Bugs in Citizen-Science Data: Robust Biodiversity AI Begins with Clean Images

Nikita Gavrilov
Gerard Schouten
Georgiana Manolache

Fontys University of Applied Sciences, ICT, The Netherlands

PROBLEM

Figure 1: Examples of iNaturalist citizen science images marked as "Research Grade"

- ❖ No criteria for image quality in *iNaturalist "Research Grade"* assessment
- ❖ Manually assessing the enormous volume of existing observations is infeasible

METHODOLOGY

To overcome the problem, we propose 2-step assessment:

1. **No-reference image quality (NR-IQA)**
2. **Semantic content evaluation with VLMs**

1. **NR-IQA** is preferred for its sensitivity to deviations from natural scene statistics, making it well-suited for noisy, large-scale datasets
2. **VLMs** automate manual assessments by detecting issues such as human presence, other taxa, or poor framing using structured prompts

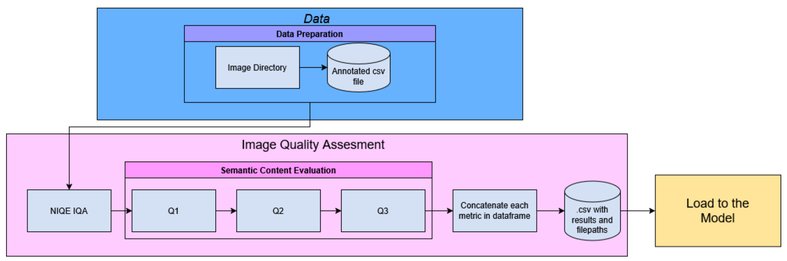

Figure 2: 2-step image quality assessment pipeline.

EMPIRICAL ANALYSIS

Figure 3: Example images of three visually similar plants species of different qualities. The last three columns showcase the image quality issues: (1) composition (blurry/angles), (2) human presence (hand, body), and (3) other species presence.

Data collection We select three visually similar plant species (see Figure 2). These species share overlapping habitats, similar floral structures, and frequently co-occur in ecosystems, making them a compelling test case. we randomly sample 600 images (200 per species) iNaturalist *"Research-Grade"* observations geographically restricted to Europe.

Pipeline setup We select NIQE algorithm (threshold=6, top 5% worst scores) and the Qwen2.5-VL-Instruct VLM (3B) or semantic filtering, using binary prompts for blurriness, human presence, and other species (see Table 1).

Experimental setup We evaluate BioCLIP (ViT-B/16) and BioTroveCLIP. We report standard metrics (e.g. Accuracy, Precision, Recall, F1), and confidence scores for both correct and incorrect predictions and analyze statistical significance with Z-tests, t-tests, and Levene's variance tests.

All experiments are conducted on a consumer desktop (AMD Ryzen 5 7600XT, 32GB RAM, Radeon RX 7600XT GPU, Python 3.12).

Table 1: Compositional reasoning with VLM prompt questions.

Quality measure		Prompt question
Composition	Q1	Does the image's blurriness or perspective prevent identification of the flower? Answer with 'Yes' or 'No' only.
Human present	Q2	Does this image show any humans or identifiable human body parts (including, but not limited to, faces, hands, fingers, arms, legs, torsos, or silhouettes)? Answer with 'Yes' or 'No' only.
Other species present	Q3	Does the image's blurriness or perspective prevent identification of the flower? Answer with 'Yes' or 'No' only.

RESULTS & DISCUSSION

Table 2: Mean accuracy (%) across three seeds.

Filtering condition	BioCLIP Acc. (%) ± SE	BioTroveCLIP Acc. (%) ± SE	# Images
Unfiltered	93.33 ± 0.76	85.22 ± 0.29	600
Higher Perceptual Quality (NIQE < 6)	93.60 ± 0.83	85.48 ± 0.35	576
Lower Perceptual Quality (NIQE > 6)	88.77 ± 2.91	79.77 ± 1.92	24
VLM: Blurry/Bad Composition	83.33 ± 16.6	70.24 ± 10.5	4
VLM: Human Presence	92.84 ± 0.25	80.99 ± 2.21	27
VLM: Other Taxa Present	91.67 ± 0.70	81.07 ± 0.31	83
VLM-Curated "Good Quality"	92.37 ± 0.29	81.54 ± 0.42	109
Combined "Ideal" (VLM Clean + NIQE < 6)	**93.69 ± 0.92**	**86.20 ± 0.48**	471
Combined "Worst-Case" (VLM Flagged + NIQE > 6)	88.89 ± 11.1	58.33 ± 12.7	3

Table 3: VLM assessments averaged across 3 seeds, N=600.

Attribute assessed	Response 'Yes' (Count)	Response 'Yes' (%)	Response 'No' (Count)	Response 'No' (%)
Composition	4.33	0.69%	594	99.28%
Human presence	27	4.50%	573	95.50%
Other species present	83	13.82%	517	86.14%

96% of images scored below the NIQE threshold of 6, indicating high perceptual quality. VLM-based filtering flagged 0.7% for poor composition, 4.5% for human presence, and 13.8% for presence of other taxa. BioCLIP consistently outperformed BioTroveCLIP, achieving 93.3% accuracy on unfiltered data versus 85.2%. Filtering had minimal impact on overall performance for both models, though extreme cases (e.g., poor NIQE and multiple taxa) reduced accuracy and confidence. *Leucanthemum vulgare* was frequently misclassified, especially under low-quality conditions, while *Bellis perennis* was consistently well-identified. Strict filtering sometimes decreased performance due to loss of useful context, suggesting both models, especially BioCLIP, are robust to moderate image imperfections.

Our findings demonstrate that while advanced vision-language models like BioCLIP and BioTroveCLIP show strong performance on biodiversity image classification, their behavior under filtered data conditions is nuanced. For high-performing models like BioCLIP, pre-filtering for ideal conditions (e.g., high NIQE scores or VLM-cleaned images) yields limited benefits, suggesting that performance bottlenecks stem more from inherent visual similarity between taxa than from image noise. Importantly, flower images used in this study are relatively high quality due to their ease of capture, so results may not generalize to more challenging taxa like insects.

FUTURE WORK

- ❖ Expand to more diverse taxa beyond plants
- ❖ Test pipeline on **dynamic subjects** (e.g., spiders, birds)
- ❖ Assess robustness to **motion blur**, occlusion, and framing issues
- ❖ Evaluate model performance on **small, fast-moving organisms**

https://github.com/wakizasher/iNaturalist_Benchmarking