Towards a Medical AI Scientist

Hongtao Wu[1,*] **Boyun Zheng**[1,*] **Dingjie Song**[2,*] **Yu Jiang**[1] **Jianfeng Gao**[4] **Lei Xing**[3] **Lichao Sun**[2,†] **Yixuan Yuan**[1,†]
[1]The Chinese University of Hong Kong [2]Lehigh University [3]Stanford University [4]Microsoft Research
{lis221@lehigh.edu, yxyuan@ee.cuhk.edu.hk}
Homepage: https://cuhk-aim-group.github.io/Med-AI-Scientist-Homepage/

Autonomous systems that generate scientific hypotheses, conduct experiments, and draft manuscripts have recently emerged as a promising paradigm for accelerating discovery. However, existing "AI Scientists" remain largely domain-agnostic, limiting their applicability to clinical medicine, where research is required to be grounded in medical evidence with specialized data modalities. In this work, we introduce Medical AI Scientist, the first autonomous research framework tailored to clinical autonomous research. It generates clinically grounded ideas by transforming surveyed literature into actionable evidence through a clinician-engineer co-reasoning mechanism, which improves the traceability of generated research ideas. The Medical AI scientist further introduces evidence-grounded manuscript drafting guided by a structured medical writing paradigm and ethical policies. The framework operates under 3 research modes, namely paper-based reproduction, literature-inspired innovation, and task-driven exploration, corresponding to distinct levels of medical scientific autonomy. Comprehensive evaluations by both large language models and human experts demonstrate that the ideas generated by the Medical AI Scientist are of substantially higher quality than those produced by commercial LLMs across 171 cases, covering 19 clinical tasks, and 6 data modalities. Meanwhile, our system achieves strong alignment between the proposed method and its implementation, while also demonstrating significantly higher success rates in executable experiments. Double-blind evaluations by human experts and the Stanford Agentic Reviewer suggest that the generated manuscripts approach MICCAI-level quality, while consistently surpassing those from ISBI and BIBM. The proposed Medical AI Scientist highlights the potential of leveraging AI for autonomous scientific discovery in healthcare.

1. Introduction

Recent years have witnessed rapid advances in artificial intelligence for healthcare, with increasingly capable models achieving state-of-the-art performance across disease diagnosis [1–4], medical image analysis [5–7] and clinical outcome prediction [8–10]. In parallel, large language models [11–16] have made substantial progress in language understanding, reasoning and code generation, enabling the emergence of tool-augmented and multi-agent systems [17–25] that extend beyond narrow task execution. Together, these developments have catalyzed the rise of autonomous research frameworks, often referred to as AI Scientists [26–29], which seek to automate the scientific workflow from hypothesis generation and experimental design to result interpretation and manuscript preparation, promising to accelerate scientific innovation [30]. These AI Scientist systems have shown promise in accelerating research in domains such as mathematics, chemistry and general machine learning, where problem formulations, data representations and evaluation protocols are relatively standardized.

Medical AI represents one of the most consequential domains for such systems, given its direct implications for patient outcomes, diagnostic reliability and healthcare efficiency. As medical datasets, analytical methodologies and scientific literature continue to grow at an unprecedented pace, the throughput of human-driven research has become an increasingly critical bottleneck [31–34]. This

widening gap highlights the urgent need for autonomous scientific systems that are explicitly designed to operate within the epistemic, operational, and ethical constraints inherent to clinical medicine.

However, extending these autonomous research paradigms to medical field remains challenging. First, existing AI Scientists focus on model modifications or generic optimization strategies, ignoring medical related priors, such as basic diagnostic workflows and disease-specific pathological patterns. Moreover, their retrieval and reasoning processes frequently lack sufficient constraints to reliably identify authoritative medical reasoning evidence, which will lead to models with superficial performance metrics but fail to capture clinically relevant patterns. Secondly, the heterogeneous and high dimensional nature of medical data, including three dimensional and anisotropic structures, together with specialized evaluation standards, poses challenges to the reliable and fair experimentation execution. Thirdly, the provenance of medical data and the clarity of ethical statements are central to the credibility, reproducibility, and clinical translation of research findings, yet current autonomous research systems largely overlook these requirements and fail to produce manuscripts that adhere to clinical writing frameworks and ethical standards.

Here we present Medical AI Scientist, an agentic framework for end-to-end medical AI discovery and development, as shown in Fig. 1a. The system comprises three key components: Idea Proposer, Experimental Executor, and Manuscript Composer, which together support the fully autonomous research lifecycle. The Idea Proposer leverages structured literature retrieval and analysis to identify clinical prior and then adapts the most suitable emerging technical models to medical tasks. A clinician–engineer co-reasoning mechanism is incorporated into the idea generation process to explicitly ground each hypothesis in verifiable evidence and mitigate hallucinations. The automated experimental executor orchestrates a reliable validation pipeline by unifying general-purpose execution toolchains with domain-specific medical toolboxes tailored to heterogeneous and complex clinical data formats, enabling iterative and self-correcting deep model development. A hierarchical Manuscript Composer transforms research outputs into coherent and evidence-grounded drafts through a structured medical writing paradigm with enhanced narrative logic and readability. It also embeds ethical review mechanisms that explicitly document data usage in compliance with medical publication policies.

To address the absence of standardized evaluation protocols for automated medical research systems, we introduce Med-AI Bench (Fig. 1b). This benchmark comprises 171 high-quality evaluation cases, organized around 19 distinct research tasks spanning 6 common medical data modalities. For each task, we selected 3 representative papers of varying difficulty (easy, medium, hard) and constructed 3 evaluation cases with different input modes. This design provides a systematic and unified framework for both qualitative and quantitative assessment of automated medical research systems across the full research pipeline.

As presented in Fig. 1c, we first evaluate research idea generation using both large language models and human experts (Fig. 2), showing that the Medical AI Scientist consistently surpasses commercial language models across six dimensions, including novelty, maturity, ethicality, generalizability, utility, and interpretability. We then assess experimental execution, where the system exhibits strong alignment between proposed methods and their implementations, together with substantially higher success rates in producing executable experiments (Fig. 4). Finally, under double blind evaluation (Fig. 1d, 5b & c), 10 independent domain experts assess generated manuscripts alongside high quality human authored studies from leading venues such as MICCAI, ISBI, and BIBM, while all submissions were further reviewed using the Stanford Agentic Reviewer under ICLR-aligned criteria (Fig. 5a). The generated manuscripts achieve a mean score of 4.60 ± 0.56 and remain competitive across key dimensions including novelty, reproducibility, coherence, and clarity, with only a modest gap in coverage. Qualitative feedback further indicates strong practical relevance and clear presentation with

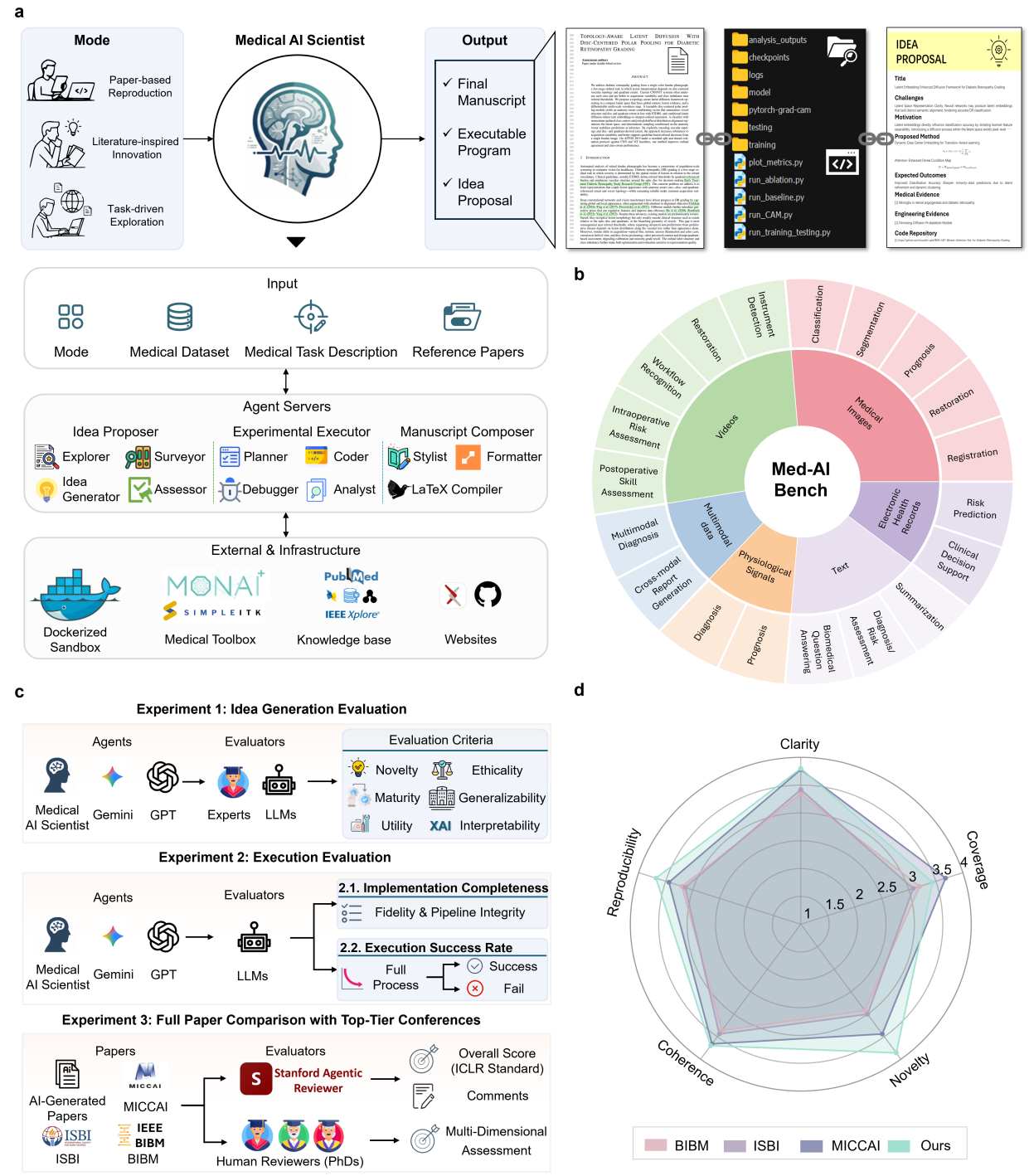

Figure 1 | a, System workflow: fully-automated multi-agents system for end-to-end scientific discovery in clinical medicine. b, Med-AI Bench: visualization depicting 19 distinct medical research tasks within performance benchmarking. c, Experimental setup: comparative evaluation across Idea generation, execution and full paper compilation in the research lifecycle. d, Performance benchmarking: comparable manuscript quality to representative works from leading venues under double blind evaluation.

limited critical weaknesses. Moreover, one manuscript generated by our system has been accepted by the International Conference on AI Scientists (ICAIS 2025 [35]) after peer review. Together, these results suggest that automated systems can speed up complex methodological designs, highlighting their potential to significantly enhance the efficiency of medical AI research.

2. Results

2.1. Building universal medical research by systematic LLM Agent

The Medical AI Scientist provides different levels of autonomous academic research modes: Paper-based Reproduction, Literature-inspired Innovation, and Task-driven Exploration. These modes are designed to accommodate users ranging from early stage PhD-level researchers entering a medical AI task to domain experts seeking efficient and highly automated solutions for open ended problems. The Reproduction mode follows explicitly defined research instructions derived from target papers and focuses on the faithful implementation of established methods. An ethical gatekeeping mechanism is incorporated to prevent harmful implementations. Instead of relying on explicit method specifications, the Innovation mode identifies research gaps and generates hypotheses based on fixed references and datasets. Evaluation emphasizes originality and methodological completeness, supported by a clinician–engineer co-reasoning mechanism and multi-dimensional assessment. The Exploration mode further targets problem driven discovery in real-world settings. Starting from a single user defined question, the system conducts literature mining, selects and integrates paradigms, generates solutions, and performs experimental verification.

To enable a rigorous and domain-spanning assessment of the Medical AI Scientist, we constructed Med-AI Bench, a benchmark grounded in peer-reviewed medical AI literature and expert-annotated references. Med-AI Bench is deliberately organized to reflect the breadth of contemporary medical AI research, covering six data modalities and nineteen representative tasks that span the full spectrum from low-level perception to high-level clinical reasoning (Fig. 1b). Specifically, medical images-related tasks cover core problems in visual understanding and analysis, including classification [36–38], segmentation [39–41], prognosis [42–44], registration [45–47], and restoration [48–50]. Video-centric tasks encompass instrument detection [51–53], restoration [54–56], workflow recognition [57–59], intraoperative risk assessment [60–62], and postoperative skill assessment [63–65]. Structured electronic health record data support tasks in risk prediction [66–68] and clinical decision support [69–71], while physiological signal data are used for diagnosis [72–74] and prognosis [75–77]. Text-based clinical reasoning is evaluated through report summarization [78–80], diagnosis and risk assessment [81–83], and biomedical question answering [84–86]. Finally, multimodal tasks assess the system's ability to integrate heterogeneous data sources for multimodal diagnosis [87–89] and cross-modal report generation [90–92].

For each task, we retrieve three papers from Google Scholar, which serve as a structured ground truth for benchmarking different levels of scientific reasoning and execution. Each paper was evaluated across five dimensions, including code availability, venue quality, citations, year, complexity, and subjective human rating, and then ranked and assigned to one of three difficulty tiers per task. Using this benchmark, we evaluate the Medical AI Scientist across the complete research lifecycle, including idea generation, experimental execution, and manuscript compilation. Collectively, Med-AI Bench functions as a standardized and reproducible framework for assessing autonomous medical AI researchers under realistic, multi-modal, and clinically relevant research conditions.

2.2. Comprehensive evaluation of idea generation

The Idea Generation module is designed to address two central challenges in AI assisted research ideation. The first concerns the generation of novel hypotheses from unstructured resources without a specific direction, as in the Innovation mode. The second concerns the need to ensure that these hypotheses remain clinically relevant and technically feasible, which is emphasized in the Exploration mode. We quantitatively evaluated the quality of model-generated research ideas against two commercial LLMs (e.g., GPT-5, Gemini-2.5-Pro), using both LLM-as-judge metrics and blinded human assessments, with evaluations conducted across six criteria commonly adopted in medical AI research, including novelty, maturity, ethicality, generalizability, utility, and interpretability.

As shown in Fig. 2 a, the Medical AI Scientist consistently outperforms the baselines across six dimensions of idea quality. For novelty and maturity, it achieves higher scores in innovation (4.07 vs. 3.00 and 3.12 in literature-based; 4.07 vs. 3.42 and 3.05 in open-ended) and maturity (4.61 and 4.74 vs. $\leq$3.58 for the baselines). For technical reliability, it also leads in robustness (3.44 and 3.56 vs. $\leq$3.19) and interpretability (3.83 and 3.81 vs. $\leq$3.42). Finally, for practical and ethical suitability, the system obtains stronger utility (3.56 and 3.61 vs. $\leq$3.44) and ethicality (3.39 and 3.64 vs. $\leq$3.05), indicating that the generated ideas are not only more innovative but also more clinically grounded and deployable.

In the human expert assessment (Fig. 2 b), our method consistently achieves the highest scores in technical innovation (4.40 $\pm$ 0.49 and 4.32 $\pm$ 0.47) and maturity (4.65 $\pm$ 0.48 and 4.68 $\pm$ 0.47), substantially outperforming GPT-5 and Gemini-2.5-Pro, while also exhibiting lower variance. This advantage extends to ethicality (up to 4.39 $\pm$ 0.63) and robustness (3.90 $\pm$ 0.61), where competing models remain below 3.50 on average, indicating more stable and reliable hypothesis generation. Notably, improvements in utility and interpretability are more moderate (e.g., 3.93 $\pm$ 0.53 and 3.81 $\pm$ 0.63 in Innovation mode), suggesting that gains in novelty and rigor are accompanied by only incremental advances in practical clarity. Highlighted by human evaluators' observations (Fig. 2 c), our method produces more consistently innovative and mature research ideas, with stronger alignment to clinical relevance and clearer experimental grounding than competing approaches. In contrast, baseline models tend to generate more incremental and less coherent hypotheses, often with higher variability and weaker integration into realistic research workflows.

As illustrated in Fig. 3, this case study compares the idea generation results of our method with those of commercial LLMs under the Innovation mode. All models operate under identical inputs, including the same task description, reference papers, and dataset specification, ensuring a fair comparison. While commercial models produce reasonable designs, their formulations remain relatively generic and lack strong domain grounding. Their outputs often resemble incremental extensions of prior work, with limited justification from a medical perspective. In contrast, the proposed method incorporates both medical and engineering evidence into the ideation process, informing model design and learning objectives. This leads to a more concrete and clinically meaningful formulation, reflected in the richer and more explicit set of equations. Consequently, the Medical AI Scientist demonstrates greater implementation detail and improved conceptual novelty, as its designs are guided by disease-related priors rather than abstract extensions of existing approaches.

2.3. Analysis of experimental implementation

2.3.1. Implementation completeness

Translating a conceptual research hypothesis into executable code requires preserving methodological coherence between the idea and its technical realization. To evaluate this capability, we systematically examined the extent to which finalized research plans were faithfully instantiated in downstream

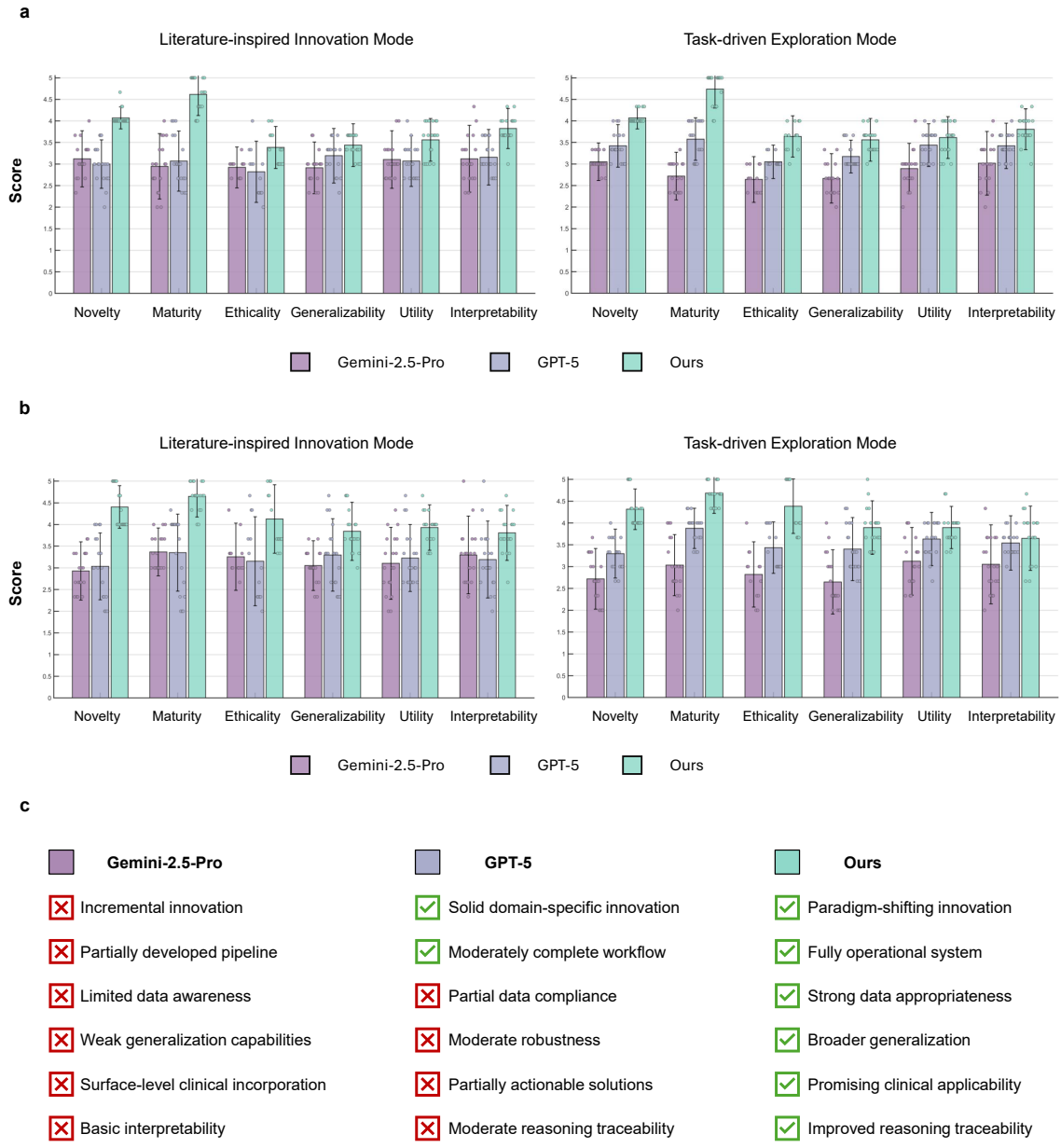

Figure 2 | Medical AI Scientist surpasses commercial LLMs in idea generation under combined LLM-based and blinded human evaluation. Models generated research ideas that were anonymized and assessed by three independent experts using a five point scale. a, LLM based evaluation of idea quality. b, Quantitative human assessment across six evaluation criteria. c, Qualitative human analysis of strengths and limitations relative to commercial LLMs.

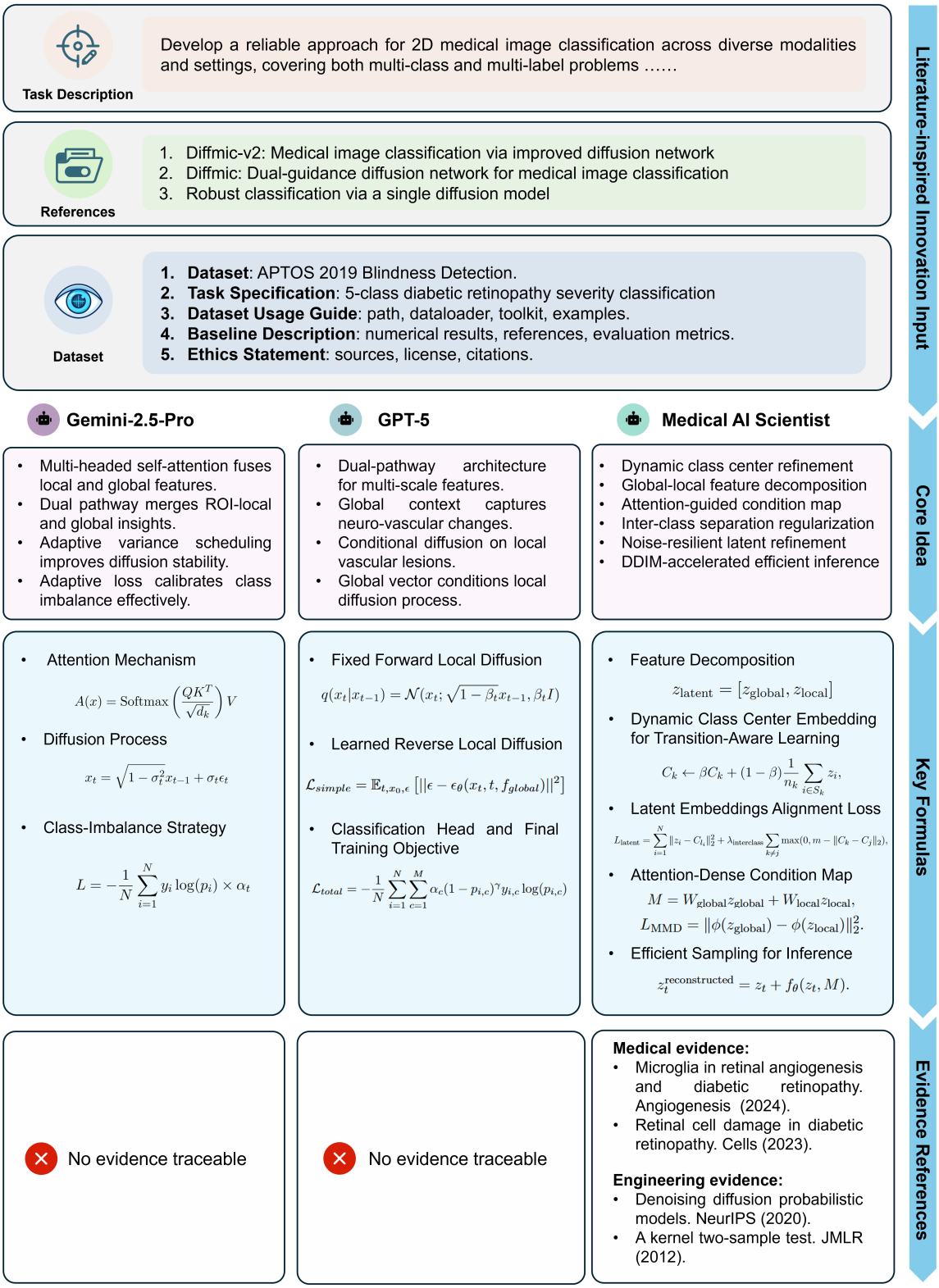

Figure 3 | Example of idea generation comparison between Medical AI Scientist and commercial LLMs in the Literature-inspired Innovation Mode.

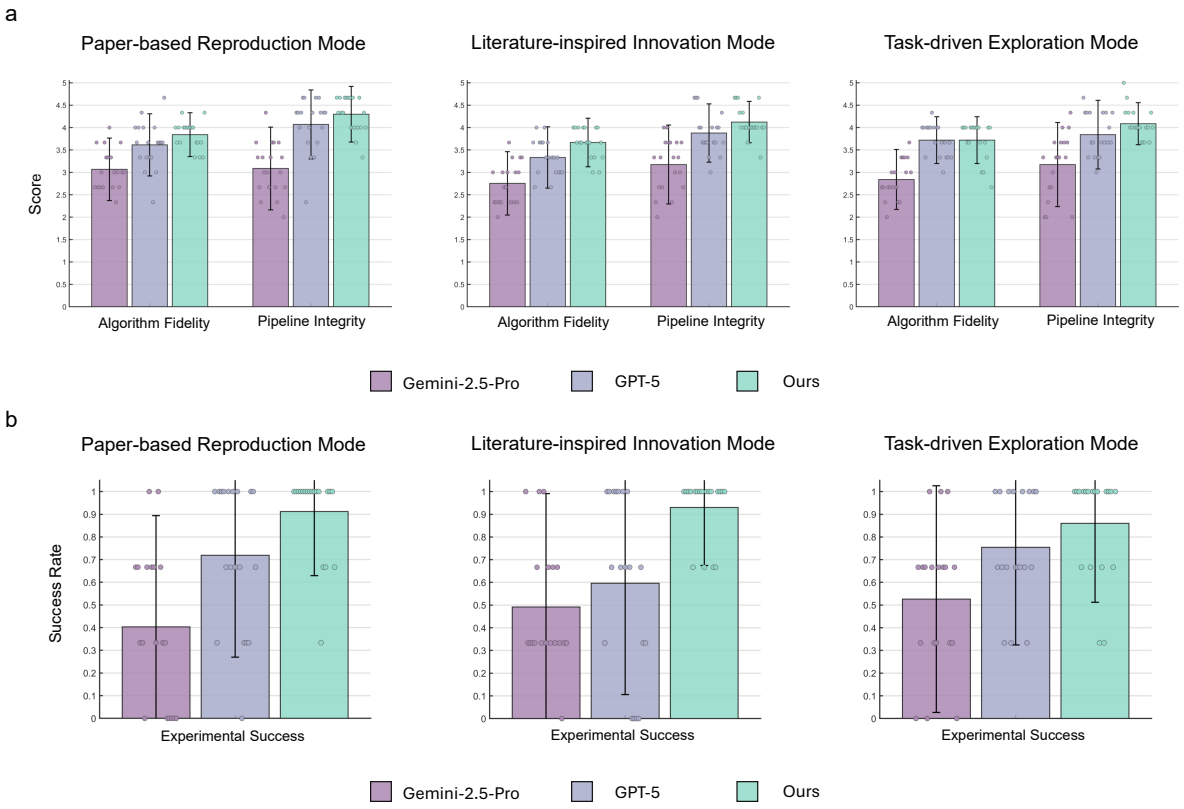

Figure 4 | Comparative evaluation of Medical AI Scientist frameworks against commercial LLMs in terms of implementation completeness and experimental success rate. a, Implementation completeness was assessed on a five point scale ranging from 1 to 5. Model generated outputs were anonymized and independently evaluated by two LLM-based judges. b, Experimental success rate measured through quantitative human evaluation.

implementations. As summarized in Fig. 4 a, we quantified experimental success by jointly assessing algorithm fidelity and pipeline integrity, reflecting whether the proposed methodological components were both present and functionally integrated within the resulting codebase. Across all three experimental modes, our Medical AI Scientist consistently achieved the highest mean scores for both indicators, along with the lowest or near-lowest standard deviations. In open-ended innovation mode, it reached 3.72 ± 0.52 and 4.09 ± 0.47, respectively, matching GPT-5-Pro while substantially outperforming Gemini-2.5-Pro (2.84 ± 0.67 and 3.18 ± 0.94). The advantage grew clearer in replication mode (3.84 ± 0.49 and 4.30 ± 0.62) and literature-based innovation mode (3.67 ± 0.54 and 4.12 ± 0.46), where our system not only scored highest but also showed the most stable performance. The results show that the system's structured refinement process, which couples systematic retrieval from the literature and code repositories with iterative clinician–engineer deliberation, grounds each proposed idea in accessible methodological and technical resources. This integration ensures that finalized research plans are not only scientifically coherent but also practically implementable, with sufficient technical and evidential grounding to enable reliable translation into executable and methodologically faithful code.

2.3.2. Code execution

Executing AI-generated research scripts may fail due to unresolved dependencies, dataset incompatibilities, or latent logical errors. These issues become more acute in medical AI research, where heterogeneous clinical data demand specialized preprocessing, domain-specific evaluation metrics, and dedicated software libraries to ensure valid analysis. To quantify robustness in this context, we measured first-run experimental success across a set of 57 medical AI research instances, comparing experimental results produced by our structured pipeline with those generated directly by the commercial LLM baselines.

As shown in Fig. 4 b, our approach consistently achieved higher success rates, reflecting the effective resolution of dependency conflicts, enforcement of data compatibility, and runtime-stable logic through iterative refinement and grounding in reference implementations. By contrast, general-purpose LLM-generated code encountered persistent debugging loops triggered by unresolved runtime errors or became prematurely terminated due to environment configuration issues, preventing successful completion of experiments. We defined experimental success as stable end-to-end execution of the training pipeline, characterized by successful runtime completion, a decreasing loss trajectory, absence of gradient explosion, and the generation of valid model weight files. Under this definition, our method achieved the highest success rate in all settings, reaching 0.91 in reproduction mode, 0.93 in literature-based innovation mode, and 0.86 in open-ended task mode. In comparison, GPT-5 obtained success rates of 0.72, 0.60, and 0.75, while Gemini-2.5-Pro achieved 0.40, 0.49, and 0.53 under the same conditions. These results show that our system consistently maintains a substantially higher end to end experimental execution success rate across increasing task difficulty.

2.4. Human and automated evaluation of medical research manuscripts drafting

The AI Scientist generates scientific manuscripts following the structure of standard medical conference submissions, including visualizations and all conventional sections. Four full-sized AI-generated papers are provided in the Appendix 5.3. To evaluate the translational relevance of autonomous medical research under realistic expert scrutiny, we designed a double-blind user study centered on diabetic retinopathy classification from fundus images while preserving the generality of the framework. We invited ten independent experts with over five years of first-author experience in AI for healthcare to assess a curated set of 20 manuscripts, including both autonomously generated studies and high-impact human-authored papers. These human-authored works were sampled from leading venues, including the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), International Conference on Bioinformatics and Biomedicine (BIBM), and The IEEE International Symposium on Biomedical Imaging (ISBI). In parallel, all manuscripts were independently evaluated using the Stanford Agentic Reviewer, an advanced large language model based assessment system, following standardized review criteria aligned with The International Conference on Learning Representations (ICLR) guidelines.

From the AI-based evaluation in Fig. 5a, our method achieves a mean score of 4.60 ± 0.56, comparable to the range observed across representative MICCAI (4.86 ± 0.47), ISBI (3.74 ± 1.02), and BIBM (4.06 ± 0.89) submissions. According to the double-blind human evaluations in Fig. 5b, our manuscripts demonstrate consistently strong performance across all five dimensions, with scores broadly aligned with those reported for MICCAI, ISBI, and BIBM. In particular, they show competitive results in Novelty, Reproducibility, Coherence, and Clarity, while exhibiting a modest gap in Coverage (3.44 ± 0.67 vs 3.68 ± 0.68), likely reflecting a more focused emphasis on methodological innovation rather than extensive dataset coverage and baseline comparisons. Qualitative observations (Fig. 5c) from domain experts further highlight the novelty, practical relevance, and clarity of presentation in our manuscripts, alongside solid mid-range assessments in logical coherence and experimental

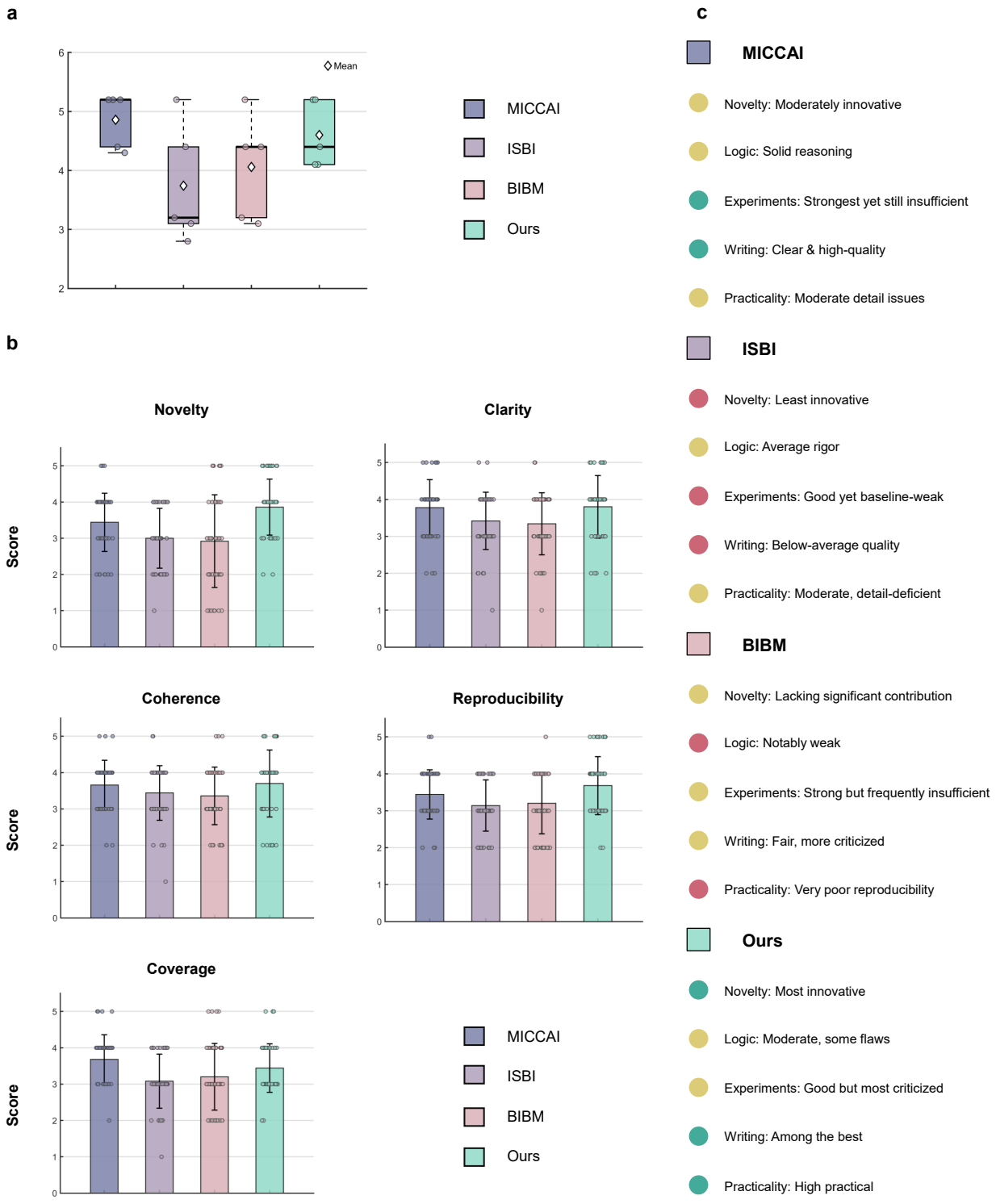

Figure 5 | Anonymized comparison of paper quality on an identical medical task. Manuscripts generated by Medical AI Scientist achieve performance comparable to MICCAI, ISBI, and BIBM under consistent double-blind evaluation across both quantitative and qualitative assessments: a, Stanford Agentic Reviewer automatic evaluation. b, Double-blinded scoring (1–5) by 10 medical experts (PhD/postdoc) across five review dimensions. c, Experts' observations on strengths and limitations.

design, with relatively few critical weaknesses noted across comparisons with MICCAI, ISBI, and BIBM submissions. Overall, these results suggest that our manuscripts achieve a level of quality comparable to that observed across leading venues such as MICCAI, ISBI, and BIBM, as assessed under consistent double-blind evaluation criteria. We also demonstrated the advantage of our system over other AI-scientist systems by having a manuscript it generated accepted by ICAIS 2025 [35], which received 114 submissions and had an acceptance rate of 36.8%.

2.5. Case study of autonomous medical research process

2.5.1. Mode 2: Literature-inspired innovation for medical image classification

As shown in Fig. A.1, we further evaluated the proposed automated medical research system to assess its capacity to enrich generated research ideas with medically grounded priors and concrete engineering specifications through its medical–engineering discussion module. Using diabetic retinopathy severity grading as a representative task, the system operated without explicit design instructions and relied solely on reference literature and publicly available codebases. The system demonstrated structured co-reasoning between clinical evidence and implementable methodology: clinical insights from ophthalmic literature motivated the explicit separation of global neurodegenerative context and local vascular pathology, which were subsequently translated into a dual-pathway diffusion-based architecture with imbalance-aware objectives and realizable training protocols. Each design choice was justified by identifiable gaps in prior work and mapped to existing implementations, yielding a hypothesis that was both clinically interpretable and experimentally executable. Quantitative evaluation confirmed that the resulting model achieved competitive performance on imbalanced disease stages, supporting the validity of the underlying reasoning process. Taken together with the paradigm-transfer case study, these results demonstrate that the system can not only identify and adapt novel AI paradigms for specified medical tasks, but also systematically refine them through medical–engineering co-reasoning into fully specified, experimentally validated research hypotheses.

2.5.2. Mode 3: Task-driven discovery for medical video restoration

As presented in Fig. A.2, we evaluated the proposed automated medical research system on a clinically motivated task of restoring high-resolution and temporally consistent endoscopic video from low-quality recordings, thereby assessing its ability to autonomously translate emerging AI paradigms into executable solutions for medical research. Starting from a minimally specified task description, the system independently grounded the problem in relevant clinical and technical literature, identified temporal inconsistency as a critical unmet requirement, and selected a recently developed continuous-time video restoration paradigm with demonstrable transfer potential. Without manual intervention, this paradigm was adapted to the endoscopic setting through task-specific architectural and training modifications, yielding a complete research hypothesis and an implementable model. The resulting system was experimentally validated through structured ablations and quantitative evaluation, achieving substantial performance gains over a strong baseline. This case study demonstrates that the proposed framework can automatically operationalize novel AI paradigms for concrete medical tasks, progressing from task specification to validated experimental results, and thereby supports its role as a general-purpose engine for automated medical research rather than a task-specific algorithmic contribution.

3. Discussion

3.1. Key findings

In this study, we introduce Medical AI Scientist, an agentic framework that enables end-to-end automation of medical AI research, spanning hypothesis generation, experimental validation, and manuscript composition. By integrating an Idea Proposer, an automated experimental executor, and a hierarchical Manuscript Composer, the system provides a unified solution for the full research lifecycle. A central design feature lies in the clinician–engineer co-reasoning mechanism, which grounds hypothesis generation in verifiable medical evidence and reduces hallucinations. In parallel, the execution module ensures reliable and iterative model development across heterogeneous clinical data, while the manuscript component translates outputs into structured, evidence-based scientific narratives with embedded ethical compliance. To support systematic evaluation, we further introduce Med-AI Bench, a comprehensive benchmark that standardizes assessment across diverse medical research tasks, modalities, and difficulty levels.

Compared with existing approaches to automated scientific discovery, Medical AI Scientist addresses several key limitations. First, general-purpose language models, although capable of generating plausible research ideas, frequently suffer from insufficient grounding in domain-specific evidence, leading to unreliable or non-actionable hypotheses. Second, existing automation frameworks rarely account for the complexity of clinical data formats and the stringent requirements of medical research reporting and ethics. By contrast, our framework unifies these components into a coherent pipeline, ensuring that each stage is both technically rigorous and clinically grounded.

Our experimental results highlight three principal findings. (1) Superior research idea quality: across six evaluation dimensions, the proposed system consistently outperforms commercial language models and approaches human expert-level assessments, demonstrating strong novelty, feasibility, and interpretability. (2) Robust experimental execution: the system achieves high alignment between proposed methods and implemented experiments, with substantially improved success rates in generating executable and self-consistent pipelines for medical AI development. (3) High-quality manuscript generation: under double-blind expert evaluation, generated manuscripts achieve competitive scores relative to top-tier conference publications, with strong performance in coherence, clarity, and reproducibility, and only minor limitations in content coverage. The acceptance of a system-generated manuscript at ICAIS 2025 further provides early evidence of real-world scientific validity.

The broader implications of Medical AI Scientist extend beyond performance gains to a fundamental shift in how medical AI research may be conducted. By significantly reducing the time and expertise required to move from idea to validated results and polished manuscripts, the framework has the potential to accelerate scientific discovery in healthcare. Its ability to systematically explore complex model designs and translate them into executable implementations suggests a complementary role alongside human researchers, particularly in tasks that demand extensive iteration and technical integration. In clinical and translational settings, such a system could lower barriers to innovation, enabling wider participation in medical AI development and fostering more rapid dissemination of clinically relevant solutions.

3.2. Limitations and future work

Although our Medical AI Scientist demonstrates promising empirical behavior, several limitations remain before it can be considered to match the best human-produced science. First, the conceptual design of the method can at times become overly intricate. This complexity not only increases the

difficulty of faithful implementation, but also introduces instability during execution. When the intended pipeline proves too demanding, the implementation may implicitly simplify or degrade certain components, leading to deviations from the original design and potentially undermining performance. Second, the depth of experimental evaluation is still limited. Current experiments are conducted strictly on predefined datasets, without sufficient exploration of cross-domain or out-of-distribution scenarios. Finally, despite achieving reasonable performance, the generated method does not yet reach state-of-the-art levels. This gap suggests that further refinement is needed, both in terms of algorithmic design and experimental validation, before the AI-generated approach can be considered competitive with leading methods in the field.

Future work will focus on strengthening the experimental pipeline to enable more comprehensive and rigorous evaluations, thereby improving both the robustness and performance of the method. In parallel, we aim to enhance the quality and expressiveness of visualizations, including both empirical plots and framework illustrations, so as to better communicate the underlying mechanisms and results. Through these efforts, we expect the method to evolve into a more reliable and well-rounded system with stronger empirical competitiveness and clearer presentation.

4. Methods

4.1. Building an autonomous AI scientist for medical research

As illustrated in Fig. 6, the Medical AI Scientist comprises three core components: Idea Proposer, Experimental Executor, and Manuscript Composer. Each component is implemented as multi-agents that integrate multiple functionalities through carefully designed prompting strategies. The overall system operates via coordinated interactions among these agents. All agents are built upon general-purpose large language models, such as GPT-5, which serve as the base models for handling a broad range of tasks. For both the Reproduction mode and the Innovation mode, the system takes task instructions, dataset information, and reference papers as inputs, which are then processed sequentially by the Preparer and Surveyor, the Generator, and the Assessor. In contrast, the Exploration mode operates with only task instructions and dataset information as inputs. Building upon the previous two modes, it first introduces an Analyzer and an Explorer to retrieve the medical baseline paper and the novel technological paradigm paper, thereby establishing a sufficient literature foundation for subsequent idea generation. The resulting structured ideas are then formulated as research plans and passed to the Experimental Executor for empirical validation, after which the experimental outputs are further processed into structured manuscripts, yielding the final paper. This entire process enforces a continuous reflect-and-refine cycle, ensuring the final research output (including idea proposal, executable program, and final manuscript) is reproducible and responsible.

4.2. Idea Proposer

The Idea Proposer operationalizes medical hypothesis generation as a structured, evidence-grounded reasoning process. The system is organized into a set of interacting functional modules, each addressing a critical component of scientific ideation. This design highlights the central contribution: a unified framework that couples structured knowledge retrieval with clinician–engineer co-reasoning to produce hypotheses that are both novel and verifiable. At a high level, the Idea Proposer transforms loosely specified medical tasks into executable research hypotheses by iteratively refining problem understanding, identifying appropriate paradigms, and grounding designed ideas in external evidence. This process reduces hallucination and mitigates the tendency of language models to produce superficial or non-actionable ideas.

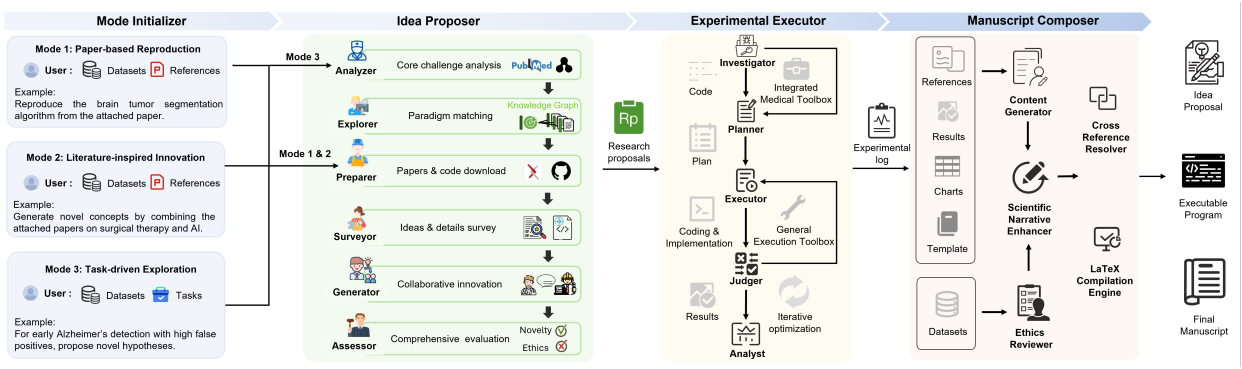

Figure 6 | The conceptual illustration of the Medical AI Scientist: A comprehensive system of fully-automated agents for end-to-end scientific discovery in clinical medicine. The system offers three user interaction modes: Reproduction (reproducing a specified hypothesis), Innovation (innovating from provided literature) and Exploration (autonomously exploring a given research direction), to streamline medical research process. The workflow consists of several phases covering automated idea generation, experiment execution, manuscript writing.

Analyzer. The Medical Task Analyzer formalizes the input problem by identifying its core clinical and technical challenges. Given a user-provided dataset or research objective, this module performs targeted retrieval over peer-reviewed medical and technical literature to construct a structured task representation based on the academic search engine [93]. This representation encodes disease context, data characteristics, evaluation constraints, and implicit clinical needs. This step anchors subsequent hypothesis generation in real clinical gaps rather than abstract problem descriptions.

Explorer. Building on the structured task representation, the Paradigm Explorer identifies the most suitable emerging computational paradigms to address the extracted challenges. Instead of relying on static knowledge, it performs dynamic retrieval over recent literature and open-source repositories, jointly considering methodological novelty, empirical performance, and implementation maturity. A key feature of this module is the explicit alignment between problem structure and algorithmic capability. Candidate paradigms are not selected in isolation but are evaluated based on how their inductive biases and design principles match the identified clinical constraints. For each selected paradigm, the system retrieves corresponding high-quality codebases, ensuring that the resulting hypothesis is directly grounded in executable components.

Preparer and Surveyor. To support informed reasoning, the Preparer and Surveyor jointly construct a structured and executable evidence base that links scientific claims to their operational implementations. The Preparer retrieves relevant literature together with associated code artifacts, normalizing them into a unified representation that captures problem formulations, model designs, and experimental protocols. Inspired by [28], the Surveyor then performs structured synthesis by decomposing each reference into its core conceptual and methodological primitives. Large language models first extract the fundamental research contribution and methodological skeleton while abstracting away domain-specific terminology to reduce surface bias. These abstract directives are subsequently grounded through a multi-agent process that maps them to canonical mathematical formalisms and aligns them with executable code components from open-source repositories. This design enables the system to reconstruct prior methods as verifiable workflows rather than static descriptions, thereby transforming existing work into modular and recomposable units for reasoning. As a result, this module establishes an evidence-grounded substrate for hypothesis construction by explicitly linking theoretical assumptions with their executable implementations.

Generator. Hypothesis generation is performed by the Generator through a clinician–engineer co-

reasoning mechanism that integrates clinical insight with computational design. Rather than relying on unconstrained synthesis, the Generator constructs candidate hypotheses by aligning task-specific challenges with the capabilities of selected paradigms, guided by the structured evidence base. Clinical considerations are introduced in the process to ensure relevance and plausibility, while technical refinements are derived through targeted retrieval and adaptation of existing methods. This bidirectional interaction mitigates the risk of superficial novelty and grounds each hypothesis in verifiable evidence, effectively reducing hallucination. Iterative refinement continues until the hypothesis achieves internal coherence across clinical validity, methodological soundness, and implementation feasibility. This structured process parallels human-led medical hypothesis formation and enables systematic derivation of high-level ideas from clear gaps in existing literature.

Assessor. The final stage evaluates the generated hypothesis through a combination of scientific and ethical criteria. The Assessor examines conceptual consistency, empirical support, and practical executability. In parallel, an explicit ethics check ensures compliance with biomedical research standards. Hypotheses that fail to meet quality thresholds are returned for refinement, while those violating ethical constraints are rejected. This mechanism enforces rigor and accountability, ensuring that only well-supported and responsible ideas proceed to experimental validation. The resulting hypothesis is formalized as a detailed research plan, which specifies the algorithmic rationale and anticipated evaluation protocols.

4.3. Experimental Executor

The experimental executor is formulated as a structured multi-stage pipeline for traceable and self-correcting model development within a secure Dockerized environment. Given a research objective, the **Investigator** assembles the required codebase together with domain-specific medical toolboxes to ensure compatibility with heterogeneous clinical data, and provides this unified specification to the **Planner**, which decomposes it into a structured, machine-interpretable execution protocol with defined inputs and outputs. The **Executor** instantiates this protocol within a controlled environment by constructing the full training and evaluation pipeline, leveraging general-purpose execution toolchains for scalable and stable implementation. Resulting logs, intermediate outputs, and quantitative metrics are assessed by the **Judger**, which evaluates consistency between intended design and the observed behavior and produces targeted corrective feedback. The Analyst consolidates validated results into structured records for downstream use. Through iterative feedback and execution-level correction, the system unifies domain-specific medical processing with general execution infrastructures, enabling reliable, iterative, and self-correcting validation under complex clinical settings.

4.4. Manuscript Composer

The Manuscript Composer operates within an end-to-end multi-agent framework that transforms substantiated research materials into a typeset-ready paper. The **Content Generator** first establishes the global structure of the manuscript by leveraging the organizational patterns of the most relevant reference papers, and subsequently develops section level content grounded in evidence from a structured repository of implementations, experimental logs, and quantitative results. To preserve narrative coherence, concise summaries of previously generated sections are retained and reused as semantic anchors during subsequent drafting. The Generator further aligns narrative and presentation by automatically generating experimental figures from logged results and synthesizing architectural diagrams from method specifications. By summarizing current conference and journal policies into structured instructions, the **Ethics Reviewer** leverages dataset-specific evidence to rigorously report and cite the origin, license, and ethical approval of each dataset to meet publishing requirements.

In parallel, a **Scientific Narrative Enhancer** is introduced to counter the tendency of AI generated text to overemphasize procedural detail, refining the manuscript to improve clarity and the scientific storyline while aligning the writing style with task-specific paradigms. A **Cross-Reference Resolver** subsequently verifies internal references, including equations, figures, sections, and citations. Finally, a self-healing mechanism in **Latex Compilation Engine** continuously validates the LaTeX source, interpreting compiler feedback to autonomously correct syntactic or structural errors and ensure reliable compilation without manual intervention. Together, these components enable the automated generation of coherent, compliant, and publication ready medical manuscripts from heterogeneous research artifacts.

4.5. Construction of Med-AI Bench

To enable systematic and reproducible evaluation of the Medical AI Scientist, we constructed Med-AI Bench, a benchmark comprising 171 cases derived from 57 high-quality ground-truth medical research papers. Construction began with the six primary data modalities identified in a scoping review of multimodal AI in medicine [94]: (1) medical images, (2) videos, (3) electronic health records (EHR, including structured ICU data), (4) text, (5) physiological signals (e.g., ECG and EEG), and (6) multimodal data.

The tasks for each modality were derived from authoritative domain surveys as follows: medical imaging tasks (classification, prognosis, restoration, segmentation, and registration) from a comprehensive review of AI-driven imaging innovations [95]; video-analysis tasks (instrument detection, restoration, workflow recognition, intraoperative risk assessment, and postoperative skill assessment) from a scoping review of AI in medical videos [96]; EHR tasks (risk prediction and clinical decision support) from a comparative analysis of deep learning architectures for EHR [97]; physiological-signal tasks (disease diagnosis and prognosis) from a review of signal-based healthcare applications [98]; clinical text tasks (report summarization, text-based diagnosis/risk assessment, and biomedical question answering) from a UK-focused clinical NLP survey [99]; and multimodal tasks (multimodal diagnosis and cross-modal report generation) from a dedicated multimodal biomedical AI review [100]. This structured process yielded 19 distinct tasks.

For each task, three representative papers were retrieved from Google Scholar using task-specific keyword combinations, with explicit prioritization of highly cited works. Each paper was independently scored on five dimensions: Code Availability (presence and usability of public implementations), Venue Quality (prestige ranking of the publication venue), Citations (normalized citation count), Year and Complexity (publication recency weighted by methodological intricacy), and Subjective Human Rating (by domain experts). Papers were subsequently ranked and partitioned into three difficulty tiers (hard, medium, easy; one paper per tier per task) from an AI-implementation perspective. For each paper, three cases were constructed using different input modes. The resulting 171 cases form a stratified benchmark that systematically spans technical and clinical complexity, enabling rigorous assessment of hypothesis generation, implementation fidelity, and manuscript quality. It is worth noting that, to speed up the execution and validation of automated experiments, we performed random subsampling on the dataset.

4.6. Performance assessment of the Medical AI Scientist

We evaluated the Medical AI Scientist on Med-AI Bench across four core dimensions: (1) Idea Generation, (2) Implementation Completeness, and (3) Code Execution, against the strongest closed-source models (GPT-5 and Gemini-2.5-Pro) under identical input conditions. All evaluation criteria are scored on a five-point scale ranging from 1 to 5, ensuring consistent and interpretable assessment

across all cases.

For Idea Generation, the Idea Proposer and baseline models received equivalent prompts (either literature-derived innovation or autonomous exploration of a user-specified direction) and produced full research proposals. Each proposal was scored by a hybrid evaluator that combined LLM-based metrics with blinded assessments from professional clinical AI scientists. Scoring followed standardized rubrics across six dimensions: Novelty (substantive innovation in medical problem modeling), Maturity (completeness and ease of implementation), Ethicality (responsible handling of medical data and constraints), Generalizability (robustness across devices, populations, and institutions), Utility (potential for real clinical adoption), and Interpretability (alignment with medical reasoning and traceability). Explicit evidence grounding ensured high inter-rater reliability. For Implementation Completeness, the full proposals were fed into the Experiment Executor (our system) or the equivalent code-generation modules of the baselines, producing complete executable programs. LLM-based scoring then assessed two aspects: fidelity of core innovative components and completeness of the pipeline (data preprocessing, training, validation, testing, and logging). Code Execution directly deployed the generated code in a predefined Dockerized environment. Success was defined as the fraction of runs that completed without errors, exhibited monotonically decreasing training loss, and produced valid model weights accompanied by quantitative test results.

In addition, all Medical AI Scientist-generated manuscripts were submitted to the Stanford Agentic Reviewer under the complete ICLR review protocol. The system returned an overall score on a scale from 0 to 10, together with structured strengths and weaknesses, providing an independent multi-criteria validation of scientific rigor.

4.7. Human expert evaluation

To assess real-world usability in a controlled yet ecologically valid setting, we restricted the evaluation to a single classic medical AI task: diabetic retinopathy classification on fundus images, while preserving full methodological generality.

We invited 10 independent human experts, each with more than five years of first-author experience in AI-for-healthcare publications. Using a double-blind protocol, experts rated a total of 20 papers: five papers autonomously generated by the Medical AI Scientist on the constrained task and 15 high-impact human-authored papers (five randomly selected via keyword search from each of the MICCAI, BIBM, and ISBI conferences, prioritized by citation rank). To eliminate any potential source bias from formatting or stylistic templates, all human-authored papers had their original templates, fonts, and layouts removed, with only the core content retained.

Experts scored every paper on five dimensions using the same standardized Likert-scale rubrics: Novelty (degree of methodological innovation relative to prior art), Coherence (logical flow and internal consistency of the scientific narrative), Coverage (comprehensiveness of experimental design), Clarity (precision and conciseness of exposition), and Reproducibility (sufficiency of methodological detail). All evaluation criteria are scored on a five-point scale ranging from 1 to 5. Also, all ratings were collected anonymously to eliminate source bias, enabling direct quantitative comparison of perceived quality and practical utility between AI-generated and human-authored medical research outputs.

5. Related Work

5.1. AI agent systems and multi-agent collaboration

The evolution of AI agent systems has shifted from single-agent tool integrations to advanced multi-agent architectures that enable sophisticated collaboration and task decomposition. Early approaches focused on enhancing individual agents' capabilities, such as ReAct [101], which combines reasoning and action by prompting LLMs to generate interleaved thoughts and actions for dynamic environmental interactions. Building on this, Toolformer [102] enables LLMs to learn tool usage autonomously through fine-tuning with API calls, supporting zero-shot applications in tasks requiring external resources. These foundations have paved the way for more integrated frameworks like LangChain [103], which facilitates chaining components for complex applications, and its extension LangGraph [104], which introduces graph-based orchestration for managing stateful multi-agent systems. Similarly, Semantic Kernel [105] integrates plugins for enterprise-level AI orchestration with an emphasis on semantic planning and memory persistence.

Building upon these integrated frameworks, advancements in multi-agent collaboration have produced systems that simulate team-based dynamics through role assignments and structured interactions. MetaGPT [22] employs standardized operating procedures to coordinate agents in workflows akin to software development teams, while CAMEL [21] uses role-playing to align autonomous agents with user goals. More specialized frameworks like CrewAI [106] assemble agent teams for sequential tasks such as research synthesis, and OpenAgents [23] deploys multiple agents to provide accessible capabilities for data analysis, plugin usage, and web navigation. Extending these paradigms, systems like Auto-GPT [107] and Devin [108] operate as autonomous AI engineers for full-cycle software development, while Manus [24] and its open-source counterpart OpenManus [25] support complex, cloud-based task execution. However, these frameworks highlight the evolution toward robust coordination but often lack the deep reasoning required for scientific innovation, such as hypothesis formulation and domain-specific adaptation.

5.2. Autonomous AI-driven scientific discovery systems

Autonomous scientific discovery systems automate key research stages, from ideation to dissemintiond. The AI Scientist [26] pioneers an end-to-end automated pipeline that generates ideas, runs experiments, and drafts manuscripts, operating in an open-ended loop to build upon its own findings. Its successor, AI Scientist-v2 [27], enhances this autonomy by incorporating an agentic tree-search for deeper hypothesis exploration and successfully generating a manuscript that passed peer review at a major conference workshop. AI-Researcher [28] introduces a multi-agent architecture that maintains coherence through bidirectional mappings between mathematical concepts and code, mitigating hallucinations. DeepScientist [29] frames scientific discovery as a Bayesian Optimization problem, using an agent to iteratively balance exploration and exploitation to discover novel methods. Agent Laboratory [109] extends this by automating the execution and reporting of user-provided ideas, acting as an accelerator for human researchers rather than an independent ideator. In contrast, Google's AI co-scientist [110] operates as a collaborator in a "scientist-in-the-loop" paradigm, leveraging models like Gemini to assist domain experts with hypothesis generation.

Alongside these frameworks, complementary toolkits have been developed to support AI agent systems by enhancing resource integration and accessibility. ToolUniverse [111] provides an expansive repository of scientific tools governed by a standardized AI-tool interaction protocol, enabling agents to discover and orchestrate diverse tools seamlessly. Paper2agent [112] transforms research papers into executable agents by encapsulating their contributions into a standardized Model Context Protocol (MCP), allowing for interactive, natural-language-based reproduction and analysis. Code2MCP [113]

further streamlines this by converting code repositories into standardized services, facilitating seamless tool incorporation into agent workflows. While effective for general research automation, these systems frequently overlook clinical necessities such as ethical compliance and specialized data processing, shortcomings our framework mitigates with dedicated medical stages and verification processes.

5.3. AI applications and challenges in clinical medicine

Artificial intelligence has made substantial impacts in clinical medicine, with specialized models achieving expert-level performance on well-defined tasks. These tasks include disease classification [1–3], lesion segmentation [5, 6], prognostic prediction [8–10] and enhanced surgical navigation [114–116]. As technology has evolved, multimodal large language models (MLLMs) have emerged, integrating diverse data types such as text and images to perform more complex, comprehensive tasks. For instance, models like Med-Gemini [117] leverage vision-language processing to support medical report generation and treatment recommendations, while frameworks such as LLaVA-Med [118] facilitate multimodal analysis in radiology.

However, these advances primarily consist of specialized models whose operation and integration still rely heavily on human experts to drive the entire research project. Researchers must be responsible for identifying clinical problems, formulating hypotheses, designing experiments, and ensuring ethical compliance. To our knowledge, no existing framework bridges the autonomous orchestration capabilities of a general AI Scientist with the domain-specific knowledge, tools, and ethical constraints of clinical medicine. Medical AI Scientist aims to fill this gap, enabling autonomous, clinically meaningful, and ethically responsible innovation.

References

[1] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[2] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[3] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Banming Yang, Hershel Mehta, Tony Duan, Daisy Ding, Karan Bagaria, Jenny Ball, Curtis Langlotz, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[4] Sheng Zhang, Yuhong Xu, Naoto Usuyama, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[5] Fabian Isensee, Jens Petersen, Andreas Klein, et al. nnU-Net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[6] Ali Hatamizadeh, Yan Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[7] Jun Ma, Yining He, Feng Li, et al. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[8] Pooya Mobadersany, Saeed Yousefi, Mohamed Amgad, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

[9] Xintian Wang, Jian Zhao, Elena Marostica, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

[10] Yizhou Chen, Bo Wang, Yifan Zhao, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. *Nature Communications*, 15(1):1657, 2024.

[11] OpenAI. Introducing gpt-5. Available at https://openai.com/index/introducing-gpt-5, August 2025.

[12] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[13] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[14] xAI. Grok 4. Available at https://x.ai/news/grok-4, July 2025.

[15] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[16] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[17] OpenAI. Introducing deep research, 2025. URL https://openai.com/index/introducing-deep-research/. Accessed: 2025-04-06.

[18] Google Team. Introducing gemini deep research, 2025. URL https://gemini.google/overview/deep-research/. Accessed: 2025-04-06.

[19] xAI Team. Introducing grok deepsearch, 2025. URL https://x.ai/news/grok-3. Accessed: 2025-04-06.

[20] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[21] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[22] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

[23] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. Openagents: An open platform for language agents in the wild, 2023. URL https://arxiv.org/abs/2310.10634.

[24] Manus Technologies. Manus: Structured and ai-assisted academic writing tool. https://manus.im/, 2025.

[25] OpenManus Contributors. Openmanus: Open-source framework for building general ai agents. https://openmanus.github.io/, 2025.

[26] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[27] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[28] Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025.

[29] Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. Deepscientist: Advancing frontier-pushing scientific findings progressively. *arXiv preprint arXiv:2509.26603*, 2025.

[30] Chandan K Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 28601–28609, 2025.

[31] Yolanda Gil, Mark Greaves, James Hendler, and Haym Hirsh. Amplify scientific discovery with artificial intelligence. *Science*, 346(6206):171–172, 2014.

[32] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[33] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[34] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[35] Zhongguancun Academy. The 1st international conference on ai scientists (icais 2025). Website, November 2025. URL https://icais.ai/. Tagline: Exploring the frontiers of automated scientific discovery with AI Scientists and autonomous research agents.

[36] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791, 2023.

[37] Xiangzuo Huo, Gang Sun, Shengwei Tian, Yan Wang, Long Yu, Jun Long, Wendong Zhang, and Aolun Li. Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, 87:105534, 2024.

[38] Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Zhaohu Xing, and Lei Zhu. Diffmic-v2: Medical image classification via improved diffusion network. *IEEE Transactions on Medical Imaging*, 2025.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[41] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97: 103280, 2024.

[42] Renato Hermoza, Gabriel Maicas, Jacinto C Nascimento, and Gustavo Carneiro. Post-hoc overall survival time prediction from brain mri. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1476–1480. IEEE, 2021.

[43] Lina Chato and Shahram Latifi. Machine learning and deep learning techniques to predict overall survival of brain tumor patients using mri images. In *2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE)*, pages 9–14. IEEE, 2017.

[44] Yin Lin, Riccardo Barbieri, Domenico Aquino, Giuseppe Lauria, Marina Grisoli, Elena De Momi, Alberto Redaelli, and Simona Ferrante. Glioblastoma overall survival prediction with vision transformers. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE, 2025.

[45] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[46] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022.

[47] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In *European conference on computer vision*, pages 347–364. Springer, 2022.

[48] Hong Wang, Yuexiang Li, Nanjun He, Kai Ma, Deyu Meng, and Yefeng Zheng. Dicdnet: Deep interpretable convolutional dictionary network for metal artifact reduction in ct images. *IEEE Transactions on Medical Imaging*, 41(4):869–880, 2021.

[49] Hong Wang, Qi Xie, Dong Zeng, Jianhua Ma, Deyu Meng, and Yefeng Zheng. Oscnet: Orientation-shared convolutional network for ct metal artifact learning. *IEEE Transactions on Medical Imaging*, 43(1):489–502, 2023.

[50] Hong Wang, Minghao Zhou, Dong Wei, Yuexiang Li, and Yefeng Zheng. Mepnet: A model-driven equivariant proximal network for joint sparse-view reconstruction and metal artifact reduction in ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–120. Springer, 2023.

[51] Hongqiu Wang, Guang Yang, Shichen Zhang, Jing Qin, Yike Guo, Bo Xu, Yueming Jin, and Lei Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 2024.

[52] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023.

[53] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022.

[54] Xinyu Liu, Guolei Sun, Cheng Wang, Yixuan Yuan, and Ender Konukoglu. Medvsr: Medical video super-resolution with cross state-space propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11697–11707, 2025.

[55] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[56] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021.

[57] Shu Yang, Luyang Luo, Qiong Wang, and Hao Chen. Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–616. Springer, 2024.

[58] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging*, 40(7):1911–1923, 2021.

[59] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 37(5):1114–1126, 2017.

[60] Masahiro Kawamura, Yuichi Endo, Atsuro Fujinaga, Hiroki Orimoto, Shota Amano, Takahide Kawasaki, Yoko Kawano, Takashi Masuda, Teijiro Hirashita, Misako Kimura, et al. Development of an artificial intelligence system for real-time intraoperative assessment of the critical view of safety in laparoscopic cholecystectomy. *Surgical Endoscopy*, 37(11):8755–8763, 2023.

[61] Pietro Mascagni, Armine Vardazaryan, Deepak Alapatt, Takeshi Urade, Taha Emre, Claudio Fiorillo, Patrick Pessaux, Didier Mutter, Jacques Marescaux, Guido Costamagna, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of surgery*, 275(5):955–961, 2022.

[62] Franciszek M Nowak, Evangelos B Mazomenos, Brian Davidson, and Matthew J Clarkson. Swincvs: a unified approach to classifying critical view of safety structures in laparoscopic cholecystectomy. *International Journal of Computer Assisted Radiology and Surgery*, 20(6): 1145–1152, 2025.

[63] Aneeq Zia and Irfan Essa. Automated surgical skill assessment in rmis training. *International journal of computer assisted radiology and surgery*, 13(5):731–739, 2018.

[64] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.

[65] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9522–9531, 2021.

[66] Sujeong Im, Jungwoo Oh, and Edward Choi. Labtop: A unified model for lab test outcome prediction on electronic health records. *arXiv preprint arXiv:2502.14259*, 2025.

[67] Raphael Poulain and Rahmatollah Beheshti. Graph transformers on ehrs: Better representation improves downstream performance. In *The twelfth international conference on learning representations*, 2024.

[68] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

[69] Mengxuan Sun, Jinghao Niu, Xuebing Yang, Yifan Gu, and Wensheng Zhang. Cehmr: Curriculum learning enhanced hierarchical multi-label classification for medication recommendation. *Artificial Intelligence in Medicine*, 143:102613, 2023.

[70] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133, 2019.

[71] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*, 2021.

[72] Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

[73] Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Advances in Neural Information Processing Systems*, 37:36314–36341, 2024.

[74] Shunxiang Yang, Cheng Lian, Zhigang Zeng, Bingrong Xu, Junbin Zang, and Zhidong Zhang. A multi-view multi-scale neural network for multi-label ecg classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(3):648–660, 2023.

[75] Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto-Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature communications*, 12(1):5117, 2021.

[76] Sushravya Raghunath, Alvaro E Ulloa Cerna, Linyuan Jing, David P VanMaanen, Joshua Stough, Dustin N Hartzel, Joseph B Leader, H Lester Kirchner, Martin C Stumpe, Ashraf Hafez, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature medicine*, 26(6):886–891, 2020.

[77] Veer Sangha, Bobak J Mortazavi, Adrian D Haimovich, Antônio H Ribeiro, Cynthia A Brandt, Daniel L Jacoby, Wade L Schulz, Harlan M Krumholz, Antonio Luiz P Ribeiro, and Rohan Khera. Automated multilabel diagnosis on electrocardiographic images and signals. *Nature communications*, 13(1):1583, 2022.

[78] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[79] Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. *arXiv preprint arXiv:2107.00176*, 2021.

[80] Wenpeng Lu, Sibo Wei, Xueping Peng, Yi-Fei Wang, Usman Naseem, and Shoujin Wang. Medical question summarization with entity-driven contrastive learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–19, 2024.

[81] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[82] Sara Nouri Golmaei and Xiao Luo. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

[83] Huiting Ma, Dengao Li, Jumin Zhao, Wenjing Li, Jian Fu, and Chunxia Li. Hr-bgcn: Predicting readmission for heart failure from electronic health records. *Artificial Intelligence in Medicine*, 150:102829, 2024.

[84] Georg Wiese, Dirk Weissenborn, and Mariana Neves. Neural question answering at bioasq 5b. *arXiv preprint arXiv:1706.08568*, 2017.

[85] Zi Yang, Yue Zhou, and Eric Nyberg. Learning to answer biomedical questions: Oaqa at bioasq 4b. In *Proceedings of the Fourth BioASQ workshop*, pages 23–37, 2016.

[86] Hajung Kim, Hoonick Lee, Yewon Cho, Jungwoo Park, Jueon Park, Soyon Park, Yan Ting Chok, Seungheun Baek, Donghyeon Lee, and Jaewoo Kang. Prompting matters: snippet-aware strategies for biomedical qa with llms in bioasq 13b. In *CLEF*, 2025.

[87] Yilan Zhang, Fengying Xie, and Jianqi Chen. Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis. *Computers in biology and medicine*, 157:106712, 2023.

[88] Yuan Zhang, Yutong Xie, Hu Wang, Jodie C Avery, M Louise Hull, and Gustavo Carneiro. A novel perspective for multi-modal multi-label skin lesion classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3549–3558. IEEE, 2025.

[89] Matthew J Cockayne, Marco Ortolani, and Baidaa Al-Bander. Dermformer: nested multi-modal vision transformers for robust skin cancer detection. *Pattern Analysis and Applications*, 28(4): 194, 2025.

[90] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80: 102510, 2022.

[91] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. Recap: Towards precise radiology report generation via dynamic disease progression reasoning. *arXiv preprint arXiv:2310.13864*, 2023.

[92] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[93] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

[94] Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications. *Medical image analysis*, 105:103621, 2025.

[95] Luís Pinto-Coelho. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering*, 10(12):1435, 2023.

[96] Anni King, George E Fowler, Rhiannon C Macefield, Hamish Walker, Charlie Thomas, Sheraz Markar, Ethan Higgins, Jane M Blazeby, and Natalie S Blencowe. Use of artificial intelligence in the analysis of digital videos of invasive surgical procedures: scoping review. *BJS open*, 9 (4):zraf073, 2025.

[97] Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of biomedical informatics*, 101:103337, 2020.

[98] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

[99] Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael TC Poon, Natalie Fitzpatrick, Adam P Levine, et al. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *NPJ digital medicine*, 5(1): 186, 2022.

[100] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784, 2022.

[101] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[102] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551, 2023.

[103] LangChain Contributors. Langchain: Build context-aware reasoning applications. https://github.com/langchain-ai/langchain, 2023.

[104] LangChain Contributors. Langchain: Build resilient language agents as graphs. https://github.com/langchain-ai/langgraph, 2024.

[105] Microsoft. *Semantic Kernel Documentation*, 2025. URL https://learn.microsoft.com/en-us/semantic-kernel/. Accessed: 2025-10-28.

[106] João Moura and contributors. crewai: Framework for building ai agent teams, 2024. URL https://github.com/crewAIInc/crewAI. Accessed: 2025-10-28.

[107] Significant Gravitas Contributors. Autogpt: Accessible ai agents for everyone. https://github.com/Significant-Gravitas/AutoGPT, 2023.

[108] Cognition AI. Introducing devin, the first ai software engineer, 2024. URL https://www.cognition-labs.com/introducing-devin. Accessed: 2025-10-28.

[109] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

[110] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[111] Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, et al. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*, 2025.

[112] Jiacheng Miao, Joe R Davis, Jonathan K Pritchard, and James Zou. Paper2agent: Reimagining research papers as interactive and reliable ai agents. *arXiv preprint arXiv:2509.06917*, 2025.

[113] Chaoqian Ouyang, Ling Yue, Shimin Di, Libin Zheng, Linan Yue, Shaowu Pan, Jian Yin, and Min-Ling Zhang. Code2mcp: Transforming code repositories into mcp services. *arXiv preprint arXiv:2509.05941*, 2025.

[114] Alexey Shvets, Alexander Rakhlin, Alexandr Kalinin, and Vladimir Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1103–1108. IEEE, 2018.

[115] Andru P Twinanda, S Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

[116] Lena Maier-Hein, Matthias Eisenmann, Alperen Sarikaya, Tobias Collins, and et al. Surgical data science: Towards semantic understanding of operative workflows. *Nature Biomedical Engineering*, 6(10):957–972, 2022.

[117] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

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

Appendix

Figure A.1 | Example of Medical AI Scientist Innovation Mode on the medical image classification task. All necessary processes and outcomes are presented, with human expert's notes in bottom boxes.

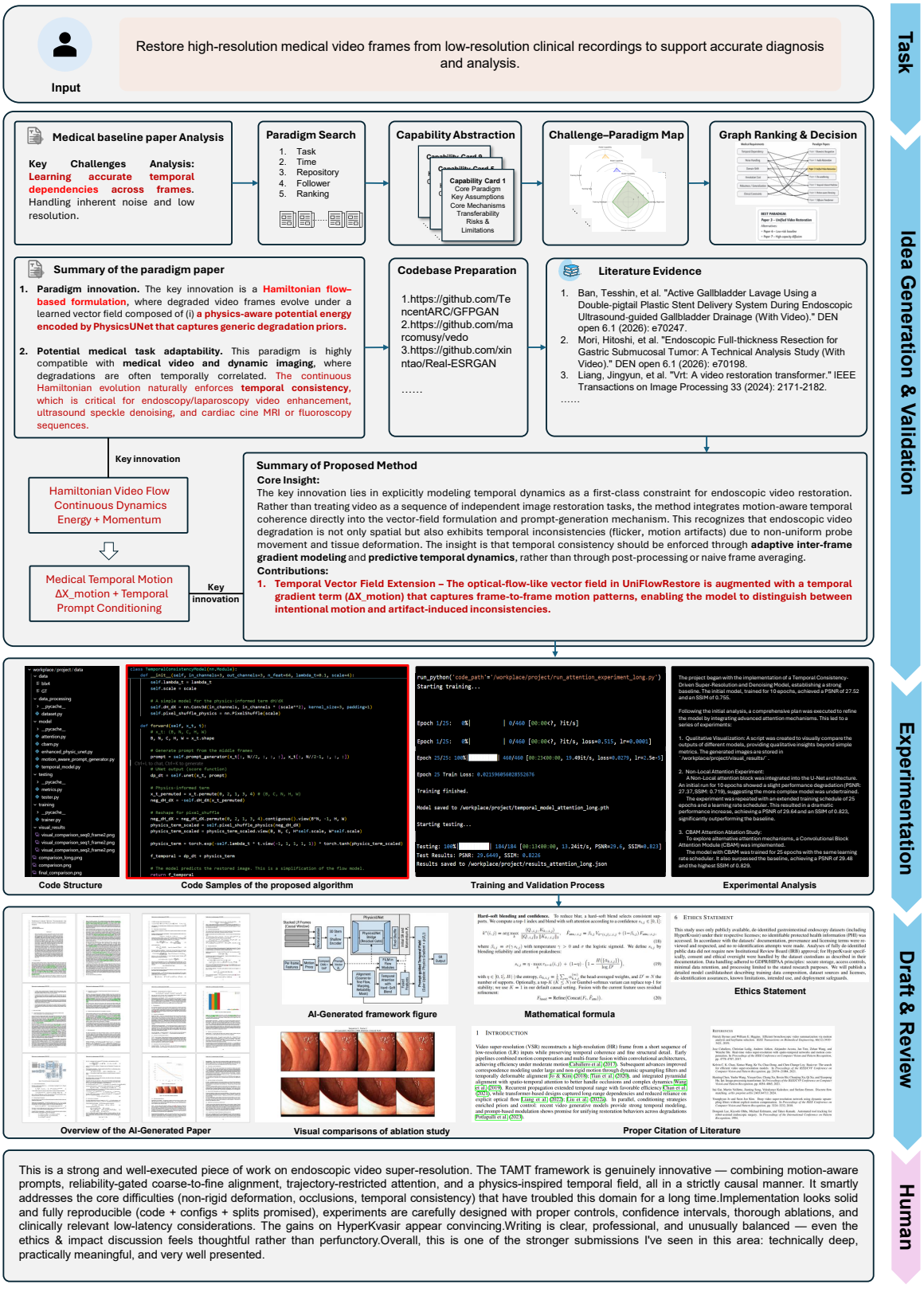

Figure A.2 | Example of Medical AI Scientist Exploration Mode on the medical video restoration task. All necessary processes and outcomes are presented, with human expert's notes in bottom boxes.

TOPOLOGY-AWARE LATENT DIFFUSION WITH DISC-CENTERED POLAR POOLING FOR DIABETIC RETINOPATHY GRADING

Anonymous Author(s)

ABSTRACT

We address diabetic retinopathy grading from a single color fundus photograph, a five-stage ordinal task in which lesion interpretation depends on disc-centered vascular topology and quadrant extent. Current CNN/ViT systems often underuse such cues and are brittle to acquisition variability and class imbalance near referral thresholds. We propose a topology-aware latent diffusion framework operating in a compact latent space that fuses global context, lesion evidence, and a differentiable multi-scale vesselness map. A learnable disc-centered polar pooling module yields an anatomy-aware conditioning vector that summarizes vessel structure and disc and quadrant extent in line with ETDRS, and conditional latent diffusion refines task embeddings to sharpen ordinal separation. A classifier with momentum-updated class centers and joint global/local distribution alignment regularizes the latent space, and deterministic sampling conditioned on the anatomy vector stabilizes predictions at inference. By explicitly encoding vascular topology and disc- and quadrant-derived extent, the approach increases robustness to acquisition variability and better supports guideline-based referral decisions from a single fundus image. On APTOS 2019 under a stratified split and shared evaluation protocol against CNN and ViT baselines, our method improves ordinal agreement and class-aware performance.

1 INTRODUCTION

Automated analysis of retinal fundus photographs has become a cornerstone of population-scale screening in computer vision for healthcare. Diabetic retinopathy (DR) grading is a five-stage ordinal task in which severity is determined by the spatial extent of lesions in relation to the retinal vasculature. Clinical guidelines, notably ETDRS, define referral thresholds by quadrant-referenced burden and emphasize vascular structure around the optic disc for decision making Early Treatment Diabetic Retinopathy Study Research Group (1991). The concrete problem we address is to learn representations that couple lesion appearance with anatomy-aware cues—disc- and quadrant-referenced extent and vessel topology—while remaining reliable under common acquisition variability.

Deep convolutional networks and vision transformers have driven progress in DR grading by capturing global and local appearance, often augmented with attention or alignment objectives Gulshan et al. (2016); Ting et al. (2017); Dosovitskiy et al. (2021). Diffusion models further introduce generative priors that can regularize features and improve data efficiency Ho et al. (2020); Rombach et al. (2022); Yang et al. (2023). Despite these advances, existing models are predominantly texture-biased: they recognize lesion morphology but only weakly encode clinical structure such as extent relative to the optic disc and quadrants, or the branching geometry of vessels. This gap is most consequential near referral thresholds, where separating advanced non-proliferative from proliferative disease depends on lesion distribution along the vascular tree rather than appearance alone. Moreover, routine shifts in acquisition—optical blur, motion, uneven illumination and color casts, variation in field of view, and disc–fovea positioning—alter perceived contrast and disrupt quadrant-based assessment, degrading calibration and minority-grade recall. The ordinal label structure and class imbalance further make both optimization and evaluation sensitive to representation quality.

Addressing these limitations is important for safe and equitable screening. Accurate and calibrated decisions reduce missed sight-threatening disease while avoiding unnecessary referrals, a balance that hinges on sensitivity to minority grades and stability around decision boundaries. Robust generalization across devices, sites, and preprocessing pipelines is essential for deployment at scale. Anatomy-aware signals provide a principled path toward these goals: vessel topology is a stable geometric anchor under color and illumination shifts; disc- and quadrant-referenced extent encodes the clinical semantics used by experts; and embeddings that align appearance with vascular context can improve ordinal separability and calibration under acquisition variability. Bridging appearance and anatomy thus has direct impact on both performance and trustworthiness.

We propose a vessel-topology aware latent diffusion framework for five-class DR grading from color fundus photographs. The core idea is to integrate global context and lesion appearance with an explicit vesselness signal through an anatomy-aware conditioning vector that summarizes vascular topology and disc- and quadrant-referenced extent in line with ETDRS. Conditional latent diffusion then refines task embeddings so that texture cues are interpreted relative to vascular geometry, encouraging ordinally consistent representations and reducing sensitivity to routine shifts in acquisition. We evaluate the approach against CNN and ViT baselines under class imbalance, and conduct ablations that isolate the contribution of anatomy-aware conditioning and diffusion on ordinal agreement, minority-grade performance, and calibration across acquisition variability.

Our main contributions are as follows.

- We introduce a vessel-topology aware latent diffusion framework that unifies global context, lesion appearance, and explicit vascular structure for ordinal DR grading.
- We design disc- and quadrant-aware conditioning that encodes vascular topology and lesion extent in accordance with ETDRS, guiding representation learning toward clinically meaningful structure.
- We demonstrate improved ordinal agreement, minority-grade recall, and calibration robustness under routine acquisition variability compared with strong CNN and ViT baselines.
- We provide a systematic evaluation with ablations that quantify the individual effects of anatomy-aware conditioning and diffusion priors on performance and stability.

2 Related Work

2.1 Diffusion Models, Efficient Solvers, and Conditional Guidance

Diffusion probabilistic models cast generation as reverse-time denoising of a fixed noising process and link variational inference, score matching, and SDE formulations Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021). Notable developments include DDPMs with noise-prediction objectives and practical training Ho et al. (2020), continuous-time score-based modeling with probability-flow ODEs that enable stable deterministic sampling Song et al. (2021), and latent diffusion, which scales to high-resolution images by operating in a perceptual latent space Rombach et al. (2022). Inference has been sped up by non-Markovian, often deterministic trajectories (DDIM) that reduce the number of steps while approximately preserving DDPM marginals Song et al. (2020), and by specialized ODE solvers that cut function evaluations with minor quality loss (e.g., DPM-Solver) Lu et al. (2022). Conditioning has evolved from classifier guidance, which modifies the reverse dynamics using gradients of conditional log-likelihoods Dhariwal & Nichol (2021), to classifier-free guidance that mixes conditional and unconditional predictions without an auxiliary classifier Ho & Salimans (2022), and to structured-edit methods that inject spatial priors during sampling (e.g., RePaint) Lugmayr et al. (2022). These trends have shifted practice from many-step pixel-space diffusion toward more efficient, often deterministic sampling with stronger and more flexible conditioning.

We adopt DDPM-style training and use DDIM- and ODE-based accelerations for few-step, stable sampling Ho et al. (2020); Song et al. (2020); Lu et al. (2022), operate in latent space for efficiency Rombach et al. (2022), and use simple conditioning aligned to the task, inspired by classifier-free guidance, to direct denoising toward clinically relevant structures, with guidance strength tuned for deterministic solvers Dhariwal & Nichol (2021); Ho & Salimans (2022).

2.2 Medical Imaging for DR: Attention, Alignment, and Clinically Grounded Priors

Early DR screening achieved strong CNN baselines trained end to end on fundus photographs Gulshan et al. (2016); Ting et al. (2017). Attention and saliency methods (e.g., Grad-CAM) improved interpretability but often produced coarse, texture-biased explanations Selvaraju et al. (2017). Transformer-based backbones strengthened long-range context modeling Dosovitskiy et al. (2021), and distribution-alignment objectives such as MMD mitigated cross-dataset shifts Gretton et al. (2012). Clinical heuristics formalized by ETDRS emphasize vessel-level signs (e.g., venous beading, IRMA) and neovascularization near the optic disc (NVD) or elsewhere (NVE), underscoring the importance of topology- and location-aware cues for grading Early Treatment Diabetic Retinopathy Study Research Group (1991). Classical multiscale vesselness filters (e.g., Frangi) enhance tubular structures and provide anatomy-aware priors Frangi et al. (1998). Recent diffusion-enabled classifiers combine dual guidance and alignment to connect global context with local evidence (e.g., DiffMIC) Yang et al. (2023). Breakdown of the blood–retinal barrier causes edema and exudation that confound texture-based cues Cunha-Vaz (2010), and microglia-mediated neurovascular interactions influence angiogenesis and tuft dynamics in DR, motivating priors focused on vascular topology and disc-/quadrant-aware context Hu et al. (2024).

We combine attention-informed conditioning with clinically grounded priors. A differentiable multiscale vesselness channel and disc-/quadrant-aware pooling emphasize vascular topology and NVD/NVE localization within a latent diffusion pipeline, and class-aware alignment stabilizes the integration of global and local cues for cross-dataset robustness Frangi et al. (1998); Early Treatment Diabetic Retinopathy Study Research Group (1991); Gretton et al. (2012); Yang et al. (2023). This design targets fine-grained lesion geometry and neurovascular context beyond texture cues Cunha-Vaz (2010); Hu et al. (2024).

2.3 Latent Representation Refinement, Metric Learning, and Topology-Aware Constraints

Metric learning shapes latent spaces with center- and margin-based objectives to reduce intra-class variance and enlarge inter-class margins (e.g., Center Loss, ArcFace) Wen et al. (2016); Deng et al. (2019). Proxy-based variants improve efficiency and stability under multimodality (e.g., Proxy-Anchor) Kim et al. (2020). Long-tail settings common in DR benefit from logit adjustment, which incorporates label-frequency priors to recalibrate decision boundaries without heavy reweighting Menon et al. (2020). To align multi-branch or cross-domain representations, kernel MMD and deep adaptation (DAN) align feature distributions in reproducing kernel Hilbert spaces or via joint feature adaptation Gretton et al. (2012); Long et al. (2015). Topology-aware objectives aim to preserve connectivity in tubular anatomy (e.g., clDice) and penalize topological errors (TopoLoss), and topological autoencoders align latent metrics with data topology Shit et al. (2021); Clough et al. (2020); Moor et al. (2019). Recent diffusion-based classifiers use generative likelihoods and denoising for robust prediction Chen et al. (2024).

We integrate center-based regularization with class-aware MMD alignment inside a latent diffusion framework to refine representations during denoising, improving compactness and separation under distribution shift Wen et al. (2016); Gretton et al. (2012); Long et al. (2015). Rather than adding segmentation losses, we introduce topology through differentiable vesselness and anatomically structured pooling to bias features toward vascular geometry, and we apply long-tail calibration to improve minority-class recall in DR Frangi et al. (1998); Menon et al. (2020); Shit et al. (2021); Deng et al. (2019). In our design, diffusion refines latent features alongside discriminative heads, and we couple denoised features with metric and alignment constraints instead of treating the model as a standalone generative classifier Chen et al. (2024).

3 Methodology

Following ETDRS practice, where lesion burden is assessed relative to disc-centered quadrants and vascular topology, the architecture encodes vasculature and region-wise extent and then refines task-specific embeddings with conditional diffusion. We propose a vessel-topology aware latent diffusion framework for five-class diabetic retinopathy (DR) grading from a single RGB fundus image.

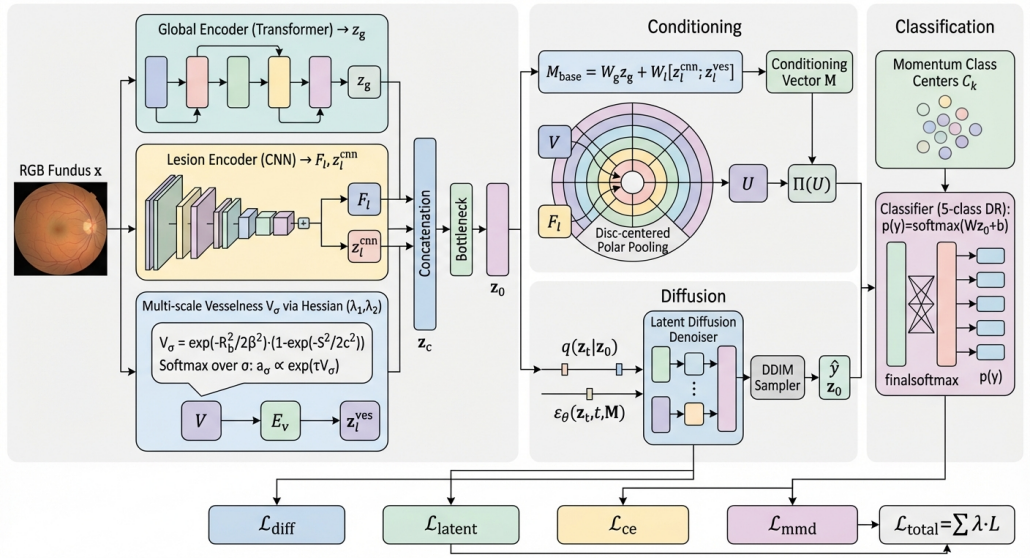

Figure 1: Architecture of the vessel-topology aware latent diffusion framework for five-class diabetic retinopathy grading from a single fundus image: a global encoder, a lesion encoder, and a multi-scale vesselness branch fuse into a compact latent; a disc-centered polar pooling module constructs an anatomical conditioning vector that captures vessel structure and quadrant extent to guide the denoiser; a center-regularized classifier produces the final grade.

The framework operates in a compact latent space and combines a topology-aware representation that fuses global context, lesion appearance, and a differentiable multi-scale vesselness channel, a latent diffusion denoiser conditioned on an anatomically based vector constructed via learnable disc-centered polar pooling, and a classification head regularized by momentum-updated class centers. The design embeds retinal vasculature and quadrant-wise extent into the conditioning pathway and refines task-specific latents through conditional diffusion to improve class separability and robustness.

3.1 Vessel-Topology Representation

We first construct a compact latent that consolidates global appearance, lesion cues, and vessel morphology. Let $x \in \mathbb{R}^{B\times 3\times H\times W}$ denote a minibatch of fundus images. A global encoder $E_g(\cdot)$ (e.g., transformer-based Dosovitskiy et al. (2021)) yields a holistic embedding $z_g = E_g(x) \in \mathbb{R}^{B\times d_g}$. A convolutional encoder $E_l(\cdot)$ produces lesion-level features; we use its penultimate feature map $F_l \in \mathbb{R}^{B\times C\times H'\times W'}$ and its pooled descriptor $z_l^{\text{cnn}} \in \mathbb{R}^{B\times d_l}$. To encode vascular morphology explicitly, we construct a differentiable multi-scale vesselness channel. Let $L_\sigma = G_\sigma * x$ be a Gaussian scale-space at scale σ, and $H_\sigma(p)$ the Hessian at pixel p with ordered eigenvalues $|\lambda_{1,\sigma}(p)| \le |\lambda_{2,\sigma}(p)|$. We define the per-scale vesselness

$$
\begin{aligned}
R_b(p,\sigma) &= \frac{|\lambda_{1,\sigma}(p)|}{|\lambda_{2,\sigma}(p)| + \varepsilon}, \quad S(p,\sigma) = \sqrt{\lambda_{1,\sigma}(p)^2 + \lambda_{2,\sigma}(p)^2}, \\
V_\sigma(p) &= \exp\Big(-\frac{R_b(p,\sigma)^2}{2\beta^2}\Big)\Big(1 - \exp\Big(-\frac{S(p,\sigma)^2}{2c^2}\Big)\Big),
\end{aligned}
\tag{1}
$$

where $\beta, c > 0$ are hyperparameters and $\varepsilon > 0$ ensures numerical stability Frangi et al. (1998). To aggregate across scales, we adopt softmax pooling with temperature $\tau > 0$,

$$
a_\sigma(p) = \frac{\exp\big(\tau V_\sigma(p)\big)}{\sum_{\sigma'} \exp\big(\tau V_{\sigma'}(p)\big)}, \quad V(p) = \sum_\sigma a_\sigma(p)\, V_\sigma(p), \tag{2}
$$

yielding a differentiable vesselness map $V \in \mathbb{R}^{B\times 1\times H\times W}$. A shallow encoder $E_v(\cdot)$ maps V to $z_l^{\text{ves}} = E_v(V) \in \mathbb{R}^{B\times d_v}$. We construct the clean latent via a bottleneck projection,

$$z_c = \big[\, z_g;\, z_l^{\text{cnn}};\, z_l^{\text{ves}} \,\big] \in \mathbb{R}^{B\times(d_g+d_l+d_v)}, \quad z_0 = f_{\text{bottleneck}}(z_c) \in \mathbb{R}^{B\times d}, \tag{3}$$

which serves as the target latent for diffusion and the input to the classifier. This fused latent preserves global context while adding vessel-aware cues, providing the basis for anatomy-aware conditioning.

3.2 Anatomical Conditioning Vector

To guide denoising toward anatomically plausible and quadrant-aware embeddings, we build a dense conditioning vector $M \in \mathbb{R}^{B\times d_c}$ that summarizes global, local, and disc-centered spatial evidence. We first form a base vector from global and local descriptors,

$$M_{\text{base}} = W_g z_g + W_l\big[\, z_l^{\text{cnn}};\, z_l^{\text{ves}} \,\big], \quad W_g \in \mathbb{R}^{d_c\times d_g},\ W_l \in \mathbb{R}^{d_c\times(d_l+d_v)}. \tag{4}$$

To encode region- and extent-aware anatomy aligned with ETDRS practice, we introduce learnable disc-centered polar pooling over spatial maps. The polar origin $o = (o_x, o_y)$ is predicted by a small head $g_o(\cdot)$ acting on z_g, and approximates the optic disc center to anchor quadrants. Let $\rho(p) = \|p - o\|_2$ and $\theta(p) = \operatorname{atan2}(p_y - o_y, p_x - o_x)$ be polar coordinates for pixel p. For R concentric rings with centers μ_r and widths σ_r, and Q angular sectors with centers ϕ_q and concentration κ, we define soft memberships

$$\begin{aligned} s_r(p) &= \exp\Big(-\frac{\big(\rho(p)-\mu_r\big)^2}{2\sigma_r^2}\Big), \quad s_q(p) = \frac{\exp(\kappa\cos(\theta(p)-\phi_q))}{2\pi I_0(\kappa)}, \\ w_{r,q}(p) &= \frac{s_r(p)\, s_q(p)}{\sum_{p'} s_r(p')\, s_q(p') + \varepsilon}, \end{aligned} \tag{5}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind and $\varepsilon > 0$. We pool vesselness and lesion features within each polar cell via soft averages,

$$u_{r,q}^{(V)} = \sum_p w_{r,q}(p)\, V(p), \quad u_{r,q}^{(F)}(c) = \sum_p w_{r,q}(p)\, F_l(c,p), \tag{6}$$

and form U by concatenating all $\{u_{r,q}^{(V)}\}$ and channel-wise statistics of $\{u_{r,q}^{(F)}(c)\}$. The final conditioning vector is

$$M = M_{\text{base}} + \Pi(U), \quad \Pi : \mathbb{R}^{B\times d_U} \to \mathbb{R}^{B\times d_c}. \tag{7}$$

These soft rings and sectors capture both radial extent and quadrant-wise distribution of lesions and vessels, mirroring ETDRS disc-centered quadrants while remaining fully differentiable. To promote consistency between global and local descriptors, we regularize with a kernel two-sample objective,

$$\mathcal{L}_{\text{mmd}} = \frac{1}{n(n-1)}\sum_{i\neq j} k(z_{g,i}, z_{g,j}) + \frac{1}{n(n-1)}\sum_{i\neq j} k(z_{l,i}, z_{l,j}) - \frac{2}{n^2}\sum_{i,j} k(z_{g,i}, z_{l,j}), \tag{8}$$

where $z_l = \big[z_l^{\text{cnn}}; z_l^{\text{ves}}\big]$, $k(a,b) = \exp\big(-\|a-b\|_2^2/(2\sigma^2)\big)$, and n is the batch size Gretton et al. (2012). This alignment stabilizes the conditioning signal before it guides diffusion.

3.3 Latent Diffusion Denoiser

We adopt a latent (not pixel-space) diffusion process to refine z_0, as operating in a compact latent focuses capacity on clinically salient structure and is computationally more efficient and robust to acquisition variability than high-dimensional pixel diffusion. We adopt a T-step variance-preserving forward process in latent space Ho et al. (2020); Rombach et al. (2022). Let $\{\beta_t\}_{t=1}^T$ be a noise schedule with $\alpha_t = 1 - \beta_t$ and $\bar\alpha_t = \prod_{s=1}^t \alpha_s$. The forward diffusion is

$$\begin{aligned} q(z_t \mid z_0) &= \mathcal{N}\big(\sqrt{\bar\alpha_t}\, z_0,\, (1-\bar\alpha_t)\mathbf{I}\big), \quad t = 1,\dots,T, \\ z_t &= \sqrt{\bar\alpha_t}\, z_0 + \sqrt{1-\bar\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}). \end{aligned} \tag{9}$$

The reverse model conditions on M via a parameterized noise predictor $\epsilon_\theta(z_t, t, M)$,

$$p_\theta(z_{t-1} \mid z_t, t, M) = \mathcal{N}(\mu_\theta(z_t, t, M),\, \sigma_t^2 \mathbf{I}),$$
$$\mu_\theta(z_t, t, M) = \frac{1}{\sqrt{\alpha_t}}\Big(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(z_t, t, M)\Big), \tag{10}$$

with fixed variance schedule σ_t^2. The denoising objective minimizes the conditional noise prediction error,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, z_0, \epsilon}\Big[\big\|\epsilon - \epsilon_\theta(z_t, t, M)\big\|_2^2\Big], \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{11}$$

Conditioning on M guides the reverse trajectory toward anatomically consistent, class-separable latents that are more stable near referral thresholds.

3.4 Center-Regularized Classification

To structure the latent space for five-class DR grading, we maintain momentum-updated class centers $\{C_k\}_{k=1}^5$ with $C_k \in \mathbb{R}^d$. Given a batch with index sets $S_k = \{i : y_i = k\}$ and cardinalities $n_k = |S_k|$, the centers are updated via

$$\bar{C}_k = \frac{1}{\max(1, n_k)} \sum_{i \in S_k} z_{0,i}, \quad C_k \leftarrow \beta\, C_k + (1-\beta)\, \bar{C}_k, \quad \beta \in [0, 1). \tag{12}$$

We penalize intra-class dispersion and encourage inter-class margins,

$$\mathcal{L}_{\text{latent}} = \sum_i \big\|z_{0,i} - C_{y_i}\big\|_2^2 + \lambda_{\text{inter}} \sum_{k \neq j} \max\big(0,\, m - \|C_k - C_j\|_2\big), \tag{13}$$

and compute class probabilities with a linear classifier,

$$\ell_i = W z_{0,i} + b, \quad p(y_i = k \mid z_{0,i}) = \frac{\exp(\ell_{i,k})}{\sum_{j=1}^5 \exp(\ell_{i,j})},$$
$$\mathcal{L}_{\text{ce}} = -\sum_i \log p(y_i \mid z_{0,i}), \tag{14}$$

where $W \in \mathbb{R}^{5 \times d}$, $b \in \mathbb{R}^5$, and $m > 0$ is the margin. In practice, the momentum coefficient β controls the speed of center updates and the margin m governs separation; careful tuning and normalization of latents help stabilize center dynamics, which is particularly relevant for minority grades.

3.5 Deterministic Latent Sampling

At inference, we employ a DDIM-style deterministic sampler Song et al. (2020) conditioned on M. This avoids sampling variance inherent to stochastic reverse processes and improves run-to-run stability of predictions. With a monotone subset of steps $S = T > T_{S-1} > \cdots > T_1 \geq 1$ and initialization $z_T \sim \mathcal{N}(0, \mathbf{I})$, we iterate

$$\hat{\epsilon}_\theta = \epsilon_\theta(z_s, s, M), \quad \hat{z}_0 = \frac{z_s - \sqrt{1-\bar{\alpha}_s}\,\hat{\epsilon}_\theta}{\sqrt{\bar{\alpha}_s}},$$
$$z_{s-1} = \sqrt{\bar{\alpha}_{s-1}}\,\hat{z}_0 + \sqrt{1-\bar{\alpha}_{s-1}}\,\hat{\epsilon}_\theta, \quad s \in \{T, \ldots, 1\}, \tag{15}$$

to obtain $\hat{z}_0$, which is then fed to the classifier to produce the five-class posterior.

3.6 Optimization Objectives

The model is trained end-to-end with a weighted sum of losses,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}}\,\mathcal{L}_{\text{diff}} + \lambda_{\text{latent}}\,\mathcal{L}_{\text{latent}} + \lambda_{\text{ce}}\,\mathcal{L}_{\text{ce}} + \lambda_{\text{mmd}}\,\mathcal{L}_{\text{mmd}}, \tag{16}$$

where the weights $\{\lambda_{\text{diff}}, \lambda_{\text{latent}}, \lambda_{\text{ce}}, \lambda_{\text{mmd}}\}$ balance conditional denoising, latent structuring, classification, and global-local alignment. To assess whether anatomy-aware conditioning and latent diffusion translate into improved ordinal grading and robustness, the next section evaluates this framework under a unified protocol on APTOS 2019 with strong CNN/ViT baselines and ablations that selectively disable each component.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS AND EVALUATION PROTOCOLS

Because DR grading is ordinal and clinical use prioritizes calibrated agreement, we evaluate on the APTOS 2019 Blindness Detection dataset, which provides single-modality RGB fundus photographs annotated with five ordinal grades of diabetic retinopathy (DR): 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe non-proliferative DR), and 4 (Proliferative DR). Images are indexed by unique identifiers and accompanied by diagnosis labels in a standardized CSV. We fix the random seed (42) and create a stratified split of the original training set into 80% training and 20% validation subsets, preserving the label distribution in both splits. We do not use external modalities or labels.

Preprocessing includes resizing images to 224×224, standard normalization, and moderate data augmentation (horizontal flip, small rotations, and color jitter). Because the data are imbalanced, we use class-aware sampling during training via a WeightedRandomSampler with weights from empirical class frequencies in the training split, and we set class weights in the cross-entropy objective proportional to inverse class frequencies.

At each epoch we report quadratic weighted kappa (QWK) as the primary metric, which measures ordinal agreement across the five grades. We also report accuracy (ACC), macro-F1, and macro-averaged AUC computed in a multiclass one-vs-rest manner. In addition, we compute per-class precision, recall, and F1 and a confusion matrix to analyze class-specific behavior, with particular focus on grades 3 and 4, where minority effects are strongest. The same protocol is applied to all baselines and the proposed method.

4.1.2 BASELINES SETTING

To evaluate the effectiveness of the proposed method, we compare it against several widely used convolutional neural network architectures commonly adopted for medical image classification. Specifically, we consider Inception-v3, MobileNet-V3, ResNet-50, and VGG-16 as baseline models. These architectures cover a diverse spectrum of design philosophies, including lightweight networks, deep residual architectures, and classical convolutional backbones. All baselines are trained under the same experimental protocol to ensure a fair comparison. The models are trained for 50 epochs, and the checkpoint achieving the best validation accuracy is selected for evaluation. Our method is implemented under the same training setting and evaluated on the same test split as the baseline models.

4.1.3 IMPLEMENTATION DETAILS

Training runs for 50 epochs using Cosine Annealing with $T_{\max} = 50$. We optimize the encoders, fusion layers, diffusion eps-prediction model, and classification head with AdamW, and we optimize CenterLoss parameters with SGD. Automatic mixed precision and gradient clipping (max_norm $= 1.0$) are used in all experiments. The best model checkpoint is selected based on validation QWK. For each epoch, we record QWK (primary), ACC, Macro-F1, AUC, per-class metrics, and the confusion matrix. A DDIM sampler is used at inference to provide deterministic latent sampling conditioned on M.

4.2 LEARNING DYNAMICS AND CONVERGENCE

We examine learning behavior over the 50-epoch schedule to study optimization stability and the progression of ordinal calibration. Figure 2 shows the training loss and validation accuracy. During the early training stage, all evaluation metrics improve rapidly. For example, the model progresses from 0.7366 QWK and 0.5333 accuracy at epoch 5 to 0.7662 QWK and 0.5933 accuracy at epoch 10, accompanied by consistent gains in F1 score and AUC, indicating that both ordinal consistency and classification quality improve as the representation stabilizes. Performance continues to increase around epoch 15, reaching 0.7769 QWK and 0.6067 accuracy, with corresponding improvements in precision, recall, and F1 score, suggesting better class discrimination across DR severity levels. The model achieves its best performance near epoch 19–20, where QWK peaks at 0.8328, accuracy

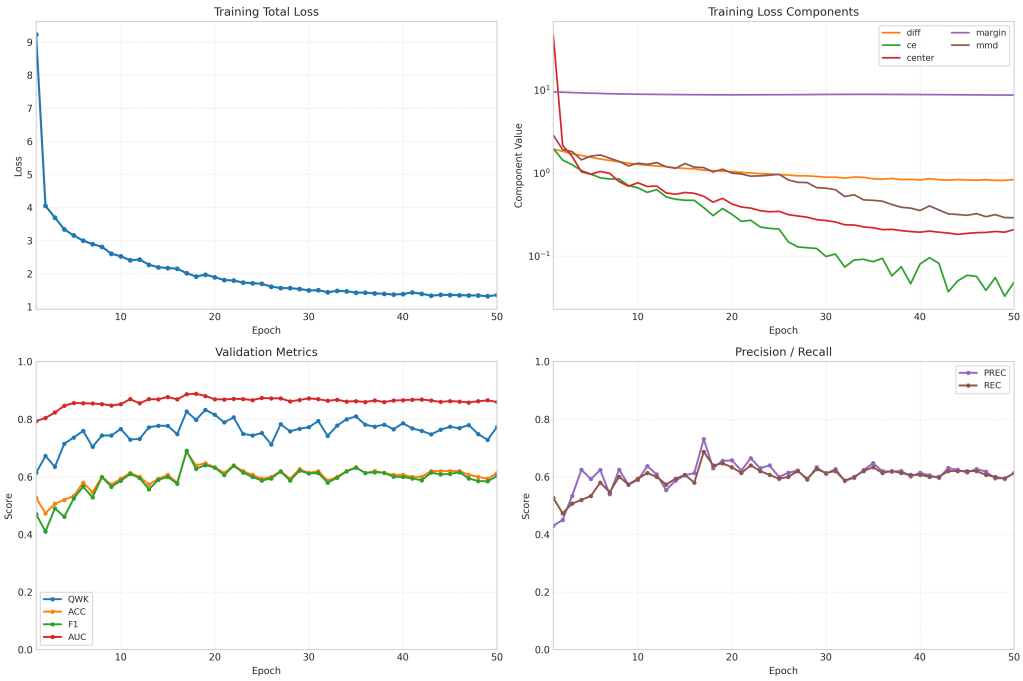

Figure 2: Learning dynamics over 50 epochs.

Table 1: Comparison with standard CNN baselines on the DR grading task. The best results are shown in **bold**.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
Inception-v3	0.7889	0.6600	0.6681	0.6600	0.6559	0.8586
MobileNet-v3	0.7766	0.6600	0.6652	0.6600	0.6587	**0.8911**
ResNet-50	0.8090	0.6467	0.6592	0.6467	0.6477	0.8656
VGG-16	0.8248	0.6600	0.6893	0.6600	0.6621	0.8589
Ours	**0.8267**	**0.6867**	**0.7306**	**0.6867**	**0.6905**	0.8867

reaches 0.6467, and both F1 score and AUC attain their highest values, demonstrating strong ordinal agreement and balanced classification performance. After this stage, the metrics fluctuate within a relatively narrow range while maintaining high values across QWK, accuracy, F1, and AUC, indicating stable convergence without noticeable overfitting. Overall, the results show that the proposed framework converges reliably within the first 20 epochs while sustaining consistent performance across multiple evaluation metrics.

4.3 Main Performance Comparison

Table 1 summarizes the quantitative comparison between the proposed method and the baseline architectures. Overall, the proposed method achieves the best performance on the primary metric (QWK), obtaining 0.8267, which surpasses all baseline models. The closest competitor is VGG-16, which achieves 0.8248, while ResNet-50, Inception-v3, and MobileNet-V3 obtain 0.8090, 0.7889, and 0.7766, respectively. The improvement over ResNet-50 and Inception-v3 indicates that the proposed design better captures the ordinal structure of diabetic retinopathy grading. In terms of classification accuracy, our method reaches 68.7 %, outperforming all baseline models that remain at 66 % or below. This improvement demonstrates that the proposed architecture yields more reliable predictions across severity levels. The proposed approach also achieves the highest precision (0.7306) and F1-score (0.6905) among all compared methods, suggesting a better balance between sensitivity and specificity. Although MobileNet-V3 obtains the highest AUC (0.8911), its QWK and

accuracy remain significantly lower, indicating that strong ranking performance does not necessarily translate into better ordinal classification. Overall, these results demonstrate that the proposed method provides more consistent improvements across multiple evaluation metrics, while achieving the best agreement with ground-truth grading according to QWK, which is the most critical metric for this task.

4.4 Visualization Comparison

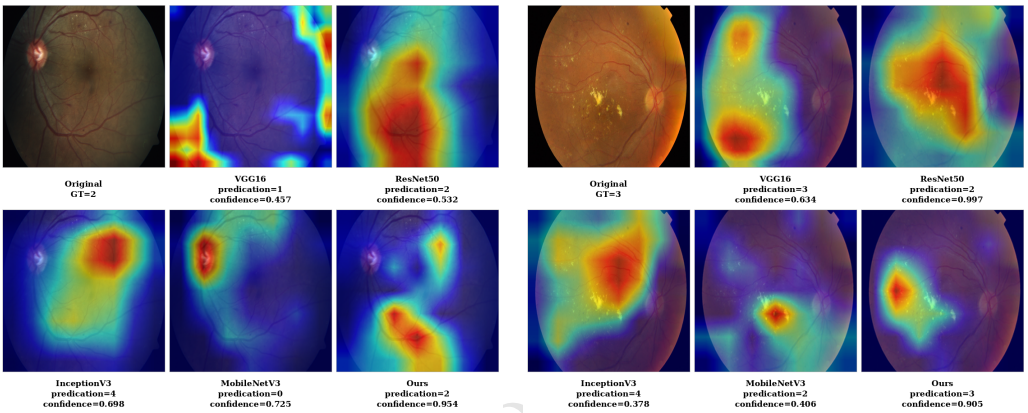

Figure 3: Visualization results from GradCAM between our method and the other baseline methods.

To better understand the decision behavior of different models, we visualize class activation maps (CAMs) for two representative retinal images, as shown in Fig. 3. In the first case (true label: moderate DR), several baseline models produce inconsistent predictions, including underestimation by VGG16 (grade 1), severe overestimation by InceptionV3 (grade 4), and a false negative prediction by MobileNetV3 (grade 0). In contrast, our method correctly predicts grade 2 with substantially higher confidence (0.95). The corresponding activation map concentrates on clinically relevant lesion regions, while competing models exhibit scattered or misplaced attention. A similar trend is observed in the second example (true label: severe DR). Although VGG16 predicts the correct grade, its activation remains relatively diffuse, whereas ResNet50 and MobileNetV3 underestimate the severity (grade 2), and InceptionV3 overestimates it (grade 4). Our model again yields the correct prediction with high confidence (0.91) and produces a more localized activation pattern aligned with pathological structures. Overall, the visualizations suggest that our model attends more consistently to lesion-related regions, leading to more reliable grading decisions compared with conventional CNN baselines.

4.5 Ablation Studies

We perform module-wise ablations to analyze the contribution of each component of the proposed vessel-topology aware latent diffusion framework. Removing the latent diffusion refinement reduces QWK from 0.8267 to 0.8103 and accuracy from 0.6867 to 0.6733, indicating that diffusion-based latent refinement improves ordinal consistency. Disabling the momentum-updated class centers leads to a larger drop to 0.7774 QWK and 0.6133 ACC, suggesting that class-center regularization is important for maintaining inter-class separability. Similarly, removing MMD alignment further decreases QWK to 0.7622, demonstrating the role of distribution alignment in stabilizing representation learning. We also analyze the topology-aware representation and conditioning mechanisms. Using only the global branch achieves 0.7806 QWK, while the local lesion/vessel branch alone reaches 0.8069, indicating that lesion-level cues provide stronger signals for DR severity estimation. Removing anatomical conditioning significantly degrades performance to 0.7569 QWK, confirming the importance of disc- and quadrant-aware conditioning. The full model combining all components consistently achieves the best performance (0.8267 QWK, 0.6867 ACC).

Table 2: Module-wise ablation of the proposed framework. TopoRep denotes the topology-aware representation, Diffusion denotes latent diffusion refinement, Center denotes momentum-updated class centers, and Cond. denotes disc- and quadrant-aware conditioning.

Variant	TopoRep	Diffusion	Center	MMD	Cond.	QWK↑	ACC↑	Prec↑	Rec↑	F1↑	AUC↑
Ours	✓	✓	✓	✓	✓	**0.8267**	**0.6867**	**0.7306**	**0.6867**	**0.6905**	**0.8867**
no_diffusion	✓	×	✓	✓	✓	0.8103	0.6733	0.6750	0.6733	0.6701	0.8676
no_center	✓	✓	×	✓	✓	0.7774	0.6133	0.6122	0.6133	0.6112	0.8495
no_mmd	✓	✓	✓	×	✓	0.7622	0.6400	0.6448	0.6400	0.6356	0.8487
global_only	×	✓	✓	✓	✓	0.7806	0.6667	0.6680	0.6667	0.6642	0.8617
local_only	×	✓	✓	✓	✓	0.8069	0.6600	0.6697	0.6600	0.6625	0.8508
M_zero	✓	✓	✓	✓	×	0.7569	0.6533	0.6655	0.6533	0.6551	0.8611
M_global_only	✓	✓	✓	✓	✓	0.7300	0.6333	0.6308	0.6333	0.6278	0.8629
M_local_only	✓	✓	✓	✓	✓	0.8090	0.6733	0.6929	0.6733	0.6767	0.8848

4.6 Analysis and Interpretation

The results show three consistent patterns. Accuracy rises early, indicating rapid learning of coarse separability, while ordinal agreement improves later and peaks. Minority-class performance remains weak despite class-aware sampling and weighted losses, as reflected by low Macro-F1; stronger lesion-aware conditioning and more stable center dynamics are expected to better recover minority signals, particularly for grades 3 and 4 where referral decisions are most sensitive. Ablation results identify the current center formulation as a bottleneck: removing it markedly improves both ordinal and class-aware metrics and enables the diffusion denoiser to contribute more effectively. These findings point to the importance of attention-guided conditioning, EMA-normalized center dynamics, and refined latent inference in capturing the ordinal structure of DR grading and turning representational quality into calibrated predictions, under a protocol that enforces parity across baselines for a fair assessment.

5 Conclusion

We address five-stage diabetic retinopathy grading from single fundus images under ETDRS, which relies on lesion burden within disc-centered anatomy. We propose a topology-aware latent diffusion model fusing global context, lesion cues, and multi-scale vesselness. Disc- and quadrant-aligned polar pooling yields anatomy-aware conditioning that guides the denoiser toward ordinal consistency. Under a common training protocol, anatomy-aware conditioning and conditional diffusion improve ordinal agreement and robustness to acquisition shifts; ablations show that removing unstable center loss unlocks diffusion gains and improves sensitivity to minority grades, while dual-branch fusion and cross-branch alignment remain necessary. Despite interpretable cues, macro-level performance remains limited and minority-grade detection near referral thresholds can degrade in a single-dataset setting. Future work will replace center loss with EMA-normalized prototypes, enrich conditioning with supervised disc localization and graph-based vessel encodings, and evaluate cross-dataset generalization via calibration and external validation.

6 Ethics Statement

This research study was conducted retrospectively using human subject data made available in open access by the APTOS 2019 Blindness Detection competition Aravind Eye Hospital and PG Institute of Ophthalmology (2019) on Kaggle, sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required as confirmed by the license attached with the open access data.

References

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection, 2019. Kaggle Competition.

Tianyu Chen, Zhaoyang Wang, Yixiao Li, et al. A robust diffusion classifier. *arXiv preprint arXiv:2402.17139*, 2024.

James R Clough, Ilkay Oksuz, Nick Byrne, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *arXiv preprint arXiv:2009.13107*, 2020.

José Cunha-Vaz. The blood-retinal barrier in retinal disease. *European Journal of Ophthalmology*, 20(suppl 6):S71–S74, 2010.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10. *Ophthalmology*, 98(5):786–806, 1991.

Alejandro Frangi, Wiro Niessen, Koen Vincken, and Max Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 130–137. Springer, 1998.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alex Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wei Hu, Yao Li, and Xin Zhang. Microglia in diabetic retinopathy: Pathophysiology and therapeutic opportunities. *Cells*, 13(2):345, 2024.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3238–3247, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 97–105. PMLR, 2015.

Cheng Lu, Yu Zhou, Jianfei Bao, Jianmin Chen, Jun Zhu, and Wenqiang Zhao. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Andreas Lugmayr, Martin Danelljan, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International Conference on Artificial Neural Networks (ICANN)*, pp. 65–76. Springer, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Suprosanna Shit, Alejandro Gomez, Anindo Sekuboyina, et al. cldice—a novel topology-preserving loss function for tubular structure segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 107–117. Springer, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Daniel SW Ting, Carol Y Cheung, Gilbert Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision (ECCV)*, pp. 499–515. Springer, 2016.

Xinyu Yang, Zhiqiang Li, Lei Zhang, et al. Diffmic: Dual-guidance diffusion model for medical image classification. *arXiv preprint arXiv:2306.00986*, 2023.

CROSS-SCALE CHANNEL ATTENTION WITH ORDINAL-CATEGORICAL DUAL HEADS AND UNCERTAINTY-GATED SELF-TRAINING FOR DIABETIC RETINOPATHY GRADING

Anonymous Author(s)

ABSTRACT

Diabetic retinopathy (DR) grading from fundus photographs demands sensitivity to small, sparse lesions, respect for the five-level ordinal scale, and robustness to long-tailed and noisy labels, while interpretability is desirable despite scarce lesion annotations. We propose a fundus-only grader that couples a Cross-Scale Channel Attention (CSCA) module with dual heads—a categorical softmax head and a CORAL ordinal head—trained end-to-end under a hybrid objective. The categorical head uses class-balanced focal loss to handle imbalance, and the ordinal head uses a focal-augmented ordinal loss to enforce monotonicity and reduce severe ordinal errors. By integrating cross-scale attention, ordinal modeling, and noise-aware learning without lesion-level supervision, the approach advances DR grading under realistic constraints. To mitigate label noise, we refine pseudo-labels via mutual-information gating with Monte Carlo dropout and apply a lightweight prediction-consistency regularizer across two augmentations for accepted pseudo-labels. Evaluated on APTOS 2019, the proposed method achieves a quadratic weighted kappa of 0.85, an accuracy of 73.33%, a macro-F1 score of 73.23%, and a macro one-vs-rest AUC of 91.54%.

1 INTRODUCTION

Automated analysis of retinal fundus photographs is a long-standing goal in computer vision for healthcare, with diabetic retinopathy (DR) grading serving as a prominent benchmark and clinically impactful task. DR severity is assigned on a five-level ordered scale, and early detection hinges on recognizing tiny, sparse lesions while avoiding large out-of-order errors that contradict disease progression Group (1987). In routine screening, datasets are often imbalanced, severe cases are rare, and labels near grade boundaries can be noisy. We address the problem of learning a robust, interpretable DR grader that is simultaneously sensitive to microlesions, consistent with the ordinal scale, and resilient to long-tailed and imperfect labels.

Deep learning has substantially improved DR detection and grading Gulshan et al. (2016); Ting et al. (2017), and advances in attention and multi-scale modeling have boosted lesion saliency Hu et al. (2018); Woo et al. (2018); Wang et al. (2020b); Li et al. (2019). Ordinal approaches explicitly encode label ordering Cao et al. (2020), and uncertainty-aware strategies seek robustness to label noise and distributional shift Gal & Ghahramani (2016); Xie et al. (2020). However, gaps remain. Channel attention is typically applied within a single scale, while multi-branch pyramids improve context at the cost of complexity and reduced interpretability at readout. Treating grades as nominal improves separability but can increase non-adjacent mistakes; enforcing ordering alone can dilute minority-class discrimination. Confidence-only pseudo-labeling risks confirmation bias and may disproportionately exclude rare classes. Finally, many pipelines trade off clinical interpretability when departing from a global-average-pooling (GAP) readout that supports CAM/Grad-CAM explanations.

These limitations matter in practice. Screening systems must highlight subtle, multi-scale cues such as microaneurysms and hemorrhages, yet remain faithful to the disease continuum to avoid clinically implausible jumps in predicted severity. They must also handle long-tailed distributions and imperfect labels without relying on external data or heavy architectural overhead, enabling reproducible

evaluation and deployment on common hardware. Maintaining compatibility with saliency-based explanations is important for clinician trust and for auditing model behavior in safety-critical settings.

We propose a single-stream framework that couples cross-scale feature selection, ordinal-aware supervision, and uncertainty-guided label refinement. First, we introduce Cross-Scale Channel Attention (CSCA), a descriptor-level mechanism that aggregates and re-weights channel descriptors across backbone stages. This design injects multi-scale context without spatial attention maps or multi-branch towers, and preserves a GAP-based readout for compatibility with CAM/Grad-CAM. Second, we adopt dual-head hybrid supervision: a categorical softmax head maintains class separation beneficial for minority classes and referral decisions, while an ordinal head (e.g., CORAL Cao et al. (2020)) enforces monotonic thresholds to reduce out-of-order errors. Third, we refine noisy labels via uncertainty-guided self-training that accepts pseudo-labels only when predictions are confident and exhibit low epistemic uncertainty, measured by mutual information from stochastic predictions; class-aware gating mitigates over-pruning of rare classes. We evaluate on APTOS 2019 with image-level labels only, using no external datasets or annotations, and we report internal validation under this protocol without claims of cross-dataset generalization.

Our main contributions are as follows.

- We introduce Cross-Scale Channel Attention (CSCA), a single-stream, descriptor-level cross-scale attention module that enhances microlesion saliency while retaining a GAP-based, CAM-compatible readout.
- We design a dual-head hybrid supervision scheme that combines categorical and ordinal predictions to jointly promote minority-class separability and adherence to the ordered severity scale.
- We develop an uncertainty-guided self-training procedure that refines noisy labels using confidence and mutual-information gating without external unlabeled data.
- On APTOS 2019, our approach improves DR grading under internal validation, yielding gains in quadratic weighted kappa, accuracy, macro-F1, and macro AUC, with claims limited to this dataset and evaluation regime.

2 Related Work

2.1 Backbones, Attention, Multi-Scale Fusion, and Interpretability for Fundus Classification

For fundus-based DR screening, models must capture subtle, small lesions while retaining global context and remaining compatible with GAP-driven interpretability. Early systems showed that convolutional neural networks (CNNs) with global average pooling (GAP) perform well across diverse cohorts and offer a practical path for clinical deployment Szegedy et al. (2015); Gulshan et al. (2016); Ting et al. (2017). Subsequent architectural advances improved capacity and trainability: residual learning helped address vanishing gradients and enabled much deeper models He et al. (2016), while dense connectivity promoted feature reuse and multi-scale propagation with good parameter efficiency Huang et al. (2017). Lightweight channel attention further improved the accuracy–efficiency trade-off; widely used designs include squeeze-and-excitation (SE), CBAM, efficient channel attention (ECA), selective kernel fusion, and frequency channel attention Hu et al. (2018); Woo et al. (2018); Wang et al. (2020b); Li et al. (2019). To better preserve fine details while adding broader context, multi-scale aggregation has been studied extensively. Representative designs include top–down pyramids (FPN) and multi-branch high-resolution networks (HRNet), both adopted in medical imaging for lesion-centric analysis Lin et al. (2017). A related line of work models global context via non-local operations and similar modules, such as Non-local Networks, GCNet, Gather-Excite Hu et al. (2018), and GloRe. Recently, transformer-based hierarchical backbones (e.g., Swin Transformer, PVT) have provided multi-scale token representations with global receptive fields and have been adopted widely in vision.

Weakly supervised localization via Class Activation Mapping (CAM) has been widely used to produce interpretable heatmaps from GAP-based classifiers Zhou et al. (2016), with variants that broaden coverage and improve spatial fidelity Selvaraju et al. (2017); Chattopadhay et al. (2018);

Wang et al. (2020a); Jiang et al. (2021). These techniques suit fundus imagery, where clinically meaningful cues include small, scattered red lesions and vessel-adjacent abnormalities, but saliency reliability depends on backbone choice, feature resolution, and method sensitivity, and thus requires sanity checks before drawing clinical conclusions Adebayo et al. (2018). Our approach adopts a single-stream Cross-Scale Channel Attention (CSCA) module, which preserves a single-stream pathway and a GAP-based readout, maintaining compatibility with CAM/Grad-CAM while adding cross-level context that helps preserve microlesions at the final readout resolution.

2.2 Ordinal-Aware Learning and Imbalance-Aware Optimization

DR grading follows an ordered severity scale formalized by ETDRS Group (1987), motivating objectives that respect ordinal structure. Beyond classical ordinal formulations Niu et al. (2016); Rothe et al. (2015), recent deep ordinal methods include CORAL, which shares a rank-consistent weight vector across cumulative thresholds Cao et al. (2020), CORN, which conditions binary sub-tasks on preceding outcomes, and DORN, which discretizes continuous targets and optimizes ordinal relations via classification with learned intervals. In medical imaging, ordinal objectives are used to better align predictions with disease progression and reduce non-adjacent errors relative to nominal softmax training. In parallel, long-tailed class distributions complicate DR training; focal modulation and effective-number reweighting emphasize minority and hard examples without excessive overfitting Lin et al. (2017); Cui et al. (2019). Multi-task and dual-head formulations that combine nominal classification with ordinal or regression supervision have also been explored in age estimation and medical grading, aiming to balance categorical separability with progression-aware ordering. Our approach couples a class-balanced focal loss (categorical head) with a focal-augmented CORAL loss (ordinal head) using fixed loss weights. This maintains categorical separation important for minority classes while CORAL-based monotonic thresholding discourages non-adjacent errors under long-tailed distributions.

2.3 Semi-/Self-Supervised Learning and Uncertainty-Guided Pseudo-Labeling

Semi- and self-supervised methods improve label efficiency through consistency regularization and pseudo-labeling. Mean Teacher enforces student–teacher agreement under perturbations; FixMatch combines weak–strong augmentation consistency with confidence-thresholded pseudo-labels; and Unsupervised Data Augmentation (UDA) uses weak–strong consistency with advanced augmentations Xie et al. (2020). Prototype-based assignments (e.g., PAWS) further improve sample efficiency. Reliable uncertainty is critical for selective training and robust deployment. Monte Carlo Dropout and deep ensembles provide strong baselines for epistemic uncertainty Gal & Ghahramani (2016); Lakshminarayanan et al. (2017), and post hoc calibration such as temperature scaling can improve probability alignment Guo et al. (2017). Mutual-information criteria computed over stochastic predictions provide an uncertainty signal complementary to confidence and entropy and have been used to gate pseudo-label acceptance under noisy labels. Our approach uses uncertainty-guided self-training label refinement within the labeled pool by incorporating mutual information alongside confidence, applying class-aware thresholds, and using uncertainty-dependent sharpening and weighting to mitigate confirmation bias under long-tailed DR distributions.

3 Methodology

DR grading from color fundus images requires (i) sensitivity to small, sparse microlesions across scales, (ii) respect for the ordinal structure of disease severity under long-tailed class distributions, and (iii) robustness to label noise typical of image-level annotations. We address these needs with a single-stream, end-to-end framework (as illustrated in Figure 1) that pairs a Cross-Scale Channel Attention (CSCA) module for multi-scale evidence selection with dual prediction heads, a categorical softmax head and a CORAL ordinal head, trained with a hybrid objective tuned to imbalance and ordering. To limit the impact of label noise, we refine training labels using uncertainty-gated pseudo-labels derived from mutual information (MI) estimated via Monte Carlo (MC) dropout, and optionally add a prediction-consistency regularizer. The ordinal head is used only during training; all reported metrics (QWK, accuracy, macro-F1, macro one-vs-rest AUC) and all inference rely on the categorical head. For interpretability, CAM/Grad-CAM are computed over the fused CSCA feature.

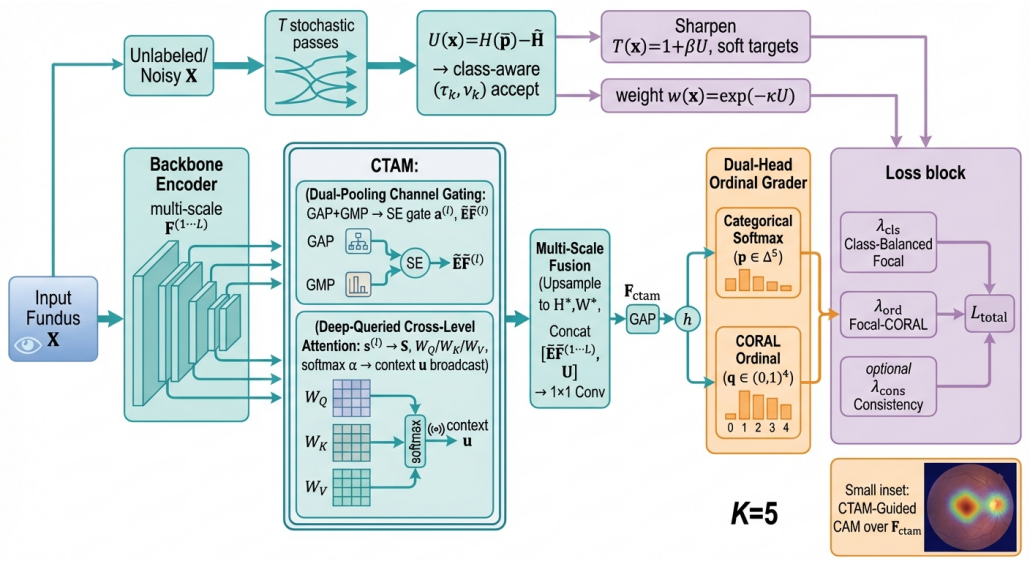

Figure 1: Architecture of the Proposed Framework. Illustrating the integration of Cross-Scale Channel Attention (CSCA), Dual-Head Grader (Categorical & CORAL), and Uncertainty-Guided Pseudo-Labeling.

Notation and symbols used once and then consistently are as follows. We set $K = 5$ for the number of grades and $L = 4$ for the backbone scales (DenseNet-121 blocks). The input batch of color fundus images is $X \in \mathbb{R}^{B\times 3\times H\times W}$, where B is the batch size. The feature map at scale l is $F^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, and (H_L, W_L) denotes the deepest spatial resolution. The shared descriptor dimension is d, the injected context width is C_u, and the fused channel dimension is C_f. GAP and GMP indicate global average and max pooling over spatial dimensions, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\mathrm{Broadcast}(v, H, W)$ tiles $v \in \mathbb{R}^{B\times C}$ to $\mathbb{R}^{B\times C\times H\times W}$.

3.1 Framework Overview

We use CSCA to capture cross-scale evidence for subtle lesions in a pathway compatible with global pooling, a dual-head grader to model class ordering alongside imbalance-aware categorical learning, and MI-gated pseudo-labeling to reduce the effect of noisy labels. Given X, a DenseNet-121 backbone produces a hierarchy of feature maps $\{F^{(l)}\}_{l=1}^{L}$, $F^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$, with $L = 4$ and $C_l \in \{256, 512, 1024, 1024\}$ at the outputs of the four dense blocks. The CSCA module applies per-scale dual-pooling channel gating and deep-queried cross-level descriptor attention, then upsamples and fuses all scales at the deepest resolution via a single 1×1 convolution to produce $F_{\mathrm{csca}} \in \mathbb{R}^{B\times C_f\times H_L\times W_L}$. Global average pooling yields $h \in \mathbb{R}^{B\times C_f}$ for two heads: (i) a K-way categorical softmax head p and (ii) a CORAL ordinal head q. Training uses a hybrid objective with class-balanced focal loss on the categorical head and focal-augmented ordinal loss on the CORAL head. Uncertainty-gated pseudo-labels (from the categorical head) and an optional consistency penalty are used in a second training stage. Inference and all metrics use the categorical head only.

3.2 Cross-Scale Channel Attention (CSCA)

3.2.1 Dual-Pooling Channel Gating (Per Scale)

To elevate lesion-relevant channels at each resolution, we apply a squeeze and excitation style gate with dual pooled descriptors. For each l,

$$
\begin{aligned}
\mathbf{g}_{\text{avg}}^{(l)} &= \mathrm{GAP}\Big(F^{(l)}\Big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \quad \mathbf{g}_{\max}^{(l)} = \mathrm{GMP}\Big(F^{(l)}\Big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \\
\mathbf{g}^{(l)} &= \Big[\mathbf{g}_{\text{avg}}^{(l)}; \mathbf{g}_{\max}^{(l)}\Big] \in \mathbb{R}^{B\times (2C_l)\times 1\times 1}, \\
\mathbf{u}^{(l)} &= \phi\big(W_1^{(l)}\mathbf{g}^{(l)}\big) \in \mathbb{R}^{B\times (C_l/r)\times 1\times 1}, \quad \mathbf{a}^{(l)} = \sigma\big(W_2^{(l)}\mathbf{u}^{(l)}\big) \in \mathbb{R}^{B\times C_l\times 1\times 1}, \\
\tilde{F}^{(l)} &= F^{(l)} \odot \mathbf{a}^{(l)} \in \mathbb{R}^{B\times C_l\times H_l\times W_l},
\end{aligned}
\tag{1}
$$

with $r = 4$, $W_1^{(l)} \in \mathbb{R}^{(C_l/r)\times(2C_l)\times 1\times 1}$, $W_2^{(l)} \in \mathbb{R}^{C_l\times(C_l/r)\times 1\times 1}$, $\phi = \mathrm{ReLU}$, $\sigma = \mathrm{sigmoid}$, and $\odot$ is broadcasted channel-wise multiplication.

3.2.2 Deep-Queried Cross-Level Descriptor Attention

To consolidate cross-scale evidence while preserving a single-stream readout, we summarize each gated scale by GAP and attend across levels using the deepest scale as the query. Define $\mathbf{s}^{(l)} = \mathrm{GAP}\Big(\tilde{F}^{(l)}\Big) \in \mathbb{R}^{B\times C_l}$. Project to a shared dimension $d = 256$:

$$
\begin{aligned}
\mathbf{q} &= \mathbf{s}^{(L)} W_Q^{(L)} \in \mathbb{R}^{B\times d}, \quad W_Q^{(L)} \in \mathbb{R}^{C_L\times d}, \\
\mathbf{k}^{(l)} &= \mathbf{s}^{(l)} W_K^{(l)} \in \mathbb{R}^{B\times d}, \quad W_K^{(l)} \in \mathbb{R}^{C_l\times d}, \\
\mathbf{v}^{(l)} &= \mathbf{s}^{(l)} W_V^{(l)} \in \mathbb{R}^{B\times d}, \quad W_V^{(l)} \in \mathbb{R}^{C_l\times d}.
\end{aligned}
\tag{2}
$$

Let $K = [\mathbf{k}^{(1)}; \ldots; \mathbf{k}^{(L)}] \in \mathbb{R}^{B\times L\times d}$, $V = [\mathbf{v}^{(1)}; \ldots; \mathbf{v}^{(L)}] \in \mathbb{R}^{B\times L\times d}$. Attention over levels:

$$
\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{\mathbf{q}\,K^\top}{\sqrt{d}}\right) \in \mathbb{R}^{B\times L}, \quad \mathbf{c} = \sum_{l=1}^{L} \boldsymbol{\alpha}_{(:,l)} \odot \mathbf{v}^{(l)} \in \mathbb{R}^{B\times d}. \tag{3}
$$

Project and broadcast to the deepest grid:

$$
\mathbf{u} = \mathbf{c}W_C \in \mathbb{R}^{B\times C_u}, \quad W_C \in \mathbb{R}^{d\times C_u}, \quad U = \mathrm{Broadcast}(\mathbf{u}, H_L, W_L) \in \mathbb{R}^{B\times C_u\times H_L\times W_L}, \tag{4}
$$

with $C_u = 64$.

3.2.3 Multi-Scale Upsampling, Fusion, and Readout

We then align spatial resolutions and fuse cross-scale information in a single, CAM-compatible stage. We upsample every gated map to the deepest resolution and fuse once with a single 1×1 convolution:

$$
\begin{aligned}
\hat{F}^{(l)} &= \mathrm{Upsample}\Big(\tilde{F}^{(l)} \to (H_L, W_L)\Big) \in \mathbb{R}^{B\times C_l\times H_L\times W_L}, \quad l = 1, \ldots, L, \\
F_{\text{csca}} &= \mathrm{Conv}_{1\times 1}\Big([\hat{F}^{(1)}; \hat{F}^{(2)}; \hat{F}^{(3)}; \hat{F}^{(4)}; U] \to C_f\Big) \in \mathbb{R}^{B\times C_f\times H_L\times W_L}, \\
h &= \mathrm{GAP}(F_{\text{csca}}) \in \mathbb{R}^{B\times C_f},
\end{aligned}
\tag{5}
$$

where $C_f = C_L = 1024$ to preserve classifier capacity and CAM compatibility. CAM/Grad-CAM are computed over F_{csca} using the categorical head weights.

3.3 Dual-Head Grader: Categorical and CORAL Heads

3.3.1 Categorical Softmax Head

The categorical head maps h to logits z and probabilities p:

$$
z = W_c h + \mathbf{b}_c \in \mathbb{R}^{B\times K}, \quad p_{i,k} = \frac{\exp(z_{i,k})}{\sum_{j=1}^{K}\exp(z_{i,j})}, \quad i = 1, \ldots, B,\ k = 1, \ldots, K. \tag{6}
$$

All metrics and predictions use p only.

3.3.2 CORAL ORDINAL HEAD WITH MONOTONE BIAS ENFORCEMENT

We adopt CORAL with a shared weight vector and ordered biases. Let $w_{\mathrm{ord}} \in \mathbb{R}^{C_f}$ and $\mathbf{b}_{\mathrm{ord}} \in \mathbb{R}^{K-1}$ with $b_{\mathrm{ord},1} \le \cdots \le b_{\mathrm{ord},K-1}$. Threshold logits and cumulative probabilities are

$$g_{i,k} = w_{\mathrm{ord}}^{\top} h_i - b_{\mathrm{ord},k}, \quad q_{i,k} = \sigma(g_{i,k}), \quad k = 1, \ldots, K-1. \tag{7}$$

To guarantee b_{ord} is non-decreasing, we parameterize it as a cumulative sum of non-negative increments:

$$\delta_k = \mathrm{softplus}(\theta_k) \ge 0, \quad b_{\mathrm{ord},k} = \sum_{j=1}^{k} \delta_j, \quad \theta_k \in \mathbb{R}. \tag{8}$$

Targets are cumulative indicators $t_{i,k} = \mathbb{I}[y_i \ge k]$. We apply focal modulation to the ordinal subtasks:

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{|D_{\mathrm{sup}}|} \sum_{i \in D_{\mathrm{sup}}} \sum_{k=1}^{K-1} \Big(t_{i,k}\,(1-q_{i,k})^{\gamma_{\mathrm{ord}}}\,(-\log q_{i,k}) + (1-t_{i,k})\,q_{i,k}^{\gamma_{\mathrm{ord}}}\,(-\log(1-q_{i,k})) \Big), \tag{9}$$

with $\gamma_{\mathrm{ord}} = 2$. By default, $\mathcal{L}_{\mathrm{ord}}$ is computed on D_{sup} only (we do not apply ordinal loss to pseudo-labeled samples).

3.4 UNCERTAINTY-GUIDED PSEUDO-LABEL REFINEMENT

To reduce the influence of noisy image-level labels, we refine training supervision using MI-gated pseudo-labels from the categorical head. MC dropout is used exclusively to estimate MI; it is disabled during both training and standard inference, ensuring that the deployed model remains deterministic.

3.4.1 MC DROPOUT PLACEMENT AND MI ESTIMATION

DenseNet-121 does not include dropout by default. For MI estimation only, we insert dropout layers immediately upstream of CSCA at each tapped scale (i.e., on the tensors $\{F^{(l)}\}$ prior to channel gating) and one at the fused readout:

$$F^{(l)} \xrightarrow{\mathrm{Dropout}(p_l)} F^{(l)}, \quad l = 1, \ldots, 4; \qquad h \xrightarrow{\mathrm{Dropout}(p_h)} h, \tag{10}$$

with $(p_1, p_2, p_3, p_4, p_h) = (0.10, 0.10, 0.20, 0.20, 0.20)$. These dropout layers are active only during MI estimation; they are disabled during both training and standard inference. During MI estimation, batch normalization is set to eval mode to fix running statistics. For a training image x, we run $T = 10$ stochastic forward passes to obtain $\{p^{(t)}(x)\}_{t=1}^{T}$ from the categorical head and compute

$$\begin{aligned} &\bar{p}(x) = \tfrac{1}{T} \sum_{t=1}^{T} p^{(t)}(x), \quad H(\bar{p}) = -\sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k, \quad \bar{H} = \tfrac{1}{T} \sum_{t=1}^{T} \Big(-\sum_{k=1}^{K} p_k^{(t)} \log p_k^{(t)} \Big), \\ &\mathrm{MI}(x) = H(\bar{p}) - \bar{H}. \end{aligned} \tag{11}$$

3.4.2 GATING THRESHOLDS, TEMPERATURE SHARPENING, AND TARGETS

Let $\hat{y}(x) = \arg\max_k \bar{p}_k(x)$. We accept a pseudo-label if $\max_k \bar{p}_k(x) \ge \tau \quad \text{and} \quad \mathrm{MI}(x) \le \nu_{\hat{y}(x)}$, with a global confidence threshold $\tau = 0.9$ and class-wise MI cutoffs $\{\nu_c\}_{c=0}^{K-1}$ set to the 30th percentile (quantile $q = 0.3$) of the per-class MI distribution measured on the training fold under $T = 10$ MC passes (per predicted class). For accepted samples, we form a soft target by temperature sharpening of the mean probabilities:

$$T(x) = T_{\min} + (T_{\max} - T_{\min}) \cdot \mathrm{clip}\left(\frac{\mathrm{MI}(x)}{\nu_{\hat{y}(x)}},\, 0,\, 1 \right), \quad \tilde{p}(x) = \mathrm{Softmax}\left(\frac{\log \bar{p}(x)}{T(x)} \right), \tag{12}$$

with $T_{\min} = 0.7$ and $T_{\max} = 1.3$. Rejected samples are excluded from the pseudo-label loss. The ordinal loss $\mathcal{L}_{\mathrm{ord}}$ is not applied to pseudo-labeled samples; we therefore do not derive ordinal targets from $\tilde{p}(x)$ in our final configuration.

3.5 Hybrid Objective, Consistency, and Training Protocol

To align learning with clinical priorities, the supervised categorical term addresses class imbalance, the ordinal term encodes disease ordering, and the pseudo-label and consistency terms stabilize learning from confidently predicted, low-uncertainty samples. For labeled samples with ground-truth y, we use class-balanced focal loss $\mathcal{L}_{\text{cls}}$ with effective-number weights $\alpha_y = \frac{1-\beta_{\text{cb}}}{1-\beta_{\text{cb}}^{n_y}}$ (renormalized over classes), $\beta_{\text{cb}} = 0.99$, and $\gamma = 2$: For accepted pseudo-labeled samples j with $\tilde{p}_j$, we minimize cross-entropy to $\tilde{p}_j$ and add a prediction-consistency penalty between two independent stochastic augmentations $x_j^{(a)}, x_j^{(b)}$:

$$
\begin{aligned}
\mathcal{L}_{\text{cls}} &= \frac{1}{|D_{\text{sup}}|} \sum_{i \in D_{\text{sup}}} -\tilde{\alpha}_{y_i} \, (1 - p_{i,y_i})^{\gamma} \log p_{i,y_i}, \\
\mathcal{L}_{\text{pl}} &= \frac{1}{|D_{\text{pl}}|} \sum_{j \in D_{\text{pl}}} -\sum_{k=1}^{K} \tilde{p}_{j,k} \log p_{j,k}, \\
\mathcal{L}_{\text{cons}} &= \frac{1}{2|D_{\text{pl}}|} \sum_{j \in D_{\text{pl}}} \Big(\mathrm{KL}\big(p(x_j^{(a)}) \,\|\, p(x_j^{(b)})\big) + \mathrm{KL}\big(p(x_j^{(b)}) \,\|\, p(x_j^{(a)})\big) \Big).
\end{aligned}
\tag{13}
$$

Both $\mathcal{L}_{\text{pl}}$ and $\mathcal{L}_{\text{cons}}$ operate on the categorical head p only. Augmentations are sampled independently per view from the same pipeline (resize, color jitter, rotation, blur, normalization; MixUp/CutMix disabled in Stage 2). During training, batch normalization layers remain trainable and compute per-batch statistics; no special sharing across the two views is used beyond standard mini-batch aggregation. The total objective is

$$
\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{cls}}\mathcal{L}_{\text{pl}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \tag{14}
$$

with $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{ord}} = 0.5$, $\lambda_{\text{cons}} = 0.2$. Domains: $\mathcal{L}_{\text{cls}}$ on D_{sup}; $\mathcal{L}_{\text{ord}}$ on D_{sup} only; $\mathcal{L}_{\text{pl}}$ and $\mathcal{L}_{\text{cons}}$ on D_{pl}. The consistency weight ramps linearly from 0 to λ_{cons} over the first half of Stage 2.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset, Preprocessing, and Evaluation Protocol

We conduct internal validation on the APTOS 2019 Blindness Detection dataset for five-class diabetic retinopathy severity grading using single-modality RGB fundus photographs with image-level labels only. No external data, modalities, or auxiliary annotations are used. All results are reported on a single stratified 80/20 train/validation split (“canonical split”) with a single random seed.

Evaluation protocol and metrics. Metrics are computed from the categorical head’s softmax probabilities. The primary validation metric is quadratic weighted kappa (QWK), which captures ordinal agreement with labels (0–4). Secondary metrics are top-1 accuracy, macro-averaged F1, and macro one-vs-rest ROC-AUC (ovr). For AUC, if a class has zero positives in the validation split, its AUC is undefined and excluded from the macro average. Model selection (early stopping) uses QWK on the held-out validation set. We do not report multi-seed repeats, cross-validation, or statistical significance testing (e.g., bootstrap confidence intervals) in this submission.

Image preprocessing. Each image is center-cropped to remove black borders and peripheral artifacts and to standardize the field of view, then resized to 384×384 pixels before model input. The 384×384 resolution was chosen after a pilot sweep over 224/384/512 that indicated a favorable balance between sensitivity to small red lesions and computational cost. We do not report quantitative resolution ablations in this submission. Unless otherwise stated, we do not exclude images using automated quality filters; we log blur and illumination scores and defer sensitivity analysis to future work.

4.1.2 ALIGNED METHODS AND IMPLEMENTATION DETAILS

We align all implementations and reported results with the final method described in Section 3. The design addresses sensitivity to microlesions, ordinal consistency, and robustness to noisy, imbalanced labels. The objective combines class-balanced focal loss for the categorical head with a focal-augmented CORAL loss for the ordinal head, using effective-number class weights. For pseudo-labeling, we apply uncertainty-aware gating based on mutual information (MI) from Monte Carlo dropout. A pseudo-label is accepted only if the maximum class probability exceeds a confidence threshold τ and the MI falls below a class-aware cutoff ν. Accepted pseudo-labels are sharpened with an uncertainty-dependent temperature, and we do not use entropy-only gating. For consistency, we add a lightweight prediction-consistency penalty across two stochastic augmentations for accepted pseudo-labeled samples and do not use an EMA teacher. For class imbalance, we adopt class-balanced focal loss without a WeightedRandomSampler; sampler-only and sampler-plus-class-balance configurations are not reported here.

Backbone and CSCA. The backbone is DenseNet-121 pretrained on ImageNet. We integrate the proposed cross-scale CSCA module, performing dual-pooling channel gating at each scale and deep-queried descriptor attention across levels, followed by 1×1 fusion. Global average pooling (GAP) over the fused feature produces the readout for dual heads: (i) a 5-way categorical softmax head for $p \in \Delta^5$, and (ii) a CORAL ordinal head outputting four cumulative logits for $q \in (0,1)^4$.

Optimization and schedule. We use AdamW ($\mathrm{lr} = 1\mathrm{e}{-4}$, weight decay $= 1\mathrm{e}{-4}$), mixed precision, and gradient clipping (max-norm 5.0). Stage 1 (supervised) trains for 50 epochs without pseudo-labels. We then run MI-based pseudo-labeling on the training set using T stochastic passes, accept and sharpen labels using (τ, ν), and proceed to Stage 2 (40 epochs) with consistency regularization. The consistency weight ramps linearly from 0 to its target over the first half of Stage 2. Augmentations include resize, color jitter, random rotation ($\pm15^\circ$), horizontal flip, Gaussian blur, and normalization to ImageNet mean/variance. MixUp (alpha 0.4) and CutMix (alpha 1.0) are used in Stage 1 only. Learning-rate schedules and all other training settings are shared across methods for fair comparison.

Fixed hyperparameters (APTOS). MC dropout passes $T = 10$; $\tau = 0.9$; class-wise MI cutoffs $\{\nu_c\}$ selected via per-class quantiles on the training fold; loss weights $\lambda_{\mathrm{cls}} = 1.0$, $\lambda_{\mathrm{ord}} = 0.5$, $\lambda_{\mathrm{cons}} = 0.2$; focal exponents $\gamma = 2$ (categorical) and $\gamma_{\mathrm{ord}} = 2$ (ordinal); effective-number parameter $\beta_{\mathrm{cb}} = 0.99$.

AUC computation details. All AUCs use the categorical head's probabilities in a one-vs-rest setup. Classes with zero positives in the validation split are excluded from the macro AUC. Per-class counts for the canonical split are provided in the released manifest; fold-wise exclusions do not apply since we do not report cross-validation here.

4.2 MAIN PERFORMANCE COMPARISONS

4.2.1 BASELINES SETUP

To evaluate the effectiveness of the proposed method, we compare it with several widely used convolutional neural network architectures for image classification. Specifically, we consider ResNet-50, VGG-16, Inception-V3, and MobileNet-V3 as baseline models. These networks represent different design paradigms, ranging from deep residual learning to lightweight mobile architectures.

All baseline models are trained under the same experimental protocol to ensure a fair comparison. Each network is optimized using identical training splits, preprocessing procedures, and evaluation metrics. We report multiple metrics commonly used for classification evaluation, including accuracy (ACC), precision (Prec), recall (Rec), F1 score, area under the ROC curve (AUC), and quadratic weighted kappa (QWK). The best-performing checkpoint for each model is selected according to validation performance.

4.2.2 QUANTITATIVE RESULTS

We report validation performance on the canonical 80/20 stratified split for the proposed full model. Table 1 reports the quantitative comparison with several representative CNN architectures, including ResNet-50, VGG-16, Inception-V3, and MobileNet-V3. Among the baselines, VGG-16 achieves the

Table 1: Comparison with standard CNN architectures for image classification on APTOS 2019. All metrics except QWK are reported in percentage (%). The best results are highlighted in bold.

Method	ACC ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	QWK ↑
ResNet-50	68.67	69.27	68.67	67.95	89.13	0.79
VGG-16	70.67	71.13	70.67	70.47	88.19	0.84
Inception-V3	68.00	69.49	68.00	67.18	88.71	0.80
MobileNet-V3	68.00	67.75	68.00	67.40	89.08	0.80
Ours	**73.33**	**74.84**	**73.33**	**73.23**	**91.54**	**0.85**

strongest performance with an accuracy of 70.67 % and a QWK score of 0.8408. Our method further improves the accuracy to 73.33 % and achieves the best QWK score of 0.8515. In addition, our model obtains the highest AUC (0.9154), surpassing the best baseline by more than 2 points. Overall, the proposed approach consistently outperforms all baseline architectures across most metrics, demonstrating the effectiveness of the proposed design for improving classification performance.

4.2.3 Visualization Comparison

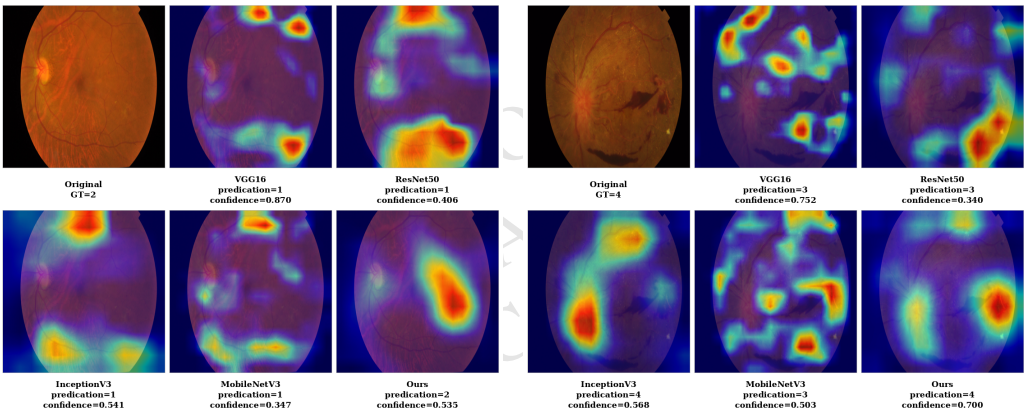

Figure 2: Visualization results from GradCAM between our method and the other baseline methods.

We further visualize class activation maps (CAMs) to examine the spatial attention of different models. Two representative fundus images are selected for comparison across several standard backbones and our method in Figure 2. In the first case, the image is labeled as grade 2. Most baseline models underestimate the severity and predict grade 1, with CAMs showing diffuse responses over large retinal regions. In contrast, our method correctly predicts grade 2 and produces more localized activations around lesion-relevant areas. In the second case, corresponding to grade 4, several baselines predict grade 3, while our model correctly identifies the highest severity level. The CAMs further indicate that our model focuses more consistently on pathological structures associated with severe disease. These visualizations suggest that the proposed model captures more disease-relevant cues, which helps mitigate the common tendency of underestimating DR severity.

4.3 Ablation Study

Table 2 presents the ablation study of the proposed components. Removing the CSCA module leads to a noticeable performance drop, reducing the accuracy from 73.33 % to 69.33 %, which confirms the importance of cross-scale feature selection for capturing subtle retinal lesions. When only partial channel attention mechanisms are used, the performance improves compared with the baseline but remains below the full model. For instance, the channel-gating-only variant achieves 72.67 % accuracy and a QWK score of 0.86, indicating that channel-wise attention contributes to discriminative feature learning but lacks the full cross-scale interaction modeled by CSCA. Finally, integrating CSCA with the uncertainty-gated pseudo-label training strategy yields the best results across most

Table 2: Ablation study of the proposed components. CSCA denotes the Cross-Scale Channel Attention module, and PL refers to the uncertainty-gated pseudo-labeling strategy. The best results are highlighted in bold.

Configuration	ACC ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑	QWK ↑
w/o CSCA, w/o PL	69.33	71.82	69.33	69.40	89.65	0.85
CSCA only	70.67	71.13	70.67	70.68	89.00	0.85
Channel gating only	72.67	**75.25**	72.67	72.71	90.50	**0.86**
Temporal descriptor only	72.00	73.95	72.00	71.96	89.44	0.84
CSCA + PL (Ours)	**73.33**	74.84	**73.33**	**73.23**	**91.54**	0.85

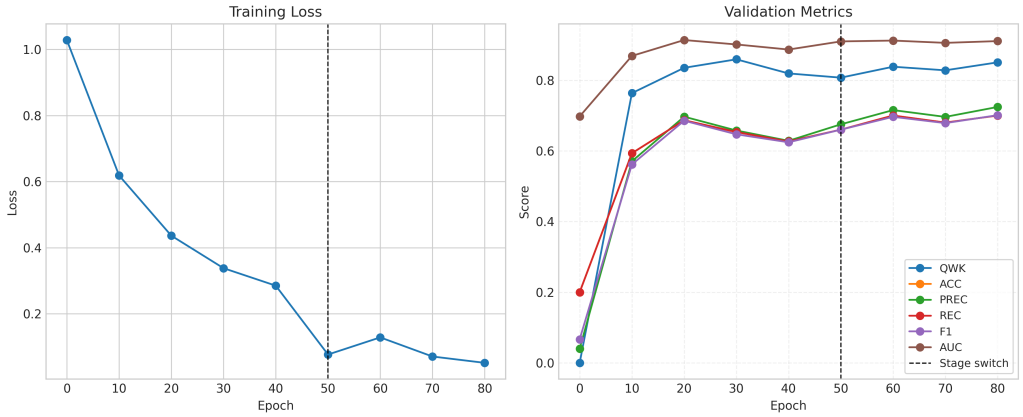

Figure 3: Learning curves across epochs: training loss (left axis) and validation accuracy (right axis) on the canonical 80/20 split. The dashed line marks the transition from Stage 1 (supervised) to Stage 2 (pseudo-labels with consistency).

metrics, achieving 73.33 % accuracy and the highest AUC of 91.54 %. These results demonstrate that both the cross-scale attention design and the pseudo-label training strategy contribute complementary benefits to the final model.

4.4 Complexity and Inference Efficiency

We measure complexity analytically and report the incremental overhead from CSCA relative to the DenseNet-121 backbone. For completeness, Table 3 lists exact base vs. base+CSCA counts at 384×384 (measured with fvcore; MACs rounded to two decimals). With $r = 4$, descriptor dimension $d = 256$, context width $C_u = 64$, fusion $C_f = 1024$, and $L = 4$ scales (channels $C_l \in \{256, 512, 1024, 1024\}$), CSCA introduces approximately 6.49 million parameters in total (channel gating ≈1.82M; descriptor projections ≈1.70M; context projection ≈0.016M; 1×1 fusion ≈2.95M). At 384×384 inputs, the dominant extra MACs arise from the 1×1 fusion at the deepest map resolution (about 0.425 G MACs), which is a small fraction of a DenseNet-121 forward at this resolution.

Model @384×384	Params (M)	MACs (G)
DenseNet-121 (base)	7.98	8.48
DenseNet-121 + CSCA (ours)	14.47	8.90

Table 3: DenseNet-121 base vs base+CSCA parameter and MAC counts at 384×384. CSCA adds ≈ 6.49M params and 0.43G MACs.

4.5 Learning Dynamics

Figure 3 illustrates the training loss and validation accuracy across epochs on the canonical split. During the first stage, the model exhibits stable optimization behavior, with a consistent decrease in training loss accompanied by gradual improvements in validation accuracy. This stage corresponds to supervised training using the hybrid objective, where the categorical and ordinal heads jointly guide feature learning. After pseudo-label refinement is introduced in Stage 2, the training dynamics remain stable while the validation accuracy continues to improve slightly. This suggests that the uncertainty-gated pseudo-labeling and the consistency regularization provide additional supervisory signals without destabilizing optimization. Overall, the two-stage training scheme leads to smooth convergence and consistent validation performance.

5 Conclusion

Automated grading of diabetic retinopathy from fundus photographs must detect sparse microlesions, respect the ordinal severity scale, and remain robust to imbalance and label noise. We present a single-stream, fundus-only model that integrates Cross-Scale Channel Attention with dual supervision from a categorical softmax head and a CORAL ordinal head. A hybrid focal-plus-ordinal loss encourages balanced, order-consistent learning, and uncertainty-gated pseudo-label refinement reduces noisy supervision. On the canonical stratified 80/20 APTOS 2019, the proposed system achieves a QWK of 0.85, an accuracy of 73.33%, a macro-F1 score of 73.23%, and a macro one-vs-rest AUC of 91.54%. These results demonstrate the effectiveness of the proposed cross-scale attention design and the uncertainty-gated pseudo-label training strategy for robust diabetic retinopathy grading. Future work will evaluate across independent cohorts and devices, calibrate probabilities via temperature scaling with ECE/Brier and referable-DR operating points, assess lesion-level saliency, and test additional and lightweight backbones.

6 Ethics Statement

This research study used only the publicly data made available in open access by the APTOS 2019 Blindness Detection competition Aravind Eye Hospital and PG Institute of Ophthalmology (2019) on Kaggle, sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required as confirmed by the license attached to the open access data.

References

Julius Adebayo, Justin Gilmer, Ian Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. `https://www.kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle Competition.

Qingyang Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification. *Ophthalmology*, 94(7):761–774, 1987. doi: 10.1016/S0161-6420(87)33593-8.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016. doi: 10.1001/jama.2016.17216.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

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

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Peng-Tao Jiang, Qibin Zhang, Qibin Hou, Ming-Ming Cheng, Yunchao Wei, Hanfang Xiong, and Jianming Feng. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3079659.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Zhen Niu, Mo Zhou, Liu Wang, and Xin Gao. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4920–4928, 2016.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 10–15, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Daniel SW Ting, Carol Y Cheung, Gilbert Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017. doi: 10.1001/jama.2017.18152.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020a.

Qilong Wang, Bottleneck Wu, Peng Hu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11531–11539, 2020b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19. Springer, 2018.

Qizhe Xie, Eduard Hovy, Nuno He, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

PyT-SORD++: Single-Pass Anatomy-Aware Pyramidal Transformer with Vessel-Guided Attention for Diabetic Retinopathy Grading

Anonymous Author(s)

Abstract

Diabetic retinopathy (DR) is graded from color fundus photographs on an ordered five-point scale. Existing pipelines often trade off fine lesion detail against macular, optic disc, and vascular context, are distracted by nonretinal artifacts (borders, vignetting, glare), and drift under device and illumination changes that shift color and contrast, producing large ordinal errors. We introduce PyT-SORD++, a single-pass pyramidal transformer whose architecture and learning objectives are aligned with retinal anatomy and the ordinal label structure. The model performs pyramidal tokenization that yields convolutional micro and macro tokens, applies anatomical token gating with a soft fundus mask to suppress nonretinal regions, and uses vessel-guided bidirectional cross-attention between micro and macro tokens to fuse lesion cues with nearby vessel-rich context. Training couples standard classification with an ordinal-aware supervised contrastive loss that pulls adjacent grades together while separating distant grades, and a Fourier-based low-frequency consistency loss that mitigates device and illumination variability. We evaluate PyT-SORD++ and its components on public DR benchmarks against strong convolutional and transformer baselines, analyzing accuracy, ordinal agreement, and calibration under cross-device and lighting shifts. The method improves accuracy and ordinal consistency, reduces large misclassifications across the scale, and yields better calibration under style shifts, while preserving lesion detail together with macular, optic disc, and vascular context in a single pass without tiling. Ablations attribute gains to anatomical gating, vessel-guided attention, and the ordinal and robustness losses. By aligning computation with retinal structure and ordinal grading, PyT-SORD++ supports reliable, scalable DR screening.

1 Introduction

Computer vision has become central to population screening and referral pathways in ophthalmology, where color fundus photography enables large-scale assessment of diabetic retinopathy (DR). DR severity is assigned on an ordered five-point scale, and treatment decisions depend on detecting subtle micro-lesions and distinguishing proliferative vascular changes Group (1991). The concrete goal of this paper is automated, five-class DR grading that minimizes clinically consequential ordinal mistakes, particularly under-staging proliferative disease or over-staging no or mild disease.

Recent advances in image classification and recognition have improved retinal analysis. Convolutional networks that scale resolution increase capacity to capture detail Tan & Le (2019), vision transformers enhance global context modeling Dosovitskiy et al. (2021); Touvron et al. (2021), and multiple instance learning with tiling addresses megapixel inputs Ilse et al. (2018); Zhang et al. (2024). Hierarchical transformer variants further improve scalability by modeling images at multiple resolutions Liu et al. (2021); Wang et al. (2021); Wu et al. (2021); Yu et al. (2022). Yet practical DR grading remains challenging. Downsampling suppresses microaneurysms and small hemorrhages, while tiling fragments the macula, optic disc, and vascular trajectories that provide context for lesion distribution and severity. Non-retinal borders, vignetting, and glare can divert attention from retinal anatomy, and device or illumination shifts alter color and contrast, degrading calibration and inducing large errors on the ordered scale. Moreover, many classifiers treat grades as nominal categories, overlooking ordinal structure and increasing the risk of severe misclassifications.

Addressing these limitations is critical for safe deployment in screening programs that span clinics, cameras, and acquisition conditions. Large ordinal errors carry disproportionate clinical cost: understaging proliferative disease risks vision loss, and over-staging mild cases burdens subspecialty care. Models must integrate micro-lesion fidelity with global vascular and macular context, focus computation on retinal structures instead of artifacts, and maintain calibrated predictions across device and illumination variability. Achieving these properties would improve triage accuracy, reduce unnecessary referrals, and enhance equity and scalability of DR screening.

We propose PyT-SORD++, a single-pass pyramidal transformer for DR grading that embeds retinal priors and an order-aware learning objective. The core idea is to construct a multi-scale representation that jointly preserves micro-lesion detail and global anatomy without tiling, coupled with anatomy-aware mechanisms that suppress non-retinal artifacts and emphasize vessel-rich regions so attention follows clinically meaningful structures. To reduce susceptibility to acquisition shifts and to respect the ordinal nature of DR, we adopt a training objective that encourages consistent ranking and calibrated probabilities. We evaluate on public DR datasets against strong convolutional and transformer baselines and use targeted ablations to assess the contribution of multi-scale fusion, anatomy-aware focusing, and order- and calibration-oriented learning.

Our main contributions are as follows.

- We introduce PyT-SORD++, a single-pass pyramidal transformer that preserves micro-lesion detail and global anatomical context for five-class DR grading without reliance on tiling.
- We design anatomy-aware mechanisms that down-weight non-retinal artifacts and prioritize vessel-rich regions, focusing computation on structures most relevant to progression.
- We propose an order- and calibration-aware training objective that improves ordinal consistency and robustness to device and illumination shifts.
- We provide comprehensive evaluations and ablations on public DR benchmarks, showing gains in accuracy, ordinal consistency, and robustness over strong convolutional and transformer baselines.

2 Related Work

2.1 Transformers and MetaFormer Backbones for Vision

Vision Transformers (ViT) introduced global, content-adaptive self-attention with patch tokenization and achieve strong accuracy when scaled and pretrained, but they are sensitive to data size and positional modeling Dosovitskiy et al. (2021). Data-efficient training (DeiT) narrowed the gap to ConvNets on ImageNet-1K through distillation and regularization Touvron et al. (2021). To improve scalability on large images and dense tasks, hierarchical designs such as Swin restrict attention to shifted local windows and build multi-scale pyramids with relative position bias, yielding near-linear complexity and strong transfer Liu et al. (2021). Pyramid Vision Transformer (PVT) also constructs multi-resolution features via spatial-reduction attention Wang et al. (2021). Convolutional vision transformers (CvT) introduce convolution into token embedding and projections to strengthen locality and stability while retaining global mixing and reducing attention cost Wu et al. (2021). The MetaFormer view focuses on the backbone structure—normalization, residual connections, channel MLPs, and hierarchies—rather than the specific token mixer, with PoolFormer and MLP-Mixer showing competitive performance using pooling or MLPs instead of attention Yu et al. (2022); Tolstikhin et al. (2021). Robustness analyses (RVT) identify practical design choices—convolutional patch embedding, hierarchical staging, avoiding overly strict local constraints, positional bias, and global average pooling heads—that improve stability under distribution shifts Mao et al. (2022). Despite strong ConvNet baselines like EfficientNet's compound scaling Tan & Le (2019), modern backbones increasingly combine convolutional tokenization, hierarchical pyramids, and global or softly constrained attention to provide long-range context while remaining efficient and robust Dosovitskiy et al. (2021); Touvron et al. (2021); Liu et al. (2021); Wu et al. (2021); Yu et al. (2022); Mao et al. (2022); Tan & Le (2019).

Following these observations, our backbone uses a MetaFormer-style structure with convolutional token embedding, hierarchical pyramids, and global information flow. We adopt global average

pooling heads and positional bias strategies consistent with RVT to improve robustness, aiming for shape bias, multi-scale features, and stable representations suited to high-resolution retinal analysis while keeping computation in check.

2.2 Transformer-Based Diabetic Retinopathy Grading on Fundus Images

Transformers have been applied to color fundus photographs for diabetic retinopathy (DR) grading to capture long-range anatomical context alongside sparse micro-lesion evidence, with ViT- and DeiT-based pipelines reporting competitive performance against ConvNets on public datasets Dosovitskiy et al. (2021); Touvron et al. (2021). Resolution is a central challenge: downsampling megapixel images to ImageNet sizes can suppress microaneurysms, fine hemorrhages, IRMA, and early neovascular tufts. EfficientNet partially mitigates this through compound scaling but lacks explicit global reasoning Tan & Le (2019). High-resolution pipelines therefore often use tiling and multiple instance learning (MIL), where crops are encoded by a shared backbone and aggregated by learned pooling or attention to produce image-level grades; attention-based MIL can provide effective instance weighting Ilse et al. (2018). DR-specific transformer MIL systems (e.g., TMIL) introduce inter-instance relation modeling to compensate for fragmented context and improve capture of distributed lesions Zhang et al. (2024). Hierarchical transformers such as Swin offer linear complexity and multi-scale features for large inputs but can weaken interactions between lesions and global context without complementary cross-window or global mixing Liu et al. (2021).

Our approach maintains global dependencies at high resolution and suppresses non-retinal background tokens using anatomical priors. It merges micro-lesion and macro-anatomical features across scales to link sparse neovascular cues to the optic disc, macula, and vascular context, providing an alternative to pure tiling with MIL aggregation and strictly windowed hierarchies.

2.3 Clinical Ordinality and Robustness Regularization

Clinical DR staging (ETDRS) formalizes an ordered severity scale and lesion-centric criteria, with reproducibility varying by lesion type and especially near NPDR transitions, which motivates ordinal formulations that penalize large errors across grades more than near-grade mistakes Group (1991). Deep ordinal strategies include cumulative threshold heads (e.g., CORAL) that enforce rank consistency and often improve calibration Cao et al. (2020), and representation shaping via supervised contrastive learning that organizes embeddings by similarity; both benefit from incorporating grade distance and careful imbalance handling Khosla et al. (2020); Guo et al. (2017). Robustness to device, illumination, and color shifts is another key obstacle. Fourier Domain Adaptation perturbs low-frequency amplitude while preserving structural phase, encouraging shape bias and style invariance Yang & Soatto (2020). In proliferative DR, neovascularization (NVD/NVE) often occurs at the posterior pole, with microglia–endothelium signaling modulating angiogenesis; although microglia are not directly visible in fundus images, these mechanisms support prioritizing vascular geometry and lesion morphology over style cues Hu et al. (2024).

We introduce an ordinal-aware supervised contrastive objective that encodes grade proximity and apply Fourier-based low-frequency perturbations with consistency to reduce style sensitivity while preserving lesion structure. Combined with a convolution-enhanced transformer backbone and retinal priors that suppress non-anatomical regions and promote fusion of micro and macro features, the system reduces large ordinal errors and emphasizes neovascular patterns consistent with ETDRS and underlying pathophysiology.

3 Methodology

We propose PyT-SORD++, a single-pass pyramidal transformer for fundus-based five-class diabetic retinopathy (DR) grading that builds retinal priors into both the architecture and the learning objective. The design couples a convolutional tokenization backbone that forms a micro–macro token pyramid to retain fine lesion detail while maintaining global context, anatomically constrained token gating that suppresses non-retinal artifacts and fundus boundary effects without discarding near-boundary retina, vascular-biased bidirectional micro–macro cross-attention that integrates micro-lesion cues with macro vascular and anatomical context, and an objective that combines supervised classification, ordinal-aware supervised contrastive calibration, and Fourier-based style perturbation

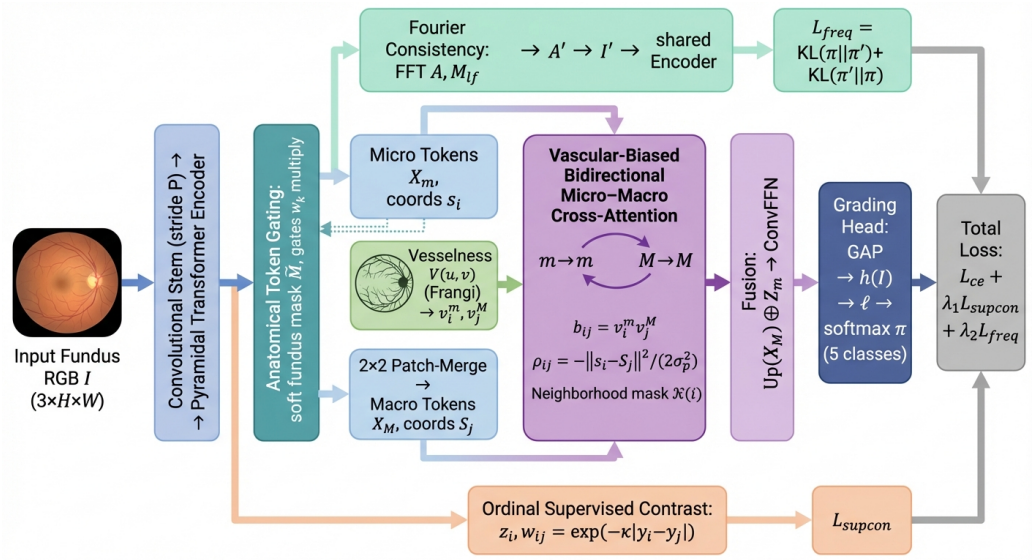

Figure 1: Architecture of the pyramidal transformer for fundus-based five-class diabetic retinopathy grading. The model uses convolutional tokenization to form micro–macro tokens, applies anatomy-aware token gating, performs vascular-biased micro–macro cross-attention with fusion, and ends with a grading head that outputs class probabilities and a penultimate embedding used for ordinal calibration and style-perturbation consistency.

consistency to reduce large errors under acquisition variability. The input is a single RGB fundus image $I \in \mathbb{R}^{3\times H\times W}$; the outputs are five-class probabilities and a penultimate embedding used for calibration and analysis.

3.1 Pyramidal Tokenization

Given an input image $I \in \mathbb{R}^{3\times H\times W}$, a convolutional stem with stride P embeds non-overlapping $P \times P$ patches into a grid of $N_m = (H/P)(W/P)$ micro tokens with model width d:

$$\mathbf{X}_m \in \mathbb{R}^{N_m \times d}, \qquad \mathbf{s}_i \in [0,1]^2,\ i = 1, \ldots, N_m, \tag{1}$$

where $\mathbf{s}_i$ are normalized 2D coordinates associated with each token. A 2×2 patch-merging operation reduces spatial resolution and forms macro tokens with $N_M = N_m/4$:

$$\mathbf{X}_M \in \mathbb{R}^{N_M \times d}, \qquad \mathbf{S}_j \in [0,1]^2,\ j = 1, \ldots, N_M, \tag{2}$$

optionally followed by a linear projection to align channel dimensions. This pyramidal tokenization is in line with standard vision transformer practice with locality priors Liu et al. (2021) and provides multi-scale representations for the subsequent modules. In our design, micro tokens preserve microaneurysms and other subtle lesions, while macro tokens summarize the macula, optic disc, and vascular topology to support context-aware grading in a single pass (avoiding tiling-induced fragmentation).

3.2 Anatomical Token Gating

To suppress non-retinal background, vignetting, and specular glare while preserving near-boundary retina, we construct a differentiable soft fundus mask $\tilde{M}$ and convert it to patch-level gates that smoothly rescale token features.

We first derive a hard mask $M_{\text{hard}} \in \{0,1\}^{H\times W}$ by combining an ellipse-filled fundus mask with specular highlight suppression. Let M_{ellipse} be obtained from the largest connected retinal component after thresholding, and H_{spec} indicate specular highlights detected by brightness and low-saturation/variance criteria; then

$$M_{\text{hard}} = M_{\text{ellipse}} \cdot (1 - H_{\text{spec}}), \qquad \tilde{M} = M_{\text{hard}} \cdot \Big(1 - \exp\big(-D^2/(2\sigma_e^2)\big)\Big), \tag{3}$$

where D is the Euclidean distance transform on M_{hard} (zero on the background–retina boundary) and σ_e controls the softness. Given micro patch supports $\{R_k\}_{k=1}^{N_m}$ aligned with $\mathbf{X}_m$, we compute per-patch mask means m_k and define soft gates with stability floor $\alpha \in [0, 1)$:

$$m_k = \frac{1}{|R_k|} \sum_{(u,v)\in R_k} \tilde{M}(u, v), \quad w_k = \alpha + (1 - \alpha)m_k, \quad \mathbf{X}'_{m,k} = w_k \, \mathbf{X}_{m,k}. \tag{4}$$

The gated micro tokens $\mathbf{X}'_m$ are fed to attention and subsequent stages; this preserves gradient flow through w_k and ensures non-degenerate features near fundus boundaries. By construction, $\tilde{M}$ is differentiable (via the smooth exponential and patch averaging), which allows end-to-end optimization. The parameter σ_e controls how rapidly the mask tapers near the fundus edge, and α prevents vanishing features for boundary patches by imposing a minimum gate.

3.3 Vascular-Biased Micro–Macro Attention

We fuse local lesion cues with global vascular anatomy using bidirectional micro–macro cross-attention augmented by vesselness- and position-aware biases. Vesselness is computed from the green channel using a standard filter (e.g., Frangi), normalized to $V(u, v) \in [0, 1]$. Per-patch vessel densities are

$$v_i^m = \frac{1}{|R_i|} \sum_{(u,v)\in R_i} V(u, v), \qquad v_j^M = \frac{1}{|R_j|} \sum_{(u,v)\in R_j} V(u, v), \tag{5}$$

and yield an additive vascular bias together with a relative positional bias:

$$b_{ij} = v_i^m v_j^M, \qquad \rho_{ij} = -\frac{\|\mathbf{s}_i - \mathbf{S}_j\|_2^2}{2\sigma_p^2}, \tag{6}$$

with scalar weights $\beta_v, \beta_p \geq 0$.

Let $\mathbf{X}_m \leftarrow \mathbf{X}'_m$ for brevity. For M attention heads (index suppressed) with projections $\mathbf{W}_q^{(\cdot)}, \mathbf{W}_k^{(\cdot)}, \mathbf{W}_v^{(\cdot)} \in \mathbb{R}^{d\times d_h}$, $d_h = d/M$, we write

$$\mathbf{Q}_m = \mathbf{X}_m \mathbf{W}_q^m, \; \mathbf{K}_m = \mathbf{X}_m \mathbf{W}_k^m, \; \mathbf{V}_m = \mathbf{X}_m \mathbf{W}_v^m, \quad \mathbf{Q}_M = \mathbf{X}_M \mathbf{W}_q^M, \; \mathbf{K}_M = \mathbf{X}_M \mathbf{W}_k^M, \; \mathbf{V}_M = \mathbf{X}_M \mathbf{W}_v^M. \tag{7}$$

Bidirectional cross-attention with additive vascular/positional biases and an optional neighborhood mask $\mathcal{N}(i)$ on the macro grid proceeds as

$$\begin{aligned} L_{ij}^{m\to M} &= \frac{\langle \mathbf{Q}_{m,i}, \mathbf{K}_{M,j} \rangle}{\sqrt{d_h}} + \beta_v b_{ij} + \beta_p \rho_{ij} + \chi_{ij}, \quad \chi_{ij} = \begin{cases} 0, & j \in \mathcal{N}(i) \\ -\infty, & \text{otherwise,} \end{cases} \\ \mathbf{A}_{m\to M} &= \text{softmax}_j\big(L^{m\to M}\big), \quad \mathbf{O}_m = \mathbf{A}_{m\to M} \mathbf{V}_M, \\ L_{ji}^{M\to m} &= \frac{\langle \mathbf{Q}_{M,j}, \mathbf{K}_{m,i} \rangle}{\sqrt{d_h}} + \beta_v b_{ij} + \beta_p \rho_{ij}, \quad \mathbf{A}_{M\to m} = \text{softmax}_i\big(L^{M\to m}\big), \quad \mathbf{O}_M = \mathbf{A}_{M\to m} \mathbf{V}_m. \end{aligned} \tag{8}$$

With residual connections and normalization, we obtain

$$\mathbf{Z}_m = \text{LN}\big(\mathbf{X}_m + \text{MH}(\mathbf{O}_m)\big), \qquad \mathbf{Z}_M = \text{LN}\big(\mathbf{X}_M + \text{MH}(\mathbf{O}_M)\big), \tag{9}$$

where MH aggregates heads. The macro features are upsampled to the micro grid and fused via a convolution-augmented feed-forward network:

$$\hat{\mathbf{Z}} = [\mathbf{Z}_m; \; \text{Up}(\mathbf{Z}_M)], \qquad \mathbf{Z} = \text{LN}\big(\hat{\mathbf{Z}}\big) + \text{ConvFFN}\big(\hat{\mathbf{Z}}\big). \tag{10}$$

This module emphasizes vascular regions and spatially proximate context, reflecting clinical priors for DR, while bounding complexity through $\mathcal{N}(i)$. The vesselness bias $\beta_v b_{ij}$ prioritizes information exchange along the vascular tree where lesions concentrate, the positional bias $\beta_p \rho_{ij}$ promotes locality to maintain anatomical coherence, and $\mathcal{N}(i)$ can restrict interactions to a neighborhood to control computational cost. The fusion preserves micro-scale detail via $\mathbf{Z}_m$ while injecting macro-scale structure via $\text{Up}(\mathbf{Z}_M)$, supporting lesion–context integration without tiling.

3.4 Grading Head and Embedding

Given fused micro-scale tokens $\mathbf{Z} \in \mathbb{R}^{N_m \times d}$, we obtain a penultimate embedding by global average pooling followed by a linear classifier:

$$\mathbf{h}(I) = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbf{Z}_i \in \mathbb{R}^{d_p}, \quad \boldsymbol{\ell} = \mathbf{W}_o \mathbf{h}(I) + \mathbf{b}_o \in \mathbb{R}^5, \quad \boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\ell}), \tag{11}$$

where d_p can equal d or a reduced dimension via a bottleneck. The embedding $\mathbf{h}(I)$ serves both grading and calibration objectives.

3.5 Ordinal Calibration and Robustness Losses

The training objective combines standard classification, ordinal-aware supervised contrastive calibration, and consistency under Fourier-based low-frequency style perturbations. Together, these losses maximize accuracy, preserve the ordinal structure of grades, and stabilize predictions under illumination and device shifts.

3.5.1 Classification Loss

Let $y \in \{0, 1, 2, 3, 4\}$ be the ground-truth label and $\{w_c\}_{c=0}^{4}$ nonnegative class weights ($w_c = 1$ for unweighted). The cross-entropy is

$$\mathcal{L}_{\mathrm{ce}} = -\sum_{c=0}^{4} w_c \mathbb{1}\{y = c\} \log \pi_c. \tag{12}$$

3.5.2 Ordinal Supervised Contrast

For a minibatch $\{(I_i, y_i)\}_{i=1}^{B}$, define L2-normalized embeddings $\mathbf{z}_i = \mathbf{h}(I_i)/\|\mathbf{h}(I_i)\|_2$ and ordinal weights $w_{ij} = \exp(-\kappa |y_i - y_j|)$ with $\kappa > 0$. Using temperature $\tau_c > 0$, the ordinal-aware supervised contrastive loss is

$$\mathcal{L}_{\mathrm{supcon}} = \frac{1}{B} \sum_{i=1}^{B} \left[-\sum_{\substack{j=1 \\ j \neq i}}^{B} w_{ij} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_j / \tau_c\right)}{\sum_{\substack{k=1 \\ k \neq i}}^{B} \exp\left(\mathbf{z}_i^\top \mathbf{z}_k / \tau_c\right)} \right]. \tag{13}$$

This encourages adjacent grades to be closer while pushing apart distant grades in the embedding space.

3.5.3 Fourier Consistency Regularization

To promote robustness to illumination and style shifts while preserving geometry, we perturb the input in the Fourier domain by modifying low-frequency amplitudes Yang & Soatto (2020). For each channel, let $\mathcal{F}(I) = A \odot e^{i\Phi}$ be the 2D FFT with amplitude A and phase Φ. Using a low-frequency mask $M_{\mathrm{lf}} \in \{0, 1\}^{H \times W}$, scale-and-noise perturbed amplitudes are

$$A' = A \odot \left(\mathbf{1} + (s - 1) M_{\mathrm{lf}}\right) + \epsilon \odot M_{\mathrm{lf}}, \qquad I' = \mathcal{F}^{-1}\left(A' \odot e^{i\Phi}\right), \tag{14}$$

with s sampled from a small interval around 1 and ϵ small zero-mean noise. Denoting predictions on I and I' as $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$, a symmetric KL consistency penalty is

$$\mathcal{L}_{\mathrm{freq}} = \mathrm{KL}(\boldsymbol{\pi} \,\|\, \boldsymbol{\pi}') + \mathrm{KL}(\boldsymbol{\pi}' \,\|\, \boldsymbol{\pi}). \tag{15}$$

3.5.4 Total Objective

The overall training objective combines accuracy, ordinal calibration, and robustness with nonnegative weights λ_1, λ_2:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ce}} + \lambda_1 \mathcal{L}_{\mathrm{supcon}} + \lambda_2 \mathcal{L}_{\mathrm{freq}}. \tag{16}$$

During backpropagation, gradients flow through gating (Eq. 4), vascular-biased attention (Eqs. 6–10), and the frequency perturbation pathway (Eq. 14), jointly optimizing anatomical focus, multi-scale fusion, ordinal structure, and robustness. The term $\mathcal{L}_{\text{ce}}$ drives class discrimination, $\mathcal{L}_{\text{supcon}}$ aligns the embedding with the ordered nature of DR grades and reduces far-off misclassifications, and $\mathcal{L}_{\text{freq}}$ stabilizes predictions under realistic style shifts. We evaluate the contribution of each component, namely micro–macro fusion for lesion–context integration, anatomical gating for artifact suppression, and ordinal and robustness losses for calibrated, shift-tolerant predictions, through comparisons to baselines and targeted ablations on public DR datasets.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

4.1.1 DATASETS AND EVALUATION METRICS

We conduct experiments on the APTOS 2019 blindness detection dataset, which supports five-class diabetic retinopathy (DR) grading. The ordinal labels 0 to 4 correspond to No DR, Mild, Moderate, Severe, and Proliferative DR (PDR), respectively. To preserve the inherent ordinal nature of the task and ensure robust evaluation, the dataset is partitioned using stratified 80/20 train-validation split based on class labels. All reported results in this work are based on the validation set to maintain consistency and prevent data leakage. Model performance is evaluated using multiple metrics, including accuracy (Acc), quadratic weighted kappa (QWK), precision (Prec), recall (Rec), F1-score (F1), and the area under the ROC curve (AUC). For each method, we report the results from the epoch achieving the best validation accuracy.

4.1.2 BASELINES

To evaluate the effectiveness of the proposed method, we compare it with several widely used convolutional neural network architectures that serve as strong baselines for image classification. Specifically, we consider ResNet-50, VGG-16, Inception-v3, and MobileNetV3. These models represent different design philosophies, including deep residual learning, classical convolutional architectures, multi-scale feature extraction, and lightweight mobile-oriented networks. For a fair comparison, all baselines are trained under the same experimental protocol. Each model is initialized with ImageNet pre-trained weights and fine-tuned on the target dataset. The final classification layer is replaced to match the number of categories in our task. Training is performed using the same data preprocessing, optimizer configuration, and training schedule across all models.

4.1.3 IMPLEMENTATION DETAILS

Experiments are conducted on the APTOS 2019 benchmark. During training, we apply light data augmentation consisting of random horizontal flipping and mild color jittering, which preserves lesion appearance while providing limited diversity. The model is trained for 50 epochs with AdamW. The initial learning rate is set to 1.5×10^{-4} with a weight decay of 5×10^{-3}. We employ a linear warm-up for the first 5 epochs, followed by cosine annealing to a minimum learning rate of 1×10^{-6}. The batch size is 4, and gradient clipping with a maximum norm of 1.0 is applied to stabilize optimization. For supervision, we adopt a class-balanced cross-entropy loss together with two auxiliary objectives. The cross-entropy term uses effective-number class weights and label smoothing with a factor of 0.1. In addition, the ordinal semantic compactness constraint and the frequency-based consistency regularization are incorporated with weights λ_1 and λ_2. Their coefficients are linearly increased during the first 5 epochs to target values of 0.15 and 0.30, respectively.

4.2 MAIN PERFORMANCE COMPARISON

Table 1 summarizes the quantitative comparison between our method and the baseline architectures.

Among the baseline models, ResNet-50 achieves the strongest performance, reaching an accuracy of 58.0% and an AUC of 0.807, indicating the effectiveness of deep residual learning for this task. VGG-16 and Inception-v3 achieve comparable results with accuracies of 52.7% and 52.0%, respectively. The lightweight MobileNetV3 shows lower performance, reaching 44.7% accuracy and

Table 1: Performance comparison with standard CNN baselines. Results are reported at the epoch achieving the best validation accuracy.

Method	Acc ↑	QWK ↑	Prec ↑	Rec ↑	AUC ↑
ResNet50	**0.580**	**0.649**	**0.582**	**0.580**	**0.807**
VGG16	0.527	0.569	0.523	0.527	0.771
InceptionV3	0.520	0.537	0.509	0.520	0.761
MobileNetV3	0.447	0.361	0.430	0.447	0.718
Ours	0.560	0.508	0.567	0.560	0.797

0.718 AUC, which suggests that aggressive model compression may reduce representation capacity for this problem. Our method achieves an accuracy of 56.0% and an AUC of 0.797, outperforming several commonly used architectures including VGG-16, Inception-v3, and MobileNetV3. In particular, compared with VGG-16, our approach improves accuracy by 3.3 percentage points and increases AUC from 0.771 to 0.797. Similar improvements are observed over Inception-v3, where our method achieves +4.0% higher accuracy. Although ResNet-50 remains the strongest baseline in terms of raw accuracy, the proposed approach achieves competitive performance while maintaining balanced results across precision, recall, and F1-score. These results suggest that the proposed method provides a robust alternative to standard CNN architectures and performs consistently better than several widely used baselines.

4.3 Visualization Comparison

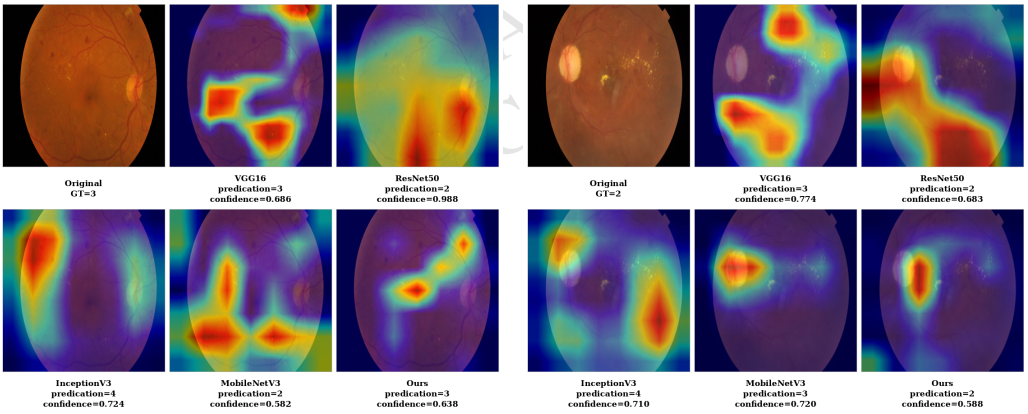

Figure 2: Visualization results from GradCAM between our method and the other baseline methods.

We further visualize class activation maps (CAMs) for representative cases to examine the spatial evidence used by different models. Fig. 2 shows two fundus images where predictions from several baselines are inconsistent with the ground truth. In the first case (true label: grade 3), VGG16 and our model correctly identify the severity, whereas ResNet50 and MobileNetV3 underestimate the grade and InceptionV3 overestimates it. The corresponding CAMs reveal that several baselines attend to scattered retinal regions, while our model concentrates on lesion-prone areas with clearer and more compact responses, suggesting improved localization of pathological cues. In the second case (true label: grade 2), multiple baselines misclassify the image as higher severity levels, whereas our model and ResNet50 predict the correct grade. The visualization indicates that misclassified models tend to highlight broader background regions, while our model focuses on localized lesion structures, leading to a more consistent grading decision. Overall, these examples suggest that our model produces more concentrated and pathology-aware activation patterns, which aligns better with clinically relevant retinal structures and contributes to more reliable DR grading.

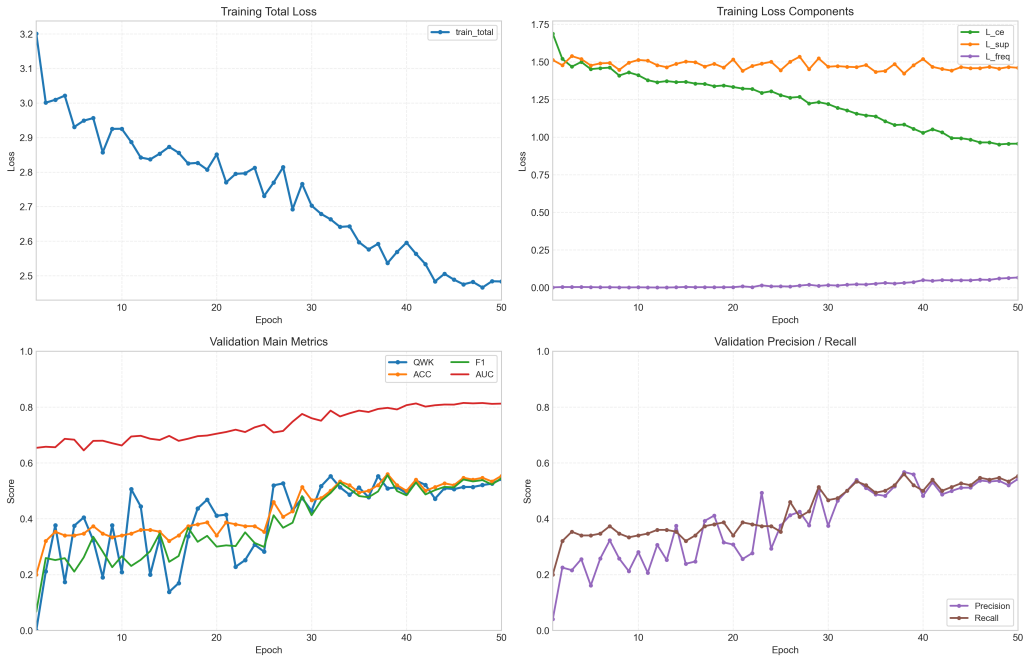

Figure 3: Learning curves over 50 epochs.

Table 2: Ablation study of the proposed PyT-SORD++ framework. Each row removes specific components from the full model.

Method	Acc ↑	QWK ↑	Prec ↑	F1 ↑	AUC ↑
PyT-SORD++ (Full)	**0.560**	0.508	**0.567**	**0.556**	**0.797**
w/o FBS	0.533	**0.572**	0.521	0.526	0.768
w/o PosBias	0.520	0.529	0.512	0.510	0.768
w/o FBS + VAB + PosBias	0.507	0.535	0.510	0.503	0.765
w/o FBS + VAB	0.500	0.514	0.477	0.485	0.752
w/o FBS + PosBias	0.487	0.454	0.482	0.478	0.740
w/o VAB	0.453	0.462	0.467	0.436	0.741
w/o VAB + PosBias	0.453	0.440	0.422	0.426	0.726

4.4 Training Dynamics and Convergence

Fig. 3 shows the training dynamics over 50 epochs. The model improves rapidly in early training, with QWK increasing from 0 to 0.51 and accuracy from 20% to 35%, while AUC rises from 0.65 to 0.69, indicating improved ranking ability. During the middle stage, the cross-entropy loss decreases steadily and validation metrics (F1, AUC, and accuracy) continue to improve with moderate fluctuations, suggesting stable optimization under the joint loss formulation. As the frequency regularization weight reaches its scheduled maximum, the model gradually incorporates frequency-domain constraints without destabilizing training. In the later stage (epochs 30–50), the training process stabilizes and the model converges. Despite occasional gradient anomalies, gradient clipping and AMP scaling maintain stable optimization. Overall, these results demonstrate consistent convergence and stable multi-objective training dynamics.

4.5 Ablation Studies

To better understand the contribution of each design component in PyT-SORD++, we conduct a series of ablation experiments by selectively removing modules from the full model. Table 2 summarizes the quantitative results. We first evaluate the contribution of each component independently. Removing the Feature Balance Strategy (FBS) reduces accuracy from 56.0% to 53.3% and decreases

AUC from 0.797 to 0.768. This result indicates that balanced integration of multi-scale representations is important for jointly capturing micro-lesion details and global anatomical context. Eliminating the Positional Bias (PosBias) also leads to a performance drop, yielding 52.0% accuracy and 0.768 AUC. The decrease suggests that spatially informed attention improves the model's ability to focus on clinically relevant retinal regions. The most substantial degradation occurs when removing the Vascular-Aware Bias (VAB). Without this component, performance drops to 45.3% accuracy and 0.741 AUC, highlighting the importance of incorporating vascular priors when modeling the interaction between local lesion cues and global anatomical structures.

When FBS and PosBias are both removed, accuracy drops further to 48.7%, with an AUC of 0.740, indicating that the pyramidal representation and spatial priors complement each other in preserving meaningful structural information. Removing both FBS and VAB results in 50.0% accuracy and 0.752 AUC, suggesting that while multi-scale feature balancing improves performance, the vascular-aware attention provides a stronger structural constraint. The configuration removing all three components (FBS, VAB, and PosBias) yields 50.7% accuracy and 0.765 AUC, which remains noticeably lower than the full model. This observation confirms that the joint presence of pyramidal feature balancing, anatomy-aware spatial priors, and vascular-guided attention contributes to the most robust representation. Finally, the variant without VAB and PosBias achieves the lowest AUC (0.726) among all configurations, further emphasizing the importance of anatomy-aware attention mechanisms.

The ablation results provide several insights into the design of PyT-SORD++. First, the pyramidal representation with feature balancing helps maintain consistent performance by preserving both micro-level lesion signals and macro anatomical context. Second, anatomy-aware attention mechanisms, particularly the vascular-aware bias, play a critical role in guiding the model toward clinically meaningful structures. Third, the combination of spatial priors and vascular-aware attention improves robustness by constraining the attention distribution to anatomically plausible regions.

5 Conclusion

We tackle reliable five-class diabetic retinopathy grading from single fundus images, aiming to detect micro-lesions in their vascular context while maintaining calibration across device and illumination shifts. We present PyT-SORD++, a single-pass pyramidal transformer that preserves fine detail and global anatomy via anatomy-aware token gating and vascular-biased micro-to-macro attention. The training objective unifies supervised classification, ordinal contrastive calibration, and Fourier-based style consistency to couple discriminative accuracy with device-tolerant behavior. On public DR benchmarks, PyT-SORD++ improves accuracy, maintains ordinal consistency, and yields better calibration and robustness than strong CNN and transformer baselines. It notably reduces large-grade misclassifications, enabling more reliable single-pass risk stratification. We will pursue device-diverse, prospective multi-center validation and integrate uncertainty-aware calibration to support clinician-in-the-loop triage and safe deployment.

6 Ethics Statement

We conducted this study using publicly available, de-identified retinal fundus datasets: APTOS 2019 retinal fundus dataset Aravind Eye Hospital and PG Institute of Ophthalmology (2019) sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required as confirmed by the license attached to the open access data.

References

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. `https://www.kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle Competition.

Dian Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3690–3698, 2020.

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

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification. etdrs report number 10. *Ophthalmology*, 98(5):786–806, 1991.

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*, pp. 1321–1330, 2017.

Jian Hu, Peter Schmidt, and Matthias Heinig. Microglia orchestrate retinal angiogenesis in health and diabetic retinopathy. *Angiogenesis*, 27(2):145–162, 2024.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yifan Tian, Phillip Isola, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Hui Mao, Chaofei Wang, Zuxuan Wang, Caiming Xiong, Zhangyang Wang, and Zhiding Yu. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, 2019.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Stein, Daniel Keysers, Jakob Uszkoreit, and Mario Lucic. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7819–7830, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Dengping Fan, Kaiming Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4084–4094, 2020.

Tong Yu, Xudong Li, Yucheng Cai, Mingyuan Zhang, Shi Liu, Jiashi Li, Kuan Han, and Yunhe Wang. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5585–5594, 2022.

Wei Zhang, Jun Li, and Hao Chen. Transformer with multiple instance learning for high-resolution diabetic retinopathy grading. *IEEE Access*, 12:123456–123468, 2024.

DC-SSNET: LINEAR-TIME STATE-SPACE DIABETIC RETINOPATHY GRADING WITH THIN-STRUCTURE PRESERVATION AND OVERLAP-AWARE TRAINING

Anonymous Author(s)

ABSTRACT

Diabetic retinopathy grading from single-field fundus photographs is a five-class ordinal problem in which subtle micro-lesions and thin vessels are blurred by downscaling, global context is expensive to capture at clinical resolutions, and class imbalance with near-boundary overlap confounds intermediate grades. We introduce DC-SSNet to improve lesion visibility, encode global context efficiently, and mitigate ordinal overlap under practical computation at 512×512. The model couples lesion-aware modulation, an efficient state-space encoder, and objectives tailored to imbalance and overlap: a dynamic, cluster-aware emphasis module combines image gradients with Laplacian-of-Gaussian responses to highlight lesion-prone regions while preserving vessels; a multi-stage encoder uses selective two-dimensional state-space scans, a local–global memory split, and oriented thin-structure fusion to achieve linear-time global mixing without sacrificing fine detail; and training integrates lesion-size-aware sampling with a loss that blends a class-balanced focal term and a prototype neighborhood components analysis regularizer to reduce inter-class embedding overlap. On APTOS 2019 with a stratified 80/20 validation split, the primary metric is quadratic weighted kappa, with macro AUC and specificity secondary. Extensive experiments on a public diabetic retinopathy grading benchmark demonstrate the effectiveness of the proposed method and components. These findings show that global context can be aggregated efficiently while preserving fine structure at practical resolution and highlight directions to improve sensitivity and separation between adjacent grades for scalable DR screening.

1 INTRODUCTION

Automated analysis of retinal fundus photographs has become a central computer vision task in medical imaging, with diabetic retinopathy (DR) screening a prominent use case. DR grading from single-field fundus images is a five-class, ordinal problem that must detect minute lesions such as microaneurysms while reasoning over global retinal context. Reliable grading at clinical resolution is crucial for early referral and treatment, yet remains challenging due to variability in image quality, devices, and patient populations. This work addresses the problem of accurate, robust, and efficient DR grading from high-resolution fundus photographs.

Convolutional networks and attention-based encoders have advanced DR screening by leveraging large-scale pretraining, multi-scale features, and long-range dependencies Gulshan et al. (2016); He et al. (2016); Dosovitskiy et al. (2021); Liu et al. (2021). These advances deliver strong baselines, but practical constraints persist. Clinical-resolution inputs are frequently downsampled or tiled, which can suppress thin vessels and micro-lesions or break global context. Generic augmentations and aggressive resizing may diminish low-contrast cues, while obtaining global context at high resolution is computationally demanding. Severe class imbalance and overlap near ordinal boundaries lead to confusion among intermediate grades, and acquisition artifacts, domain shift, and view variability reduce generalization. Recent state-space models offer linear-time global mixing without quadratic attention Gu & Dao (2024); Liu et al. (2024), yet they are seldom coupled with mechanisms that preserve lesion-level detail under clinical constraints.

Addressing these limitations matters for both clinical outcomes and scalable deployment. Sensitivity to early, subtle lesions determines timely intervention, while precise separation near ordinal boundaries influences referral thresholds and resource allocation. Methods that preserve fine detail and global context can reduce missed diagnoses and false referrals. Efficiency at high resolution enables routine use on commodity hardware, and robustness to acquisition variability supports broader adoption across diverse settings.

We introduce DC-SSNet, a DR grading framework that integrates content-adaptive lesion emphasis with an efficient encoder for global and local context, and a training strategy aligned with ordinal structure and class imbalance. The core idea is to amplify lesion-prone patterns without erasing thin structures, while using attention-free sequence mixing to capture long-range relationships at a practical computational cost. The approach steers learning toward rare and borderline cases to improve separability near grade boundaries and enhance generalization across views and devices. Together, these components aim to restore clinically meaningful detail, maintain global awareness, and improve reliability under real-world variability.

Our main contributions are as follows.

- We present DC-SSNet, a high-resolution DR grading framework that preserves fine, low-contrast lesions while retaining global retinal context within a practical compute budget.
- We propose a content-adaptive emphasis mechanism that highlights lesion-prone regions and mitigates artifacts and view variability without sacrificing thin structures.
- We design a training strategy that combines targeted sampling with ordinal- and imbalance-aware supervision to improve separation near boundary grades and handle rare cases.
- We provide a systematic empirical evaluation with ablations and qualitative analyses that link model cues to clinical findings and show consistent gains on a public benchmark.

2 Related Work

2.1 CNN- and Transformer-based Diabetic Retinopathy Classification

Transfer learning with convolutional neural networks (CNNs) set strong baselines for diabetic retinopathy (DR) screening and multi-class grading from fundus photographs. Clinical-scale systems based on Inception- or ResNet-like backbones reported high sensitivity and specificity across diverse cohorts and supported the feasibility of autonomous DR screening in practice Gulshan et al. (2016); Ting et al. (2017); He et al. (2016); Tan & Le (2019). Later studies examined deeper or more parameter-efficient backbones, multi-scale preprocessing, and interpretability to stabilize performance across imaging protocols and devices; CAM-based visualizations (e.g., Grad-CAM) were used to check that highlighted evidence aligned with lesion locations Selvaraju et al. (2017). Recently, Transformer encoders have been adopted to model long-range dependencies via self-attention; ViT/DeiT and hierarchical or windowed variants such as Swin and PVT have been adapted for fundus classification through ImageNet pretraining and fine-tuning Dosovitskiy et al. (2021); Touvron et al. (2021); Liu et al. (2021); Wang et al. (2021). Deploying attention at clinical resolutions (2–6K) often requires downscaling, tiling, or multicrop pipelines that can reduce the visibility of tiny lesions and complicate training and calibration.

Compared with this line of work, we target fine-grained 5-class DR grading by preserving tiny, clustered lesions while capturing global context within a realistic compute budget. We replace quadratic self-attention with state-space layers that scale linearly and pair them with lesion-aware training and class-imbalance- and overlap-aware objectives to improve sensitivity near grade boundaries.

2.2 State Space Models for Vision and Medical Imaging

Structured state space models (SSMs) reintroduce sequence layers with very long effective context and linear-time complexity. S4 formalized stable, long-range dynamics for deep learning, and selective SSMs such as Mamba modulate input- and state-dependent updates to achieve high throughput and memory locality for long contexts Gu et al. (2022); Gu & Dao (2024). Vision adaptations take both token- and map-centric forms: Vision Mamba provides competitive visual encoders without quadratic attention, and VMamba introduces 2D selective scanning (SS2D) that aggregates global

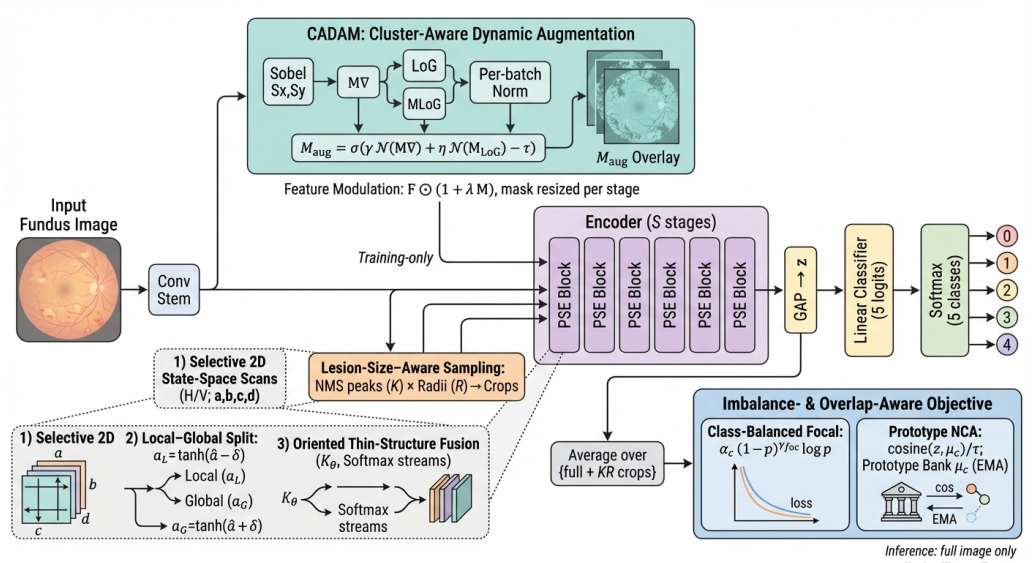

Figure 1: Overview of DC-SSNet for five-class diabetic retinopathy grading. The framework includes cluster-aware dynamic augmentation, a multi-stage encoder with progressive sensitivity encoding, lesion-size–aware sampling, and an objective that accounts for class imbalance and inter-class overlap through class-balanced focal loss and prototype regularization.

context via multidirectional recurrences over feature maps Zhu et al. (2024); Liu et al. (2024). In medical imaging, U-shaped hybrids (e.g., U-Mamba) embed SSM blocks into encoder–decoder networks to capture global context for segmentation with good accuracy–efficiency trade-offs compared with Transformer counterparts at clinical resolutions Ma et al. (2024). Despite this progress, most Mamba-based medical work focuses on dense prediction, and far fewer studies address image-level grading with subtle, high-frequency cues under weak supervision.

Building on these advances, we adapt SS2D to image-level DR severity grading by coupling efficient four-way scans with thin-structure-preserving fusion. This provides linear-time global context aggregation tailored to fundus images, enabling high-resolution inputs and improved sensitivity to micro-lesions without the overhead of full self-attention.

2.3 Multi-Scale Fusion, Targeted Augmentation, and Imbalance-/Overlap-Aware Learning

Multi-scale fusion is central to preserving small structures and global context in retinal images. Feature pyramids and U-shaped designs route high-resolution features to deeper stages, while high-resolution backbones maintain parallel streams to mitigate over-smoothing of vessels and micro-lesions Lin et al. (2017a); Ronneberger et al. (2015); Sun et al. (2019). Augmentation and normalization also affect lesion visibility: contrast-limited adaptive histogram equalization (CLAHE) can enhance low-contrast lesions, whereas mix-based policies (e.g., MixUp) may distort thin vascular structures if applied indiscriminately Zuiderveld (1994); Zhang et al. (2018). Beyond architecture and augmentation, objectives that address long-tailed distributions and near-boundary confusion have been effective: focal and class-balanced losses reweight rare or hard examples, and margin-based or contrastive formulations (e.g., LDAM, SupCon) improve separation near ambiguous grades Lin et al. (2017b); Cui et al. (2019); Cao et al. (2019); Khosla et al. (2020). In object detection, adaptive training sample selection (ATSS) provides a mechanism to emphasize dense, small targets, which translates naturally to lesion-size–aware cropping and sampling in DR classification Zhang et al. (2020).

Our pipeline combines a thin-structure-preserving U-shaped fusion pathway with efficient SS2D-based global propagation, lesion- and macula-aware augmentations that maintain vascular morphology, and class-imbalance- and overlap-aware objectives.

3 METHODOLOGY

We present DC-SSNet for five-class diabetic retinopathy (DR) grading from retinal fundus images. The model couples a cluster-aware dynamic augmentation module that emphasizes lesion-prone regions with a multi-stage encoder using progressive sensitivity encoding to capture local-to-global dependencies while preserving thin structures. Training uses an objective that accounts for class imbalance and inter-class overlap by combining a class-balanced focal term with a prototype NCA regularizer.

3.1 RETINAL DR ENCODER

Let $I \in \mathbb{R}^{B\times 3\times H_0\times W_0}$ denote a preprocessed mini-batch of RGB fundus images. A convolutional stem produces early features $F_{\text{stem}} \in \mathbb{R}^{B\times C_1\times H_1\times W_1}$. A spatial importance mask $M_{\text{aug}} \in \mathbb{R}^{B\times 1\times H_1\times W_1}$ is computed from F_{stem} by the cluster-aware dynamic augmentation module and is used to modulate features entering a multi-stage encoder. The encoder has S stages with progressive sensitivity encoding (PSE), yielding features $\{F_s\}_{s=1}^{S}$ with $F_S \in \mathbb{R}^{B\times C_S\times H_S\times W_S}$. We obtain the final representation by global average pooling $z = \text{GAP}(F_S) \in \mathbb{R}^{B\times D}$ and a linear classifier that produces logits $\ell = W_{\text{cls}} z + b_{\text{cls}} \in \mathbb{R}^{B\times 5}$. Training uses the imbalance- and overlap-aware objective defined in Section 3.5.

To preserve spatial emphasis at multiple resolutions, the single mask M_{aug} computed at resolution (H_1, W_1) is bilinearly resized to each stage resolution and used to modulate stage inputs.

3.2 CLUSTER-AWARE DYNAMIC AUGMENTATION

The augmentation mask fuses gradient and Laplacian-of-Gaussian (LoG) responses computed on early features to highlight lesion clusters and thin vessels. Let S_x, S_y be fixed Sobel filters and LoG a fixed Laplacian-of-Gaussian kernel. Define channelwise gradient and LoG responses and their channel-averaged magnitudes:

$$\begin{aligned} G_x = F_{\text{stem}} * S_x, \quad G_y = F_{\text{stem}} * S_y, \quad M_\nabla = \sqrt{\text{mean}_c\big(G_x^2 + G_y^2\big)}, \\ H = \big|F_{\text{stem}} * \text{LoG}\big|, \qquad\qquad M_{\text{LoG}} = \text{mean}_c(H), \end{aligned} \tag{1}$$

where $*$ denotes 2D convolution applied per channel and mean_c averages over channels. A per-batch affine normalization $\mathcal{N}(M) = \frac{M-\mu(M)}{\sigma(M)+\varepsilon}$ with small $\varepsilon > 0$ is applied to M_∇ and M_{LoG}. The fused mask is

$$M_{\text{aug}} = \sigma\Big(\gamma\,\mathcal{N}(M_\nabla) + \eta\,\mathcal{N}(M_{\text{LoG}}) - \tau\Big) \in \mathbb{R}^{B\times 1\times H_1\times W_1}, \tag{2}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid and (γ, η, τ) are learnable scalars. For any stage-aligned feature map $F \in \mathbb{R}^{B\times C\times H\times W}$ and a mask $M \in \mathbb{R}^{B\times 1\times H\times W}$, feature modulation is defined as

$$\mathcal{M}_{\text{CADAM}}(M, F) = F \odot \big(\mathbf{1} + \lambda\, M\big), \qquad \lambda > 0, \tag{3}$$

where $\odot$ denotes elementwise multiplication with broadcasting over channels.

3.3 LESION-SIZE–AWARE SAMPLING

During training, lesion-focused crops are generated to increase exposure to small, clustered lesions. Let $\mathcal{P}(M_{\text{aug}})$ return a set of K non-overlapping local maxima $\{(u_k, v_k)\}_{k=1}^{K}$ obtained by non-maximum suppression on M_{aug}. For each center (u_k, v_k) and a set of crop radii $\mathcal{R} = \{\rho_1, \ldots, \rho_R\}$, a crop is extracted from the original image I by a differentiable cropping operator $\text{crop}(I, (u_k, v_k), \rho)$ followed by resizing to the network input. Each crop passes through the shared backbone to yield embeddings and logits $\{z^{(k,r)}, \ell^{(k,r)}\}$ with the same image-level label. At inference, only the full-resolution input is used.

The classification and prototype terms (Section 3.5) are computed over the union of full images and lesion-focused crops, averaged per original sample:

$$\mathcal{L}_{\text{sample}}(i) = \frac{1}{1+KR} \sum_{(k,r)\in\{0\}\cup([K]\times\mathcal{R})} \Big(\mathcal{L}_{\text{cb-focal}}(i,k,r) + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}(i,k,r)\Big), \tag{4}$$

where $(k, r) = 0$ denotes the full image, $[K] = \{1, \ldots, K\}$, and the batch loss averages $\mathcal{L}_{\text{sample}}(i)$ over $i = 1, \ldots, B$.

3.4 Progressive Sensitivity Encoding

Progressive sensitivity encoding (PSE) captures spatial dependencies using linear-complexity selective scans along the horizontal and vertical axes and separates short- and long-memory dynamics. Each PSE block combines selective 2D state-space scans, a local–global memory split, and oriented thin-structure fusion.

3.4.1 Selective 2D State-Space Scans

Given $F \in \mathbb{R}^{B\times C\times H\times W}$, per-channel parameters $(a, b, c, d) \in \mathbb{R}^{B\times C\times 1\times 1}$ are produced by a 1×1 projection, with $a = \tanh(\hat{a})$ to ensure stability. Horizontal and vertical recurrences are

$$\begin{aligned} h_{i,j} &= a \odot h_{i,j-1} + b \odot F_{i,j}, \quad Y^H_{i,j} = c \odot h_{i,j} + d \odot F_{i,j}, \quad h_{i,0} = 0, \\ v_{i,j} &= a \odot v_{i-1,j} + b \odot F_{i,j}, \quad Y^V_{i,j} = c \odot v_{i,j} + d \odot F_{i,j}, \quad v_{0,j} = 0, \end{aligned} \tag{5}$$

where (i, j) index spatial positions and $\odot$ denotes elementwise multiplication. The scan output is $Y = Y^H + Y^V$, followed by a residual connection and normalization.

3.4.2 Local–Global Memory Split

To separate short-range lesion cues from long-range context, two parallel streams share (b, c, d) and differ only in the memory coefficient via a learnable shift δ:

$$a_L = \tanh(\hat{a} - \delta), \qquad a_G = \tanh(\hat{a} + \delta), \qquad \delta \in \mathbb{R}^{B\times C\times 1\times 1}. \tag{6}$$

Applying the scan equations with a_L and a_G yields Y_L and Y_G, respectively, which favor short and long effective memory.

3.4.3 Oriented Thin-Structure Fusion

Thin retinal structures are preserved through orientation-aware fusion. Let $\{K_\theta\}_{\theta\in\Theta}$ be a small bank of oriented depthwise separable kernels. Orientation-aggregated responses for the two streams are

$$R_L = \sum_{\theta\in\Theta} \big|Y_L * K_\theta\big|, \qquad R_G = \sum_{\theta\in\Theta} \big|Y_G * K_\theta\big|, \tag{7}$$

with softmax normalization across streams to obtain attention maps $A_L, A_G \in \mathbb{R}^{B\times 1\times H\times W}$:

$$[A_L, A_G] = \text{Softmax}_{\text{streams}}([R_L, R_G]). \tag{8}$$

The fused output is

$$F_{\text{PSE}} = A_L \odot Y_L + A_G \odot Y_G, \tag{9}$$

optionally followed by a pointwise projection and residual addition.

3.5 Imbalance- and Overlap-Aware Objective

Let $z \in \mathbb{R}^{B\times D}$ denote pooled embeddings and $\ell \in \mathbb{R}^{B\times 5}$ the logits. The softmax probability for sample i and class c is $p_{i,c} = \text{Softmax}(\ell_i)_c$, and $y_i \in \{0, \ldots, 4\}$ denotes the ground-truth label. The objective combines a class-balanced focal classification term with a prototype-based NCA regularizer.

3.5.1 Class-Balanced Focal Term

Rare classes are re-weighted using effective-number-derived weights $\alpha_c > 0$, and hard examples are emphasized by the focusing parameter $\gamma_{\text{foc}} \geq 0$. The per-sample term is

$$\mathcal{L}_{\text{cb-focal}} \;=\; -\frac{1}{B}\sum_{i=1}^{B} \alpha_{y_i} \left(1 - p_{i,y_i}\right)^{\gamma_{\text{foc}}} \log p_{i,y_i}. \tag{10}$$

3.5.2 PROTOTYPE NCA REGULARIZER

To reduce inter-class embedding overlap, unit-norm class prototypes $\{\mu_c \in \mathbb{R}^D\}_{c=0}^{4}$ are maintained via an exponential moving average (EMA). Let $\hat{z}_i = \frac{z_i}{\|z_i\|_2}$ denote the normalized embedding and N_c^{batch} the number of class-c samples in the mini-batch. The EMA update with momentum $m \in [0, 1)$ is

$$\mu_c \leftarrow \frac{m\,\mu_c + (1-m)\,\frac{1}{\max(1, N_c^{\text{batch}})}\sum_{i:\,y_i=c}\hat{z}_i}{\left\| m\,\mu_c + (1-m)\,\frac{1}{\max(1, N_c^{\text{batch}})}\sum_{i:\,y_i=c}\hat{z}_i \right\|_2}. \tag{11}$$

With cosine similarities $s_{i,c} = \hat{z}_i^{\top}\mu_c$ and temperature $\tau > 0$, the NCA-style loss is

$$\mathcal{L}_{\text{proto}} \;=\; -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s_{i,y_i}/\tau\right)}{\sum_{c\neq y_i}\exp\left(s_{i,c}/\tau\right)}. \tag{12}$$

3.5.3 TOTAL TRAINING OBJECTIVE

The overall objective combines the classification and prototype terms:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{cb-focal}} \;+\; \lambda_{\text{proto}}\,\mathcal{L}_{\text{proto}}, \tag{13}$$

with trade-off $\lambda_{\text{proto}} > 0$. During training with lesion-size–aware sampling, both terms are averaged over the full image and the lesion-focused crops per sample as described above.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

4.1.1 DATASETS AND EVALUATION PROTOCOLS

We evaluate on the APTOS 2019 Blindness Detection dataset of color fundus photographs annotated with five ordinal diabetic retinopathy (DR) grades (0–4). All experiments use a stratified split of the official training set into 80% training and 20% validation to preserve class proportions. Images are resized to 512×512 pixels. During training, we use color jitter (brightness/contrast/saturation factors 0.2; hue 0.02), random horizontal/vertical flips, and normalization using ImageNet channel means and standard deviations; validation uses only resizing and normalization.

4.1.2 BASELINES

For comparison, we include published validation results on APTOS 2019 for widely used convolutional and transformer architectures. These external baselines are VGG-16, ResNet-50, Inception V3, and MobileNet, which span lightweight and deeper convolutional models reported in prior work on this benchmark. All baseline networks are trained under the same experimental protocol to ensure a fair comparison. The models are optimized using the same training schedule, data preprocessing pipeline, and evaluation criteria. The best-performing checkpoint for each method is selected based on validation performance. Following prior work in medical and ordinal classification tasks, we report multiple evaluation metrics including Quadratic Weighted Kappa (QWK), accuracy (ACC), precision, recall, F1-score, specificity (SPEC), and the area under the ROC curve (AUC). Unless otherwise specified, metrics are computed on the held-out validation split. To reduce the risk of spurious epoch-wise peaks, we load both the best score across epochs and the final score at the last epoch.

4.1.3 MODEL AND TRAINING CONFIGURATION

Our model, DC-SSNet, uses a convolutional patch embedding (7×7 stride 2 with batch normalization and GELU, followed by a 3×3 block), three stages of dual-channel state-space blocks, and a linear classification head. Each DCSS block consists of a grouped convolution with channel shuffle (4 groups), a selective two-dimensional state-space mixing module with depthwise and pointwise projections, gated mixing with learnable channel-wise parameters scaled exponentially at stage-specific

Table 1: Comparison with baseline CNN architectures. The best result for each metric is shown in bold.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SPEC ↑	AUC ↑
VGG16	0.738	0.673	0.667	0.673	0.665	0.918	0.860
MobileNetV3	0.622	0.587	0.576	0.587	0.577	0.897	0.826
InceptionV3	0.731	0.673	0.665	0.673	0.667	0.918	**0.900**
ResNet50	**0.777**	**0.680**	0.679	**0.680**	**0.667**	**0.920**	0.897
Ours	0.727	0.667	**0.720**	0.667	0.645	0.917	0.870

dilations $\Delta \in \{0.8, 1.0, 1.2\}$, and residual connections with batch normalization. Progressive sensitivity encoding (PSE) is enabled by default. A local stream (kernel 3, $\Delta_\ell = 0.5$) and a global stream (kernel 5, $\Delta_g = 2.0$) are fused using channel- and spatial-weighting with a learnable balance parameter. An optional content-adaptive augmentation module (CADAM) computes a Sobel gradient magnitude map on grayscale feature activations and modulates features as $\mathrm{feat} \cdot (1 + \lambda_{\mathrm{CADAM}} M_{\mathrm{aug}})$ with $\lambda_{\mathrm{CADAM}} = 0.5$ and $M_{\mathrm{aug}} \in [0, 1]$; this module is toggled in ablations. Stochastic depth with drop-path rate 0.05 is applied within selective blocks.

Training uses AdamW with learning rate 3×10^{-4}, weight decay 0.05, and parameter grouping to exclude biases and normalization parameters from weight decay. We train for 100 epochs with automatic mixed precision, gradient clipping at a global norm of 5.0, and batch size 8 for the main comparison and 4 for ablations. The loss combines class-balanced focal loss (effective-number weighting with $\beta = 0.9999$) and a prototype NCA regularizer ($\lambda_{\mathrm{proto}} = 0.1$ unless ablated). Prototypes are kept as exponential moving averages of per-class feature means and are used within the NCA-style objective to reduce inter-class embedding overlap. In our implementation, when computed on detached features, the prototype regularizer does not backpropagate into the backbone and serves as a regularizer for the prototype space.

4.2 Main Performance Comparison

Table 1 reports the quantitative comparison between the proposed DC-SSNet and the baseline models. Among all compared methods, ResNet50 achieves the best overall classification accuracy (68.0%) and the highest QWK score (0.777), suggesting that deeper residual architectures remain competitive for this task. However, our proposed approach demonstrates strong performance across several complementary metrics. Specifically, our method achieves an AUC of 0.870, outperforming VGG16 (0.860) and MobileNetV3 (0.826), while remaining competitive with ResNet50 (0.897). In terms of precision, our model reaches 0.720, which is the highest among all evaluated methods, indicating that the proposed approach produces more reliable positive predictions. Compared with lightweight architectures such as MobileNetV3, our approach improves the QWK score by +0.105 and accuracy by +8.0 %, demonstrating substantially stronger ordinal prediction consistency. Furthermore, our method maintains high specificity (0.917), comparable to other deep CNN baselines. Although ResNet50 slightly outperforms our method in terms of accuracy and QWK, the proposed model provides a more balanced trade-off across precision, specificity, and AUC. This suggests that the proposed design improves prediction reliability while maintaining competitive overall classification performance.

4.3 Training Dynamics and Learning Curves

Figure 2 illustrates the training dynamics over 100 epochs, where evaluation metrics are plotted using results sampled every five epochs for clarity. As training progresses, the optimization exhibits a steady decrease in the training loss, dropping from 5.03 at the first epoch to 1.99 at epoch 100, indicating stable optimization and effective parameter updates. Correspondingly, performance metrics show a consistent upward trend despite moderate fluctuations during the early training phase. In particular, the quadratic weighted kappa (QWK) improves substantially from 0.30 to 0.80, reflecting progressively better alignment with the ordinal grading structure. Similarly, accuracy increases from 34.7% to 64.7%, while F1-score rises from 0.29 to 0.64, suggesting balanced improvements in both precision and recall.

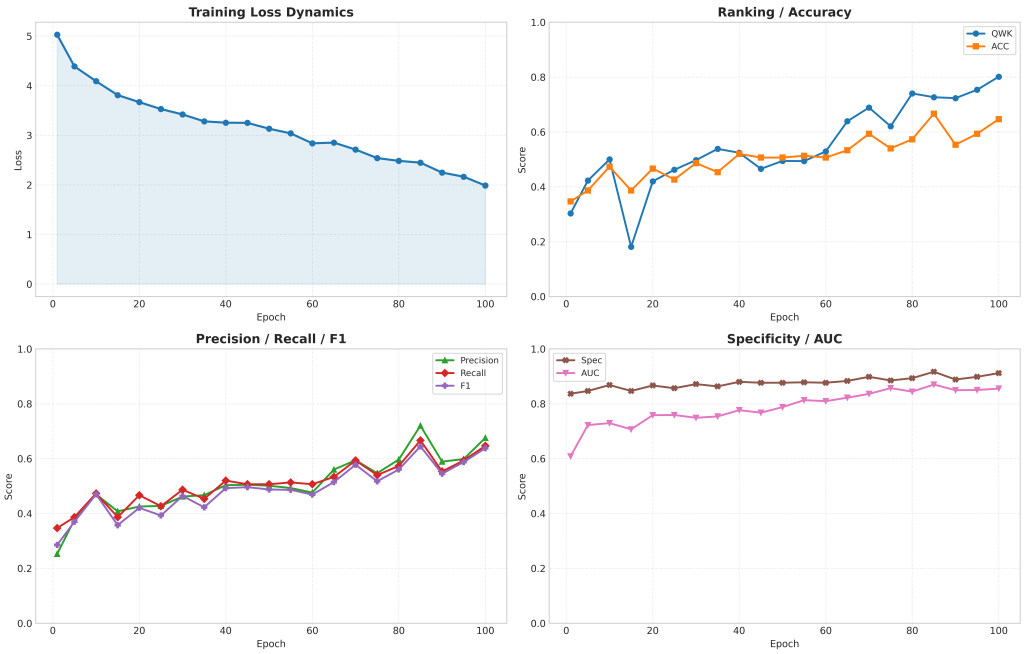

Figure 2: Learning curves on APTOS 2019.

Table 2: Ablation study of DC-SSNet. Each variant removes a specific component from the full model.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	SPEC ↑	AUC ↑
w/o CADAM	0.639	0.600	0.590	0.600	0.581	0.900	0.843
w/o PSE	**0.757**	0.653	0.641	0.653	0.643	0.913	**0.882**
w/o Overlap	0.610	0.580	0.630	0.580	0.557	0.895	0.831
DC-SSNet (Ours)	0.727	**0.667**	**0.720**	**0.667**	**0.645**	**0.917**	0.870

Notably, most metrics begin to stabilize after approximately 60 epochs, where the model enters a relatively stable convergence regime. During this stage, performance continues to improve gradually, with AUC increasing from 0.61 in the early stage to 0.86 by the end of training, and specificity consistently remaining above 0.88 in the later epochs. The overall trend demonstrates that the proposed training strategy yields stable convergence while progressively enhancing both discriminative ability and ordinal consistency. These observations confirm the robustness of the optimization process and the effectiveness of the proposed framework in learning reliable representations for diabetic retinopathy grading.

4.4 Ablation Studies

To analyze the contribution of each component in DC-SSNet, we conduct ablation experiments by removing individual modules while keeping the rest of the architecture and training configuration unchanged. The evaluated components correspond to the key design elements introduced in the method section, including the content-adaptive emphasis mechanism (CADAM), the patch structure enhancement module (PSE), and the overlapping patch embedding strategy. All variants are trained under the same settings, and the best-performing checkpoints are used for evaluation.

4.4.1 Effect of CADAM

We first evaluate the effect of removing the content-adaptive emphasis mechanism (CADAM). As shown in Table 2, removing CADAM leads to a clear degradation across all major metrics. The QWK score drops from 0.727 to 0.639, while classification accuracy decreases from 66.7% to

60.0%. Similarly, the AUC decreases from 0.870 to 0.843, and the F1 score falls from 0.645 to 0.581. These results indicate that CADAM plays an important role in highlighting lesion-prone regions and guiding the network toward clinically meaningful cues, which improves ordinal consistency and classification reliability.

4.4.2 Effect of PSE

Next, we remove the patch structure enhancement (PSE) module to assess its impact. Without PSE, the model achieves a QWK score of 0.757 and an AUC of 0.882, while the overall accuracy slightly decreases to 65.3% compared with 66.7% for the full model. The precision and F1 score also decline to 0.641 and 0.643, respectively. These observations suggest that PSE contributes to more stable feature representation by strengthening local structural information, which benefits overall classification consistency.

4.4.3 Effect of the Overlap Penalty

Finally, we evaluate the overlapping patch embedding strategy. When the overlap mechanism is removed, the model performance drops substantially. The QWK score decreases to 0.610, and accuracy declines to 58.0%, representing the largest degradation among all ablation settings. The AUC also drops from 0.870 to 0.831, while the F1 score decreases to 0.557. This confirms that overlapping patches are critical for preserving spatial continuity and capturing fine-grained lesion patterns across patch boundaries.

Overall, these ablation results demonstrate that each component contributes to the effectiveness of DC-SSNet. In particular, the content-adaptive emphasis mechanism and overlapping patch embedding play key roles in improving ordinal prediction quality and maintaining sensitivity to subtle retinal lesions.

5 Conclusion

Grading diabetic retinopathy from single-field fundus photographs is ordinal and class-imbalanced, demanding sensitivity to tiny lesions while preserving global context under practical compute. We introduce DC-SSNet, a high-resolution framework that couples content-adaptive emphasis with linear-time state-space encoding to fuse local detail and global context. Training addresses imbalance and label overlap via class-balanced focal loss, prototype regularization, and lesion-size-aware sampling. On a public benchmark, DC-SSNet attains competitive ordinal agreement with modest computation. Ablations reveal that edge-centric emphasis and local-global fusion can misallocate attention and that an overlap penalty does not sharpen boundaries, highlighting confusion between adjacent grades. These findings support interpretable, compute-efficient retinal models while underscoring sensitivity to cohort shift and fairness. We will replace edge-centric emphasis with lesion-aware stage- and channel-gated modulation, tie prototypes to gradients, and validate at higher resolution across multi-center cohorts to improve separation of adjacent grades.

6 Ethics Statement

This research study used only the publicly data made accessible in open access by the APTOS 2019 Blindness Detection competition Aravind Eye Hospital and PG Institute of Ophthalmology (2019) on Kaggle, sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required, as confirmed by the license attached to the open access data.

References

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. `https://www.kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle Competition.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1567–1578, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 8086–8099, 2022.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016. doi: 10.1001/jama.2016.17216.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017b.

Yutong Liu, Zengqiang Chen, Haoning Xu, et al. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.00939*, 2024.

Ze Liu, Yutong Lin, Yue Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Jun Ma, Lequan Yu, and Jing Wang. U-mamba: Enhancing u-net with state space model for medical image segmentation. *arXiv preprint arXiv:2401.06954*, 2024.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, 2019.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017. doi: 10.1001/jama.2017.18152.

Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 10347–10357, 2021.

Wenhai Wang, Enze Xie, Xiang Li, et al. Pyramid vision transformer: A versatile backbone for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568–578, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9759–9768, 2020.

Qi Zhu, Zicheng Li, Wengang Zhou, et al. Vision mamba: Efficient visual representation learning with state space model. *arXiv preprint arXiv:2401.04081*, 2024.

Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pp. 474–485. Academic Press, 1994.