions such as tool invocation or state transitions [37]. Extensions include feedback driven refinement [26], adaptive task decomposition [19], and explicit search over reasoning trajectories [36]. While highly effective, these architectures still face challenges in generalization, compositional reasoning, and decision making under uncertainty, motivating our design of causal and disentangled agents.

A.2. Knowledge Extraction for medical application

A central challenge is extracting the relevant expert knowledge required to guide model decisions. Such knowledge extraction often relies on domain experts, medical annotators, or specialized deep learning pipelines. For example, lesion- or structure-specific cues frequently require fine-tuned object-detection models such as YOLO [22], SAM [15], or Detectron2 [33] for lesion localization, or U-Net [24] for vessel segmentation, each typically trained on expert-annotated datasets of at least 500 images per knowledge attribute. This dependence on fine-grained, pixel-level, or bounding-box annotations limits the scalability of knowledge integration to broader medical tasks where such resources are scarce or prohibitively expensive. Therefore, a promising next direction is to develop an agentic framework called ***Agentic Causal Disen- tanglement (CANDICE)***, capable of autonomously constructing these pipelines, retrieving domain-specific cues., thereby significantly reducing manual human involvement.

B. Reinforcement Learning–Based Optimization

Beyond code refinement, we employ a reinforcement learning agent to tune key detection and preprocessing parameters (e.g., YOLO confidence thresholds, segmentation post-processing filters). When annotated validation data are available, the agent observes the current parameter value as the state, selects actions from $\{-0.05, 0, +0.05\}$, and receives the resulting IoU as the reward. Updates follow a standard Q-learning schedule with learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.9$. Additional implementation details are provided in final suplimentary materials.

C. Discussion of EchoNet-Dynamic system's Result

CANDICE's results do not surpass the hand-engineered EchoNet-Dynamic system, which benefits from extensive video-specific optimization and years of domain-tailored refinement. Importantly, CANDICE provides **substantial reductions in human intervention workload**. While we did not perform a formal measurement, the improvement in external MAE demonstrates enhanced domain generalization, and beat-level aggregation reduces the required human review effort by an estimated $\sim 70\%$.

D. Analysis and Discussion

In this section, we analyze the behavior of CANDICE beyond aggregate performance metrics and discuss the roles played by individual agents, the generality of the framework, and its relationship to prior reasoning paradigms such as Chain of Thought (CoT). Our goal is to clarify *why* CANDICE works, not merely *that* it works.

D.1. Importance of CGA Agent

The Code Generation Agent (CGA) plays a critical but often underappreciated role in the CANDICE framework. While the CRA and CDA are responsible for reasoning and planning, respectively, the CGA ensures that these abstract decisions are grounded in executable, verifiable computations. Our experiments show that this grounding is essential for both robustness and interpretability.

Without the CGA, reasoning outputs remain symbolic or textual artifacts that may appear coherent but fail silently when applied to real inputs. This failure mode is particularly problematic under domain shift, where assumptions encoded in reasoning chains may not hold. By contrast, the CGA enforces executability: every decision pathway selected by the CDA must correspond to a concrete program whose behavior can be observed and validated.

The tool use evaluation in Table 4 (main paper) highlights this effect quantitatively. Compared to single shot LLM code generation and Toolformer style agents, the CGA achieves higher final execution rates with fewer correction iterations. More importantly, execution failures become explicit signals that can be used by the CDA to revise plans, rather than latent errors that propagate unnoticed.

From a causal perspective, the CGA acts as a *grounding intervention*. It prevents the system from relying on spurious symbolic reasoning by forcing alignment between abstract knowledge and observable computation. This property is especially valuable in safety-critical or high stakes settings, but it is equally important for scientific validity: it enables precise error attribution and systematic debugging of agentic behavior.

D.2. Importance of CDA Agent

The Causal Disentanglement Agent (CDA) is the core decision making component of CANDICE and the primary source of performance gains observed across tasks. Ablation results in Table 3 (main paper) demonstrate that removing or simplifying the CDA leads to substantial degradation in Tail F1 and success rate, even when all other components are retained.

The key contribution of the CDA is not merely planning, but *selective intervention*. Prior agentic systems often apply reasoning or tool use uniformly across inputs, leading to unnecessary computation and increased error rates. The CDA instead learns to differentiate between inputs that benefit from knowledge intensive reasoning and those that are best handled by domain-invariant statistical models.

This selectivity is crucial for resolving the long standing conflict between domain generalization and long tailed learning. Head classes benefit from smooth, invariant decision boundaries, while tail classes require sharp, knowledge-guided distinctions. By dynamically routing inputs through different pathways, the CDA prevents these objectives from interfering with one another.

Conceptually, the CDA transforms the learning problem from a single global optimization into a collection of local, context dependent decisions. This perspective aligns with decision-theoretic views of intelligence and suggests that robustness under distribution shift may fundamentally require agentic control rather than monolithic predictors.

D.3. Model Agnostic Nature

An important property of CANDICE is its model agnostic nature. The framework does not assume a specific backbone architecture, modality, or training objective. Instead, it operates at the level of decision orchestration, making it compatible with a wide range of base models, including vision encoders, sequence models, and multimodal systems.

This property is empirically supported by the diversity of tasks evaluated in Section F.2. Despite substantial differences in input structure and label semantics, CANDICE consistently improves robustness and tail performance. These gains cannot be attributed to architectural specialization, but rather to the agentic CCKI principle that governs when and how knowledge is integrated.

From a practical standpoint, model agnosticism makes CANDICE easier to deploy and extend. Existing systems can be augmented with agentic causal disentanglement without retraining core models from scratch. From a scientific standpoint, it suggests that the benefits of CANDICE stem from structural properties of decision-making rather than domain-specific heuristics.

D.4. Constraints based CoT vs. CANDICE

Recent work has proposed constraining Chain of Thought (CoT) reasoning to improve reliability and reduce hallucination. While these approaches share superficial similarities with CANDICE, they differ fundamentally in scope and mechanism.

Constraints-based CoT methods operate within a single reasoning trace, enforcing syntactic or semantic validity of intermediate steps. They do not alter the underlying decision structure of the model, nor do they provide a mechanism for resolving conflicts between competing objectives such as robustness and tail sensitivity.

CANDICE, by contrast, treats reasoning as one component of a broader causal decision process. Reasoning outputs are not ends in themselves, but inputs to a planner (CDA) that decides whether, when, and how they should influence predictions. Moreover, CANDICE grounds reasoning through executable programs, something that CoT based methods do not address.

In this sense, CANDICE subsumes constrained CoT as a special case: reasoning can be constrained, but it is never unconditional. This distinction explains why CANDICE achieves consistent gains under domain shift, whereas CoT based methods often fail to generalize beyond their training distributions.

Method	Accuracy (%)	Reasoning F1	Decision Consistency
LLM (Zero Shot)	63.4	58.7	61.2
RAG + LLM	70.5	65.2	68.0
ReAct	72.9	67.8	70.1
CANDICE (Ours)	**78.3**	**74.5**	**77.6**

Table 1. Evaluation of CANDICE on a language-only clinical reasoning task.

D.5. Runtime: One-Time Agentic Construction vs. Deployment Inference (General, with a DR Example)

General view. Our framework separates runtime into (i) a *one-time* **agentic system construction** phase and (ii) an *application-dependent* **deployment inference** phase. The construction phase is dominated by the agent/HIRL loop that proposes candidate knowledge pipelines, validates them against a fixed DL-only reference using task-specific metrics, and retains only candidates that provide measurable gain. Its cost depends on the input modality and resolution, the target operating point (e.g., tail sensitivity vs. overall accuracy), the agent stopping criteria (budget, maximum iterations, improvement threshold ϵ), and the availability/quality of domain knowledge and knowledge-extraction tools. Importantly, candidates that do not improve over the DL-only baseline are discarded and do not contribute to deployment-time overhead.

Symptom	Key Observations and Diagnostic Relevance
Microaneurysms	Tiny red capillary dilations in the retina; the earliest sign of Mild NPDR. Their progression correlates with disease severity [10, 29, 32].
Haemorrhages	Includes dot/blot and flame-shaped types indicating microvascular leakage. Severe NPDR is marked by >20 hemorrhages in all quadrants; risk of PDR rises to $\sim$ 50% within a year [11, 17, 20, 29].
Hard Exudates	Lipid-rich deposits from chronic leakage, often in/near the macula. Indicative of risk for Diabetic Macular Edema (DME), a major cause of vision loss [11, 20, 27].
Cotton Wool Spots	Fluffy white retinal lesions caused by nerve fiber layer infarctions. Signify retinal ischemia in Moderate to Severe NPDR [10, 20, 27].
Subhyaloid Haemorrhages	Boat- or D-shaped hemorrhages between retina and hyaloid face, typically from ruptured neovascular vessels. Hallmark of Proliferative DR [9, 27, 35].
Neovascularization	Fragile vessel growth on optic disc (NVD) or retina (NVE). Defining trait of PDR. High-risk cases without treatment face $\sim$ 50% vision loss within 5 years [11, 17, 27].

Table 2. Clinical signs of DR and their diagnostic significance.

DR example (system construction). In our DR 5-stage grading instantiation, the agent explores knowledge pipelines built from multiple candidate biomarkers/diagnostic cues and selects a subset that improves target-domain tail metrics. In this setting, the agent required ≈ 25 iterations to converge to the final system. The one-time construction cost therefore includes: (i) training the ViT backbone on the source domain, (ii) fine-tuning the YOLOv11 biomarker detector used for knowledge extraction, and (iii) repeated evaluation of candidate knowledge cues during the agent loop. The wall-clock cost scales with the number of evaluated candidates per iteration and the frequency of re-training vs. re-using cached model checkpoints; however, this cost is amortized across all future inferences once the final pipeline is fixed.

Deployment inference cost (general form + DR example). After construction, deployment latency depends only on the *final selected pipeline* and the compute resources available. For a single input, we can express end-to-end latency as

$$T_{\text{sys}} = T_{\text{DL}} + p_{\text{KL}}\, T_{\text{KX}} + T_{\text{route}},$$

where T_{DL} is DL inference (e.g., ViT forward pass), T_{KX} is the inference time of the retained knowledge-extraction tool(s) (e.g., YOLOv11 detection plus feature post-processing), p_{KL} is the fraction of samples routed to the knowledge branch (coverage), and T_{route} is negligible. In the DR pipeline, T_{KX} is dominated by YOLOv11 biomarker detection at the chosen input resolution; for reference, official Ultralytics benchmarks report YOLO11 detection latency (TensorRT on an NVIDIA T4) of $\{1.5, 2.5, 4.7, 6.2, 11.3\}$ ms for {n,s,m,l,x} models at 640px input resolution, respectively. The overall deployment cost is therefore controlled by the selected YOLO11 variant, the image resolution, and the learned routing coverage p_{KL}, while the one-time agentic search overhead does not affect per-sample inference once the system is deployed.

E. Experimental Setup

E.1. Dataset Details

We categorize datasets as: **public** (e.g., EyePACS, APTOS, MESSIDOR, MESSIDOR2, EchoNet-Dynamic) and **restricted/private** (e.g., PCH/UNC rs-fMRI; MISC ECG). For restricted datasets, we provide: IRB/DUA status, de-identification summary, and a center-level datasheet describing acquisition differences.

Public Dataset Details: Messidor [1] is a DR dataset consisting of 1,200 images collected from three French clinical centers. The images were captured using a Topcon TRC-NW6 non-mydriatic retinograph with a 45-degree field of view.

The Indian Diabetic Retinopathy Image Dataset (IDRiD) [18] contains 516 color fundus images, divided into 413 training images and 103 testing images. Released as part of the ISBI 2018 Challenge, it is the first DR grading dataset representative of the Indian population.

APTOS [5] is another DR grading dataset representative of the Indian population. The retinal images were collected at the Aravind Eye Hospital and include 1,857 DR images and 1,805 non-DR images across five DR severity grades.

EyePACS [8] is a large-scale DR grading dataset consisting of high-resolution retinal images collected under diverse imaging conditions and is representative of the American population.

EchoNet-Dynamic [3] is a large-scale echocardiography dataset released by the Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI). The dataset consists of 10,030 apical four-chamber echocardiogram videos collected from 3,282 patients at Stanford University Medical Center. Each video is annotated with expert-derived measurements, including left ventricular ejection fraction (LVEF), as well as additional clinical labels. EchoNet-Dynamic is designed to support research in cardiac function assessment and video-based medical image analysis.

Private Dataset Details: PCH/UNC rs-fMRI Dataset. The PCH/UNC rs-fMRI dataset is a restricted, multi-center resting-state functional magnetic resonance imaging (rs-fMRI) dataset collected from Phoenix Children's Hospital (PCH) and the University of North Carolina (UNC). The dataset comprises rs-fMRI scans from pediatric patients with drug-resistant epilepsy and is used for seizure onset

zone (SOZ) localization. Imaging data were acquired under institution-specific protocols and scanners, introducing realistic cross-center domain shifts. All data are fully deidentified and accessed under Institutional Review Board (IRB) approval and data use agreements (DUA). In our experiments, models were trained on the PCH cohort and evaluated on the unseen UNC cohort to assess cross-center generalization without fine-tuning.

MISC ECG Dataset. The MISC ECG dataset is a restricted clinical electrocardiography dataset collected from the Mayo Integrated Stress Center (MISC). It consists of Exercise Stress Electrocardiogram (ESE) recordings acquired from patients undergoing clinical evaluation for coronary artery disease (CAD). The dataset includes multilead ECG signals captured during graded exercise protocols, along with expert clinical annotations. Due to patient privacy considerations, access to the dataset is governed by IRB approval and institutional data-sharing agreements. In this work, models were trained on ECG data from 726 subjects and evaluated on 227 held-out cases following the same split protocol as prior clinical studies.

E.1.1. Exact split tables and sample counts

For all experiment configurations, we adopt a standardized data-splitting protocol in which each dataset is partitioned into 60% training, 20% validation, and 20% testing subsets, with splits performed at the subject level to prevent subject leakage and stratified by class where applicable. For Single-Domain Generalization (SDG), models are trained on a single dataset (domain) and evaluated on held-out test splits from other datasets, with the validation split drawn from the training domain. For Multi-Domain Generalization (MDG), training data are pooled from multiple source domains using a unified data-loading library, and performance is evaluated on an unseen target domain that is excluded entirely from training and validation. e adopt DomainBed because it eliminates inconsistent split strategies, reduces evaluation bias, and enables fair comparison with prior domain generalization methods. By adhering to this standardized protocol, our results reflect true generalization performance rather than dataset-specific tuning To explicitly assess cross-center generalization in selected applications, models are trained on data from one clinical center and tested on data from a different center within the same dataset, without fine-tuning. Exact sample counts, per-class distributions, and center-wise splits for train/validation/test are reported in the corresponding application and dataset details.

Domain shift magnitude reporting. We report quantitative shift measures for each source–target pair using KL divergence, JS divergence, Hellinger distance, total variation (TV), and Wasserstein-1 distance (W1). Overall, shifts between APTOS and EyePACS are relatively low, indicating similar data distributions, while shifts involving Messidor-1 tend to be higher, particularly for KL and W1, reflecting more substantial distributional differences. Messidor-2 generally exhibits smaller shifts relative to APTOS and EyePACS but moderate shifts when compared to Messidor-1. These metrics highlight the asymmetric nature of domain shifts across DR datasets, which is important for assessing model generalization across sources and targets.

Table 3. Domain Shift Magnitude Between DR Datasets

Source→Target	KL	JS	Hellinger	TV	W1
APTOS→EyePACS	0.1541	0.0346	0.1875	0.2419	0.6009
APTOS→Messidor-1	1.6756	0.0584	0.2680	0.1849	0.2195
APTOS→Messidor-2	0.0828	0.0174	0.1334	0.1375	0.3565
EyePACS→APTOS	0.1313	0.0346	0.1875	0.2419	0.6009
EyePACS→Messidor-1	0.5968	0.0731	0.2829	0.2983	0.6996
EyePACS→Messidor-2	0.0636	0.0168	0.1302	0.1613	0.2467
Messidor-1→APTOS	0.2381	0.0584	0.2680	0.1849	0.2195
Messidor-1→EyePACS	0.3865	0.0731	0.2829	0.2983	0.6996
Messidor-1→Messidor-2	0.2371	0.0461	0.2260	0.1803	0.4529
Messidor-2→APTOS	0.0634	0.0174	0.1334	0.1375	0.3565
Messidor-2→EyePACS	0.0731	0.0168	0.1302	0.1613	0.2467
Messidor-2→Messidor-1	0.4764	0.0461	0.2260	0.1803	0.4529

E.1.2. Variance and significance

All tables will report mean$\pm$std over S seeds (default $S = 3$) for: Accuracy, Macro F1, Tail F1, and DG Gap.

Ethics Statement

This work proposes a general algorithmic framework and does not constitute a deployable clinical decision system. All datasets used are publicly available or approved for research use and were handled in accordance with their original licenses. CANDICE is intended to support, not replace, human experts, and improper deployment without validation or oversight could lead to harm. We emphasize the need for rigorous external validation, transparency, and human-in-the-loop safeguards when applying agentic systems in high-stakes settings.

Limitations

CANDICE has several limitations. Our evaluation focuses on healthcare tasks, which limits direct claims about performance in non-medical domains. The framework relies on the availability and quality of trusted knowledge sources; poorly curated or outdated knowledge can degrade reasoning quality. CANDICE also introduces additional inference-time overhead due to retrieval, planning, and execution steps, which may increase latency compared to single-pass models. Finally, while ablations provide insight into agent roles, more fine-grained causal diagnostics for agent decisions remain an open challenge.

F. Ablation Study

F.1. Ablation Study I: Evaluating Knowledge and Knowledge Models performance

Ablation Study I: Evaluating Symbolic Knowledge for Selecting $h_{\mathrm{K}}(x)$. To determine which symbolic knowledge features are suitable for constructing the knowledge hypothesis $h_{\mathrm{K}}(x)$, we evaluate two clinically motivated feature families on APTOS: (1) lesion biomarkers (exudates, hard hemorrhages, soft hemorrhages, cotton wool spots), and (2) retinal vein morphology (tortuosity, caliber, branching angles). The goal is to test whether these symbolic features provide domain-stable discriminative power consistent with the requirement that $P(K \mid Y)$ remains approximately invariant across imaging centers. We train five standard classifiers on each feature set and report performance in Table 4.

Across all models, lesion-only features yield substantially higher accuracy and F1-score than the combined lesion-plus-vein feature set. Gradient Boosting with lesion biomarkers achieves the best results (accuracy 0.8465, F1 0.8412), indicating that lesion-level symbolic cues form a clean, well-separated representation for DR stages. In contrast, adding vein morphology consistently degrades performance for every classifier, suggesting that these features introduce domain-sensitive variability rather than causal invariants. This behavior aligns with our theoretical framework: only knowledge features with low class-conditional domain divergence are appropriate for $h_{\mathrm{K}}(x)$ for accurate diagnosis, while domain-unstable features increase the effective discrepancy term and weaken rare-class guarantees. Based on this ablation, we select **Gradient Boosting on lesion biomarkers only** as the canonical knowledge classifier $h_{\mathrm{K}}(x)$ used in our hypothesis pool. This choice provides stable, interpretable, and clinically grounded decision boundaries that integrate reliably with deep hypotheses $h_{\mathrm{D}}(x)$.

Model	Feature Set	Acc	F1	Prec	Rec	AUC
Logistic Reg.	Lesions only	0.7732	0.7322	0.59	0.49	0.74
Random Forest	Lesions only	0.8169	0.8115	0.82	0.80	0.81
SVM	Lesions only	0.7814	0.7432	0.59	0.50	0.76
Grad. Boost.	Lesions only	**0.8465**	**0.8412**	**0.82**	**0.76**	**0.84**
KNN	Lesions only	0.7814	0.7896	0.63	0.56	0.77
Logistic Reg.	Lesions + vein	0.6424	0.6019	0.25	0.33	0.58
Random Forest	Lesions + vein	0.7384	0.7038	0.55	0.47	0.70
SVM	Lesions + vein	0.6556	0.6083	0.26	0.34	0.58
Grad. Boost.	Lesions + vein	0.7252	0.7389	0.51	0.44	0.69
KNN	Lesions + vein	0.6987	0.6369	0.43	0.44	0.66

Table 4. **Ablation on symbolic lesion biomarkers with and without retinal vein features on APTOS.** Lesion-only features provide the strongest and most stable performance across models; adding vein morphology degrades accuracy and F1.

Condition	# Deep $h_{\mathrm{D}}(x)$	# KL $h_{\mathrm{K}}(x)$	AUC (%)
A1: 5-class $h_{\mathrm{D}}(x)$ + 5-class $h_{\mathrm{K}}(x)$	1	1	83.24 ± 0.60
A3: binary $h_{\mathrm{D}}(x)$ + binary $h_{\mathrm{K}}(x)$	5	5	81.49 ± 0.30
A2: binary $h_{\mathrm{K}}(x)$ + 5-class $h_{\mathrm{D}}(x)$	1	5	$\mathbf{84.65 \pm 0.30}$
A4: binary $h_{\mathrm{D}}(x)$ + 5-class $h_{\mathrm{K}}(x)$	5	1	78.05 ± 0.76
A5: 5-class $h_{\mathrm{D}}(x)$ only	1	0	78.74 ± 0.98
A6: 5-class $h_{\mathrm{K}}(x)$ only	0	1	80.63 ± 0.13

Table 5. Ablation of different hypothesis pool composition on APTOS (5-class DR classification).

F.2. Ablation Study II: Effect of Hypothesis Pool Composition by CDA agent

Ablation Study II: Effect of Hypothesis Pool Composition. We evaluate how the composition of the CCKI hypothesis pool $\mathcal{H}$ influences in-domain performance on APTOS. All experiments use the standard 5-class Diabetic Retinopathy (DR) classification setting (stages 0 4). The hypothesis pool contains two types of models:

- **Knowledge-guided hypotheses** $h_{\mathrm{K}}(x)$ implemented using Gradient Boosting over a fixed 10-dimensional clinical feature vector $\mathcal{K}$ (as per Ablation Study I).
- **Deep-learning hypotheses** $h_{\mathrm{D}}(x)$ implemented as ViT-based image classifiers fine-tuned for DR grading.

Across all settings, the clinical feature vector $\mathcal{K}$ remains unchanged; we vary: (i) the number of $h_{\mathrm{K}}(x)$ and $h_{\mathrm{D}}(x)$ in $\mathcal{H}$, and (ii) the prediction granularity of each hypothesis (5-class vs. binary one-vs-rest). Six configurations are evaluated (Table 5).

Discussion. Condition A2 delivers the highest accuracy because the 5-class deep hypothesis provides holistic visual representation, while the five binary $h_{\mathrm{K}}(x)$ models contribute class-specific clinical cues that improve fine-grained discrimination. Large binary-only mixtures (A3, A4) perform worse due to overlapping or contradictory decision boundaries within the models decisions. Single-family baselines (A5, A6) underperform as they lack complementary perspectives.

G. Appendix Figures

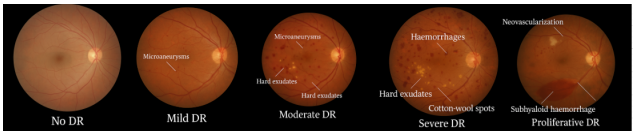

Figure 1. Fundus images showing Diabetic Retinopathy progression: from No DR to Proliferative DR, highlighting key lesions at each stage [14].

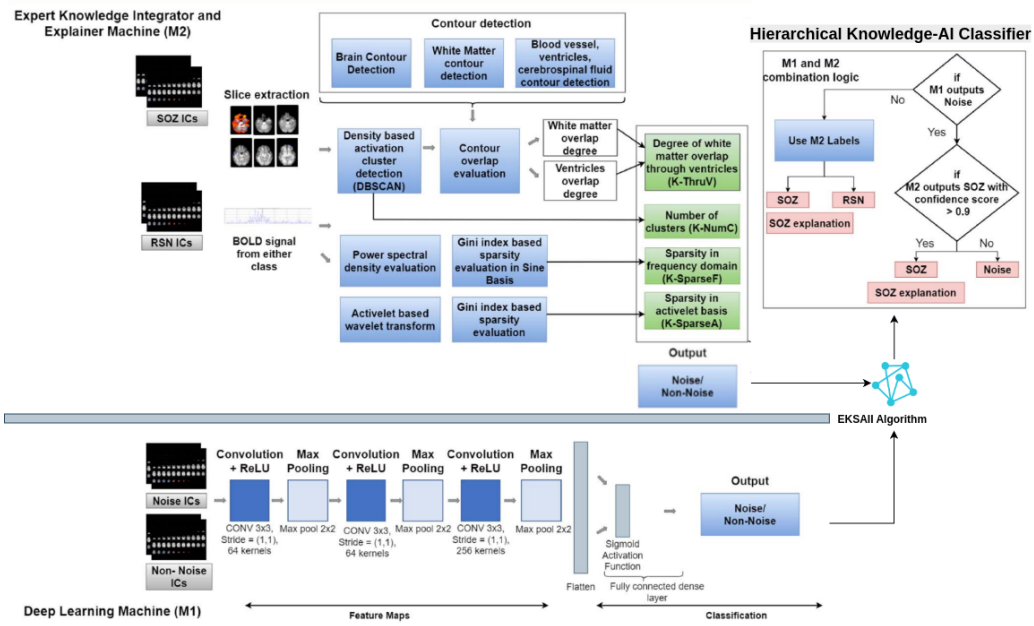

Figure 2. **DeepXSOZ: A Hybrid Knowledge-AI Architecture for Seizure Onset Zone (SOZ) Localization.** The framework employs a bipartite training architecture. During inference, the final SOZ classification is determined by integrating the labels from both M_{DL} and M_{CCKI} via confidence scores, yielding a final, integrated, and explainable diagnostic result.

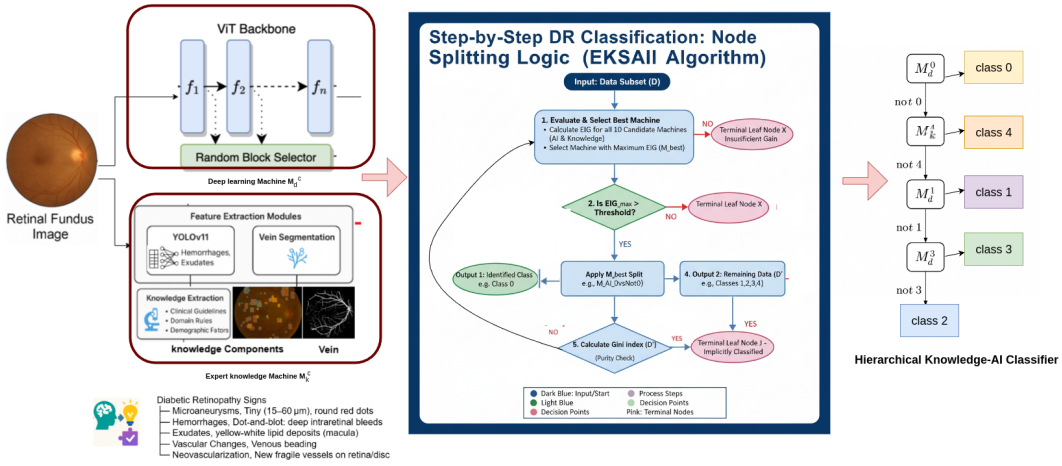

Figure 3. **Hierarchical Knowledge-AI Integration Framework for Diabetic Retinopathy (DR) Classification.** The system integrates a **Deep Learning Machine** (M_d, ViT backbone) and an **Expert Knowledge Machine** (M_k, clinical features/guidelines) within a decision tree. The **CCKI algorithm** iteratively selects the optimal binary classifier (maximum Entropy Imbalance Gain, EIG) for node splitting.

Metric	Definition	Computation / Formula	Who Computes	Auto	Human
Replanning Triggered (%)	Fraction of samples where the agent detects failure or uncertainty and initiates a new plan	$\frac{\#\text{samples with replanning}}{\#\text{total samples}} \times 100$	System logger (CDA)	✓	
Recovery Success (%)	Percentage of replanned cases successfully solved after replanning	$\frac{\#\text{successful recoveries}}{\#\text{replanned cases}} \times 100$	Evaluation script	✓	
Avg. Steps to Recovery	Average number of actions required after failure to reach a valid solution	Mean number of agent/tool calls after first failure	Execution trace analyzer	✓	
Faithful Steps (%)	Proportion of reasoning steps supported by retrieved evidence	$\frac{\#\text{evidence supported steps}}{\#\text{total reasoning steps}} \times 100$	NLI based verifier	✓	△
Unsupported Claims ↓	Avg. number of reasoning claims without evidence	Mean count of unsupported steps per sample	Trace validation script	✓	
Hallucination Rate ↓	Percentage of outputs containing factually incorrect claims	$\frac{\#\text{hallucinated outputs}}{\#\text{total outputs}} \times 100$	Verifier + audit	△	✓
Expert Agreement (%)	Agreement between agent reasoning and expert judgment	$\frac{\#\text{expert approved outputs}}{\#\text{evaluated outputs}} \times 100$	Domain experts		✓
Reasoning Stability	Consistency of reasoning under minor input perturbations	Average similarity (e.g., Jaccard / tree edit distance) across runs	Stability analysis script	✓	
Error Attribution	Distribution of failure sources across reasoning, retrieval, perception, and tools	Categorical classification of failure causes	Mixed (rules + audit)	△	△
First Run Success (%)	Percentage of executions succeeding without correction	$\frac{\#\text{first run successes}}{\#\text{total executions}} \times 100$	Execution logs	✓	
Avg. Fix Iterations ↓	Average number of correction loops after failure	Mean number of fix cycles per failed execution	Execution trace analyzer	✓	
Final Execution Rate (%)	Percentage of tasks succeeding after all corrections	$\frac{\#\text{eventually successful runs}}{\#\text{total runs}} \times 100$	Execution monitor	✓	
Runtime Overhead	Additional runtime introduced by agent reasoning	$T_{\text{agent}} T_{\text{baseline}}$	System profiler	✓	

Auto indicates fully automated computation. △ denotes partial automation with human validation on a subset of samples.

Table 6. Definition and computation protocol for agent centric evaluation metrics used in this work.

Agent	Role in CANDICE	How it is made (inputs → outputs)	How it works (core steps)	Comparator baselines and empirical wins (from your tables)
CRA	Grounded reasoning and causal explanation (no final prediction).	Trusted corpus retrieval → atomic facts → class-conditional reasoning artifacts + evidence links.	(1) Retrieve trusted docs [16] (2) Decompose into atomic factual units [23] (3) Build class-conditional causal chains [4, 30] (4) Verify step faithfulness; flag unsupported steps [13].	**Baselines:** LLM (No Retrieval), RAG (No Reasoning) [16], ReAct [37]. **Wins:** Table 1 (main paper): Faithful 85.7 vs 73.5 (ReAct) / 68.9 (RAG) / 55.2 (LLM); Hallucination 4.5 vs 9.8 / 12.4 / 18.5; Expert 90.2 vs 79.0 / 74.6 / 61.3. Robust degradation under knowledge ablation (Table 2 (main paper)).
CDA	Constraint-aware planning; selects latent pathway z and orchestrates agents/tools.	CRA artifacts + environment constraints $(X, K, D) \rightarrow$ pathway choice z (agent order, tool/model selection, hypothesis instantiation).	(1) Assess input availability/extractability (2) Estimate class uncertainty + domain sensitivity (3) Choose class-conditional pathway z to preserve tail fidelity under shift (4) Allocate steps/hypotheses under budget.	**Baselines:** Fixed order, Random order, Greedy (single-step), No CDA (heuristic only). **Wins:** Table 3 (main paper): Accuracy 81.1 vs 76.2 (No CDA); Tail F1 70.1 vs 64.5; Success 84.7 vs 77.0; Avg. Steps 2.2 vs 3.0 (fewer steps with higher success).
CGA	Executable realization; translates specs into verifiable code and repairs failures.	CRA specs + CDA plan → runnable pipeline code + execution logs + repaired code (if needed).	(1) Generate program from specs [6] (2) Execute and validate tool outputs (3) Diagnose failures and revise (self-debug / repair) [7, 12] (4) Return executable pipeline + trace.	**Baselines:** Single-shot LLM code generation [6], Toolformer-style agent [25], Human-written code (upper bound). **Wins:** Table 4 (main paper): First-run 90.1 vs 78.5 (Toolformer) / 68.3 (single-shot); Avg. Fix 1.2 vs 1.9 / 2.7; Final Exec 96.3 vs 91.2 / 84.1.

Table 7. Summary of CANDICE agents: responsibilities, construction, operational steps, and empirical wins relative to comparator baselines (using values from Tables 1,2,3, and 4 (main paper).

Without Document RAG CoT Prompt (clinical only, no class prediction)

Role. Retina specialist describing clinical grading criteria for diabetic retinopathy (DR) on color fundus photography.

Classes (5).
1. No DR
2. Mild NPDR
3. Moderate NPDR
4. Severe NPDR
5. Proliferative DR (PDR)

Important constraints.
- Do NOT assign or predict a final class for this image.
- Output must be clinical findings only (no lay explanations).
- Do not hallucinate lesions. If not clearly visible, label "uncertain" or "not assessable".
- If image quality/field of view prevents assessment of a criterion, explicitly state "not assessable".
- No treatment, prognosis, or medical advice.

Prompt template.

Without Document RAG CoT Prompt (clinical only, no class prediction) (continued)

```
ROLE: You are a retina specialist describing clinical grading criteria for diabetic
↪  retinopathy (DR) on color fundus photography.

CLASSES (5):
1) No DR
2) Mild NPDR
3) Moderate NPDR
4) Severe NPDR
5) Proliferative DR (PDR)

IMPORTANT CONSTRAINTS:
   Do NOT assign or predict a final class for this image.
   Output must be clinical findings only (no lay explanations).
   Do not hallucinate lesions. If not clearly visible, label "uncertain" or "not
   ↪  assessable".
   If image quality/field  of  view prevents assessment of a criterion, explicitly
   ↪  state "not assessable".
   No treatment, prognosis, or medical advice.

INPUT:
   Disease: Diabetic Retinopathy (DR)
   Image: (attached fundus photo)

TASK (think step  by  step internally, but DO NOT reveal private chain  of
↪  thought):
1) Image adequacy: report focus, illumination, artifacts, and whether macula + optic
↪  disc + quadrants are assessable.
2) Extract ONLY observable findings in this image (lesion inventory):
      Microaneurysms (MA)
      Intraretinal hemorrhages (dot/blot/flame), approximate distribution by
      ↪  quadrant
      Hard exudates (and proximity to fovea)
      Cotton  wool spots (CWS)
      Venous beading (VB)
      IRMA
      Neovascularization (NVD/NVE)
      Pre  retinal hemorrhage / vitreous hemorrhage
      Fibrovascular proliferation / tractional signs (if visible)
    A) Required/defining findings for that class
   B) Exclusion findings (what would rule it out or push to a different class)
   C) For THIS image: mark each defining finding as one of:
         Present
         Absent
         Uncertain
         Not assessable
   D) What additional confirmation a doctor would seek if uncertain (e.g., wider
   ↪  field, OCT for DME, FA, repeat photo)

GRADING ANCHORS (use clinically standard cues):
   Mild NPDR: MA only.
   Moderate NPDR: more than MA  only but not severe; may have
   ↪  hemorrhages/exudates/CWS; mild VB/IRMA possible.
   Severe NPDR: 4  2  1 rule (any one):

   PDR: NVD/NVE and/or pre  retinal/vitreous hemorrhage; fibrovascular
   ↪  proliferation.
```

Without Document RAG CoT Prompt (clinical only, no class prediction) (continued)

```
OUTPUT FORMAT (STRICT):
[Image Adequacy]
...

[Lesion Inventory (visible only)]
   MA:
   Hemorrhages:
   Hard exudates:
   CWS:
   VB:
   IRMA:
   NVD/NVE:
   Pre  /vitreous hemorrhage:
   Fibrovascular/tractional cues:

[Per  Class Doctor Checklists (NO final grade)]
(Class 1) No DR
   Defining findings:
   Exclusions:
   This image (present/absent/uncertain/not assessable):
   Additional confirmation if needed:

(Class 2) Mild NPDR
...

(Class 3) Moderate NPDR
...

(Class 4) Severe NPDR
...

(Class 5) PDR
...
```

Without Document RAG ,ToT Prompt (clinical only, branches, no class prediction)

Role. Retina specialist. Provide a per class clinical decision checklist for DR severity using a Tree of Thought structure.

Constraints.

- Do NOT assign/predict a final class.
- Clinical criteria only. No lay language.
- No hallucination: if not clearly visible, mark “uncertain” or “not assessable”.
- No treatment/prognosis.

Prompt template.

Without Document RAG ,ToT Prompt (clinical only, branches, no class prediction) (continued)

```
ROLE: Retina specialist. Provide a per  class clinical decision checklist for DR
↪  severity using a Tree  of  Thought structure.

CLASSES (5):
1) No DR
2) Mild NPDR
3) Moderate NPDR
4) Severe NPDR
5) PDR

CONSTRAINTS:
   Do NOT assign/predict a final class.
   Clinical criteria only. No lay language.
   No hallucination: if not clearly visible, mark "uncertain" or "not assessable".
   No treatment/prognosis.

INPUT:
   Disease: DR
   Image: (attached fundus photo)

TREE  OF  THOUGHT PROCEDURE:
Think in multiple branches internally, then output ONLY the structured branch
↪  summaries.

Step 1) Image adequacy: focus, illumination, artifacts, and coverage (macula, disc,
↪  quadrants).

Step 2) Lesion inventory (visible only): MA, hemorrhages (by quadrant), hard
↪  exudates (foveal proximity),
        CWS, VB, IRMA, NVD/NVE, pre  /vitreous hemorrhage, fibrovascular/tractional
        ↪  signs.

Step 3) Build 4 branches that cover all severity criteria, WITHOUT concluding a
↪  final class:
   Branch A: "Red lesion burden" (MA + hemorrhages extent; quadrant distribution)
   Branch B: "Ischemia markers" (CWS + VB + IRMA; explicitly map to 4  2  1
   ↪  components)
   Branch C: "Proliferation screen" (NVD/NVE; pre  retinal/vitreous hemorrhage;
   ↪  fibrovascular cues)
   Branch D: "Quality/confounders" (artifacts, blur, poor field; mimics)

Each branch outputs:
   What findings are assessed (criteria)
   For THIS image: present/absent/uncertain/not assessable
   What additional evidence would be needed for confident assessment

Step 4) Convert branches into a per  class checklist table (text  only):
For each class (1  5):
   Defining criteria (clinical)
   "Image evidence status" for each criterion (present/absent/uncertain/not
   ↪  assessable)
   Exclusion triggers (findings that would push to another class)
   Additional confirmation if needed
```

With Document RAG ToT Prompt (primary reference grounded; no class prediction)

Role. Retina specialist. Provide a per class clinical decision checklist for DR severity using a Tree of Thought structure.

Evidence requirement (published reference document).

- Use the following clinically curated published reference as the **PRIMARY** source for per class criteria and severity anchors:
- *Shukla UV, Tripathy K. Diabetic Retinopathy. [Updated 2023 Aug 25]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan . (NBK560805)*
- If any criterion is not explicitly stated in the reference, label it as `(supplemental)` and keep it minimal.

Prompt template.

With Document RAG ToT Prompt (primary reference grounded; no class prediction) (continued)

```
ROLE: Retina specialist. Provide a per  class clinical decision checklist for DR
↪  severity using a Tree  of  Thought structure.

EVIDENCE REQUIREMENT (PUBLISHED REFERENCE DOCUMENT):
   Use the following proven, clinically curated published reference as the PRIMARY
   ↪  source for the per  class criteria and severity anchors:
  Shukla UV, Tripathy K. Diabetic Retinopathy. [Updated 2023 Aug 25]. In: StatPearls
   ↪  [Internet].
  Treasure Island (FL): StatPearls Publishing; 2025 Jan  . (NBK560805)
   When producing each checklist item, prefer definitions explicitly stated in the
   ↪  reference
  (e.g., International Clinical DR Severity Scale / ETDRS  related definitions).
   If you include any criterion not explicitly stated in the reference, label it
   ↪  "(supplemental)" and keep it minimal.

CLASSES (5):
1) No DR
2) Mild NPDR
3) Moderate NPDR
4) Severe NPDR
5) PDR

CONSTRAINTS:
   Do NOT assign/predict a final class.
   Clinical criteria only. No lay language.
   No hallucination: if not clearly visible, mark "uncertain" or "not assessable".
   No treatment/prognosis.

INPUT:
   Disease: DR
   Image: (attached fundus photo)
   Reference: StatPearls NBK560805 (as above; assume it is available to you)

TREE  OF  THOUGHT PROCEDURE:
Think in multiple branches internally, then output ONLY the structured branch
↪  summaries.

Step 1) Image adequacy: focus, illumination, artifacts, and coverage (macula, disc,
↪  quadrants).

Step 2) Lesion inventory (visible only): MA, hemorrhages (by quadrant), hard
↪  exudates (foveal proximity),
        CWS, VB, IRMA, NVD/NVE, pre  /vitreous hemorrhage, fibrovascular/tractional
        ↪  signs.

Step 3) Build 4 branches that cover all severity criteria, WITHOUT concluding a
↪  final class:
   Branch A: "Red lesion burden" (MA + hemorrhages extent; quadrant distribution)
   Branch B: "Ischemia markers" (CWS + VB + IRMA; explicitly map to 4  2  1
   ↪  components per reference)
   Branch C: "Proliferation screen" (NVD/NVE; pre  retinal/vitreous hemorrhage;
   ↪  fibrovascular cues)
   Branch D: "Quality/confounders" (artifacts, blur, poor field; mimics)
```

With Document RAG ToT Prompt (primary reference grounded; no class prediction) (continued)

```
Each branch outputs:
   What findings are assessed (criteria; aligned to the reference)
   For THIS image: present/absent/uncertain/not assessable
   What additional evidence would be needed for confident assessment

Step 4) Convert branches into a per  class checklist table (text  only):
For each class (1  5):
   Defining criteria (clinical; grounded in the reference)
   "Image evidence status" for each criterion (present/absent/uncertain/not
    ↪  assessable)
   Exclusion triggers (findings that would push to another class)
   Additional confirmation if needed
OUTPUT FORMAT:
[Lesion Inventory]
...

[Branches]
(Branch A) ...
(Branch B) ...
(Branch C) ...
(Branch D) ...

[Per  Class Doctor Checklists (NO final grade)]
(Class 1) No DR
.....

(Class 2) Mild NPDR
...

(Class 3) Moderate NPDR
...

(Class 4) Severe NPDR
...

(Class 5) PDR
...
```

CDA Prompt (Block 1/3: Role + Parts A C; NO final grade)

Role & objectives.

```
ROLE: You are a Senior Retina Specialist + Causal Machine Learning Engineer.
 ↪  Your objective is to:
(1) decompose a fundus image into a structured, engineering  ready feature set
 ↪  for a Diabetic Retinopathy (DR) grading system,
(2) output per  class evidence using Binary Gates + Non  Binary Gradients
 ↪  (Excluded/Possible/Uncertain),
(3) perform CDA  style causal considerations for confounding and domain
 ↪  generalization,
(4) design a "tree of machines" (hierarchical classifiers) including BOTH
 ↪  knowledge classifiers (rule  based) and deep learning classifiers,
(5) provide a coding  ready implementation plan (schemas + module pipeline +
 ↪  node interfaces) so the next agent can write code.

DR CLASSES (5):
1) No DR
2) Mild NPDR (MA  only)
3) Moderate NPDR (more than MA  only; not Severe)
4) Severe NPDR (4  2  1 rule; no NV)
5) PDR (NV and/or pre  /vitreous hemorrhage; fibrovascular/tractional signs)

NON  NEGOTIABLE RULES:
   DO NOT assign or predict a final class for this image.
   DO NOT output "Class X", "final grade", or any single  class decision.
   Every class outcome must be expressed as: EXCLUDED / POSSIBLE / UNCERTAIN.
   All gates/nodes must output: True / False / Unknown (never a final grade).
   Output must be clinical + technical only (no lay language).
   Do not hallucinate lesions. If unclear due to blur/FOV/artifacts, label
    ↪  "Uncertain" or "Not Assessable".
   If the field  of  view prevents assessing a criterion, explicitly state "Not
    ↪  Assessable".
   No treatment/prognosis/advice.
   Follow the internal reasoning path, but DO NOT reveal chain  of  thought.
    ↪  Output only the structured sections below.
   Precision rule: if a lesion is ~70% likely but blurry, mark it "Uncertain".

INTERNAL REASONING PATH (DO INTERNALLY, DO NOT OUTPUT):
1) Visual Evidence Extraction: Scan the image for 5  class anchors; visible vs
 ↪  obscured.
2) Differential Logic: For each class, decide strict exclusions (Binary Gate) vs
 ↪  insufficient separation (Non  Binary Gradient).
3) System Requirements: What must the system measure to automate this?
4) Tree  of  Machines Planning: Hierarchical binary/multiclass gates using
 ↪  knowledge rules + feature  ML + deep models.

INPUT:
   Disease: Diabetic Retinopathy (DR)
   Image: (attached fundus photo)
```

CDA Prompt (Block 1/3: Role + Parts A C; NO final grade) (continued)

Structured output spec (Parts A to C).

```
=========================================================
PART A: LESION INVENTORY (VISIBLE ONLY)
=========================================================
A1) Image Adequacy (STRICT):
   focus: Good/Fair/Poor
   illumination: Good/Fair/Poor
   artifacts: [list]
   peripheral_visibility/FOV: Adequate/Limited
   quadrant_assessability: {Q1: Assessable/NotAssessable, Q2:..., Q3:...,
   ↪  Q4:...}
   disc_visible: Yes/No/Uncertain
   macula_visible: Yes/No/Uncertain

A2) Findings (status belong {Present, Absent, Uncertain, NotAssessable}; include
↪  1 line note each):
   MA:
   Hemorrhages (dot/blot/flame):
   Hard Exudates (HE) + fovea proximity:
   Cotton  Wool Spots (CWS):
   Venous Beading (VB):
   IRMA:
   Neovascularization (NV: NVD/NVE):
   Pre  retinal / vitreous hemorrhage:
   Fibrovascular / tractional signs:

=========================================================
PART B: PER  CLASS IMPORTANT FEATURE SET
=========================================================
For each Class 1  5 provide:
B1) Key Distinguishing Features (minimum clinical requirements)
B2) Must  Quantify (presence vs count vs quadrant distribution vs
↪  disc/macula/fovea location)
B3) Upgrade Trigger (single finding that moves into this class or above)
B4) Minimum machine  observable signals required (what detectors/segmenters must
↪  output)

=========================================================
PART C: PER  CLASS REASONING (BINARY + NON  BINARY)
=========================================================
For each Class 1  5 output:

C1) Binary Gates (hard checks; gate_status {True, False, Unknown}):
   Provide 3  6 gates per class.
   Each gate must include:
  {gate_name, gate_definition, required_inputs, gate_status_for_this_image,
   ↪  evidence_from_PartA}

C2) Non  Binary Gradients (graded signals; level {Low, Medium, High} or
↪  numeric):
   Provide 3  6 signals per class:
  {signal_name, definition, required_inputs, level_for_this_image,
   ↪  evidence_from_PartA}

C3) Status for this class (REQUIRED; NOT a final grade):
   status_for_this_class: EXCLUDED / POSSIBLE / UNCERTAIN
   explanation: 2  4 bullets referencing C1/C2 and assessability limits
```

CDA Prompt (Block 2/3: Parts D to E; NO final grade)

```
============================================================
PART D: KNOWLEDGE SUFFICIENCY TEST
============================================================
For each Class 1  5:
   sufficiency: Sufficient / PartiallySufficient / Insufficient
   why: 2  4 bullets (must cite missing quadrants/blur/artifacts if relevant)
   missing_information: [exact missing counts/locations/visibility]
   recommended_additional_imaging/tests:
     UWF/Wide  field (purpose)
     FA (purpose: IRMA vs NV; leakage; nonperfusion)
     OCT (purpose: macular edema evaluation if relevant to feature extraction; note
      ↪  not equivalent to DR class)

============================================================
PART E: CAUSAL DISCOVERY DECISION (CDA)
============================================================
E1) Confounders & measurement variables:
   camera_type, site, illumination, blur, FOV, compression, artifact_presence,
    ↪  grader_noise
E2) Causal goal:
   Ensure model learns Lesion to Grade rather than ImageQuality/Camera to Grade
E3) Decision:
   causal_ml: NOT_NEEDED / OPTIONAL / RECOMMENDED
   justification: 3  6 bullets tied to Part A limitations and domain shift risks
E4) Minimal text DAG (nodes + arrows):
   U(systemic severity) to L(lesions) to S(severity)
   Q(measurement: blur/illum/FOV/camera) to Y(pixels)
   L to Y
   Q masks L (missingness/measurement error)
   grader_noise → labels
```

CDA Prompt (Block 2/3: Parts D to E; NO final grade) (continued)

```
=====================================================
PART F: CODING PLAN & LOGIC ENGINE (ENGINEERING  READY)
=====================================================
F1) Data Schemas (JSON  like; MUST use exact field names):
QualityReport:
{illumination_score: float, blur_score: float, fov_score: float, artifact_flags:
↪  [str],
 quadrant_visibility: {Q1: bool, Q2: bool, Q3: bool, Q4: bool}}

Anatomy:
{disc_center: [x,y]|null, fovea_center: [x,y]|null, disc_visible: bool,
↪  macula_visible: bool,
 quadrant_masks: {Q1: ..., Q2: ..., Q3: ..., Q4: ...}}

LesionDetections:
{MA: [{x: float, y: float, conf: float}],
 IRH: [{bbox: [x1,y1,x2,y2], subtype: "dot"|"blot"|"flame"|"unknown", conf: float}],
 HE: [{mask_or_bbox: ..., area_px: float, conf: float}],
 CWS: [{bbox_or_mask: ..., conf: float}],
 VB: [{segment_id: str, score: float, conf: float}],
 IRMA: [{bbox_or_mask: ..., conf: float}],
 NV: [{type: "NVD"|"NVE"|"unknown", bbox_or_mask: ..., conf: float}],
 PR_VH: [{bbox_or_mask: ..., conf: float}]}

DerivedFeatures:
{MA_count_total: int, IRH_count_total: int, HE_area_total: float,
 MA_by_quadrant: {Q1: int, Q2: int, Q3: int, Q4: int, not_assessable: bool},
 IRH_by_quadrant: {Q1: int, Q2: int, Q3: int, Q4: int, not_assessable: bool},
 VB_quadrants: int|"not_assessable", IRMA_quadrants: int|"not_assessable",
 severe_421_flags: {heme_4q: bool|"not_assessable", vb_2q: bool|"not_assessable",
 ↪  irma_1q: bool|"not_assessable"},
 pdr_flags: {nv_present: bool|"not_assessable", pr_vh_present:
 ↪  bool|"not_assessable"}}

EvidenceChecklistPerClass:
{class_id: int,
 binary_gates: [{gate_name: str, status: "True"|"False"|"Unknown", evidence: str}],
 nonbinary_signals: [{signal_name: str, level: "Low"|"Medium"|"High"|float,
 ↪  evidence: str}],
 status: "EXCLUDED"|"POSSIBLE"|"UNCERTAIN",
 notes: [str]}

F2) Module Pipeline (M1  M7; include inputs→outputs→method):
M1 preprocess(image)  >img_norm
M2 quality(img_norm)  >QualityReport
M3 anatomy(img_norm)  >Anatomy
M4 quadrant_map(Anatomy)  >quadrant_masks + quadrant_visibility
M5 lesions(img_norm,Anatomy)  >LesionDetections
M6 aggregate(LesionDetections,QualityReport,Anatomy)  >DerivedFeatures
M7 logic_engine(DerivedFeatures,LesionDetections,QualityReport)
↪  >EvidenceChecklistPerClass[1..5]
(IMPORTANT: M7 outputs only per  class statuses; never a final grade.)

F3) Pseudocode (MUST NOT RETURN A CLASS):
pipeline_infer(image)   > {QualityReport, Anatomy, LesionDetections,
↪  DerivedFeatures, EvidenceChecklistPerClass}
```

CDA Prompt (Block 2/3: Parts D to E; NO final grade) (continued)

```
=======================================================
PART G: TREE  OF  MACHINES PLAN (HIERARCHICAL CLASSIFIERS)
=======================================================
Goal: Build a hierarchy of "machines" where each node can be implemented as:
(1) Knowledge classifier (rule gate from DerivedFeatures) and
(2) Learning classifier (feature  ML and/or deep model),
and outputs only gate decisions (True/False/Unknown), never a final class.

G0) Planning Steps (MUST INCLUDE):
1) Define node targets (binary/ternary/multiclass) aligned with clinical anchors:
      NoDR vs AnyDR
      PDR vs not  PDR
      Severe (4  2  1) vs not  Severe
      Mild (MA  only) vs More  than  mild
      Moderate consistency check (optional)
2) For each node, define:
      required features + assessability prerequisites
      Unknown conditions (when data insufficient)
      knowledge rule version
      learning version (feature  ML + deep)
3) Decide training data needs per node:
      Are image-level labels sufficient, or are lesion-level annotations required?
4) Decide calibration:
      thresholds to output Unknown under poor quality/low confidence
5) Compose node orchestration:
      Run nodes in order, aggregate node outputs to EvidenceChecklistPerClass only
       ↪  (no final grade)

G1) Decision Tree Topology Table (REQUIRED): (include the specified columns)
G2) Node interfaces (coding  ready):
node_k(derived_features, quality_report, lesions)  >{decision, confidence,
↪  unknown_reason, evidence_used}
run_tree(image)
↪  >{QualityReport,DerivedFeatures,EvidenceChecklistPerClass,node_outputs,uncertainty_report}
(IMPORTANT: run_tree MUST NOT output a final grade.)

=======================================================
OUTPUT FORMAT REQUIREMENTS:
   Use tables for Part E3 and Part G1.
   Use structured lists for schemas and modules.
   Do NOT output a final grade.
=======================================================
```

References

[1] Michael D. Abràmoff, Ying Lou, Ali Erginay, Will Clarida, Ryan Amelon, James C. Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206, 2016. 3

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[3] Stanford AIMI. Echonet-dynamic dataset. https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a, 2020. Accessed: 2025-11-11. 3

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 39–48, 2016. 8

[5] APTOS. Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection, 2019. Accessed: 20 February 2022. 3

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 8

[7] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 8

[8] EyePACS. Kaggle eyepacs dataset. https://paperswithcode.com/dataset/kaggle-eyepacs, 2023. Accessed: 20 February 2023. 3

[9] American Association for Pediatric Ophthalmology and Strabismus. Proliferative diabetic retinopathy. https://aapos.org/glossary/proliferative-diabetic-retinopathy, 2023. 3

[10] Robert N. Frank. Diabetic retinopathy. *New England Journal of Medicine*, 350(1):48–58, 2004. 3

[11] ETDRS Research Group. Grading diabetic retinopathy and estimating its progression. *Ophthalmology*, 98(5):786–806, 1991. 3

[12] Rahul Gupta, Shubham Pal, Aditya Kanade, and Shirish Shevade. Deepfix: Fixing common c language errors by deep learning. In *Proceedings of the International Conference on Software Engineering (ICSE)*, pages 134–145, 2017. 8

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tianyu Yu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 8

[14] Tom Kauppi et al. The aptos 2019 blindness detection dataset. Kaggle, 2019. https://www.kaggle.com/c/aptos2019-blindness-detection. 5

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Tete Rolland, Kaiming Fu, Yuxin Cai, Aditya Tejani, Ishan Misra, Piotr Dollár, and Ross Girshick. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8

[17] American Academy of Ophthalmology. Diabetic retinopathy preferred practice pattern, 2023. https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp, 2023. 3

[18] Prasanna Porwal, Sachin Pachade, Rohan Kamble, Manesh Kokare, Gopal Deshmukh, Vinayak Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. 3

[19] Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. Adapt: As-needed decomposition and planning with language models. *arXiv preprint arXiv:2311.05772*, 2023. 1

[20] StatPearls Publishing. Diabetic retinopathy. https://www.ncbi.nlm.nih.gov/books/NBK560805/, 2024. 3

[21] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. 1

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1

[23] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, 2020. 8

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351:234–241, 2015. 1

[25] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. 1, 8

[26] Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. 1

[27] Unnati V. Shukla and Koushik Tripathy. Diabetic retinopathy, 2025. Updated August 25, 2023, https://www.ncbi.nlm.nih.gov/books/NBK560805/. 3

[28] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. 1

[29] R. Singh, K. Ramasamy, C. Abraham, V. Gupta, and A. Gupta. Diabetic retinopathy: An update. *Indian Journal of Ophthalmology*, 56(3):179–188, 2008. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636123/. 3

[30] Oyvind Tafjord and Peter Clark. Proofwriter: Generating implications, proofs, and abductive explanations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 3621–3637, 2021. 8

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 1

[32] Charles P. Wilkinson, Frederick L. Ferris, Ronald E. Klein, Peter P. Lee, Carl-David Agardh, Mark Davis, and Hans-Peter Hammes. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. 3

[33] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019. Software available from https://github.com/facebookresearch/detectron2. 1

[34] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback. In *Advances in Neural Information Processing Systems*, 2023. 1

[35] Myron Yanoff and Jay S. Duker. *Ophthalmology*. Elsevier, 5th edition, 2019. 3

[36] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. 1

[37] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, et al. React: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 1, 8