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

Nikita Gavrilov, Gerard Schouten, Georgiana Manolache
Fontys University of Applied Sciences, Eindhoven, The Netherlands
Correspondence: g.manolache@fontys.nl

Abstract

Despite bold claims that AI will accelerate scientific discovery, domains like climate change research still face challenges in learning from real-world data. We propose a data preprocessing pipeline that addresses a key bottleneck in biodiversity monitoring: the lack of standardized image quality control in large-scale species datasets. As climate change drives shifts in ecosystems, accurate species identification is critical. Yet citizen science images, though rich in species diversity, are often noisy and inconsistent. We systematically filters such data using classical heuristics and Vision-Language Model (VLM)-based image quality assessment to detect poor composition, human presence, and multiple-species interference. Zero-shot benchmarks with state-of-the-art biodiversity fine-tuned foundation models on filtered datasets of visually similar plant species demonstrate that data quality significantly affects AI reliability. With this work, we highlight a core limitation in biodiversity AI and encourage broader exploration of quality-related bottlenecks in biodiversity monitoring. Code is available at the project website[1].

1 Introduction

Advancements in AI are set to play a pivotal role in biodiversity conservation and ecological management, especially as data from open citizen science platforms continues to grow. Among these, one of the most prominent sources of raw biodiversity data is iNaturalist [25], a platform that hosts in-situ observations across an exceptionally broad taxonomic spectrum. Each observation typically includes one or more photographs, the date and time of the encounter, geographic coordinates, and optional metadata such as life stage or observed behavior. This rich set of user-generated observations has emerged as a valuable asset for the development, evaluation, and deployment of machine learning (ML) systems. Numerous studies already demonstrate the utility of iNaturalist's *Research Grade* data for species identification [47, 51, 53]. To qualify as *Research Grade*, two or more experienced iNaturalist community members—naturalists, biologists, or citizen scientists—must agree on the species label assigned to an observation [26] (see Appendix B for details). Despite its immense value, iNaturalist data is not immediately AI-ready. Crucially, iNaturalist applies no criteria for image quality in its assessment. As illustrated in Figure 1, *Research Grade* observations may include harvested specimens (top-left), partial views or non-representative angles (bottom-left), or images that are heavily blurred or poorly exposed (right). While the *Research Grade* label implies taxonomic confidence, the corresponding images often vary dramatically in visual fidelity—posing a major challenge for biodiversity AI systems that depend on consistent, high-quality visual input. Both researchers and community members have called for incorporating image quality into the *Research Grade* standard [33]. However, manually assessing the enormous volume of existing observations is infeasible, underlining the need for automated approaches to filter low-quality images in biodiversity datasets.

[1] https://github.com/wakizasher/iNaturalist_Benchmarking

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Tackling Climate Change with Machine Learning.

Figure 1: Examples of iNaturalist *Research Grade* low-quality images associated with *Bellis perennis*: (top-left) harvested; (bottom-left) non-representative; (right) blurry.

In this paper, we introduce BIOCLEANSE, a data preprocessing pipeline designed to improve the visual quality of biodiversity datasets for AI applications. BIOCLEANSE extends iNaturalist's *Research Grade* data with three complementary image quality filters: (1) poor composition (e.g., blur, under/overexposure), (2) human presence (e.g., hands, body), and (3) multiple-species interference. To detect these, we combine traditional image quality heuristics with modern Vision-Language Model (VLM)-based image quality assessment (IQA). We evaluate BIOCLEANSE on a zero-shot benchmark using state-of-the-art biodiversity biodiversity fine-tuned models applied to three visually similar plant species. Our results show that automated image curation significantly improves model reliability and scientific inference in biodiversity AI. Our main contributions are:

- BIOCLEANSE: a ready-to-use, open-source image quality preprocessing pipeline for biodiversity datasets;
- A new benchmark demonstrating the impact of image quality on species identification for visually similar plant species.

The remainder of the paper is organized as follows: Section 2 introduces the BIOCLEANSE pipeline and reviews related work. Section 3 presents the benchmark design, experimental results, and discussion. Section 4 outlines our conclusions and future directions. Additional information, implementation details, extended results, and data access instructions are provided in the Appendix.

2 BioCleanse vs. other IQA related work

Image quality assessment (IQA) methods vary across domains, as summarized in Table 1 and detailed in Appendix D. In biodiversity, subjective approaches such as manual annotation have been used [36], while medical imaging often relies on objective [15] or full-reference techniques like PSNR and SSIM [5]. No-reference methods, such as pyBRISQUE [4] and UIQA [11], estimate image quality using only the input image—without requiring a reference—making them especially well-suited for uncontrolled, real-world datasets like those from citizen science platforms. These methods typically rely on statistical patterns found in natural scenes, and are efficient for large-scale automated filtering. More recently, Vision-Language Models (VLMs) like IQAGPT [12] and DepictQA [56] have enabled semantic-aware image quality filtering, though their application in biodiversity remains limited.

To address this, we introduce BIOCLEANSE, a modular pipeline for curating high-quality biodiversity image datasets through three stages: (1) data preparation, (2) no-reference IQA using NIQE [39], and (3) semantic filtering via Vision-Language Models (VLMs). NIQE is preferred for its sensitivity to deviations from natural scene statistics, making it well-suited for noisy, large-scale datasets [4, 11]. VLMs automate manual assessments [36] by detecting issues such as human presence, other taxa, or poor framing using structured prompts (see Table 2). The pipeline is model-agnostic and supports any VLMs [3, 31, 34] (see Appendix E).

Table 1: IQA-related work.

Technique	Domain	Study	IQA	Target task
Subjective	Biodiversity	(2023) Luccioni et al. [36]	Manual assessment	Biodiversity reporting
Objective	Medical	(2016) Chow et al. [15]	IQA on ultrasonic images	Diagnostic imaging
Full-reference	Medical	(2024) Breger et al. [5]	PSNR and SSIM	Tumor segmentation
No-reference	Biodiversity	(2022) Billotte [4]	pyBRISQUE	Species classification
No-reference	Biodiversity	(2023) Chen et al. [11]	UIQA	Underwater object recognition
VLM	Medical	(2023) Chen et al. [12]	IQAGPT	Radiology report generation
VLM	General	(2024) You et al. [56]	DepictQA	General-purpose image retrieval

Table 2: Quality measures employed with VLM prompt questions. Questions are according to the manual assessment of image composition from Luccioni et al. [36].

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.

3 Benchmarks

We employ a quantitative experimental design to evaluate how iNaturalist image quality impacts the zero-shot classification performance of state-of-the-art biodiversity CLIP models. Central to this process is the use of the BIOCLEANSE pipeline, which curates image subsets based on both statistical and semantic quality assessments before benchmarking. All ethical concerns are addressed in Appendix A.

Data collection We focus on three visually similar plant species as illustrated in Figure 2 (see more datasets and benchmarks in Appendix C). These species share overlapping habitats, similar floral structures, and frequently co-occur in ecosystems, making them a compelling test case for evaluating how image quality influences model performance in visually ambiguous scenarios [44]. From approximately 114K *Research-Grade* iNaturalist observations geographically restricted to Europe, we randomly sample 600 images (200 per species).

Curation pipeline Image quality is assessed using the NIQE algorithm (threshold=6, top 5% worst scores) and the Qwen2.5-VL-Instruct VLM (3B) [49] for semantic filtering, using binary prompts for blurriness, human presence, and other species (see Table 2). Based on these assessments, we compose filtered subsets as shown in Table 4.

Experimental setup We evaluate two state-of-the-art biodiversity fine-tuned foundation models BIOCLIP (ViT-B/16) and BIOTROVECLIP (see Appendix G). We report standard metrics (e.g.

Figure 2: Example images of *Bellis perennis*, *Matricaria chamomilla*, *Leucanthemum vulgare* of different qualities. The last three columns showcase the BIOCLEANSE image quality issues: (1) composition (blurry/angles), (2) human presence (hand, body), and (3) other species presence.

Table 3: Qwen VLM assessments for image attributes, averaged over 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%

Table 4: 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

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).

Results Using the BioCleanse pipeline, 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 (see Table 3 and detailed results in Appendix H). As shown in Table 4, 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 BioTrove-CLIP 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. In contrast, BioTroveCLIP showed greater sensitivity to difficult conditions and a tendency toward overconfident misclassifications, highlighting differences in robustness and confidence calibration. 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. Despite limitations, automated pipelines like BioCleanse offer a scalable and consistent alternative to subjective manual filtering, making them a valuable tool for preparing large-scale citizen science datasets for biodiversity AI.

4 Conclusion

We presented BioCleanse, an open-source preprocessing pipeline for filtering biodiversity images using both NIQE and VLMs IQA, enabling automated removal of common issues such as poor composition, human presence, and multiple taxa. Applied to iNaturalist's *Research Grade* data, and evaluated on three visually similar plant species using two state-of-the-art CLIP-style foundation models, BioCleanse proved valuable not for uniformly improving accuracy, but for exposing model-specific sensitivities under real-world conditions. However, our study is limited by its narrow focus on flowering plants in similar visual conditions, a relatively small sample size (600 images), and the exclusion of other life stages or diverse habitats (e.g., underwater environments), which may affect generalizability and pipeline robustness. We invite the biodiversity and machine learning communities to employ BioCleanse on additional species, building more robust, generalizable AI-ready biodiversity data, and ultimately accelerate ecological research and conservation efforts.

References

[1] Jong-Won Baek, Jung-Il Kim, and Chang-Bae Kim. "Deep learning-based image classification of sea turtles using object detection and instance segmentation models". In: *PloS one* 19.11 (2024), e0313323.

[2] Jinze Bai et al. "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond". In: (2023).

[3] Yuxiao Bai et al. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. 2023. arXiv: 2311.12032 [cs.CV].

[4] Jackie Billotte. "A Pipeline for Assessing the Quality of Images and Metadata from Crowd-Sourced Databases". In: *bioRxiv : the preprint server for biology* (2022). DOI: 10.1101/2022.04.29.490112. eprint: https://www.biorxiv.org/content/early/2022/11/29/2022.04.29.490112.full.pdf.

[5] Anna Breger et al. "A study of why we need to reassess full reference image quality assessment with medical images". In: *Journal of imaging informatics in medicine* (2024). URL: https://api.semanticscholar.org/CorpusID:270094590.

[6] BugGuide.Net contributors. *BugGuide.Net: Identification, Images, & Information for Insects, Spiders & Their Kin for the United States Canada*. https://bugguide.net/node/view/15740. Accessed: 2025-04-22. 2025. URL: https://bugguide.net/node/view/15740.

[7] Corey T. Callaghan et al. "The Benefits of Contributing to the Citizen Science Platform iNaturalist as an Identifier". In: *PLOS Biology* 20.11 (2022-11), pp. 1–6. DOI: 10.1371/journal.pbio.3001843. URL: https://doi.org/10.1371/journal.pbio.3001843.

[8] Cj Campbell et al. "Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science". In: *BioScience* (July 2023), biad051. DOI: 10.1093/biosci/biad051.

[9] Yuqin Cao et al. "Subjective and Objective Audio-Visual Quality Assessment for User Generated Content". In: *IEEE Transactions on Image Processing* 32 (2023), pp. 3847–3861. DOI: 10.1109/TIP.2023.3290528.

[10] Zhong Cao et al. "Fine-grained image classification on bats using VGG16-CBAM: a practical example with 7 horseshoe bats taxa (CHIROPTERA: Rhinolophidae: Rhinolophus) from Southern China". In: *Frontiers in Zoology* 21.1 (2024), p. 10.

[11] Tianhai Chen et al. *Underwater Image Quality Assessment method basedon Color Space Multi-feature Fusion*. Aug. 2023. DOI: 10.21203/rs.3.rs-3226222/v1.

[12] Zhihao Chen et al. *IQAGPT: Image Quality Assessment with Vision-language and ChatGPT Models*. 2023. arXiv: 2312.15663 [cs.CV]. URL: https://arxiv.org/abs/2312.15663.

[13] Shivani Chiranjeevi et al. "Deep learning powered real-time identification of insects using citizen science data". In: *arXiv preprint arXiv:2306.02507* (2023).

[14] Min Goo Choi, Jung Hoon Jung, and Jae Wook Jeon. "No-Reference Image Quality Assessment using Blur and Noise". In: *International Journal of Electrical and Computer Engineering* 3 (2009), pp. 184–188. URL: https://api.semanticscholar.org/CorpusID:10894292.

[15] Li Sze Chow and Raveendran Paramesran. "Review of Medical Image Quality Assessment". In: *Biomedical Signal Processing and Control* 27 (2016), pp. 145–154. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2016.02.006. URL: https://www.sciencedirect.com/science/article/pii/S1746809416300180.

[16] Copernicus Global Land Service. *S2Maps – Sentinel-2 Cloudless Maps*. Accessed: 2025-02-29. 2023. URL: https://s2maps.eu.

[17] eBird. *eBird: An online database of bird distribution and abundance*. https://www.ebird.org. Cornell Lab of Ornithology, Ithaca, New York. Accessed: April 22, 2025. 2021.

[18] Encyclopedia of Life. *Encyclopedia of Life*. Accessed: 2025-02-29. 2025. URL: https://eol.org.

[19] Stephen E Fick and Robert J Hijmans. "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12 (2017), pp. 4302–4315. DOI: 10.1002/joc.5086.

[20] M Foellmer. *Harnessing the full potential of iNaturalist and other databases*. 2022.

[21] Zahra Gharaee et al. “A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 43593–43619.

[22] gus$_{f}$*ordyce*. *Tips for taking photographs (important for plants!)* 2024. URL: https://www.inaturalist.org/posts/95047-tips-for-taking-photographs-important-for-plants.

[23] Rich Hatfield, Jeremy Kerr, and Max Larrivée. *Xerces Society - Bumble Bee Watch*. https://www.bumblebeewatch.org/. Accessed: 2025-04-22. 2024. URL: https://doi.org/10.15468/t4rau8.

[24] *How to Make an Observation*. https://help.inaturalist.org/en/support/solutions/articles/151000192921-how-to-make-an-observation. Accessed: 2025-04-17.

[25] iNaturalist. https://www.inaturalist.org/. Accessed via inaturalist.org on 2025-03-26. 2025.

[26] iNaturalist. *What is the Data Quality Assessment and how do observations qualify to become "Research Grade"?* https://shorturl.at/15lQa. Accessed via help.inaturalist.org on 2025-03-26. 2025.

[27] Jung-Il Kim, Jong-Won Baek, and Chang-Bae Kim. “Hierarchical image classification using transfer learning to improve deep learning model performance for amazon parrots”. In: *Scientific Reports* 15.1 (2025), p. 3790.

[28] Nikita Kisel et al. *Flaws of ImageNet, Computer Vision’s Favourite Dataset*. 2024. arXiv: 2412.00076 [cs.CV]. URL: https://arxiv.org/abs/2412.00076.

[29] Shujun Lang et al. “A Full-Reference Image Quality Assessment Method via Deep Meta-Learning and Conformer”. In: *IEEE Transactions on Broadcasting* 70.1 (2024), pp. 316–324. DOI: 10.1109/TBC.2023.3308349.

[30] A. Lemos, W. Caminhas, and F. Gomide. “Multivariable Gaussian Evolving Fuzzy Modeling System”. In: *IEEE Transactions on Fuzzy Systems* 19 (2011), pp. 91–104. DOI: 10.1109/TFUZZ.2010.2087381.

[31] Haotian Liu et al. “Visual Instruction Tuning”. In: *arXiv preprint arXiv:2304.08485*. 2023.

[32] Haotian Liu et al. “Visual Instruction Tuning”. In: *ArXiv* abs/2304.08485 (2023). DOI: 10.48550/arXiv.2304.08485.

[33] Eduard López-Guillén et al. “Strengths and challenges of using iNaturalist in plant research with focus on data quality”. In: *Diversity* 16.1 (2024), p. 42.

[34] Haoyu Lu et al. *DeepSeek-VL: Towards Real-World Vision-Language Understanding*. 2024. arXiv: 2403.05525 [cs.AI].

[35] Alexandra Sasha Luccioni and David Rolnick. *Bugs in the Data: How ImageNet Misrepresents Biodiversity*. 2022. arXiv: 2208.11695 [cs.CV]. URL: https://arxiv.org/abs/2208.11695.

[36] Alexandra Sasha Luccioni and David Rolnick. “Bugs in the data: How ImageNet misrepresents biodiversity”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 14382–14390.

[37] mickley. *Creating High Quality iNaturalist Observations*. 2024. URL: https://www.inaturalist.org/posts/80155-creating-high-quality-inaturalist-observations.

[38] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-Reference Image Quality Assessment in the Spatial Domain”. In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708. DOI: 10.1109/TIP.2012.2214050.

[39] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. “Making a “Completely Blind” Image Quality Analyzer”. In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212. DOI: 10.1109/LSP.2012.2227726.

[40] Catarina Pinho et al. “Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: Podarcis) as a case study”. In: *Zoological Journal of the Linnean Society* 198.1 (Oct. 2022), pp. 184–201. ISSN: 0024-4082. DOI: https://doi.org/10.1093/zoolinnean/zlac08.

[41] Ashirbani Saha and Q. M. Jonathan Wu. "Full-reference image quality assessment by combining global and local distortion measures". In: *CoRR* abs/1412.5488 (2014). arXiv: 1412.5488. URL: http://arxiv.org/abs/1412.5488.

[42] Srikumar Sastry et al. "TaxaBind: A unified embedding space for ecological applications". In: *arXiv preprint arXiv:2411.00683* (2024).

[43] Gerard Schouten, Bas S. H. T. Michielsen, and Barbara Gravendeel. "Data-centric AI approach for automated wildflower monitoring". In: *PLOS ONE* 19.9 (Sept. 2024). DOI: 10.1371/journal.pone.0302958.

[44] Gerard Schouten and Bart Wernaart. *Moral Design and Green Technology*. Leiden, The Netherlands: Wageningen Academic, 2025. ISBN: 978-90-04-73077-9. DOI: 10.1163/9789004730779. URL: https://brill.com/view/title/71071.

[45] Brian J Spiesman et al. "Assessing the potential for deep learning and computer vision to identify bumble bee species from images". In: *Scientific reports* 11.1 (2021), p. 7580.

[46] stevem4560. *Photograph quality and image selection*. 2024. URL: https://www.inaturalist.org/posts/95047-tips-for-taking-photographs-important-for-plants.

[47] Samuel Stevens et al. "Bioclip: A vision foundation model for the tree of life". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 19412–19424.

[48] D. Swainson Sujana, D. Peter Augustine, and D. Sheefa Ruby Grace. "Full Reference Image Quality Assessment (FR-IQA) of Pre-processed Structural Magnetic Resonance Images". In: *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*. Vol. 1. 2024, pp. 1–5. DOI: 10.1109/InC460750.2024.10649151.

[49] Qwen Team. *Qwen2.5-VL-3B-Instruct: An Edge-Scale Vision-Language Model*. Instruction-tuned 3B model with image/video and agentic capabilities. 2025. URL: https://huggingface.co/Qwen/Qwen2.5%E2%80%91VL%E2%80%913B%E2%80%91Instruct.

[50] Grant Van Horn et al. "Benchmarking representation learning for natural world image collections". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12884–12893.

[51] Edward Vendrow et al. "INQUIRE: A Natural World Text-to-Image Retrieval Benchmark". In: *NeurIPS* (2024).

[52] *What is the Data Quality Assessment and how do observations qualify to become "Research Grade"?* https://help.inaturalist.org/en/support/solutions/articles/151000169936-what-is-the-data-quality-assessment-and-how-do-observations-qualify-to-become-research-grade-. Accessed: 2024-06-30.

[53] Chih-Hsuan Yang et al. "BioTrove: A Large Curated Image Dataset Enabling AI for Biodiversity". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 102101–102120.

[54] Yuxuan Yang, Zhichun Lei, and Changlu Li. "No-Reference Image Quality Assessment Combining Swin-Transformer and Natural Scene Statistics". In: *Sensors* 24.5221 (16 2024). ISSN: 1424-8220. DOI: 10.3390/s24165221. URL: https://www.mdpi.com/1424-8220/24/16/5221.

[55] Hui Wu Yang Yang Chang Liu and Dingguo Yu. "A quality assessment algorithm for no-reference images based on transferlearning". In: *Peer Computer Science* (2025).

[56] Zhiyuan You et al. *Depicting Beyond Scores: Advancing Image Quality Assessment through Multi-modal Language Models*. 2024. arXiv: 2312.08962 [cs.CV]. URL: https://arxiv.org/abs/2312.08962.

[57] Zhuobin Yuan et al. "No- Reference Image Quality Assessment for Intelligent Sensing Applications". In: *NAECON 2024 - IEEE National Aerospace and Electronics Conference* (2024), pp. 185–189. URL: https://api.semanticscholar.org/CorpusID:272722267.

[58] Lin Zhang, Lei Zhang, and Alan C. Bovik. "A Feature-Enriched Completely Blind Image Quality Evaluator". In: *IEEE Transactions on Image Processing* 24.8 (2015), pp. 2579–2591. DOI: 10.1109/TIP.2015.2426416.

A Ethical concerns

License Only images released under a Creative Commons (CC) license are included, ensuring that the dataset is openly available for public research and non-commercial use.

Offensive content Although iNaturalist rely on community contributions and expert moderation to uphold data quality and appropriateness, we acknowledge that our dataset may occasionally feature unsettling content such as predation, roadkill, or other scenes from the raw realities of nature. Rather than sanitize these instances, we retain them to preserve the ecological authenticity of species interactions and the unfiltered diversity of real-world observations.

Privacy We strictly exclude all personally identifiable information (PII) from the metadata associated with the dataset, ensuring that fields such as observer names and email addresses are removed. However, we acknowledge that in rare cases, PII may still be visible within the image content itself; for example, faces of individuals, vehicle license plates, distinctive property features, or GPS location markers embedded in the media. While such occurrences are unintended and infrequent, users of the dataset should be aware of this residual risk when analyzing or displaying images.

Responsible use Models trained on this data should not be used for unlawful wildlife tracking or poaching; we provide the data to support conservation efforts and ecological research.

Geoprivacy We do not include any geolocation metadata in our dataset. This aligns with best practices for safeguarding both ecological integrity and user privacy.

Large Vision-Language models (LVLMs) privacy leakage LVLMs are trained on broad, web-scale data and may have inadvertently memorized sensitive content—such as faces, license plates, or location indicators—embedded in images. Although our pipeline does not involve fine-tuning and uses only publicly accessible images (e.g., iNaturalist), we acknowledge that using such models can still pose residual risks. Specifically, VLMs may surface private or identifying visual features during inference, even if that information was not explicitly included in the dataset.

B Quality assurance in iNaturalist

Observations on iNaturalist are categorized into "Casual", "Needs ID", or "Research Grade", reflecting varying levels of data quality and the progression through the community identification process. The "Research Grade" designation signifies that an observation has achieved a level of data quality and taxonomic consensus deemed suitable for scientific research. These observations are frequently incorporated into major online biodiversity databases, such as the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia, underscoring their utility in a broader scientific context. The term "Research Grade" is widely recognized and utilized within the scientific community and among citizen scientists, despite ongoing discussions regarding its precise implications for absolute data accuracy. It functions as a key filter for researchers seeking higher quality citizen science data [52].

Before an observation can achieve "Research Grade" status, it must be classified as "Verifiable". This initial classification establishes a baseline level of data quality essential for subsequent community assessment. A verifiable observation must include specific metadata to provide sufficent context for identificaiton and validation. These include: a recorded date of observation, georeferenced coordinates (latitude and logitude) and accompanying photos or sound recordings as empirical evidence of the observed organism. While an observation can be created without media, it is a strict requiremnt for verifiability and progression to Research Grade.

The enforcement of these basic data completeness requirements is a fundamental building block of iNaturalist's quality control process. Without a date, location, and supporting media, an observation cannot enter the "Needs ID" queue, let alone progress to "Research Grade". This structured data input ensures that observations provide sufficient contextual information for accurate identification and validation by the community. This initial baseline of data quality is crucial for the subsequent steps of the Research Grade process and for the utility of the data when incorporated into external databases like GBIF. It is also important for users to provide clear, multiple photos from different angles to aid identification, as poor quality or insufficient imagery can hinder an observation's progression to Research Grade [24].

Following initial verifiability, observations enter the "Needs ID" status, where they await input from the iNaturalist community. This phase is central to the platform's crowd sourced quality control. Registered users actively participate by adding identifications to observations, aiming to confirm or refine taxonomic assignments. An observation typically progresses to Research Grade when a robust consensus is achieved among identifiers. Specifically, more than two-thirds (2/3) of identifiers must agree on a species-level identification or a finer taxonomic resolution [26]. This rule implies a minimum of two agreeing identifications for a secies-level ID to be considered for Research Grade.

This 2/3 agreement rule serves as a core mechanism for quality control, leveraging the collective knowledge of the user base. However, the effectiveness of this crowdsourced expertise is subject to certain limitations. While the system aims for broad consensus, the quality of this consensus is heavily reliant on a relatively small cohort of expert or highly active identifiers. A significant portion of iNaturalist's 2.5 million users primarily contribute observations, with less than 1% focusing solely on identifications, and only 7% engaging in both [7]. Furthermore, a small subset of highly active users is responsible for the vast majority of identifications. This dependence on a concentrated group of identifiers suggest a potential vulnerability: if these key individuals exhibit biases or are less active in specific taxonomic groups or geographical regions, the accuracy and speed of Research Grade assignment could be affected. This indicates that "Research Grade" primarily reflects community agreement, not an absolute guarantee of expert validation for every individual observation, although experienced identifiers do vet most records.

It is important to understand that while "Research Grade" is explicitly linked to scientific use, it does not guarantee absolute accuracy or flawlessness. Multiple sources indicate that the quality grade itself may be an inadequate proxy for accuracy, emphasizing the continued importance of expert verification. This understanding is critical for researchers, as it implies that Research Grade observations, while having undergone significant vetting, are not exempt from further quality control or expert veiew in specific research applications. The designation primarily indicates that an observation has met iNaturalist's internal criteria for verifiability and community consensus, making it potentially suitable for research, but not necessarily validated for every specific scientific application. This encourages critical engagement with the data rather than uncritical acceptance.

B.1 iNaturalist Image Quality

Artificial intelligence, specifically computer vision (CV), plays an increasingly integral role in the iNaturalist identificaiton process, significantly impacting efficency and accuracy. iNaturalist incorporates an automated species identification tool that provides identification suggestions to users based on uploaded photos or sounds. The CV model is trained on a vast dataset of existing iNaturalist observations, comprising photos and their associated taxa. It generates a ranked list of potential taxa, with the most probable suggestion presented at the top. If the model is less certain about a specific species, it often provides a broader taxonomic suggestion, such as genus or family. The CV also leverages location data, prioritizing matches expected to occur nearby, although it can also perform global searches if local matches are not strong.

The integration of computer vision significantly streamlines the identification process and has contributed to a rapid decrease in the time observations take to reach Research Grade [8]. However, this technological acceleration also presents potential challenges. Concerns have been raised regarding the accuracy of CV suggestions, with some users reporting instances where suggestions for certain species appear to be "getting worse". There is also a recognized risk that new or inexperienced users might "blindly agree" with CV suggestions without critical assessment, potentially leading to inaccurate data entering the Research Grade pool. This situation highlights a trade-off between the efficiency gained through automation and the maintenance of data accuracy, underscoring the continued necessity of human verification and "course correction" by experienced identifiers. The quality of Research Grade observations is particularly crucial, as these data are used to train the CV model, creating a feedback loop where incorrectly identified Research Grade data can negatively impact the CV's future performance.

The dual nature of iNaturalist's "Research Grade" status—a community consensus metric that does not guarantee absolute scientific accuracy, coupled with the known limitations of its internal computer vision system and the inherent variability of citizen science data—creates a significant need for external, objective image quality assessment. This need extends beyond merely making pictures "look good"; it is fundamentally about enhancing the scientific utility and reliability of the vast dataset.

This could involve identifying observations that are visually ambiguous despite having achieved community consensus, or conversely, prioritizing "Needs ID" observations that possess high visual quality but are awaiting expert taxonomic review.

C Existing biodiversity datasets & benchmarks

Table 5: Biodiversity datasets using *Research Grade* iNaturalist observations as main data source.

Datasets	Images	Species	Annotations	Source
BIOTROVE [53]	161.9M	366.6K	common, scientific terms, taxonomic hierarchies	iNaturalist
TREEOFLIFE-10M [47]	10.4M	454.1K	common, scientific terms, taxonomic hierarchies	iNaturalist, Encyclopedia of Life (EOL)[18], BIOSCAN-1M[21]
INAT2024 [51]	4.9M	9K	common, scientific term, taxonomic hierarchies, location	iNaturalist
TAXABIND-8K [42]	8.8K	2.2K	common, scientific term, taxonomic hierarchies, location, environmental features, audio recordings, satellite imagery	iNaturalist, iNat2021[50], Santinel-2[16], WorldClim-2.1[19]

Table 6: Existing benchmarks of visually difficult to distinguish species. Our new benchmark is described in Section 3 Data collection.

Taxon	Benchmark	Images	Species	Annotations	Source
Aves	**AMAZON PARROTS** [27]	14K	35	scientific terms	iNaturalist, eBird [17], Google Images
Insecta	**BUMBLE BEES** [45] (not publicly available)	89K	36	scientific terms	iNaturalist, Bumble Bee Watch [23], BugGuide [6]
	CONFOUNDING SPECIES [13] (not publicly available)	100	10	scientific term	iNaturalist
Mammalia	**CHIROPTERA RHINOLOPHIDAE RHINOLOPHUS** [10]	293	7	scientific terms	personal collection during field surveys
Reptilia	**SEA TURTLES** [1] (not publicly available)	6.9K	36	vernacular, scientific terms	Internet
	SQUAMATA LACERTIDAE PODARCIS [40]	4.0K	9	scientific terms	personal collection during field surveys

D IQA related work

D.1 Subjective and objective IQA

Subjective and objective IQA [9] plays a critical role in various domains, including biodiversity monitoring and medical imaging, where accurate interpretation of visual data is essential. In biodiversity research, subjective IQA often involves human observers evaluating the perceptual quality of images used for species identification or habitat analysis. This approach can be particularly useful when automated metrics fail to capture nuances that are visually clear to experts. For example, [36] critically examines ImageNet-1k [28]. This paper specifically looks at 13K wild animal images from 269 classes. The key aspect of this study is that they involved expert annotators; for instance, the images of primates were annotated by two postdoctoral researchers in primate biology and one zoo keeper specializing in primates [35]. The interest in objective image quality assessment (IQA) has been growing at an accelerated pace over the past decade. This field is dedicated to developing automatic methods that can predict the subjective quality of image, effectively mimicking human perception without requiring human intervention for every evaluation. In the field of medicine, objective IQA is used to evaluate the performance of images captured by medical equipment such as MRI, CT and X-ray systems. The work of Li Sze Chow suggest the usage of Objective IQA to

ensure that diagnostic images maintain sufficient clarity and detail for accurate interpretation [15]. In biodiversity context the objective IQA is used for assessment of images that are suffer from unique types of degradation such as images that are made underwater. The work of Tianhai Chen introduce a specialized, intelligent system that evaluates underwater image quality by thinking about how humans see those images and by looking for very specific features that reveal common underwater problems such as murkiness and color distortion [11].

D.2 Full-reference IQA (FR-IQA)

Full-reference image quality assessment (FR-IQA) [29] is a widely utilized approach in image processing that evaluates the perceptual quality of a distorted or test image by comparing it with a reference image, which is assumed to be undistorted [41]. This method is particularly relevant in fields where high fidelity and accuracy are essential, such as medical imaging. FR-IQA techniques often rely on metrics like Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and more advanced models that incorporate low-level and high-level image. In the medical context, FR-IQA can be used to measure the effectiveness of different reprocessing steps. In the work of D. Swainson Sujana a perfect version of image ("ground truth") was used to improve the quality of medical images, specifically sMRI (structural Magnetic Resonance Imaging), before they are used to train deep learning AI models for diagnosis [48]. In their work they state that while deep learning AI is great for medical diagnosis, especially with radiology images, they need huge amount of high-quality data to learn effectively, that is why additional preprocessing steps are necessary.

On the other hand, The study of Anna Breger warns that blindly applying full-reference image quality techniques and can lead to flawed conclusions [5]. The research provides a comprehensive collection of examples where PSNR and SSIM fail to accurately assess the quality of various real-world medical images.

In the context of citizen science biodiversity data FR-IQA methods such as PSNR and SSIM, face extreme limitations and are near-totaly inapplicable. The primary limitation is the fundamental absence of reference images. Citizen science platforms like iNaturalist accumulate images captured by valunteers in uncontrolled, real-world environments [55]. Each photograph represents a unique instance of an organism at a specific moment in time, under unique lighting conditions, with a unique background, and subject to various unpredictable factors like motion, focus, and occlusion. For such "in-the-wild" captures, a "pristine" or "original" reference version of that exact observation simply does not exist and cannot be created retrospectively. Given these profound limitations, the focus for assessing citizen science biodiversity image quality must necessarily shift to paradigms that do not require a reference image such as no-reference image quality assessment.

D.3 No-reference IQA (NR-IQA)

On the other hand, no-reference image quality assessment (NR-IQA) offers an alternative when a pristine reference image is unavailable. It is also commonly referred to as Blind IQA (BIQA), involving methods that evaluate the quality of an image using only the information conained within the image itself, without access to any corresponding "ground truth" image [54]. One of the most popular and widely used NR-IQA models is Natural Scene Statistics (NSS). They are based on the observation that high-quality natural images exhibit certain statistical properties. Distoritons are assumed to dirupt these inherent statistics in predictable ways [54].

NR-IQA NSS base models such as NIQE [39] and BRISQUE [38] are common for different computer vision tasks - for example, segmentation, classification, and object detection in intelligent sensing systems [57]. Moreover, digital photography and camera applications benefit from the NR-IQA techniques. In the work of Min Goo Choi, the researchers focused on two common image problems: blur and noise. The study's experiment showed that their proposed method has a high correlation with human judgment and requires very little computational effort, making it efficient to use [14].

D.3.1 Natural image quality evaluator (NIQE)

NIQE aims to be "opinion-unaware," meaning it does not require training on human-rated distorted images. It extracts a set of local features (derived from image patches) that model salient statistical properties found in natural scenes. These features from a test image are fitted to a Multivariate Gaussian (MVG) model [30]. The quality of the test image is then determined by measuring the

distance (e.g., Mahalanobis distance) between its MVG model and a pre-learned MVG model derived from a corpus of high-quality, pristine natural images. A smaller NIQE score typically indicates better perceptual quality [58].

D.4 Visual language models (VLMs)

Vision Language Models (VLMs) such as Qwen-VL [2] and LLaVA [32] have emerged as powerful tools for image quality analysis across a range of domains, including biodiversity and medicine. These models combine the capabilities of vision encoders with large language models (LLMs), enabling them to interpret visual content while generating descriptive, linguistically rich responses. In the medical domain, the research of Zhihao Chen proposes a novel AI system called IQAGPT that leverages the power of large vision-language models and language models to automatically assess the quality of medical CT images. It combines a VLM that "captions" image quality with ChatGPT to generate both numerical quality scores and detailed textual reports [12].

In broader applications, VLMs have been adapted for general-purpose image quality assessment using datasets that span multiple domains. For example, DepictQA leverages Multi-modal Large Language Models (MLLMs), which can understand both images and text. It explains what is wrong with the image (e.g., "blurry background," "noisy shadows"), and it can explain why one image might be better or worse than another, mimicking human reasoning. The results showed that DepictQA performs better than traditional score-based IQA methods on several benchmark tests [56].

D.5 iNaturalist IQA

iNaturalist employs several strategies to ensure the quality of uploaded images, which are crucial for accurate species recognition and research-grade observations. One of the primary methods involves encouraging users to submit high-quality photographs that are well-lit and sharply focused on the subject, particularly important for distinguishing morphological features in plants and animals [37]. The platform also provides guidelines for users on best practices for taking and submitting photos, including avoiding blurry images and ensuring that key diagnostic features are visible [22][46]. These user-oriented recommendations serve as a preliminary filter for image quality before upload. The work of Sarah J. Ackland identifies the problem of data quality on iNaturalist platform. One of the issues is low quality media: photos or other uploaded evidence might be unclear or insufficient. The results underscore the vital role of expert verification when using citizen science data for media quality assessment. However, there are some limitations and challenges. For example, it takes a lot of effort to manually check each record. Moreover, the individual doing the verification might introduce their own subjective biases [8].

However, with enormous amounts of natural data images the platform lack the transparent quality control. Jackie Billotte attempted to create more reliable iNaturalist public data for ecological research and education [20]. His work provides a protocol for the quality assessment of downloaded observations from iNaturalist. The image quality assessment is performed with NR-IQA BRISQUE algorithm on the dataset containing spiders (Araneae) [4].

While existing NR-IQA methods and VLMs are valuable tools for accessing image quality, their application to biodiversity citizen science data remains unexplored. The BIOCLEANSE pipeline addresses these gaps by integrating NR-IQA for quantitative quality control with VLM-driven semantic content evaluation, creating a scalable, reference-free framework. This hybrid approach leverages the strengths of both paradigms while mitigating their individual limitations.

E BioCleanse pipeline

The BIOCLEANSE pipeline comprises sequential stages: (1) data preparation, (2) quantitative quality assessment, and (3) semantic content filtering (see Figure 3).

E.1 Quantitative quality assessment

Our BIOCLEANSE pipeline uses a NR-IQA technique. Natural, high-quality images of nature, like flowers, follow predictable statistical patters. When an image is distorted (blurry, noise, or

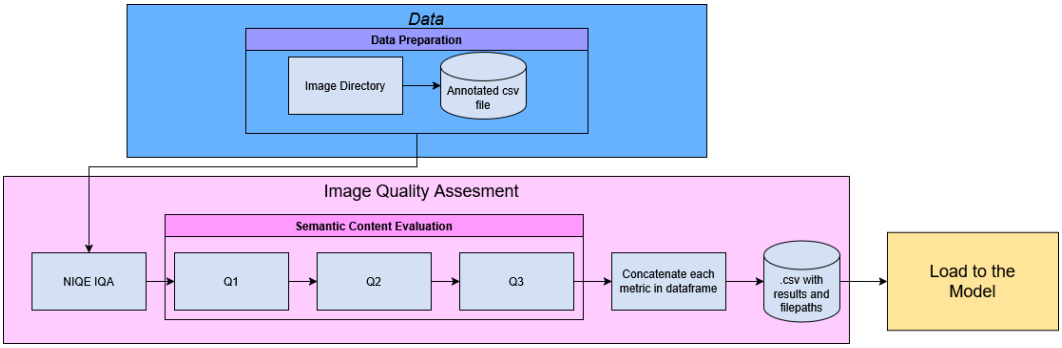

Figure 3: Workflow for BIOCLEANSE. The piplene assesses the quality of the images using NIQE and QWEN VLM. The model requires image directory (can be both local or in the cloud) with annotated .csv file which contain the "ground truth" for each observation as well as file path. Before data is loaded to the model, the statistic image assessment and semantic content evaluation take place to filter out poor quality images.

compressed), these natural patterns are disrupted, allowing an NR-IQA algorithm to identify it as a lower quality.

For our pipeline example we used NIQE algorithm. Unlike some other NR-IQA methods like BRISQUE that are trained on images with human-assigned quality scores, NIQE is opinion-unaware making it a suitable choice for domain specific evaluation of images. It builds its understanding of "good" quality directly from the statistical patterns found in a wide variety of natural images, without specific training on distorted examples or human preferences. This makes NIQE robust solution, especially in biodiversity context where the types of distortions may not be well-represented in existing training datasets.

A crucial aspect of quality control with NIQE is setting a quality threshold for the images. Images with a NIQE score above this threshold should be considered not suitable for further analysis and should be filtered out. This threshold is not fixed and it can be adjusted by the user. This flexibility allows researchers to customize the quality based on their specific research needs. For instance, in our pipeline a threshold of 6 was set. Analysis of the NIQE score distribution, visualized through box plot, revealed that scores of 6 or higher were above the 95th percentile. This statistical representation indicated that these images represented the lowest quality outliers in the dataset.

E.2 Semantic content filtering

The final stage of the BIOCLEANSE pipeline uses a modular approach to content-based filtering, using the capabilities of VLMs. This design ensures flexibility, allowing for the integration of any suitable VLM to perform detailed semantic assessments of image quality. Although the QWEN 2.5 VL 3B model was used in this prototype, the architecture supports integration of other VLMs, such as LLaVA, DeepSeek VL, or custom-trained models, depending on specific project requirements and computational resources.

It is important to note that, for the purpose of this pipeline, these VLMs are not fine-tuned on biodiversity-specific datasets. Their broad understanding allows them to work effectively on various images without requiring specific adaptation for each new domain. This approach allows flexibility of the pipeline and reduces the computational overhead associated with fine-tuning.

According [36], the experts propose to validate images according to quality measures described in the Table 2. Images are queried with "Are humans present in the image?"; images where human presence is detected are excluded to ensure the focus remains on the biodiversity subject. Images are further assessed with "Is the image too blurry or low quality to allow identification?". This complements the NIQE assessment by adding a semantic understanding of blurriness or low resolution in the context of species identification. Images flagged as too poor for identification by the VLM are removed. This VLM-driven filtering step aims to remove images with distracting elements or those

where the primary subject is not clearly visible, thereby improving the image dataset for training biodiversity-focused AI models.

F Data collection

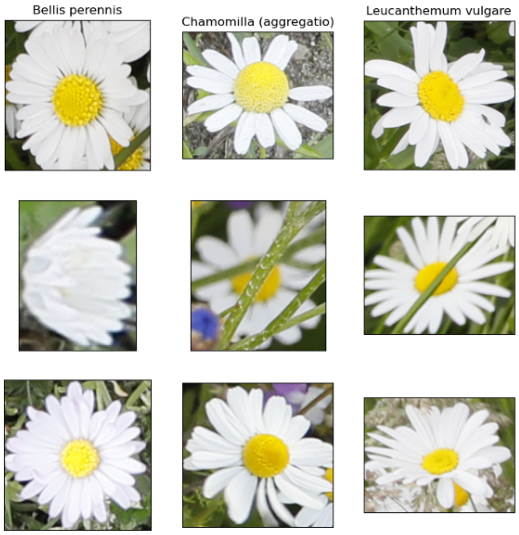

Figure 4: Selected plant: *Bellis perennis*, *Chamomilla (aggregatio)* or *Matricaria chamomilla*, and *Leucanthemum vulgare*. Photos are crops randomly sampled from EWD [43].

We select three flowering plant species from the *Asteraceae* family: *Bellis perennis* (*Common Daisy*), *Matricaria chamomilla*, and *Leucanthemum vulgare* (*Ox-eye Daisy*). Example images for each species of different qualities are illustrated in Figure 2. These species were selected due to their morphological similarities, frequent co-occurrence in European ecosystems, and their potential to challenge AI identification models [44]. Observational data, specifically *Research Grade* records, were sourced exclusively from the iNaturalist platform, with data acquisition geographically constrained to Europe to maintain a degree of environmental consistency. Metadata for each observation, including image URLs, unique identifiers, quality grade, observation timestamps, geolocation data, user identifiers, and scientific names, were collected. From an initial pool of 114K *Research Grade* observations, 3-distinct random subsets comprising 600 images each (200 per species) were collected.

G Models

Vision foundation models (VFMs) have demonstrated remarkable success across diverse applications in computer vision. Their capacity to learn generalizable representations from extensive datasets makes them promising for complex tasks in biodiversity observation. This section reviews key VFMs, specifically CLIP, BIOCLIP [47], and BIOTROVE-CLIP [53], highlighting their relevance to automating and enhancing biodiversity research.

OpenAI's CLIP (Contrastive Language-Image Pre-training) learns to associate images with their corresponding textual descriptions in an unsupervised manner. Its architecture comprises an image encoder (e.g., RESNET, VIT) and a text encoder (e.g., Transformer), jointly trained to maximize the similarity between embeddings of matching image-text pairs. Trained on a massive dataset of internet-sourced image-text pairs, CLIP generalizes to unseen images and descriptions.

CLIP's zero-shot capabilities hold significant potential for biodiversity observation, enabling image classification based on textual descriptions of species without explicit training. This is particularly useful given the challenges of discovering new species and the impracticality of collecting large labeled datasets for every species. However, CLIP's reliance on internet-sourced data may introduce

biases, and its capacity to handle fine-grained distinctions between visually similar species can be limited.

BIOCLIP [47] is specifically designed for biological image analysis. Recognizing that general-purpose VFMs like CLIP may not be optimally suited to the nuances of biological data, BIOCLIP aims to learn representations attuned to the specific characteristics of biological imagery. It is trained on a dataset of biological images with associated textual descriptions, emphasizing the model's ability to learn fine-grained visual features and understand semantic relationships between biological entities. BIOCLIP has demonstrated improved performance compared to general-purpose VFMs on biological image classification tasks and is designed to better capture the hierarchical structure of biological taxonomy.

BIOTROVE-CLIP [53] is a family of vision-language foundation models developed using the BIOTROVE-40M dataset, a large-scale collection of biodiversity images from the iNaturalist platform. BIOTROVE-CLIP aims to leverage this community-sourced data, addressing its inherent challenges. The BIOTROVE-TRAIN dataset is larger and more taxonomically diverse than many other biological image datasets.

BIOTROVE-CLIP addresses the challenges of noisy iNaturalist data, which exhibits variations in image quality, annotation accuracy, and taxonomic consistency, by utilizing expert-verified subsets, developing robust training strategies, and leveraging textual information. BIOTROVE-CLIP demonstrates the potential of leveraging large-scale, community-sourced data for biodiversity assessment. Its strong generalization performance across diverse datasets highlights its ability to learn robust representations of biological diversity. However, BIOTROVE-CLIP also faces challenges, including potential biases, difficulties in handling fine-grained distinctions, and the need for continued research into improving accuracy and reliability when training on noisy data.

H Extended results

H.1 Code

All source code, preprocessing scripts, and data needed for full reproducibility are publicly available at our project repository: github.com/wakizasher/iNaturalist_Benchmarking.

H.2 VLM-based image content assessment

The Qwen2.5-VL-Instruct VLM was employed to evaluate semantic content attributes for each of the 600-image seed datasets. For the Seed 42 dataset (Table 3), human presence was detected in 4.5% (27 images) of the observations. Images flagged as "Blurry/Unidentifiable" constituted a very small fraction, at 0.6% (4 images). The presence of "Other Taxa" was more common, identified in 12.0% (72 images) of this seed.

In the Seed 123 dataset, similar proportions were observed: human presence was noted in 4.1% (25 images) , while "Blurry/Unidentifiable" images were again minimal at 0.3% (2 images). "Other Taxa Present" were identified in 13.1% (79 images) of the Seed 123 observations (Table 3).

Table 7: Qwen VLM assessments for image attributes (seed 42, N=600).

Attribute Assessed	Response 'Yes' (Count)	Response 'Yes' (%)	Response 'No' (Count)	Response 'No' (%)
Composition	4	0.6%	596	99.3%
Human Presence	27	4.5%	573	95.5%
Other Taxa Present	72	12.0%	528	88.0%

For the Seed 456 dataset, human presence was detected in 4.5% (27 images). The "Blurry/Unidentifiable" characteristic was slightly more frequent in this seed at 1.16% (7 images). "Other Taxa Present" was the most prevalent of the flagged attributes, identified in 16.3% (98 images) of the Seed 456 images (Table 5). These VLM assessments formed the basis for several subsequent filtering conditions aimed at understanding model sensitivity to these specific content characteristics.

Table 8: Qwen VLM assessments for image attributes (seed 123, N=600)

Attribute Assessed	Response 'Yes' (Count)	Response 'Yes' (%)	Response 'No' (Count)	Response 'No' (%)
Composition	2	0.3%	593	99.7%
Human Presence	27	4.5%	573	95.5%
Other Taxa Present	79	13.16%	521	86.83%

Table 9: Qwen VLM assessments for image attributes (seed 456, N=600)

Attribute Assessed	Response 'Yes' (Count)	Response 'Yes' (%)	Response 'No' (Count)	Response 'No' (%)
Composition	7	1.16%	593	98.83%
Human Presence	27	4.5%	573	95.5%
Other Taxa Present	98	16.3%	502	83.6%

Table 10: NIQE assessments for image attributes seed 42 (N=600)

Attribute Assessed	NIQE > 6 (Count)	NIQE > 6 (%)	NIQE < 6 (Count)	NIQE < 6 (%)
Evaluation	30	5%	570	95%

Table 11: NIQE sssessments for image attributes seed 123 (N=600)

Attribute Assessed	NIQE > 6 (Count)	NIQE > 6 (%)	NIQE < 6 (Count)	NIQE < 6 (%)
Evaluation	29	4.83%	571	95.16%

Table 12: NIQE assessments for image attributes seed 456 (N=600)

Attribute Assessed	NIQE > 6 (Count)	NIQE > 6 (%)	NIQE < 6 (Count)	NIQE < 6 (%)
Evaluation	12	2%	588	98%