

# TeleDerm SnapCheck: Automating Dermatology Image Quality Gating for Safer Teledermatology Triage

## Abstract

### Background

Teledermatology programs rely on patient-submitted photos that frequently suffer from blur, shadowing, or framing errors, undermining downstream clinical decision support. Manual screening is labor-intensive and inconsistently applied.

### Objective

We test whether an automated dermatology image quality gate can lower urgent-miss risk for vision-language triage models while maintaining clinical throughput.

### Methods

We augmented public dermoscopy archives (ISIC, HAM10000, Derm7pt) with synthetic degradations spanning blur, contrast, lighting, obstruction, and framing defects. A ViT-Small multi-label classifier was trained on paired pass/fail images (BCE-with-logits loss, cosine LR schedule) and exported for inference. Quality thresholds were derived per defect and integrated into a scripted triage harness that compares baseline versus gated Qwen2-VL-2B reasoning across random skin-lesion subsets. Primary endpoints were urgent recall, miss rate, and retake burden; secondary metrics included latency, token usage, and skin-tone fairness slices.

### Results

The quality model reached macro AUROC 0.968 and macro average precision 0.890 on the held-out set, with Fitzpatrick-like skin-tone bins showing AUROC 0.957-0.972. In a 721-image triage batch, the baseline VLM achieved 0.38 overall accuracy and 0.65 urgent recall. Introducing the prototype gate triggered retakes in 50% of encounters and suppressed urgent recall to 0.31, doubling the urgent miss rate (0.69 vs 0.35). Latency and token usage were unchanged because the gate re-used cached prompts.

### Conclusions

Automated quality gating is technically feasible and reproducible, yet the initial thresholds are over-conservative for clinical deployment. Tune-once gating cannot be accepted without prospective calibration to keep retakes manageable and prevent harm from deferred urgent lesions.

### Keywords

teledermatology; dermatology image quality; vision-language models; fairness; clinical decision support

## Introduction

Teledermatology has expanded rapidly under access pressures, but poor patient-generated imagery remains a key failure point. Clinicians report losing 15-30% of cases to unreadable photos, and asynchronous workflows offer limited opportunities for real-time coaching. Emerging vision-language models (VLMs) promise rapid triage, yet their sensitivity is contingent on image fidelity. Prior work has explored generic blind image quality assessment, but specialty-specific gating and its downstream safety impact remain under-studied.

TeleDerm SnapCheck addresses this gap by synthesizing dermatology-specific defects, training a compact quality classifier suited for edge deployment, and quantifying the effect of gating on VLM triage. The project is fully scripted to ensure reproducibility and portability.

## Methods

### Data Sources and Augmentation

We curated public dermoscopy datasets (ISIC 2020, HAM10000, Derm7pt) and generated paired pass/fail crops. The augmentation engine injects blur, motion blur, exposure shifts, contrast shifts, shadow occlusion, obstructions, cropping/framing errors, and resolution downscaling. Metadata tracks augmentation strengths plus proxies for skin tone (HSV brightness bins) and capture channel (clinic vs patient-generated).

### Quality Model Training

A ViT-Small (patch16, 224 px) backbone initialized with ImageNet weights was fine-tuned using multi-label BCE loss. Training used batch size 64, AdamW (lr=5e-5, weight decay 0.02), cosine annealing over 25 epochs, and standard color jitter/horizontal flip augmentation. Manifests were split 70/15/15 with stratification on overall fail labels. Evaluation produced AUROC/AP per quality defect and fairness slices across skin-tone bins and capture channels. Best checkpoints were saved as `models/snapcheck\_quality.pt` and an ONNX export (planned) for deployment.

### Thresholding and Gating

Per-defect probability thresholds were selected from validation curves and persisted in `reports/diqa/thresholds.json`. Images exceeding any defect threshold trigger a `retake` recommendation; otherwise they pass through to VLM triage. Thresholds emphasized sensitivity to catastrophic failures (blur, obstruction), recognizing the retake burden as a tunable parameter.

### Triage Simulation Harness

We configured a Qwen2-VL-2B backend via Hugging Face, generating differential diagnoses and triage labels (reassurance, routine, urgent). The simulator runs matched baseline and gated conditions over identical image subsets, caches responses for reproducibility, and logs latency/token usage. Metrics include triage accuracy, urgent recall, urgent miss rate, and retake rate, with confusion matrices saved to `reports/triage/confusion/`.

### Statistical Analysis

Primary comparisons rely on descriptive statistics due to the pilot sample size. Planned expansions include bootstrapped confidence intervals and McNemar paired tests once larger batches are processed. Fairness gaps (>0.05 AUROC difference) are flagged for follow-up calibration.

## Results

### Quality Model Performance

The ViT-SnapCheck classifier achieved macro AUROC 0.968 and macro AP 0.890 on the held-out test set. Skin-tone fairness slices indicated AUROC 0.957 (bin 2), 0.972 (bin 3), and 0.967 (bin 4); precision for clinic-captured images was 0.877. Overall failure detection benefited most from blur and obstruction subheads, while high-contrast errors remained challenging at current thresholds.

### Triage Impact

In the 721-image evaluation, the ungated VLM produced triage accuracy 0.38 with an urgent miss rate of 0.35. The SnapCheck gate recommended retakes for 50% of cases. The gated pipeline delivered an urgent recall of 0.31 and an urgent miss rate of 0.69. Latency (mean 4.90 s) and token usage (~741 tokens) were identical across conditions, given the reuse of cached prompts when retakes were triggered.

### Failure Analysis

Inspection of confusion matrices and reasoning traces highlighted that urgent lesions filtered by the gate lacked immediate re-triage pathways. Without a retake image, the system defaults to recommending deferral, which inflates miss rate. Additionally, thresholds tuned for high sensitivity treat moderate contrast shifts as failures, contributing to the 70% retake rate. No catastrophic VLM hallucinations were observed in pass-through cases.

## Discussion

The early prototype establishes a reproducible toolkit for integrating DIQA with VLM triage, but the initial gating policy is not yet clinically acceptable. High retake rates risk overburdening patients, while urgent deferrals could delay care. These results underscore the need for calibrated thresholds, potentially governed by utility-aware optimization or reinforcement learning that trades off miss risk against retake burden.

Future iterations should (1) expand evaluation to hundreds of encounters, (2) incorporate real patient-uploaded data with manual quality labels, (3) study clinician-in-the-loop retake workflows, and (4) benchmark alternative DIQA architectures as ablations. Pairing automated retake prompts with guided capture instructions may convert deferrals into higher-quality resubmissions, improving overall safety.

### Limitations

Current findings rely on synthetic degradations and a small triage batch. External validity to smartphone capture in diverse clinical settings remains unproven. Latency estimates exclude real-time retake interactions. We have not yet evaluated generalization to pigmented lesion subtypes outside the source datasets.

## Conclusion

TeleDerm SnapCheck demonstrates that specialty-aware image quality gating can be operationalized alongside VLM triage. Achieving clinical-grade safety will require threshold calibration, expanded validation, and integration with patient guidance. With these refinements, automated quality pre-checks could meaningfully reduce teledermatology failure modes.

## Acknowledgments

We acknowledge open-source contributors to ISIC, HAM10000, Derm7pt, and the timm and Hugging Face communities.

## Data and Code Availability

All scripts, configurations, and checkpoints are available in this repository. Public datasets are accessible via their respective licenses; synthetic augmentations will be released under an open data agreement pending institutional approval.