RetinAI Report

Maheeka Pandita
UF BME
Gainesville, FL

***Abstract*—In this paper we will the implementation of the RetinAI project. This is a project that utilizes two Diabetic Retinopathy datasets along with Vision Language Models (VLMs) to create a screening tool. By tuning a LLaVA model on these datasets, the final project will produce classification and explanations for these medical images.**

***Index Terms*—Diabetic Retinopathy, Vision Language Models, LLaVA, Medical Imaging, Deep Learning, Transfer Learning**

I. Problem Summary

Diabetic Retinopathy is a leading cause behind preventable blindness. This is a significant challenge for public health, and one of the main factors is the very subtle early onset indicators. This is a major issue in areas where healthcare is limited or ares without proper medical facilities. This can lead to delayed diagnosis, resulting in unrepairable eye damage.

To solve this issue, the goal of this project is to use VLMs, specifically the use of LLaVA, which is specifically tailored for medical image datasets. This method will allow for medical professionals to gain a fast and secondary opinion for their initial classification, providing an additional line of defense against preventable blindness.

A. Diabetic Retinopathy Impact

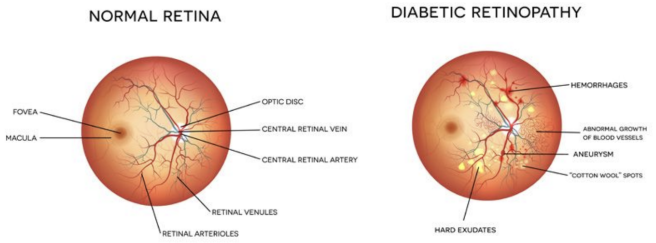


Fig. 1. Diabetic Retinopathy Image [11]

Recent epidemiological research shows that Diabetic Retinopathy is not just a medical issue but a significant socioeconomic crisis. It is currently the leading cause of blindness among working age adults (aged 20–65) globally, a demographic critical to economic stability [6,7].

1) The Burden on the Working Class: Unlike other ocular diseases such as cataracts or macular degeneration which primarily affect the elderly generation, DR disproportionately impacts economically active individuals.

- Employment issues: Vision impairment in this age group directly affects employment opportunities and productivity, increasing the risk of poverty for affected families. In 2021 alone, an estimated 1.33 million people were blind due to DR, with the heaviest burden concentrated in low and middle income regions such as South Asia and Latin America [6].
- Healthcare Disparity: Research indicates a screening blind spot for this demographic. While prevention programs often target children or the elderly, this age group often lack prioritized access to screening until symptoms become severe and irreversible [7].

2) Human Cost/Quality of Life: Beyond the economic metrics, more importantly, the human impact of DR is immense. The condition degrades the functional vision required for aspects of daily life.

- Daily Living: As retinal damage progresses, patients struggle with essential tasks such as reading, driving, and recognizing faces. This leads to a loss of autonomy in their life [8].
- Mental Health: The correlation between DR severity and psychological distress is well documented, as many patients with advanced DR exhibit significantly higher rates of anxiety and social isolation, often as a result of fear of total blindness and the inability to engage in social interactions [9].

By automating the screening process, RetinAI directly addresses these disparities. It provides a solution for settings where the majority of these cases go undiagnosed, potentially preventing up to 95% of vision loss cases through timely detection [10].

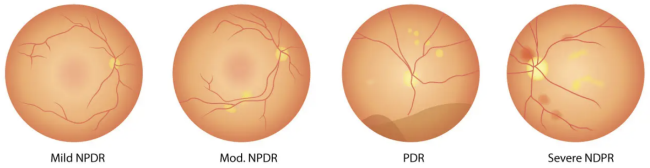


Fig. 2. Grading Scale [12]

II. Image Acquisition

The datasets utilized in this project (APTOS and IDRiD) consist of images captured via digital fundus photography. This is a specialized form of medical imaging that uses a low power microscope attached to a camera to photograph the interior surface of the eye, including the retina, optic disc, and macula [13].

A. Mechanism and Protocol

The acquisition process works on the method of ophthalmoscopy. A high intensity flash illuminates the retina through the pupil, and the reflected light is captured by the sensor.

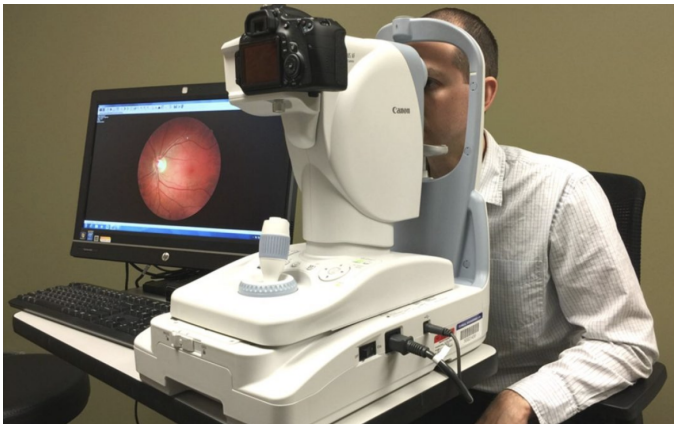


Fig. 3. Digital Fundus Photography [15]

- Mydriasis: To obtain high quality images like those in the IDRiD dataset, patients often undergo mydriasis (also called pupil dilation) using eye drops. This maximizes the field of view (typically 30° to 50°) and allows for clearer visualization of peripheral lesions [14].
- Non Mydriatic Screening: In mass screening contexts (such as in the APTOS dataset), non mydriatic cameras are often used. These rely on infrared focusing systems to capture images without dilation. While more accessible, this method frequently results in darker images or artifacts if the patient blinks or moves, contributing to the noise that is more present in the APTOS dataset.

B. Impact on Model Training

The hardware used for acquisition introduces significant variance known as domain shift. Different camera manufacturers use different sensor types and color processing algorithms. As noted in the Ethical Concerns section, this hardware bias can affect model performance. For instance, the IDRiD dataset was captured using high res clinical cameras, while APTOS contains images from diverse sources, resulting in the variations in lighting and aspect ratio that our must be normalized in the preprocessing pipeline.

III. Related Work

Recent advancements in automated Diabetic Retinopathy detection have largely moved from traditional machine learning to deep learning, and more recently, to Vision Language Models (VLMs). This section compares the approach for RetinAI to this evolving landscape.

A. Traditional Deep Learning Approaches

The standard for DR detection has historically been dominated by Convolutional Neural Networks (CNNs). Works such as identifying DR stages using ResNet and VGG architectures have achieved high classification accuracy by treating the problem as a multi class categorization task [1].

At the same time, recent efforts have employed Vision Transformers and Federated Learning to address privacy concerns while improving the feature extraction capabilities [2].

While these models achieve strong performance in classification metrics, they only provide a diagnostic grade, but lack the capability to generate explanations or describe specific lesions that lead to diagnosis.

RetinAI differs significantly by utilizing a VLM architecture, which allows for the model to output not just a classification label, but also a reasoning explanation, which allows interpretability for healthcare providers and patients.

B. Vision Language Models in Medicine

The use of VLMs has enabled models to process medical imagery alongside textual instructions and descriptions. A primary baseline in this domain is LLaVA-Med [3], which built off of the preexisting LLaVA architecture to the medical field using a curriculum learning approach on large scale biomedical image caption pairs.

While LLaVA-Med represents the baseline for general biomedical visual question/answering, it does require massive computational resources and extensive pretraining on bio data.

In contrast, this project demonstrates a more accessible approach by employing 4-bit quantization and LoRA on the standard LLaVA-1.5 model. As a result, this project achieves similar performance without the need for the massive resource load used in LLaVA-Med.

C. Explainable DR Diagnosis

Recently, works such as RetinalGPT [4] and XDR-LVLM [5] have started to explore the specific application of VLMs to ophthalmology. These both are designed to provide diagnostic reports and have set high benchmarks.

RetinAI advances this through its unique training methodology. Unlike standard transfer learning which typically moves from large generic datasets to small specific

ones, this project implements a two step process.

First, the model is trained on the smaller, high quality IDRiD dataset before generalizing to the larger, noisier APTOS dataset. This allows the model to learn fine grained lesion features in a controlled environment before applying that knowledge to large scale screening tasks, a strategy distinct from the direct fine tuning methods in RetinalGPT [4] and XDR-LVLM [5].

IV. Datasets

This project utilizes two datasets:

1) APTOS 2019 Blindness Detection
2) IDRiD (Indian Diabetic Retinopathy Dataset)

A. APTOS

This dataset is a large scale dataset of retinal fundus images via a prior Kaggle competition.

Dataset Info:

- About 10 GB of images and masks
- 3,662 training images
- 1,928 testing images
- Diagnosis scale from 0-4
- Images in .png format
- Labels in a .csv file, with corresponding image ID

This dataset will be utilized for training the final classification model. The diversity in the data and the larger size are great to help with generalization on unseen images, which is essential for this project.

This dataset will also be used for benchmarking, and will be further fine tuned to this task via additional training from the IDRiD dataset.

B. IDRiD

This is a smaller dataset that contains more detailed fundus images from a Grand Challenge via IEEE.

Dataset Info:

- 413 training images
- 103 testing images
- Images in .jpg format
- Diagnosis scale from 0-4
- Pixel level segmentation masks for specific retinal lesions

The purpose of this dataset is to cut down on the training time for the final dataset. It will allow for faster and more efficient initial training of the model. This will allow for faster and easier initial prototyping, before working on the larger dataset.

Also due to the pixel level segmentation masks that are in this dataset, it will allow for the VLM to generate feature based explanations in the predictions and summary that are specific to certain lesions such as hemorrhages.

C. Access Methods

The APTOS dataset was downloaded via Kaggle, by registering for a competition. The dataset was downloaded using the Kaggle API or via this link:
https://www.kaggle.com/competitions/aptos2019-blindness-detection.

The IDRiD dataset was downloaded via the IEEE Dataport. This can be directly downloaded via the IEEE webpage after creating an account via this link:
https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid.

D. Preprocessing

The main issue for preprocessing of these two datasets, is the large amount of data with a diagnosis of Grade 0, i.e. no Diabetic Retinopathy. As a result, the classification model can become biased, learning the classification of non Retinopathy very well, but performing poorly in those cases that have very small features that are positive for Diabetic Retinopathy, which are of the utmost clinical importance.

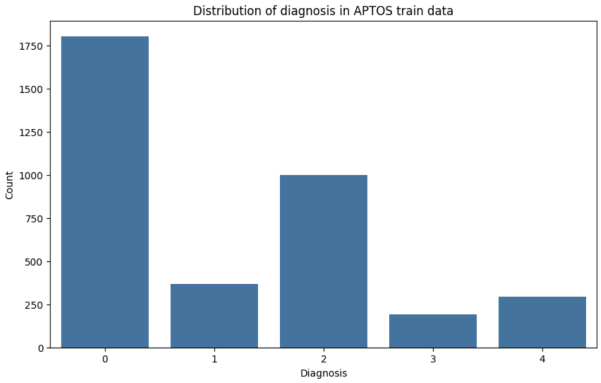


Fig. 4. Distribution of Diagnosis for APTOS

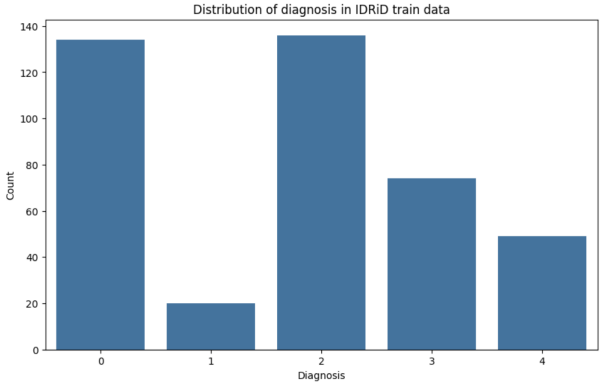


Fig. 5. Distribution of Diagnosis for IDRiD

There are a few methods that can be used to overcome this issue

1) Weighted loss functions to penalize incorrect classifications more harshly
2) SMOTE, syntheticaly creating new images of minority diagnosis grades
3) Data augmentation by creating different versions of preexisting images via rotations or mirroring

Another potential issue would be the variation in the images that are in the datasets.

Certain images are of different quality or lighting, and between the two datasets, the images are not necessarily the same size. This can cause the model to learn features that are not directly applicable to the actual classification, causing an incorrect classification.

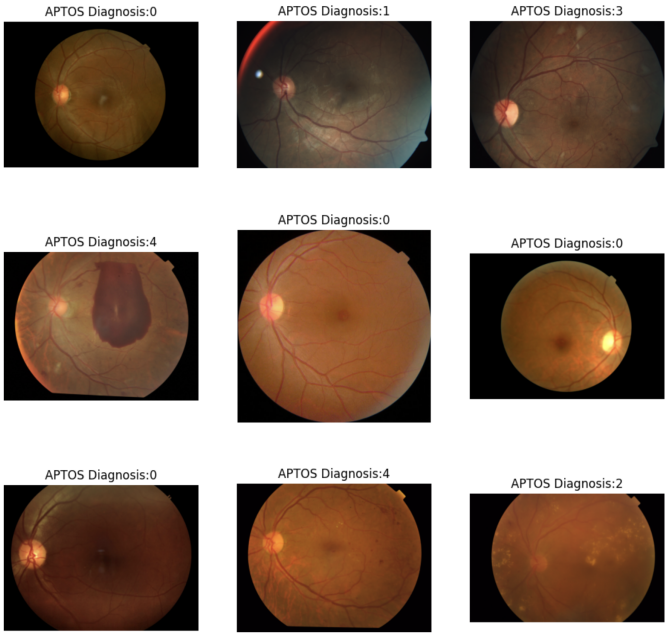


Fig. 6. Image Differences from APTOS Dataset

This can be solved by creating a preprocessing pipeline that does the following:

1) Resizes the images to a uniform dimension
2) Normalize pixel range
3) Potential data augmentation if necessary

E. Current Work

At this point, the current implementation has mainly concerned the training of LLaVA on the IDRID dataset. This has been a challenge mainly due to resources. Implementation was attempted using a M2 Max Macbook Pro, which upon initial research should have been decent for this initial training. However, upon testing system memory became a critical issue.

To solve this use of Google Collab was attempted and that was still having issues. The solution was to use Hipergator and SLURM requests, which proved to be a great solution.

The initial training of the model was done following a project from another course (BME6938), as personal experience with training LLaVA was extremely limited. Using this framework and basic code architecture, a model was created that has average performance. However this is not an issue as this will be used to then train on the APTOS dataset. The purpose of the IDRID dataset training is purely to simplify that much larger task, and to pretrain the LLaVA model on this type of data.

A point to note, is the original proposition of using LLaVA-Med, however this very quickly became something out of the scope of my level of expertise, as just getting access to the model was not possible with my experience and timeframe.

The current state of the project is that the APTOS training has been implemented. This is a more basic model, primarely due to the large size of this dataset, and the difficulty in acquiring resources to run the codes, as well as the time it takes to run. The goal of this dataset is to be an improvement upon the IDRID dataset due to the increased number of images that are used to train. At the same time it is much simpler to train as a lot of the initial training has been completed via the IDRID dataset, simplifying the work and tuning for this model.

F. Frameworks and Libraries

The implementation of this project centers on fine tuning a pre trained VLM .

The project relies on a conda environment, defined in the environment.yml file, to ensure reproducibility. Key libraries include:

- Core Architecture: PyTorch and the Hugging Face transformers library are the core frameworks for loading, fine-tuning, and running the VLM.
- Performance & Memory: Accelerate and bitsandbytes are used to enable a more effcient 4-bit quantized training on NVIDIA GPUs, drastically reducing memory usage.
- VLM Architecture: The LLaVA library is included to provide the custom model code required to run the llava-hf/llava-1.5-7b-hf architecture.
- Data Handling: pandas is used for managing the CSV label files, while Pillow (PIL) and OpenCV are used for loading and processing images.
- Interface: Gradio is the selecte library for building the final interactive user interface.

Training Setup:

- Hardware: All of the training will be done on Hipergator via the A100 GPU.
- Model: The model to be trained is llava-hf/llava-1.5-7b-hf. This is a powerful VLM that will be fine tuned for the Retinopathy imaging task.

- Two Phase Training Strategy: To prevent overfitting on the small IDRiD dataset, a two phase strategy will be used:
  1) The model will first be tuned on the small, high quality IDRiD dataset (413 images) for only one epoch. The goal of this phase is not to achieve high accuracy, but to teach the model the basic vocabulary and visual features for Diabetic Retinopathy.
  2) The model saved from Phase 1 will then be used as the starting point for a second round of tuning on the larger APTOS dataset for 3-5 epochs, depending on performance. This step will allow the model to generalize its knowledge and will be the primary source of its final accuracy.
- Key Hyperparameters (IDRID):
  - model_id: llava-hf/llava-1.5-7b-hf
  - quantization: 4-bit (load_in_4bit=True)
  - precision: bf16=True (optimal for A100 GPUs)
  - per_device_train_batch_size: 1
  - gradient_accumulation_steps: 16 (to simulate an effective batch size of 16)
  - optimizer: adamw_8bit
  - learning_rate: 2e-5
  - gradient_checkpointing: True (to further conserve GPU memory)
- Key Hyperparameters (APTOS initial model):
  - Quantization: 4-bit
  - Compute Dtype: torch.float16
  - LoRA Rank (r): 16
  - LoRA Alpha: 32
  - LoRA Dropout: 0.05
  - LoRA Target Modules: [”q_proj”, ”k_proj”, ”v_proj”, ”o_proj”, ”gate_proj”, ”up_proj”, ”down_proj”]
  - LoRA Bias: ”none”
  - Batch Size (Per Device): 1
  - Gradient Accumulation Steps: 16
  - Number of Epochs: 5
  - Learning Rate: 2e-4
  - FP16 Training: True
  - Gradient Checkpointing: True
- Key Hyperparameters (APTOS final model):
  -
  - Quantization: 4-bit
  - Compute Dtype: torch.float16
  - LoRA Rank (r): 64
  - LoRA Alpha: 128
  - LoRA Dropout: 0.05
  - LoRA Target Modules: [”q_proj”, ”k_proj”, ”v_proj”, ”o_proj”, ”gate_proj”, ”up_proj”, ”down_proj”]
  - LoRA Bias: ”none”
  - Batch Size (Per Device): 1
  - Gradient Accumulation Steps: 16
  - Number of Epochs: 6
  - Learning Rate: 2e-4
  - FP16 Training: True
  - Gradient Checkpointing: True

Modules for Reproducibility: The project is structured to be fully reproducible by using the following key components:

- environment.yml: Conda environment file that has all required packages and their exact versions, ensuring the software environment can be perfectly replicated.
- src/idrid_setup.py: These scripts convert the raw downloaded datasets and CSV files into the .jsonl format required by the training script.
- src/train_idrid.py and src/test_idrid.py: These scripts contain training and testing processes including all hyperparameters and settings for the tuning process.
- Slurm Folder: The SLURM submission scripts allow for others to determine the resource needs for replication of this project.

G. Ethical Concerns

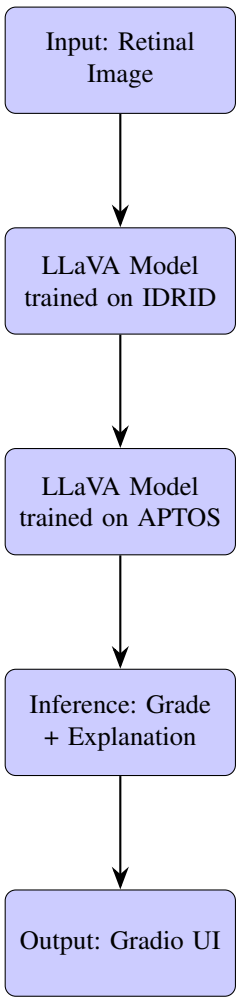
As this project is related to the field of medicine, there is concern about patient confidentiality and ethicality. All data sets for this project were supplied by legitimate competitions, one via Kaggle and the other from IEEE.

All images in these datasets are intended for public research, and any identifiable information is removed (names, clinics, etc).

In terms of ethical bias, one of the most significant challenges is that the IDRiD dataset may not be representative of all ethnic groups because it is a dataset from Indian patients. Therefore, it may not perform as well on certain populations due to factors such as eye color or cameras that were used to take these images.

Finally, the use of this tool is not a replacement for a medical opinion. It is purely a support system or a second opinion. The project may not correctly predict a valid diagnosis, and can cause false positives, leading to patient stress and costs. Or even worse, it could cause false negatives, and if treated as a primary diagnostic tool, can cause missed treatment.

V. Architecture



This project is based around the LLaVA VLM. This hybrid model uses a vision encoder such as CLIP with a LLM allowing for processing of both images and text. The Med portion of LLaVA means that it is trained on medical images and performs better on analysis of these images.

The implementation will use PyTorch for the training and will utilize the Hugging Face library for the model tuning.

The final interface will be created via Gradio. In this interface the user, in this case a healthcare provider, can input a retinal image. The image will then be processed and the model will output a predicted grade for the diagnosis, along with a short description of the reasoning for that diagnosis such as presence of certain lesions.

VI. Gradio Interface

The interface will be created to be as simple and interpretable as possible. The steps for use are as follows:

1) User uploads a single retinal images to the interface
2) The user will click upload, and the interface will preform the analysis
3) Once the analysis is complete, the interface will display two items:
    - Predicted diagnostic grade (eg: Predicted Grade - 2 (moderate))
    - A generated explanation of why that grade was given via features in the image

This implementation is simple to use, and does not have sharp learning curve, it is the same as uploading an image to any other software. The UI clearly displays the results, and is simple to understand and follow for both the healthcare provider and for any patient.

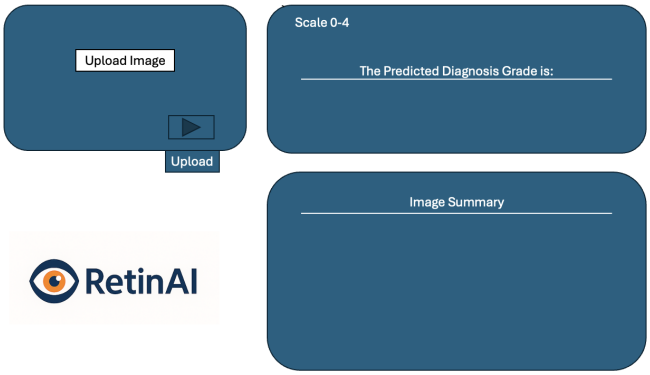


Fig. 7. Sample Gradio Interface

The current stage of the implementation is as follows. There is a Gradio interface that uses the current IDRID trained LLaVA model. The interface consists of multiple sections. The user will upload an image and the model will classify that image and provide some information based off of the grade. If the image is not visible or there is no detection, the output will ask the user to upload a different image.

1) Description of grades and meanings
2) Image upload section
3) Instructions section
4) Predicted grade
5) Recommendation based off of grade
6) Output of model and description of image

The current interface is in basic design stage, it will be improved visually later with the logo and other features at a later date.

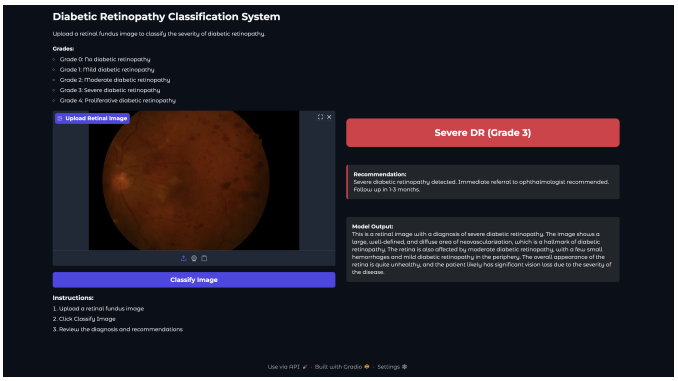


Fig. 8. Initial Gradio Interface

The most current interface has kept a variation of the initial UI, as it was simple to use and understand. A few tweaks have been made, such as the inclusion of the RetinAI logo, and some color adjustments to better represent the logo.

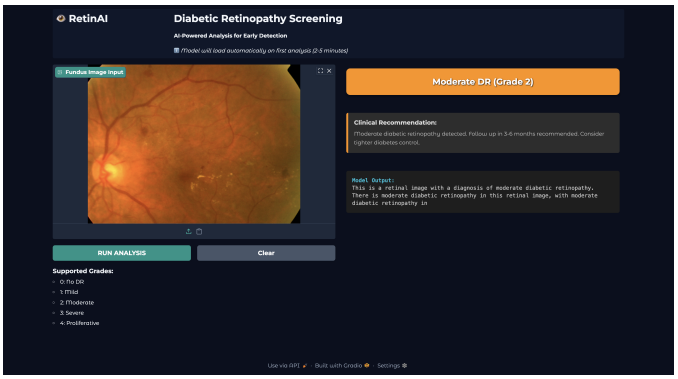


Fig. 9. Most Recent Interface

VII. HCI Considerations

The UI design for RetinAI has a main focus on usability, interpretability, and safety. The interface is designed to reduce the cognitive load while ensuring appropriate trust in the AI's second opinion.

A. Usability and Workflow Efficiency

Medical professionals often operate in high pressure environments where software complexity can lead to small, but critical errors. To address this, RetinAI adopts a simple and easy to follow interface:

- Linear Interaction: The system enforces a sequential process (Upload → Analyze → Review). This prevents user confusion and ensures that all necessary inputs are present before processing.
- Low Learning Curve: By mimicking standard file upload interfaces, the design minimizes the need for specialized training, anyone who has used a computer to upload a file can use this system.
- Immediate Feedback: The interface provides real time validation of inputs. If an uploaded image is non compliant or undetectable, the system halts execution and prompts the user for a new input.

B. Explainability and Trust Calibration

Trust in medical AI is often an issueas models provide a diagnosis without a proper explanation. RetinAI addresses this issue via the following:

- Text Justification: Unlike traditional CNNs that output only a numeric grade, RetinAI's VLM model generates a text based reasoning. This allows the clinician to verify the AI's logic against their own observations, transforming the tool from a replacement to a helpful assistant.
- Transparent Uncertainty: By explicitly stating that the output is a Predicted Grade rather than a definitive diagnosis, the interface encourages the clinician to question the result. This framing helps prevent bias where users over rely on the system, and blindly accept the output.

C. Error Management and Safety

Given the risks associated with false negatives in screening tools, the HCI design includes specific tools to prevent such risks:

- Human involvment: The interface is explicitly framed as a support system, and not a direct answer, enforcing the role of the human expert as the final decision maker. This aligns with the idea that AI is not a replacement to the clinician.
- Visual Hierarchy: Critical information such as the grade is separated from supplementary information (the explanation). This allows for rapid scanning while preserving access to deeper details when necessary.

VIII. Results

Currently the model is only trained on the IDRID dataset, and as a result the performance is not extremely high. However, this is important as we do not want the model to fully learn the IDRID dataset, it is purely just for initially training the LLaVA model on Diabetic Retinopathy data.

In the future, the model will be trained on the APTOS dataset, which is significanly larger than the IDRID dataset, and will have higher constraints for the performance. The current performance is as follows.

The initial fine-tuning on the 413 image IDRiD dataset gives a promising baseline accuracy of 82.52%. Critically, the model produced 103 out of 103 valid predictions, demonstrating that it successfully learned the classification task on this small dataset. This is a bit concerning as

TABLE I
BASELINE MODEL PERFORMANCE ON IDRID TEST SET

Class	precision	recall	f1-score	support
No DR	0.94	0.88	0.91	34
Mild DR	0.29	0.40	0.33	5
Moderate DR	0.84	0.84	0.84	32
Severe DR	0.79	0.79	0.79	19
Proliferative DR	0.85	0.85	0.85	13
accuracy			0.83	103
macro avg	0.74	0.75	0.74	103
weighted avg	0.83	0.83	0.83	103

overfitting is an issue that may be faced.

The classification report shows strong performance on the easiest No Diabetic Retinopathy and most severe cases. As expected, the model struggled most with the more subtle Mild DR, which is a topic that will hopefully be adressed via the APTOS dataset.

These results are excellent for a baseline model. They confirm that the model is learning the correct visual features of diabetic retinopathy and is not just simply overfitting. This successfully trained model now serves as an ideal starting point for the transfer learning on the larger and more diverse APTOS dataset.

TABLE II
INITIAL MODEL PERFORMANCE ON APTOS 2019 TEST SET

Class	Precision	Recall	F1-Score	Support
No DR	0.97	0.99	0.98	1805
Mild DR	0.59	0.52	0.55	370
Moderate DR	0.68	0.86	0.76	999
Severe DR	0.50	0.01	0.01	193
Proliferative DR	0.57	0.46	0.51	295
Accuracy			0.81	3662
Macro Avg	0.66	0.57	0.56	3662
Weighted Avg	0.80	0.81	0.79	3662

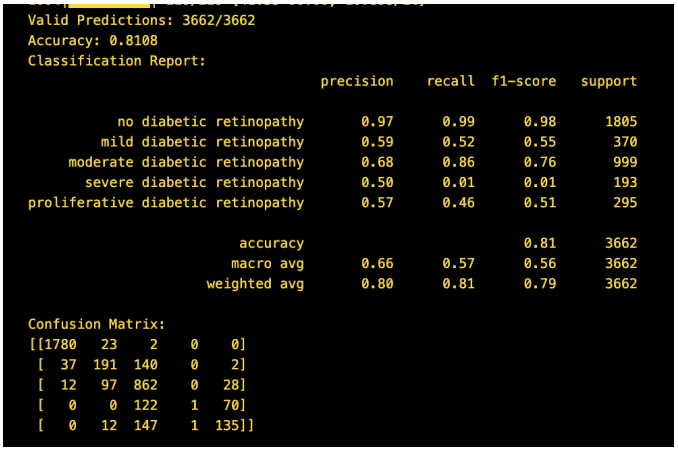


Fig. 10. Screenshot of Initial Metrics

The Phase 2 fine-tuning on the APTOS 2019 dataset yielded a functional model with strong screening capabilities, though specific challenges remain in distinguishing severe disease grades.

Updated tuning of the model has given this result

TABLE III
UPDATED MODEL PERFORMANCE ON APTOS 2019 TEST SET

Class	Precision	Recall	F1-Score	Support
No DR	0.97	0.99	0.98	1805
Mild DR	0.59	0.51	0.55	370
Moderate DR	0.67	0.83	0.74	999
Severe DR	0.62	0.48	0.54	193
Proliferative DR	0.54	0.42	0.47	295
Accuracy			0.82	3662
Macro Avg	0.68	0.65	0.66	3662
Weighted Avg	0.81	0.82	0.81	3662

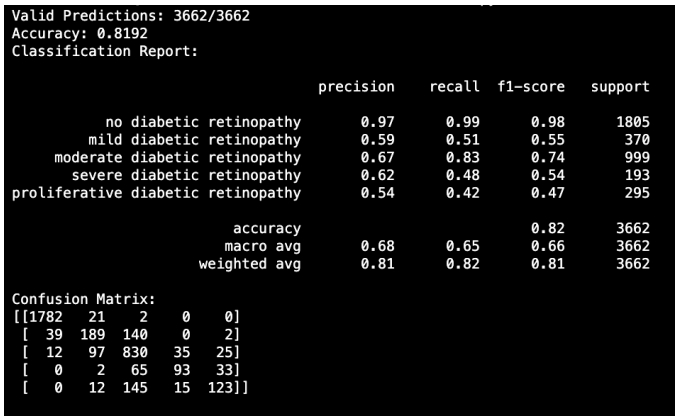


Fig. 11. Screenshot of Updated Metrics

As seen in the updated metrics, the main improvement was in the sever retinopathy. Grades of moderate and proliferative both had a slight decrease, but the major increase in performance for severe offsets this decrease.

A. Strengths and Successes

- 100% Valid Predictions: The model successfully generated a valid, parse able diagnosis string for every single image in the test set. This confirms that the teacher forcing prompt method effectively solved the earlier issue of empty or malformed outputs.
- High Overall Accuracy: The model achieved an overall accuracy of 81.92%, which is a strong baseline for a this medical classification task.
- Excellent Screening Capability: The model performed exceptionally well on the No DR category, achieving an F1 score of 0.98. With a recall of 0.99 for this class, the model is highly reliable at identifying healthy patients, which is the primary requirement for a screening tool.

Despite the high overall accuracy, the model exhibits specific behaviors regarding disease severity grading, particularly in the trade off between sensitivity and specificity for the intermediate classes.

- Improved Sensitivity for Severe DR (Class 3): Previous versions of the model struggled to detect Severe DR

(Recall ≈ 0.01), but the current adjustments to the tuning strategy has significantly corrected this imbalance. The model now achieves a recall of 0.48 for this class, correctly identifying 93 out of 193 cases. This improvement, however does comes with a slight trade off. The recall for adjacent classes (Moderate and Proliferative) decreased very slightly, as the model has become more strict in flagging potential severe cases rather than defaulting to the majority classes.

- Off by One Errors: The confusion matrix continues to show a pattern of off by one grade errors. For example, the majority of misclassified Severe DR cases were predicted as either Moderate or Proliferative, rather than No DR. This indicates that the model still struggles to delineate the precise fine grained boundaries that distinguish the severe stage from its immediate neighbors.

IX. Challenges

This is a unique project as it is using a VLM instead of a traditional classification model, which can only classify the images based on the grade scale. This project aims to continue from that, and also provide a description of why that diagnosis was given to that image.

An important challenge will be the high computational cost for this project. This will require the use of a powerful computer, such as Hipergator to do the training. The APTOS dataset training is currently going well, and is just requiring some tuning to improve the performance on certain classes of the grading.

Another issue is from the way VLMs work, and the fact that they can sometimes generate false text, that may not be true. This is an issue due to the nature of this project. As a medical project, ensuring that the model is outputting correct information is of the highest priority. If this information is going to be used in a medical setting, patients must be under the impression that the model is preforming correctly, and is actually giving correct diagnosis, in order to assist the healthcare provider. Luckly due to the amount of additional information in the IDRiD dataset regarding pixel level annotations, this will hopefully not be an issue, as the model can be trained specifically to look for these fine features.

X. Future Work

While RetinAI has demonstrated the viability of using general purpose VLMs for specialized medical screening, several additions remains for optimization and clinical translation.

A. Improved Data Augmentation

To further solve the class imbalance currently going on, future tuning will move beyond simple geometric augmentations (rotations/flips). The use of Generative Adversarial Networks (GANs) or diffusion models specifically trained on retinal fundus images to synthesize high quality examples of rare severity grades could provide the model with a larger variety of lesion configurations, helping to resolve the remaining confusion between Grade 3 and its neighbors.

B. Visual Grounding and Localization

Currently, the model provides a text based explanation. The next step is to implement Visual Grounding. By tuning the model to output bounding box coordinates alongside the diagnosis, the interface could potentially visually highlight the exact regions of interest on the patient's scan. This would significantly add to its utility, allowing clinicians to instantly verify the specific features in the prediction.

C. Clinical Validation Study

Finally, a technical benchmark on a test set is insufficient for medical safety. The ultimate phase of this project would involve a blinded reader study, where ophthalmologists grade a set of fundus images, both with and without the assistance of RetinAI. This would allow to quantitatively measure the tool's impact on decision making speed and accuracy in a with real clinical data.

XI. AI Reflection

This project is using a VLM, this is not a replacement for a human and cannot be the sole use in a diagnosis. The model is purely a check and second opinion for the user. The model is trained on limited data, and cannot always be generalized to all populations. As a result the model may not preform well on populations not highly represented in the IDRiD and APTOS datasets.

In terms of environmental concerns, training these large models is very energy consuming, and utilizes a lot of resources, however by using a smaller dataset to pretrain, this will hopefully be limited.

XII. Conclusion

This project successfully demonstrated the ability to fine tune large scale Vision Language Models for specialized medical screening.

By using a two step training methodology, transitioning from the high fidelity IDRiD dataset to the large scale APTOS dataset, this project achieved an accuracy of 82% and a critical improvement in detecting Severe Diabetic Retinopathy (Recall 0.48).

This current approach advances the field by moving beyond general grade classification methods to providing explainable reasoning. This allows for greater trust between the AI system and healthcare providers.

While challenges remain regarding class imbalance and computational requirements, RetinAI represents a significant step toward accessible and high level screening tools for preventing blindness in underserved populations.

The performance as well as the source of the data display that this system is not just a theoretical model, but a proper foundation for future clinical deployment that with proper tuning, testing, and clinical trials can potentially be applicable in health care settings.

XIII. Timeline

TABLE IV
Project Timeline and Expected Outcomes

Week	Focus	Expected Outcome
Oct 21 – Oct 28	Data Preprocessing & Baseline Setup	Working data loaders for both APTOS and IDRiD datasets. Completed EDA.
Oct 29 – Nov 5	Initial Model Training & UI Prototype	LLaVA fine tuning on the IDRiD dataset. Basic Gradio UI prototype is built.
Nov 6 – Nov 19	Full Training & Evaluation	Model fine tuned on the APTOS dataset. Initial performance metrics established.
Nov 20 – Nov 27	Model Tuning & UI Integration	Model performance is improved through hyperparameter tuning and refinement of text prompts. The best model is integrated into the Gradio UI for a functional prototype.
Nov 28 – Dec 1	Final Testing, Demo Prep & Report	testing of the application is complete. The final presentation and technical report are prepared and finalized.

REFERENCES

[1] Asima Shazia, Fida Hussain Dahri, A. Ali, M. Adnan, Asif Ali Laghari, and T. Nawaz, "Automated Early Diabetic Retinopathy Detection Using a Deep Hybrid Model," vol. 1, no. 1, pp. 71–83, Nov. 2024, doi: https://doi.org/10.62762/tetai.2024.305743 (accessed Dec. 05, 2025)

[2] M. Chetoui and M. A. Akhloufi, "Federated Learning for Diabetic Retinopathy Detection Using Vision Transformers," BioMedInformatics, vol. 3, no. 4, pp. 948–961, Nov. 2023, doi: https://doi.org/10.3390/biomedinformatics3040058. (accessed Dec. 05, 2025)

[3] C. Li et al., "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," arXiv.org, Jun. 01, 2023. https://arxiv.org/abs/2306.00890 (accessed Dec. 05, 2025)

[4] W. Zhu et al., "RetinalGPT: A Retinal Clinical Preference Conversational Assistant Powered by Large Vision-Language Models," arXiv.org, 2025. https://arxiv.org/abs/2503.03987 (accessed Dec. 05, 2025).

[5] M. Ito, K. Tanaka, K. Matsuda, and A. Nakayama, "XDR-LVLM: An Explainable Vision-Language Large Model for Diabetic Retinopathy Diagnosis," arXiv.org, 2025. https://arxiv.org/abs/2508.15168 (accessed Dec. 05, 2025).

[6] Y. Meng et al., "Global, Regional, and National Burden of Blindness due to Diabetic Retinopathy, 1990–2021," Ophthalmology and Therapy, Aug. 2025, doi: https://doi.org/10.1007/s40123-025-01230-y. (accessed Dec. 05, 2025).

[7] Y. Meng et al., "Global, Regional, and National Epidemiology of Vision Impairment due to Diabetic Retinopathy Among Working-Age Population, 1990–2021," Journal of Diabetes, vol. 17, no. 7, Jul. 2025, doi: https://doi.org/10.1111/1753-0407.70121. (accessed Dec. 05, 2025).

[8] "Senior Home Care: Diabetic Retinopathy and Daily Life for Seniors," Home Helpers® Home Care, Jul. 26, 2024. https://homehelpershomecare.com/sc-valley/community-blog/2023/december/how-does-diabetic-retinopathy-impact-daily-life/ (accessed Dec. 05, 2025).

[9] G. Saitakis, D. Roukas, E. Hatziagelaki, V. Efstathiou, P. Theodossiadis, and E. Rizos, "Evaluation of Quality of Life and Emotional Disturbances in Patients with Diabetic Retinopathy," European Journal of Investigation in Health, Psychology and Education, vol. 13, no. 11, pp. 2516–2528, Nov. 2023, doi: https://doi.org/10.3390/ejihpe13110175. (accessed Dec. 05, 2025).

[10] "Diabetic Retinopathy: A Call for Global Action - The International Agency for the Prevention of Blindness," The International Agency for the Prevention of Blindness, Jun. 12, 2024. https://www.iapb.org/learn/resources/diabetic-retinopathy-a-call-for-global-action/ (accessed Dec. 05, 2025).

[11] D. R. R. R. P. FRCS et al., "Understanding Diabetic Retinopathy and how to reverse it," Neoretina Blog, Dec. 11, 2018. https://neoretina.com/blog/diabetic-retinopathy-can-it-be-reversed/ (accessed Dec. 05, 2025).

[12] "The 4 Stages of Diabetic Retinopathy," Pacific Eye Care Doctors of Optometry, 2025. https://www.pacificeyecare.net/eyecare-services/eye-disease-management/diabetic-retinopathy/the-4-stages-of-diabetic-retinopathy/ (accessed Dec. 05, 2025).

[13] K. Cockerham, "Kimberly Cockerham, MD," Kimberly Cockerham, MD, Jan. 06, 2025. https://cockerhammd.com/posts/dr-cockerham-utilizes-new-retina-scanner-to-detect-retinal-diseases/ (accessed Dec. 06, 2025).

[14] N. Panwar et al., "Fundus Photography in the 21st Century—A Review of Recent Technological Advances and Their Implications for Worldwide Healthcare," Telemedicine and e-Health, vol. 22, no. 3, pp. 198–208, Mar. 2016, doi: https://doi.org/10.1089/tmj.2015.0068. (accessed Dec. 06, 2025

[15] V. S. Academy, "Digital Fundus Photography- A Tool for Diagnosis," Vision Science Academy, Jun. 30, 2020. https://visionscienceacademy.org/digital-fundus-photography-a-tool-for-diagnosis/ (accessed Dec. 05, 2025).