arXiv:submit/8168866 [cs.CV] 2 Oct 2026

Class Balancing and External Transfer in Memory-Constrained Retinal Vision-Language Adaptation

Yuvraj Verma

yuvrajverma282004@gmail.com

Abstract

Class imbalance and implementation discrepancies complicate interpretation of retinal vision-language adaptation on consumer hardware. We investigate inverse-class-frequency sampling in a corrected, grade-only adaptation of Qwen2-VL-2B and audit the archived project that motivates the experiment. After excluding local severity-conflict groups and exact training/test overlap and deduplicating training images, we use 322 training and 83 validation images from the Indian Diabetic Retinopathy Image Dataset (IDRiD). Ordinary and inverse-frequency sampling are compared at three paired seeds with matched response-only supervision, image preprocessing and optimizer budgets. Every final checkpoint is evaluated on all 400 RetinaMNIST+ test images without external tuning. Balanced-minus-unbalanced validation balanced accuracy changes by +1.7 percentage points on average (sample standard deviation 1.9); overall accuracy changes by -13.7 points. The external balanced-accuracy difference is +2.8 points (standard deviation 3.1), with an accuracy difference of -13.8 points. Peak reserved graphics memory is 3.684 GiB. The audit also identifies a confounded historical resolution comparison, prompt-token supervision and synthetic label-derived explanations. Image interventions demonstrate dependence on supplied images, without establishing lesion grounding. We provide raw responses, split and model hashes, per-grade metrics and uncertainty estimates. The evidence supports a reproducibility and sampling study; three seeds, small class supports, unknown patient clustering and absent clinician adjudication limit generalization and preclude clinical-readiness claims.

Keywords: diabetic retinopathy; vision-language models; class imbalance; external evaluation; reproducibility; parameter-efficient adaptation.

1 Introduction

Diabetic retinopathy grading from fundus photographs is an established medical image-analysis task. IDRiD supplies an ordinal grading benchmark and a separate lesion-segmentation collection [1, 2]. General vision-language models (VLMs) make it possible to generate natural-language answers alongside severity predictions. Parameter-efficient adaptation, including LoRA and four-bit QLoRA [3, 4], brings some experiments within consumer GPU budgets. Feasibility, however, does not establish that the resulting text describes visible findings or that aggregate grading accuracy represents minority-class performance.

This paper studies those distinctions through an audit and a new controlled sampling experiment. The original project adapted Qwen2-VL-2B [5] using synthetic answers generated from severity and

edema labels and grade-derived lesion priors. The current study asks four questions: (1) do the implemented preprocessing and loss match the recorded scientific claims; (2) what do the predictions establish after examining class imbalance, duplicate images and textual consistency; (3) how sensitive are archived models to image interventions, inference resolution and an external benchmark; and (4) does inverse-frequency sampling improve balanced accuracy in a matched, grade-only adaptation study across three paired training seeds?

The empirical contributions are:

1. A matched comparison of ordinary and inverse-frequency sampling across three paired seeds, with all six final checkpoints evaluated on a fixed IDRiD validation split and a separate RetinaMNIST+ test benchmark.

2. A reproducibility audit linking local image identity, target provenance, actual input grids and supervised tokens to the interpretation of archived results.

3. A joint analysis of sampling trade-offs, per-grade failures and external transfer under a measured consumer-GPU memory budget, retaining unfavorable results and raw generated responses.

The scope is a controlled empirical study of this pipeline. The adaptation architecture and balancing rule are established methods. Image interventions supply diagnostics of image dependence; they do not measure lesion-level causal faithfulness. Comparisons among the six candidates isolate sampling within the rebuilt pipeline, while comparisons with archived adapters cannot isolate its other corrections.

2 Related Work

Retinal grading and benchmark provenance. The IDRiD dataset and challenge established publicly available grading and segmentation resources [1, 2]. Their presence does not imply that a given local grading record has a verified lesion-mask correspondence. RetinaMNIST derives an ordinal retinal task from DeepDRiD and provides standardized partitions; MedMNIST+ adds larger image sizes [6, 7]. We use its 224-pixel version as a cross-dataset benchmark, acknowledging its substantial preprocessing and resolution differences from the IDRiD images.

Visual concepts and shortcut evaluation. VLM-GCR uses lesion concepts and graph reasoning for DR, with an explicit concept-pseudolabeling mechanism and benchmark evaluation [8]. Corbetta et al. evaluate synthetic acquisition confounders, artifact removal and inversion across retinal and mammography models [9]. Thus, neither text generation nor a generic image-ablation experiment is new. Our scope is narrower: auditing the implementation and evidence claims of small-hardware adaptation with label-derived text, including duplicate identity, actual supervised tokens, and independent grade–rationale outputs.

Class balancing and compact adaptation. Class balancing is an established training tool. Cui et al. develop effective-number-based loss reweighting [10]; our study instead evaluates simple inverse-frequency sampling with replacement. It does not propose a new balancing rule. Broader

medical-image artifact robustness has also been studied [11], further limiting any claim that image perturbation alone constitutes a new contribution.

Qwen2-VL supports dynamic image resolution [5]; this capability can be defeated by an upstream collator that resizes images again. LoRA learns low-rank changes to frozen projections, while QLoRA combines adapters with quantized backbone weights [3, 4]. We distinguish the configuration requested by the training script from the tensor grid actually passed to the model.

3 Experimental Setup

3.1 Local data and synthetic targets

The local grading manifest has 516 images in the canonical 413 training / 103 test directories. Severity labels range from 0 to 4. Training class counts are 134, 20, 136, 74 and 49; test counts are 34, 5, 32, 19 and 13. Directory names and image identifiers jointly identify records because filenames repeat between partitions. This prevents filename collisions, but is insufficient to rule out duplicate image content.

The project generated four question–answer pairs per image, giving 2,064 pairs (1,652 training and 412 test). The question types concern grade identification, lesions, clinical reasoning and recommendations. Every local lesion vector is marked `grade_prior`; every QA record is marked `gemini`. The generator received structured labels and priors rather than the fundus photograph. These targets therefore cannot be treated as independent lesion observations, clinician explanations, or expert assessment of a patient's management. Per-record generator model identifiers were not retained, limiting exact regeneration. We do not use synthetic-reference text similarity as an explanation-quality endpoint.

3.2 Archived models and computational setting

Two adapters are available. Training logs identify three epochs / 618 optimizer steps and four epochs / 824 steps over the same reported 1,652 QA records. Both logs print a 512-pixel collator fallback. The script uses rank and scaling 16, zero dropout, adapters in vision and language attention and MLP modules, batch size one, accumulation eight, an eight-bit AdamW optimizer, and bfloat16 activations. The logged trainable parameter count is 28,950,528. The four-epoch training log records approximately 89.4 minutes and 3.52 GiB peak reserved CUDA memory on a laptop RTX 4050. These are historical observations, not a memory guarantee for a corrected 768-pixel training run.

New inference uses the existing WSL environment, four-bit model loading, greedy decoding, and synchronized CUDA timing. Versions, prompt text, input grids, raw generated answers and checkpoint identifiers are retained separately for each run. No paid API baseline is used. Missing complete historical baseline predictions are regenerated rather than reconstructed from the five examples in the archived summary.

4 Methods

Figure 1 summarizes the fixed data, adaptation and evaluation protocol.

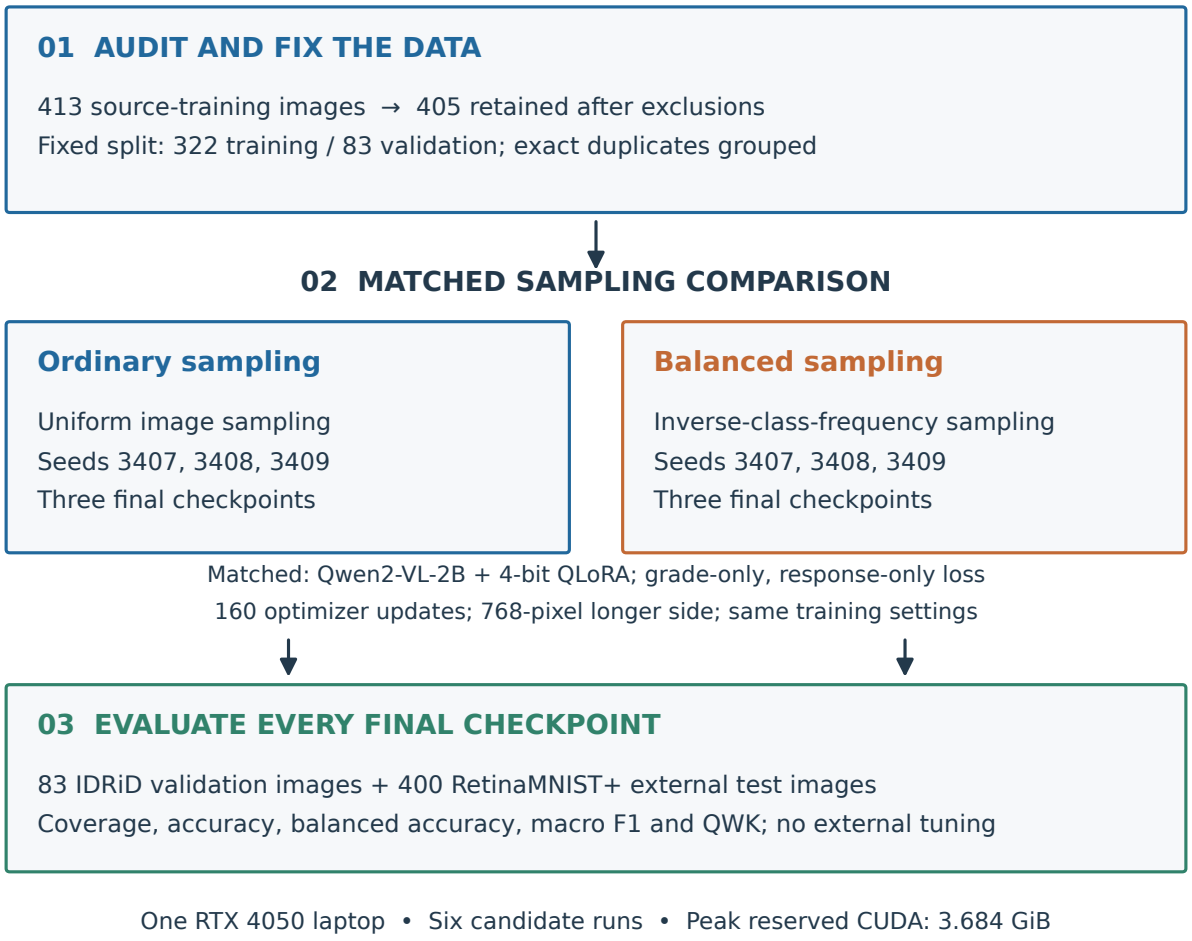


Figure 1: Controlled study design. All six final checkpoints are evaluated under the same protocol. Data cleaning removes exact image overlap and severity conflicts in the local corpus; patient-level independence remains unknown. BA: balanced accuracy.

4.1 Prospective grade-only balancing study

Before candidate training, we specify three paired random seeds (3407, 3408, 3409), each comparing ordinary sampling with inverse-class-frequency sampling on the same 322 cleaned training images. Both strategies use grade-only targets, four requested epochs, a 768-pixel longer side, explicit response-only loss and collator `resize=max` to preserve the requested longer side. Other settings match across strategies: rank/scaling 16, zero dropout, batch one, accumulation eight, learning rate 2×10^{-4}, 3% warmup, linear decay and eight-bit AdamW. Floating model parameters and activations use bfloat16; frozen quantized weights remain four-bit. The final-budget checkpoint rule applies to every run. The primary contrast is paired-seed validation balanced accuracy; external transfer is secondary. No seed, hyperparameter or ensemble is selected from test performance. Reserved CUDA allocation is capped at 5.0 GiB, below 5.5 decimal GB.

This contrast tests sampling within the rebuilt grade-only pipeline. The experiment definition is written locally before candidate training and its hash is recorded in each run configuration; it is not

a public preregistration. It cannot isolate separate effects of cleaning, supervision type, masking and training resolution relative to the archived models. Failed runs and protocol changes must be recorded. Three seeds do not establish general superiority.

Sampling distribution and primary contrast. Let n_k be the number of training images of grade k, with $(n_0, \ldots, n_4) = (106, 14, 106, 58, 38)$ and $N = 322$. Ordinary training uses shuffled images without replacement within each epoch. Balanced training makes N draws with replacement from image weights $w_i = 1/n_{y_i}$, giving

$$p_i = \frac{1/n_{y_i}}{\sum_{j=1}^{N} 1/n_{y_j}} = \frac{1}{5n_{y_i}}. \tag{1}$$

Each grade therefore receives probability $1/5$ per balanced draw. Its expected draw count is $322/5 = 64.4$ per nominal epoch, rather than the fixed ordinary count n_k. This is an expectation of the sampler, not an observed count or an increase in independent images. Replacement changes repetition and unique-image exposure intentionally; the paired comparison holds the optimizer budget fixed.

For evaluation support m_k and correct-grade count $c_{s,k}$ at seed s, balanced accuracy is $\mathrm{BA}_s = \frac{1}{5}\sum_{k=0}^{4} c_{s,k}/m_k$. The primary descriptive contrast is

$$\overline{\Delta}_{\mathrm{BA}} = \frac{1}{3} \sum_{s \in \{3407, 3408, 3409\}} \left(\mathrm{BA}_s^{\mathrm{balanced}} - \mathrm{BA}_s^{\mathrm{ordinary}} \right). \tag{2}$$

We report its three component differences and sample standard deviation. The evaluation images are reused across seeds; the seeds are training replicates, not additional patients. No significance test or population confidence interval is attached to this three-seed mean. External contrasts use the same definitions as secondary descriptive outcomes.

4.2 Image identity and label consistency

We calculate SHA256 hashes of all local grading-image files. Duplicate groups are inspected for split membership, severity labels and edema labels. This identifies exact file duplication; it does not establish patient-level independence for nonidentical images. Findings are attributed to the local copy because we have not authenticated each file against a newly downloaded official release.

The principal retrospective analysis preserves the original 103 test records. A sensitivity analysis excludes the cross-partition duplicate. We do not silently deduplicate a reported benchmark. A new training split excludes conflicting severity groups and cross-partition overlap and retains one member of each remaining identical-image group. This split is used for new candidates; it is not an independent validation set for the archived adapters, which already saw the original training records. Validation supports for grades 0–4 are 27, 4, 27, 15 and 10, respectively; the small mild-disease support limits precision of class-balanced comparisons.

Figure 2 visualizes the class supports used in the controlled study.

4.3 Inspecting grids and supervised tokens

Using the saved processor and current installed collator, we form identical conversations with a known grade answer and images preprocessed to longer sides of 512 and 768 pixels. We record

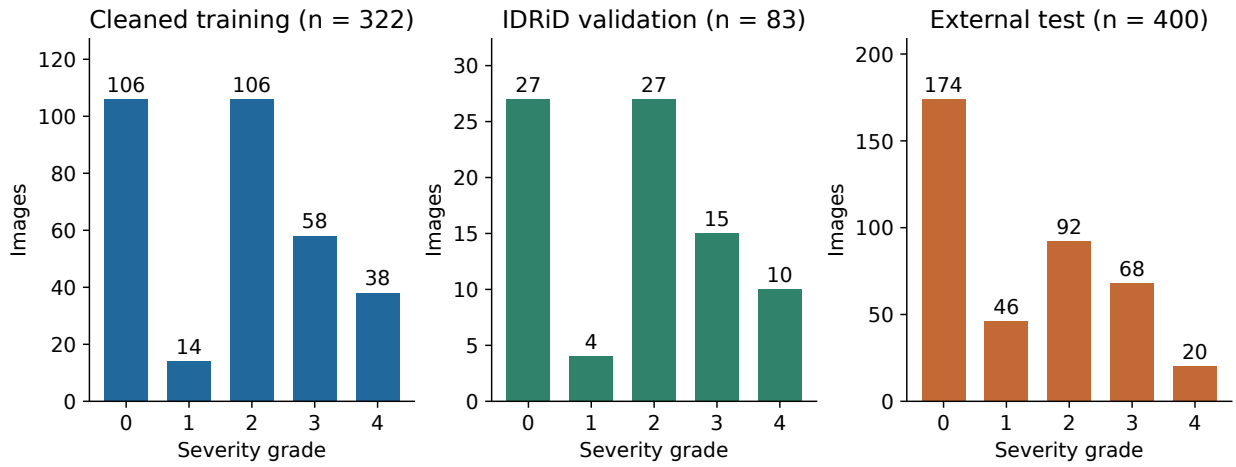


Figure 2: Recorded severity supports for cleaned training, validation and the external test. Each panel has its own count scale. Only four mild-disease images are available in validation; repeating their evaluation across seeds does not increase the number of independent cases.

the image grid and decode the tokens whose labels are not the ignore index. We then repeat with explicit `resize="max"` and response-only supervision enabled. The historical logs corroborate the fallback, but the absence of historical batch dumps prevents proving every historical tensor from present-day inspection alone.

For Qwen2-VL's two-by-two patch merging, a grid $[t, h, w]$ corresponds to approximately $thw/4$ visual tokens. For the inspected photograph, both historical-default inputs produce grid $[1, 24, 36]$, or 216 visual tokens. Explicit `resize=max` gives $[1, 36, 54]$, or 486 tokens, at 768 pixels. Historical labels include system, user and assistant text, excluding image and padding tokens; the historical loss must be described over those unmasked conversation tokens, rather than as answer-only loss.

4.4 Grade parsing and uncertainty

We parse explicit grade numbers first, anchored leading integers second, and unambiguous named severities third. Incidental numbers and the substring "proliferative" within "non-proliferative" are not sufficient. Contradictory explicit grades and severity names are flagged. Failures count as errors in exact-match accuracy. Five-class macro F1 and balanced accuracy are computed with all five classes fixed; failures count as missed true-class predictions. Quadratic weighted kappa (QWK) is reported only among valid parsed answers, with the corresponding denominator. An abstaining model must not be interpreted as having perfect agreement on a small covered subset.

For confusion matrix O, independent-marginal expectation E, and $w_{ij} = (i - j)^2/16$, QWK is

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}. \tag{3}$$

We report Wilson 95% intervals for accuracy and class recalls. Other metric intervals and paired differences use 5,000 image-level bootstrap draws with seed 20261001. These intervals condition on the fitted checkpoint and evaluation procedure; they do not quantify training-seed variation, patient clustering, pretraining overlap or model-selection bias. Undefined bootstrap statistics are omitted from that statistic's percentile calculation.

4.5 Image interventions

The four-epoch adapter is evaluated on every IDRiD test record with the original grade question: "What is the diabetic retinopathy severity grade (0-4) shown in this fundus image? Answer with the number and its name." The maximum generated length is 256 tokens. Conditions are the real image, an all-black image of identical dimensions, and a deterministic image selected from a different-grade test record. Labels construct the offline replacement intervention only; they are never supplied to the model. Inputs use 768-pixel longer-side preprocessing. Replacement images can be reused, so bootstrap summaries for this condition are conditional on the fixed mapping and do not remove donor dependence. We report agreement with both the original and replacement labels.

These controls establish sensitivity to supplied image content if predictions change systematically. An all-black image is out of distribution. Replacement-label agreement may reflect pathology, acquisition correlates or other visual features; it is not a lesion-localization or causal-faithfulness measure.

4.6 Inference resolution and external evaluation

Each fixed archived adapter is evaluated at both 512 and 768 pixels on all 103 IDRiD test records, with identical prompts and decoding. Within-checkpoint contrasts test inference preprocessing, not training resolution. Between-checkpoint contrasts remain confounded by training duration and potentially undocumented historical settings.

For external evaluation, the official RetinaMNIST+ 224-pixel archive is downloaded and checked against its official MD5 checksum. All 400 test images are evaluated without adaptation, validation tuning, threshold selection, or upscaling. The base and four-epoch adapter use the same prompt, preprocessing and decoding. This is a low-resolution cross-dataset transfer experiment, not an independent clinical validation study. We cannot determine pretraining exposure or patient clustering from the supplied arrays.

4.7 Grade–rationale consistency

The historical evaluator asked for a grade and a rationale in independent conversations; the rationale question did not supply the predicted grade. We inspect all archived rationale outputs for one unambiguous asserted severity name and compare it with the grade response. Mentions of future progression, excluded severity thresholds and immediate negation are excluded. We report coverage and disagreements among covered responses. Unclassified rationales are not counted as consistent. This is a linguistic-consistency diagnostic; it does not evaluate whether a mentioned lesion is present or whether a diagnosis is clinically correct. No ophthalmologist adjudicated these outputs.

4.8 Conventional visual baseline

We extract frozen ResNet-18 features using torchvision's official ImageNet-1K V1 initialization [12]. IDRiD images preserve aspect ratio at a 768-pixel longer side; RetinaMNIST remains at native 224 pixels. Inputs use ImageNet mean/std normalization and bfloat16 GPU inference. Global average pooling gives a 512-dimensional representation. A standardized, class-balanced multinomial logistic-regression head is fit on the 322 cleaned training images with $C = 1$, an L2 penalty and the L-BFGS solver (maximum 2,000 iterations). Standardization is fit only on training features. No

regularization selection is performed on validation or external data. The head's float64 CPU fitting is distinct from GPU bfloat16 inference. This is a contextual comparator, not a state-of-the-art retinal classifier or a matched architecture ablation.

5 Results

5.1 Matched class-balancing experiment

Figure 3 shows every paired-seed accuracy and balanced-accuracy difference. The graph displays measured seed effects, without treating three seeds as a population confidence interval. All six

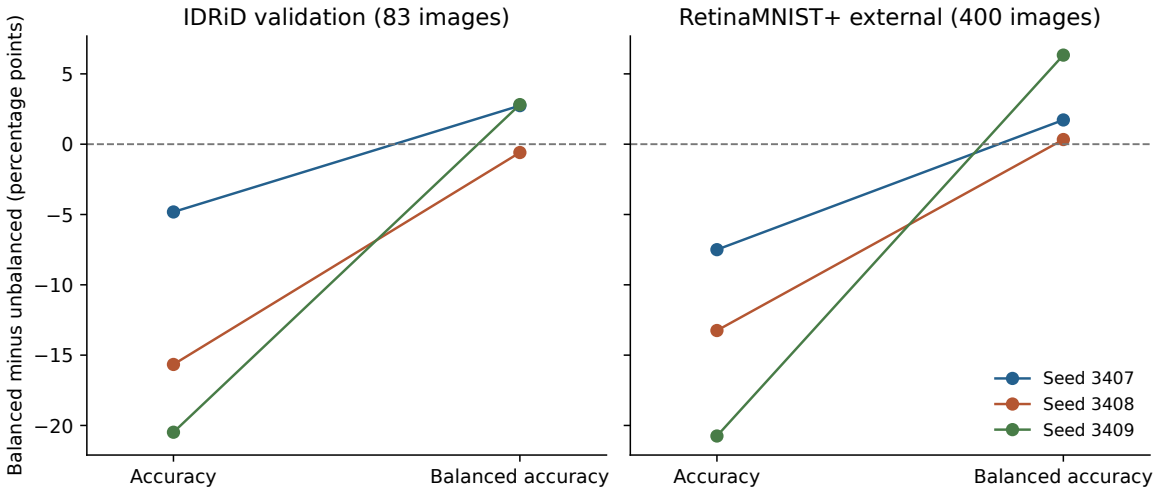


Figure 3: Balanced minus ordinary sampling for all three prespecified seeds. Overall accuracy falls for every seed on both datasets; balanced-accuracy effects are smaller and vary across seeds. Each color identifies one paired training seed.

candidate runs complete the fixed 160-optimizer-step budget, with four requested epochs (trainer-reported final epoch 3.92). All final checkpoints are evaluated; no best-seed selection or external-data tuning is performed. Table 1 reports each seed on both datasets.

The conservative parser covers 2898/2898 candidate responses. Failures count as errors for accuracy and class recall; QWK conditions on parsed responses.

Validation balanced-minus-unbalanced balanced accuracy changes by +1.7 percentage points on average (sample SD 1.9; paired-seed differences +2.7, -0.6, +2.8). The corresponding accuracy difference is -13.7 points (SD 8.0), macro F1 difference -0.024 (SD 0.024), and QWK difference +0.017 (SD 0.032). These are descriptive seed statistics, not confidence intervals or a significance test.

External balanced-minus-unbalanced balanced accuracy changes by +2.8 percentage points on average (sample SD 3.1; paired-seed differences +1.7, +0.3, +6.3). The corresponding accuracy difference is -13.8 points (SD 6.6), macro F1 difference +0.014 (SD 0.033), and QWK difference -0.103 (SD 0.052). These are descriptive seed statistics, not confidence intervals or a significance test.

Table 1: Every candidate seed. U: ordinary sampling; B: inverse-frequency sampling. Validation uses 83 images; external evaluation uses all 400 RetinaMNIST+ test images. BA: balanced accuracy.

Dataset / seed	Sampling	Acc. (%)	BA (%)	Macro F1	QWK
Validation / 3407	U	62.7	45.3	0.445	0.766
Validation / 3407	B	57.8	48.0	0.443	0.819
Validation / 3408	U	66.3	51.9	0.492	0.783
Validation / 3408	B	50.6	51.3	0.470	0.793
Validation / 3409	U	67.5	49.5	0.487	0.799
Validation / 3409	B	47.0	52.3	0.438	0.789
External / 3407	U	47.0	33.4	0.295	0.521
External / 3407	B	39.5	35.2	0.272	0.466
External / 3408	U	44.5	30.6	0.256	0.495
External / 3408	B	31.2	30.9	0.285	0.400
External / 3409	U	47.5	28.5	0.240	0.481
External / 3409	B	26.8	34.9	0.277	0.323

For the four mild-disease validation images, ordinary sampling correctly grades 0, 0, 0 images across the three seeds; balanced sampling correctly grades 0, 2, 2. These repeated evaluations involve the same four images and do not provide twelve independent mild-disease observations.

Peak reserved CUDA allocation is 3.684 GiB across these runs; training takes 37.2–38.1 minutes per run. Evaluation and model initialization time are excluded. Raw parse coverage, per-grade recalls, image-bootstrap intervals and all confusion matrices are supplied for each run.

Figures 4 and 5 show the absolute score trade-offs and the per-grade recall profiles for every candidate seed.

The three paired seeds expose a metric-dependent sampling effect rather than universal superiority. Comparisons with archived adapters additionally change targets, cleaning, loss masking, image preprocessing and update count. They cannot attribute differences to any one of those changes. No comparison here demonstrates clinical utility or lesion grounding.

5.2 Data and implementation audit

Figure 6 visualizes the distribution of errors for the first prespecified seed (3407), rather than selecting the best-performing seed. There are six pairs of byte-identical local files (Table 2). Two within-training groups and one cross-partition group have conflicting severity labels. Another within-training group has the same severity but conflicting edema labels. The local manifest is checked against the local source CSV labels, and saved evaluation labels are checked against that manifest. This establishes consistency within the local artifacts, not authentication of the upstream release.

Removing severity-conflict groups and the cross-partition training image, then deduplicating remaining training groups, leaves 405 source-training images. The fixed stratified split contains 322 training and 83 validation images. Patient-level independence remains unknown. The 78 available segmentation originals have no exact byte or decoded-pixel identity matches to the grading corpus; their masks are not joined by filename.

Batch inspection confirms both the extra resize and supervised prompt text. Consequently, the

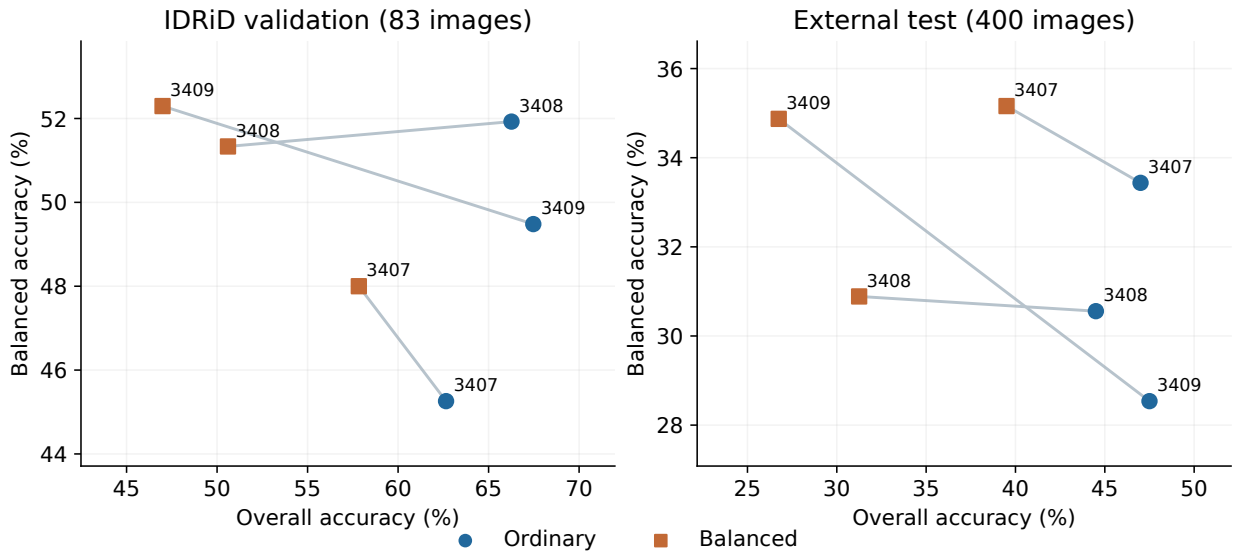


Figure 4: Absolute accuracy and balanced accuracy for all six candidate checkpoints on both evaluation datasets. Connecting segments join the ordinary and balanced conditions at the same training seed; labels identify seeds. Axis ranges differ between datasets. These points summarize fixed runs, not population confidence regions.

Table 2: Exact duplicate groups in the local grading corpus. These findings are not attributed to the upstream release without independent provenance verification.

Images	Grades	Partition	Other discrepancy
021 / 391	1 / 2	train/train	–
028 / 327	4 / 3	train/train	–
045 / 333	2 / 2	train/train	edema 0 / 1
073 / 375	2 / 2	train/train	–
079 / 279	1 / 1	train/train	–
118 / 064	0 / 3	train/test	split overlap

historical comparison cannot support a causal training-resolution claim. The corrected collator produces the requested distinct token grids and masks the example prompt, but these corrections alone do not establish better model performance.

5.3 Retrospective grading and minority classes

The complete historical adapted predictions reproduce 68/103 correct grades, QWK 0.763 (image-bootstrap 95% interval 0.665–0.846), balanced accuracy 0.475 and macro F1 0.468. No prediction is grade 1. Table 3 gives recalls and support. Removing test 064, the exact cross-partition duplicate, gives 68/102 correct (66.7%). This sensitivity analysis does not establish that overlap or conflicting labels had no effect.

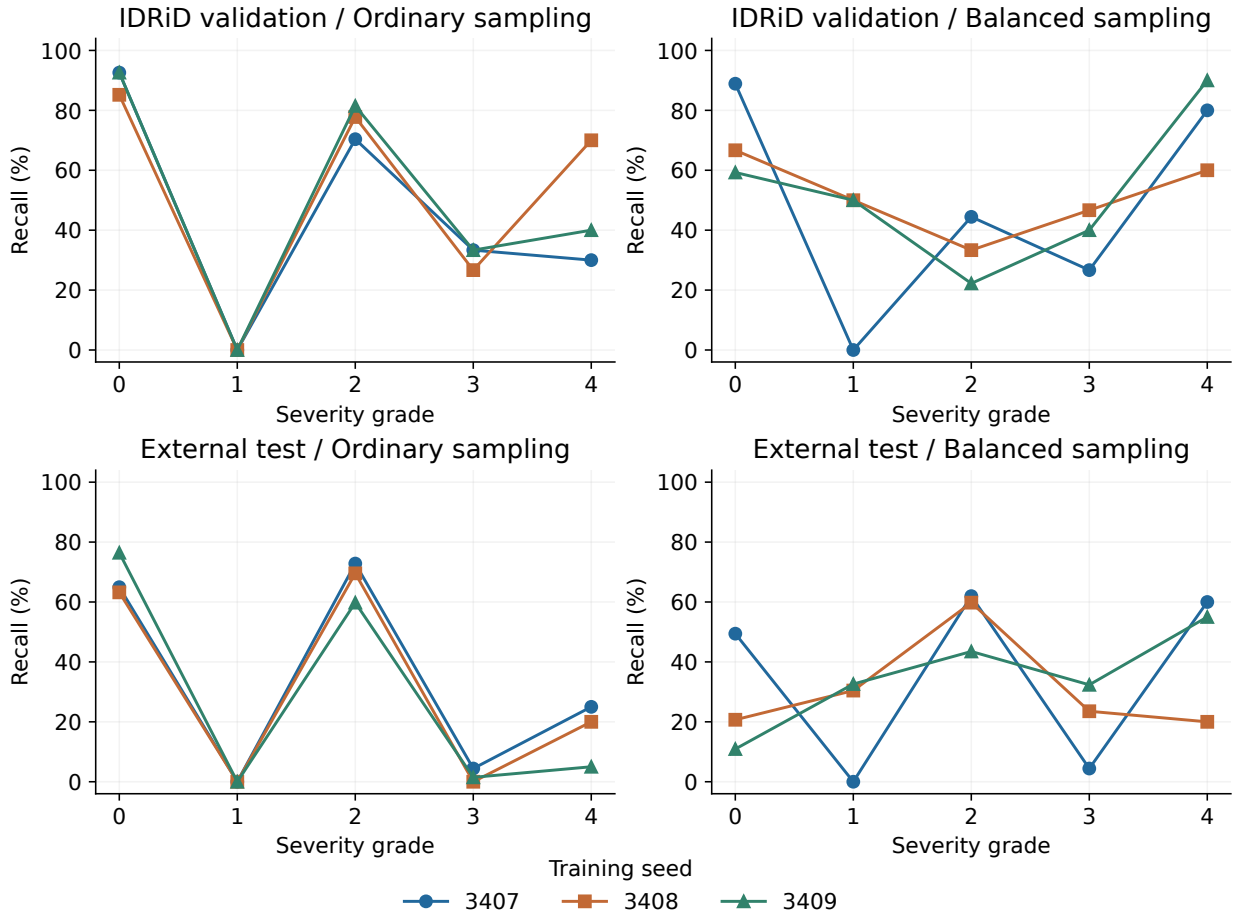


Figure 5: Per-grade recall for all prespecified seeds under each sampling strategy. Connecting lines guide the eye across ordinal grades; they do not interpolate outcomes. The panels use identical recall scales. Validation supports are 27, 4, 27, 15 and 10; external supports are 174, 46, 92, 68 and 20. The curves describe fixed checkpoints and do not quantify patient-level uncertainty.

Table 3: Historical adapted-model class recalls; intervals are image-level Wilson intervals.

Grade	Support	Recall (%)	95% CI (%)
0	34	97.1	85.1–99.5
1	5	0.0	0.0–43.4
2	32	78.1	61.2–89.0
3	19	31.6	15.4–54.0
4	13	30.8	12.7–57.6

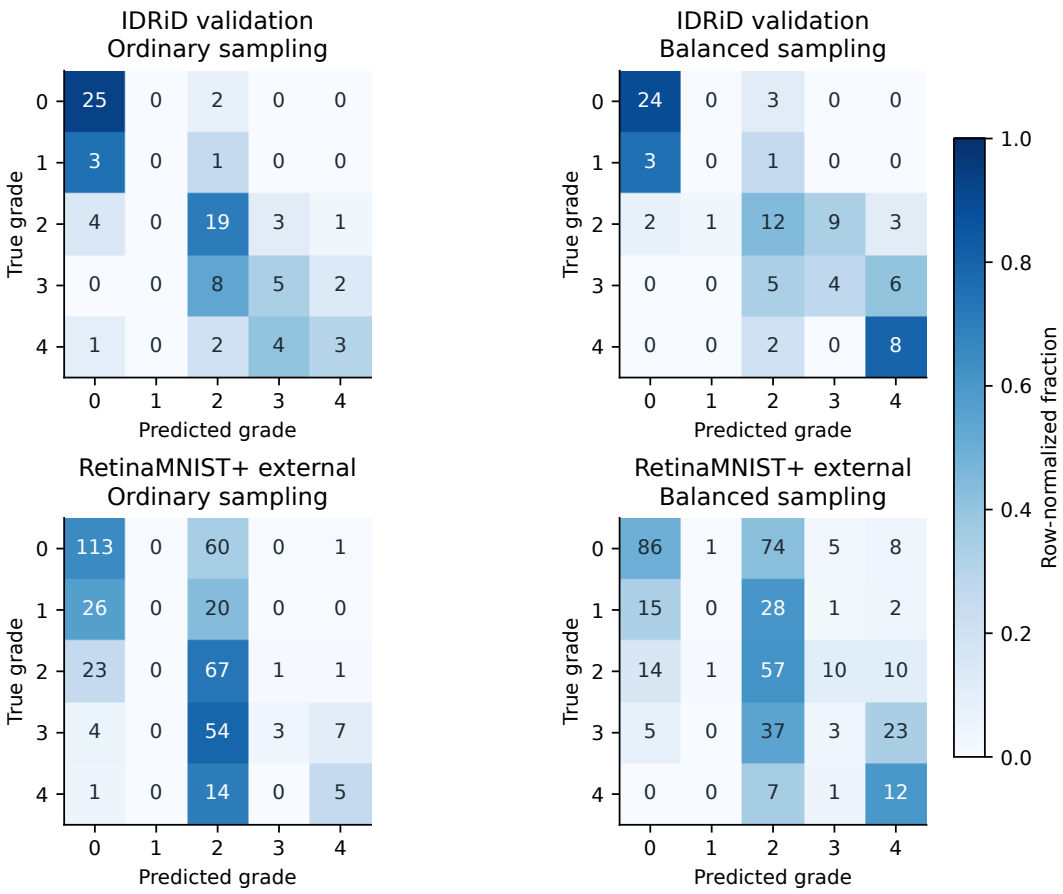


Figure 6: Confusion matrices for seed 3407, the first prespecified seed. Cell labels are image counts; color encodes the row-normalized fraction. Rows are true grades and columns are predicted grades. All seeds remain reported in Table 1 and the raw results. These panels are examples of the fixed-seed behavior, not ensembles or independent clinical validation.

5.4 Image interventions

The new real-image run reproduces all 103 archived grade predictions. Blank inputs yield grade 0 in all cases, giving 34/103 correct (33.0%), balanced accuracy 0.200 and QWK 0. The real-minus-blank accuracy difference is 33.0 percentage points, with a paired image-bootstrap interval of 23.3–42.7 points. Predictions change in 58/103 cases. With different-grade replacements, original-label accuracy is 18/103 (17.5%), while replacement-label agreement is 62/103 (60.2%). Predictions change from the real-image condition in 75/103 cases.

Figure 7 summarizes the image intervention scores against the original image labels.

The model therefore responds to the supplied image under these controls. The same model's absent mild predictions and low severe/proliferative recalls show why image dependence alone is insufficient evidence of a clinically adequate grader.

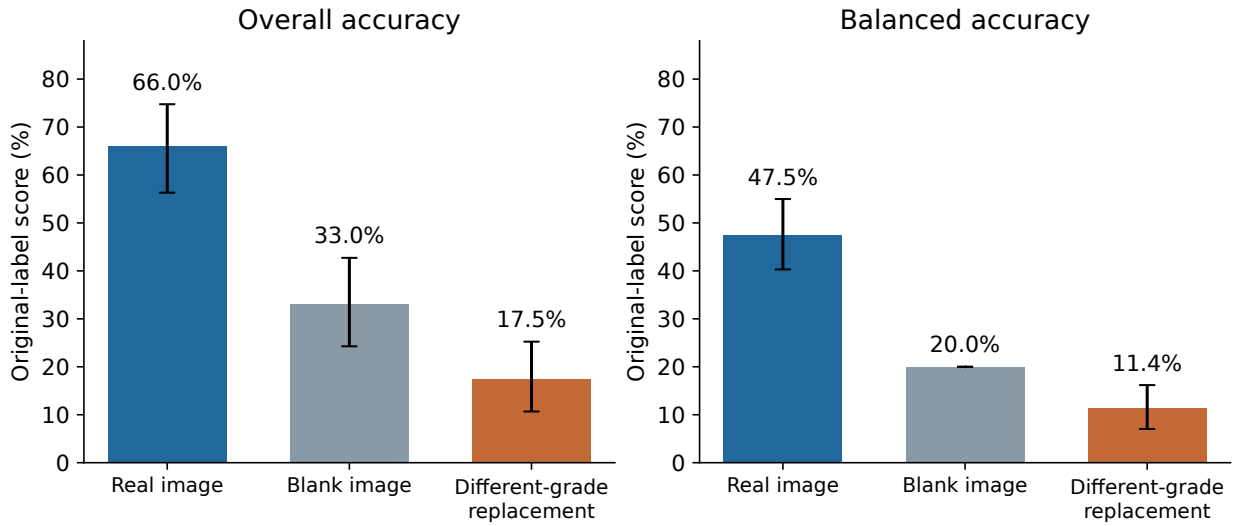


Figure 7: Archived four-epoch adapter scores on the same 103 IDRiD records using real images, blank images and different-grade replacement images. Bars use the original labels in all conditions; replacement-label agreement is a separate endpoint reported in the text. Whiskers are image-bootstrap 95% intervals conditional on the fitted checkpoint and fixed donor mapping. Blank images are out of distribution, and these controls demonstrate image sensitivity without establishing lesion grounding.

5.5 Inference resolution and external transfer

Figure 8 presents paired inference-resolution contrasts, and Figure 9 compares external class recall with the conventional visual baseline. Within the four-epoch checkpoint, 768-minus-512 inference

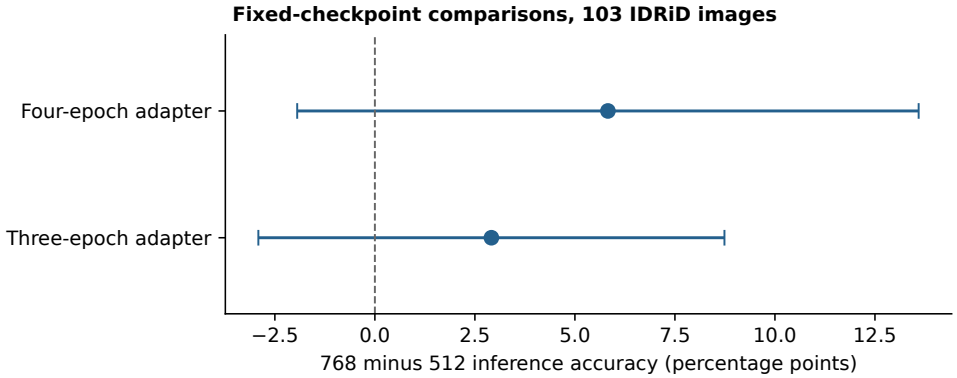


Figure 8: Fixed-checkpoint inference comparisons on 103 IDRiD images. Points are measured accuracy differences; bars are paired image-bootstrap 95% intervals. Both intervals include zero. These comparisons do not estimate a training-resolution effect.

accuracy changes by 5.8 percentage points (paired image-bootstrap 95% interval -1.9 to 13.6). The three-epoch contrast is 2.9 points (-2.9 to 8.7). Both intervals include zero. These measurements do not establish a general inference-resolution benefit and do not estimate a training-resolution effect.

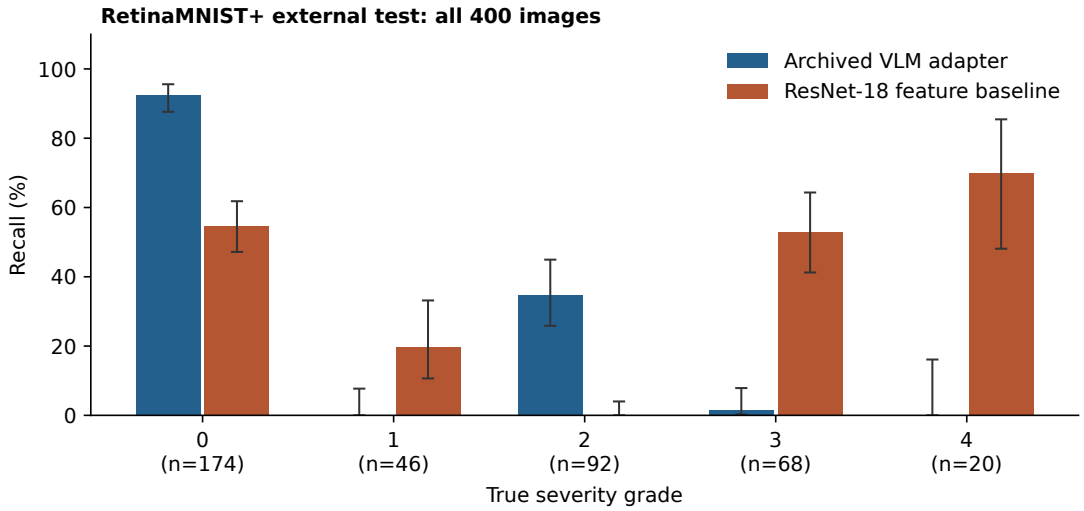


Figure 9: External class recalls for the archived adapter and frozen ResNet-18 feature baseline on all 400 RetinaMNIST+ test images. Bars show recall and whiskers show image-level Wilson 95% intervals; supports are given below each grade. Architectures and learning procedures differ. Higher aggregate adapter accuracy does not establish superiority across grades.

Table 4: Fixed-checkpoint IDRiD inference comparisons. Training logs report a 512-pixel collator default for both adapters. Acc. and BA are percentages; F1 is macro F1; all 103 responses per setting are parsed.

Setting	Acc.	BA	F1	QWK
3 epochs / 512	54.4	37.4	0.325	0.622
3 epochs / 768	57.3	39.3	0.342	0.643
4 epochs / 512	60.2	43.3	0.416	0.702
4 epochs / 768	66.0	47.5	0.468	0.763

Table 5: External RetinaMNIST+ test evaluation at native 224 pixels. All 400 responses per setting are parsed. Acc. and BA are percentages; F1 is macro F1. No external training or threshold tuning is performed. The conventional feature baseline has a different architecture and learning procedure.

Setting	Acc.	BA	F1	QWK
Base VLM	23.0	20.0	0.075	-0.003
Archived adapter	48.5	25.8	0.210	0.483
ResNet-18 + head	38.5	39.4	0.288	0.403

On all 400 external test images, the archived adapter obtains 48.5% accuracy and 25.8% balanced accuracy. Its paired accuracy difference relative to the base is 25.5 points (image-bootstrap 95% interval 18.8–32.2). The adapter predicts no grade-1 case; grade-3 recall is 1/68 and grade-4 recall 0/20. This limits interpretation of the aggregate external gain. Acquisition and resolution differences are not isolated, and these are not clinical validation results.

The conventional feature baseline has lower overall external accuracy (38.5%) but higher balanced accuracy (39.4%) and macro F1 (0.288). The archived VLM therefore does not dominate this simple visual comparator across metrics. Architecture and learning procedure differ, so this is a contextual comparison, not a matched mechanism test.

5.6 Rationale consistency

Context-filtered named-severity extraction covers 49/103 rationale answers (47.6%) and flags 22/49 disagreements (44.9% of covered answers). These are linguistic flags rather than clinician-adjudicated errors. An explicit example is test 001: the grade answer is grade 4, while the rationale states that the image is classified as severe NPDR (grade 3). Other outputs describe future progression to proliferative disease or compare a grade against mild/severe criteria; those mentions cannot be counted as asserted grades. Unclassified answers are not counted as consistent. The independent-query design makes contradictions possible, and our diagnostic does not establish the clinical validity of either answer.

6 Discussion

6.1 What the evidence supports

The controlled study establishes a sampling trade-off within this particular pipeline. Validation balanced accuracy rises in two of three paired seeds and falls in one, while validation and external accuracy fall in every seed. External QWK also falls in every seed. Thus, the small average balanced-accuracy increase does not justify a general improvement claim. The contextual frozen-CNN baseline has higher external balanced accuracy than all six new candidates, although architectures and learning procedures differ.

The audit additionally reproduces the archived adapted grades and their sensitivity to image interventions. Corrective findings concern the gap between intended configuration and actual batch behavior, and between aggregate grading agreement and minority-class coverage. The local duplicate-label groups show that directory separation alone is insufficient for data auditing. Context inspection of rationale text demonstrates why a lexical consistency count must be checked before it becomes a scientific result.

These observations support an audit contribution. They do not support a new best-performing DR algorithm, faithful lesion explanations, or clinical readiness. Comparable shortcut and concept-supervision studies already exist [9, 8]; any novelty claim must be confined to the measured integration of artifact provenance, implementation checks, and this particular adaptation setting.

6.2 Corrected pipeline and sampling trade-offs

The rebuilt pipeline uses grade-only targets, a deduplicated image-grouped training/validation split, explicit preservation of the requested image dimensions and explicit response-only loss. It retains raw predictions and protects checkpoints from overwriting. Removing generated management advice from grading supervision avoids treating unsupported recommendations as desired outputs. The matched sampling study evaluates this corrected pipeline; it does not isolate the individual benefits of its corrections.

A future matched comparison should isolate grade-only versus synthetic multitask targets at the same update and image-exposure budgets and assess additional backbones and datasets. Our conventional visual comparator differs in architecture and learning procedure. Validation choices must be made within the training-source partition, with external tests kept frozen. The reused IDRiD test remains retrospective. For a lesion-grounding study, verified image-mask correspondence or new independent annotations are required; IDRiD's existing mask classes do not provide neovascularization masks. Clinical utility requires a separate clinical study.

6.3 Limitations

This is a small study of one adapted VLM family. The controlled experiment measures variation across only three training seeds; it does not quantify uncertainty across architectures, sites or patient cohorts. The IDRiD test partition was previously used in development and includes one exact local historical-training overlap. The rebuilt split removes that overlap from candidate training, but cannot erase historical test use. Patient identities and clustering are unavailable. Hash identity finds exact duplicates but can miss differently encoded, cropped or repeated-patient photographs. The observed duplicate defects have not been authenticated against a new upstream release.

The historical processor/collator reconstruction uses the currently installed environment; logs corroborate the fallback but do not capture every original tensor or library revision. Synthetic-target generator identities were not logged per record. The original grading parser and independent rationale prompts also limit interpretation of older summary metrics.

Blank images are out of distribution; replacement images can be reused, and their labels influence intervention construction. The controls therefore cannot establish lesion-specific causation. Rationale extraction is a limited linguistic diagnostic with incomplete coverage and no clinical adjudication. RetinaMNIST's low resolution, different acquisition distribution, unknown pretraining exposure and unavailable patient metadata limit claims about external performance. Image-bootstrap intervals do not capture these broader sources of uncertainty.

7 Conclusion

A memory-constrained retinal VLM can exhibit reproducible image-dependent grading while aggregate scores conceal minority-class failures and implementation discrepancies. Our audit identifies those discrepancies and supplies a corrected, grade-only pipeline with a matched sampling experiment. Sampling effects must be interpreted across accuracy, class-balanced metrics, training seeds and external transfer rather than reduced to a single favorable score. The study supports reproducible evaluation of this setting; it does not establish a clinically valid grading method, lesion-grounded explanations or a new state of the art.

Data, Code and Reproducibility

The audit uses local copies of public IDRiD and RetinaMNIST+ data under their respective access terms. RetinaMNIST+ is distributed under CC BY 4.0. Licensed images and model weights are not included in the paper package. The accompanying code records file hashes, split exclusions, raw outputs, grids, prompt text, software versions, random seeds, metric definitions and bootstrap summaries. The archived IDRiD project is available at https://github.com/vermayuvraj/idrid-vqa; newly created audit code is supplied with this draft, and no claim is made that it has already been published online. Exact synthetic-target regeneration is unavailable because generator identities were not retained per record.

Statements and Declarations

Funding. The research was self-funded by the author. No external research grant or sponsorship was received.

Ethics and Use of Generative Tools

No new patient recruitment or clinical intervention was conducted. Models and synthetic outputs are research artifacts and must not be used for diagnosis, triage or treatment. Synthetic label-conditioned text may contain unsupported findings and recommendations; it was not reviewed by clinicians. An AI coding assistant contributed to literature discovery, artifact auditing, implementation, statistical analysis and manuscript drafting. The human author is responsible for verifying and approving claims, citations and code before publication. No clinician review is asserted by this draft.

References

[1] P. Porwal et al., "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," *Data*, 3(3), 25, 2018. https://doi.org/10.3390/data3030025.

[2] P. Porwal et al., "IDRiD: Diabetic Retinopathy—Segmentation and Grading Challenge," *Medical Image Analysis*, 59, 101561, 2020. https://doi.org/10.1016/j.media.2019.101561.

[3] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022. https://arxiv.org/abs/2106.09685.

[4] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," NeurIPS, 2023. https://arxiv.org/abs/2305.14314.

[5] P. Wang et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," 2024. https://arxiv.org/abs/2409.12191.

[6] J. Yang et al., "MedMNIST v2—A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, 10, 41, 2023. https://doi.org/10.1038/s41597-022-01721-8.

[7] MedMNIST authors, "MedMNIST+ dataset release and RetinaMNIST metadata," official project documentation, accessed 1 October 2026. https://github.com/MedMNIST/MedMNIST.

[8] Q. Xu et al., "Vision-Language Models Guided Graph Concept Reasoning for Interpretable Diabetic Retinopathy Diagnosis," *Proceedings of AAAI*, 40(32), 27314–27322, 2026. https://doi.org/10.1609/aaai.v40i32.39948.

[9] V. Corbetta et al., "Beyond Clean Test Sets: Spurious Correlations in Medical Vision-language Models and the Role of Concept Supervision," MICCAI, 2026. https://papers.miccai.org/miccai-2026/0098-Paper5420.html.

[10] Y. Cui et al., "Class-Balanced Loss Based on Effective Number of Samples," CVPR, 2019. https://arxiv.org/abs/1901.05555.

[11] Z. Cheng et al., "Understanding the robustness of vision-language models to medical image artefacts," *npj Digital Medicine*, 8, 727, 2025. https://doi.org/10.1038/s41746-025-02108-w.

[12] K. He et al., "Deep Residual Learning for Image Recognition," CVPR, 2016. https://arxiv.org/abs/1512.03385.