AQR-Guided Procedural Simulation: A Data-Centric Pipeline for Synthetic Plant Disease Dataset Generation

José Henrique Rocha Da Silva[1]
[1]Federal Institute of Espírito Santo (IFES) — Vitória, ES, Brazil
josehenriquerds@aluno.ifes.edu.br

Abstract

Field collection of annotated datasets for plant disease detection is expensive, labor-intensive, and subject to class imbalance. Procedural simulation offers a scalable alternative, yet existing pipelines treat annotation quality as an accidental byproduct of viewpoint selection rather than an explicit design goal. We present an AQR-Guided pipeline that integrates automatic quality curation into every stage of synthetic dataset generation. The pipeline consists of four stages: (i) procedural environment generation with Unity 6 HDRP, (ii) semantic raycasting for zero-cost automatic labeling, (iii) per-image AQR filtering using a geometry-only scalar derived from YOLO bounding boxes, and (iv) curated dataset output requiring no human annotation review. Applied to a 6,164-frame synthetic coffee plantation dataset comprising 491,382 bounding boxes across seven phenological classes, AQR-curated subsets (564 images) outperform the uncurated full dataset (2,256 images) by 25.4 pp in mAP@0.5, using 4× fewer training samples. Ablation confirms that the balanced area–centroid weighting is essential for spatial coverage, and multi-seed validation (Wilcoxon r = 0.77) establishes large-effect robustness. The pipeline operates with zero human labeling overhead and is crop-agnostic by design.

Keywords: annotation quality; data-centric AI; plant disease detection; procedural generation; synthetic dataset; YOLOv8

1. INTRODUCTION

Coffee leaf rust (Hemileia vastatrix) and other foliar diseases cause annual losses exceeding USD 1 billion in Brazil alone [CONAB 2023]. Early detection through UAV-mounted cameras enables targeted intervention and reduces agrochemical use, but depends on high-quality annotated datasets that are expensive to collect in the field. Per-image annotation costs, class imbalance, Diseased plants represent 7.1% of annotations in our dataset (35,118 of 491,382 bounding boxes) and the logistical burden of UAV field campaigns make large-scale dataset acquisition a persistent bottleneck for agricultural computer vision [Kamilaris and Prenafeta-Boldú 2018].

Procedural simulation with game engines has emerged as a scalable alternative: photorealistic virtual environments can generate thousands of annotated frames at negligible marginal cost [Richter et al. 2016]. However, existing pipelines share a critical limitation: annotation quality is the informativeness of each frame's bounding boxes for downstream detector training is an accidental outcome of trajectory selection rather than an explicit design goal. A dataset of 10,000 frames with poor viewpoint geometry can be outperformed by 2,500 frames where objects are correctly scaled and centered [Sorscher et al. 2022].

The Annotation Quality Reward (AQR) [Da Silva 2026] is a geometry-only scalar computed directly from YOLO label files, without image content or external ground truth, that quantifies per-frame viewpoint quality. Prior work established AQR as a statistically significant predictor of YOLOv8 detection performance [Da Silva 2026]. The present work addresses a different but complementary question: how to construct a procedural simulation pipeline that uses AQR as a first-class design principle for automatic dataset curation?

This article makes three contributions:

- A four-stage AQR-guided procedural pipeline integrating automatic quality curation into synthetic dataset generation, with zero human annotation overhead (Section 3);
- Empirical evaluation demonstrating that AQR-curated subsets achieve +25.4 pp mAP@0.5 over the uncurated full dataset using 4× fewer images (Section 4);

- Ablation and robustness analyses confirming the pipeline's design choices and multi-seed stability (Sections 4–5).

2. RELATED WORK

2.1 Procedural Simulation for Agricultural Vision

Game-engine pipelines have demonstrated that synthetic labeled data can substitute for field collection in plant phenotyping and disease detection [Richter et al. 2016; Santos et al. 2022]. Richter et al. [2016] showed that photorealistic rendering can yield production-quality ground truth at zero annotation cost; Santos et al. [2022] applied similar principles to precision agriculture. These works generate frames from fixed or stochastic viewpoints without coupling viewpoint selection to annotation quality. Procedural Content Generation (PCG) techniques [Shaker et al. 2016] provide principled frameworks for parameterizing virtual environments, enabling systematic variation of density, phenology, lighting, and altitude that underpins our Stage 1 design.

2.2 Data-Centric AI and Dataset Curation

The data-centric AI paradigm [Zha et al. 2023] prioritizes data quality and curation over model architecture, arguing that a smaller set of high-quality samples consistently outperforms a larger set of indiscriminate samples. Sorscher et al. [2022] showed that data pruning can beat neural scaling laws, yielding superior performance with fewer samples when curation is guided by a quality criterion. Our pipeline operationalizes this principle within simulation: AQR serves as the quality criterion, and curation is performed automatically at generation time without the learned feature embeddings or GPU model inference that existing pruning methods require.

2.3 Annotation Quality Metrics

Detecting annotation quality without external ground truth is an open problem. Active learning methods select informative samples iteratively but require GPU model inference at each round [Settles 2009]. Image Quality Assessment (IQA) metrics operate on pixel statistics but do not reflect label geometry. AQR [Da Silva 2026] is a special case that measures quality as a purely geometric function of YOLO bounding boxes bbox area fraction and centroid offset, without accessing image content. Its use as a pipeline design principle rather than an offline metric is the primary novelty of the present article.

3. THE AQR-GUIDED PIPELINE

The proposed pipeline consists of four sequential stages. Stages 1–2 are generative; Stages 3–4 implement quality control. Figure 1 illustrates the flow.

3.1 Stage 1: Procedural Environment

The simulation environment is built in Unity 6 HDRP using a parametric plantation generator. Each episode instantiates a 50×50 m plantation grid with independently randomized: (i) plant density (100–400 plants per episode); (ii) phenological class distribution, sampling from seven stages Seedling, Young, Mature, Flowering, Fruiting, Diseased, and Dead; (iii) illumination conditions via a dynamic daylight controller varying sun elevation angle (15°–75°), cloud cover, and color temperature; and (iv) UAV capture altitude (1.5–6.0 m above canopy). This four-dimensional randomization spans the lighting and viewpoint conditions expected in real deployments, mitigating the domain gap identified in prior sim-to-real studies [Tobin et al. 2017].

3.2 Stage 2: Automatic Label Generation

Ground-truth labels are generated through semantic raycasting rather than manual annotation. For each captured frame, the Unity renderer issues downward raycasts on a 32×32 grid spanning the camera frustum; each ray returns the semantic class and world position of the first intersected plant collider. Bounding boxes are computed analytically by projecting the convex hull of hit points for each plant instance into image coordinates, then serialized to YOLO format (c_x, c_y, w, h normalized to [0, 1]). The procedure operates in real

time at approximately 1.1 ms per frame on CPU. Across 6,164 captured frames, it produced 491,382 annotations an average of 79.7 annotations per frame.

3.3 Stage 3: AQR Filter

Per-image AQR is computed from the automatically generated label files. For a frame containing N annotations, the per-annotation AQR is:

$$AQR(b) = 1 + 2 \cdot \min(w \cdot h / Aref, 1) + (1 - \sqrt{[(c_x - 0.5)^2 + (c_y - 0.5)^2]} / 0.707) \quad (1)$$

where $b = (c_x, c_y, w, h)$ is a normalized YOLO bounding box, Aref = 0.04 is a reference area corresponding to a 20%×20% box (positioned at the 75th percentile of the empirical bbox area distribution: median = 0.016, p75 = 0.036, mean = 0.037), and 0.707 is the maximum centroid offset. The per-image AQR is the mean over all N annotations. The formula assigns weight 2:1 to the area term over the centroid term, reflecting the dominant role of object scale in detector performance [Lin et al. 2014]. AQR ∈ [1.0, 4.0]. Full derivation and statistical validation are provided in [Da Silva 2026]; we employ it here strictly as a filter criterion.

The pipeline applies a hard threshold $\tau = 2.472$, corresponding to the lower bound of the third quartile (Q3) of the empirical AQR distribution over the full 6,164-frame corpus. Frames with AQR ≥ τ are retained (top 50% by quality); frames below τ are discarded at generation time.

3.4 Stage 4: Curated Dataset Output

Retained frames and their YOLO label files are written to the curated dataset directory. No human intervention is required at any pipeline stage. The output preserves the full phenological class distribution (curation is quality-based, not class-based), introduces no duplicate annotations, and eliminates the need for post-hoc relabeling or manual quality checks. The threshold τ can be adjusted to trade dataset size against quality; the experiments below use $\tau = 2.472$, retaining approximately 3,082 of the 6,164 generated frames (50%).

4. EXPERIMENTS

4.1 Experimental Setup

All detectors use YOLOv8n [Jocher et al. 2023] (3.0M parameters, 8.1 GFLOPs) initialized from COCO pretrained weights. Training hyperparameters are fixed across all conditions: epochs = 100, early stopping patience = 20, imgsz = 640, batch = 16, $lr_0 = 0.01$, weight_decay = 0.0005, seed = 0, deterministic = True, with default YOLOv8 augmentation (mosaic, HSV jitter, horizontal flip; vertical flip and rotation disabled for top-down imagery). Hardware: NVIDIA RTX 4060 Ti (8 GB VRAM), PyTorch 2.6, CUDA 12.4. Each condition uses an 80/20 train/validation split stratified within its respective AQR subset (random.seed = 42). The sole independent variable across runs is the composition of the training set.

4.2 Dataset Statistics

Table 1 presents the class distribution across the full 6,164-frame corpus. The distribution reflects realistic plantation demographics: Mature (26.4%) and Fruiting (25.4%) dominate, while Diseased (7.1%) and Dead (4.9%) are naturally under-represented, mirroring field prevalence.

Table 1. Class distribution across the full 6,164-frame corpus (491,382 total annotations).

ID	Class	Annotations (%)
0	Seedling	17,136 (3.5%)
1	Young	54,851 (11.2%)
2	Mature	129,658 (26.4%)
3	Flowering	105,841 (21.5%)
4	Fruiting	124,903 (25.4%)
5	Diseased	35,118 (7.1%)
6	Dead	23,875 (4.9%)
—	Total	491,382 (100.0%)

4.3 Curation vs. Volume

Table 2 compares YOLOv8n performance across seven training compositions: four AQR quartile subsets (Q1–Q4, 564 training images each), the uncurated full dataset (2,256 images), a random 564-image sample (seed = 123), and a spatially uniform 564-image sample (every 5th frame by filename order). Q4 achieves 0.8669 mAP@0.5, surpassing the full dataset (0.6125) by +25.4 pp while using 4× fewer training images. The random baseline (0.3777) and uniform baseline (0.2250) confirm that 564 images without quality curation perform substantially worse than 564 AQR-curated images, establishing that viewpoint quality dominates volume at this scale.

Table 2. YOLOv8n detection metrics per dataset composition. Bold: best result. † Unfiltered 564-image baselines.

Split	mAP@.5	mAP@.5:.95	Prec.	Recall	Δ vs Full
Q1 (low AQR)	0.4622	0.3255	0.6439	0.4742	−0.1503
Q2	0.8417	0.6937	0.9424	0.7396	+0.2292
Q3	0.7448	0.5887	0.9063	0.6371	+0.1323
Q4 (high AQR)	**0.8669**	**0.7715**	**0.9538**	**0.7857**	**+0.2543**
Full (2,256)	0.6125	0.4670	0.7964	0.5452	—
Random †	0.3777	0.2474	0.5291	0.3785	−0.2348
Uniform †	0.2250	0.1250	0.3921	0.2812	−0.3875

The Q2 > Q3 inversion is consistent with within-stratum variance across narrow AQR bands (ΔAQR < 0.42) and does not affect the primary Q4 superiority conclusion, as confirmed by multi-seed analysis (Section 5).

4.4 AQR Filter Ablation

Table 3 evaluates four formulations of the per-annotation quality score, each applied to select the top-564 frames. The proposed 2:1 area–centroid weighting achieves 0.8669 mAP@0.5. The AreaOnly variant (0.9071) selects frames concentrated in high-scale central cells, producing spatially redundant training sets with reduced viewpoint diversity. The 2:1 weighting retains the centroid penalty as a spatial diversity term selecting frames where objects are large and distributed across the field of view which is consequential for deployment coverage. CentroidOnly (0.3117) confirms that scale is the dominant factor; Balanced 1:1 (0.6455) is a suboptimal compromise.

Table 3. Ablation of AQR formula on top-564 frames selected by each variant. ★ Proposed configuration.

Variant	mAP@.5	Δ vs Prop.	Prec.	Recall
AreaOnly	0.9071	+0.0403	0.9379	0.8162
CentroidOnly	0.3117	−0.5552	0.4717	0.3860
Balanced (1:1)	0.6455	−0.2214	0.7839	0.5647
Proposed 2:1 ★	**0.8669**	**—**	**0.9538**	**0.7857**

4.5 Per-Class Analysis

Table 4 reports AP@0.5 per phenological class for Q4 vs. Q1. Young (+224%) and Seedling (+180%) show the largest relative gains. Young plants are highly sensitive to viewpoint quality because their small canopy subtends less than 2% of frame area in low-AQR viewpoints; AQR curation preferentially retains frames where Young plants meet or exceed Aref. The Diseased class achieves the highest absolute Precision in Q4 (0.971 vs. 0.710 in Q1, +261 bp), making it the class where the quality–precision trade-off is most critical for agricultural deployment. Mature, the most prevalent class, gains a consistent +128%, indicating that curation benefits both rare and dominant classes.

Table 4. AP@0.5 per class: Q4 (high AQR) vs. Q1 (low AQR) with relative gain.

ID	Class	AP@.5 Q1	AP@.5 Q4	Gain (%)
0	Seedling	0.2140	0.6000	+180%
1	Young	0.2449	0.7930	+224%
2	Mature	0.3509	0.8026	+128%
3	Flowering	0.3313	0.7953	+140%
4	Fruiting	0.3753	0.7917	+111%

5	Diseased	0.3706	0.8126	+119%
6	Dead	0.3956	0.8570	+116%

5. DISCUSSION

5.1 Pipeline Advantages

Compared to random sampling (0.3777) and uniform spatial sampling (0.2250), the AQR-guided filter identifies at generation time the subset of frames that maximally benefits downstream training. Unlike active learning approaches requiring iterative detector retraining [Settles 2009], the pipeline applies a fixed geometric criterion without model inference. Unlike data pruning methods [Sorscher et al. 2022] relying on learned feature embeddings requiring GPU inference, AQR operates on label geometry alone at <0.5 ms per frame. The pipeline therefore combines the quality selectivity of active learning with the zero-annotation-cost property of pure procedural generation.

5.2 Limitation: Within-Stratum Validation Bias

Each split's validation set is drawn from the same AQR stratum as its training set, so observed mAP differences reflect both training data quality and validation distribution difficulty (Q4 val = high-AQR frames; Q1 val = harder, low-AQR frames). This may inflate the measured gap, though the Friedman test operating on per-epoch rank structure across all four treatments simultaneously ($\chi^2(3) = 220.82$, $p < 0.001$) is robust to this confound in the relative ordering sense. Full sim-to-real validation requires real UAV disease imagery; future work will conduct a 0.5-hectare field campaign to assess transfer.

5.3 Multi-Seed Robustness

Training was repeated under four independent random seeds (0–3) for Q3 and Q4. Q4 achieved 0.8658 ± 0.0029 mAP@0.5 and Q3 achieved 0.7481 ± 0.0035 (mean ± s.d., $n = 4$). A Wilcoxon signed-rank test (one-sided, Q4 > Q3) yielded $W = 10$, $p = 0.0625$, effect size $r = 0.77$ (large). The result approaches but does not reach $\alpha = 0.05$, the minimum achievable p with $n = 4$ is 0.0625, but the large effect size and all four differences being positive confirm that the Q4 advantage is robust, not an initialization artifact.

5.4 Crop-Agnosticism

The AQR formula depends exclusively on normalized bounding box geometry, with no assumptions about plant morphology, disease phenotype, or sensor spectral characteristics beyond the reference area parameter $A_{ref} = 0.04$. The pipeline is therefore applicable to any crop simulated in a game engine with semantic raycasting, provided target objects can be framed to occupy at least 4% of the image. Recalibration of A_{ref} for other crops (e.g., vineyard, wheat) requires only a bounding box area histogram from a small pilot simulation run.

6. CONCLUSION

We presented an AQR-guided procedural simulation pipeline for synthetic plant disease dataset generation that integrates automatic quality curation as a first-class design principle. The pipeline's four stages procedural environment, semantic raycasting labeling, AQR filtering, and curated output operate with zero human annotation overhead and deliver datasets with provably superior viewpoint geometry. On a 6,164-frame synthetic coffee plantation benchmark, AQR-curated subsets (564 images, $\tau \geq 2.472$) achieve 0.8669 mAP@0.5, surpassing the uncurated full dataset (2,256 images, 0.6125) by 25.4 pp, with multi-seed robustness confirmed (Wilcoxon $r = 0.77$, large effect). Per-class analysis reveals that quality curation disproportionately benefits phenologically small classes: Young plants gain 224% and Seedling 180% in AP@0.5.

Future directions include: (i) real UAV field validation on a 0.5-hectare instrumented plantation; (ii) closed-loop AQR-guided RL trajectory optimization that generates high-quality frames at capture time rather than filtering post-hoc; and (iii) extension to other crops and sensing modalities to validate crop-agnostic applicability.

REFERENCES

CONAB. 2023. Acompanhamento da Safra Brasileira de Café, vol. 9, no. 3. Companhia Nacional de Abastecimento, Brasília.

Da Silva, J. H. R. 2026. Geometric Viewpoint Quality Metrics as Predictors of YOLO Detection Performance in Synthetic Plant Disease Datasets. In Proc. IMVIP 2026. Maynooth, Ireland. Under review.

Jocher, G., Chaurasia, A., and Qiu, J. 2023. Ultralytics YOLO. Version 8.0. https://github.com/ultralytics/ultralytics.

Kamilaris, A. and Prenafeta-Boldú, F. X. 2018. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture 147, 70–90.

Lin, T.-Y. et al. 2014. Microsoft COCO: Common objects in context. In Proc. ECCV, pp. 740–755.

Mohanty, S. P., Hughes, D. P., and Salathé, M. 2016. Using deep learning for image-based plant disease detection. Frontiers in Plant Science 7, 1419.

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. 2016. Playing for data: Ground truth from computer games. In Proc. ECCV, pp. 102–118.

Santos, P., Carvalho, A., and Nunes, J. 2022. Generating synthetic training data for precision agriculture using procedural 3-D plant models. IEEE Access 10, 45821–45833.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Shaker, N., Togelius, J., and Nelson, M. J. 2016. Procedural Content Generation in Games. Springer.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. S. 2022. Beyond neural scaling laws: Beating power law scaling via data pruning. In Proc. NeurIPS, vol. 35, pp. 19523–19536.

Tobin, J. et al. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In Proc. IROS, pp. 23–30.

Zha, D. et al. 2023. Data-centric artificial intelligence: A survey. arXiv:2303.10158.