

Ground Validation of Model-Tier Inference for Earth Observation on a Jetson Orin Nano

Nomos Research

hello@nomos.run <https://nomos.run/research/orin-ground-validation/>

NOMOS-TR-2026-02, version 1.0 24 August 2026

Preprint / technical report. Not peer reviewed.

Abstract

Before evaluating an adaptive onboard inference controller, its model choices must form a real quality/resource frontier and its hardware measurements must be stable under the intended workload. We test those premises on an 8 GB Jetson Orin Nano in its 15 W configuration. Two ImageNet-initialized MobileNetV3 tiers are fine-tuned for eight epochs directly on the board using 16,200 EuroSAT RGB training images, with 5,400 images each for validation and test under a fixed longitude-based split. Training takes 118.22 and 193.72 seconds; the selected checkpoints reach 90.315% and 91.333% accuracy on the held-out test set. Target-built TensorRT 10.3 FP16 engines then process the test set three times in each of two workload modes. Under continuous, preloaded inference, the small and large tiers reach 90.463% at 6.506 mJ/image and 91.389% at 8.258 mJ/image. The large tier gains 0.926 accuracy points for 26.9% more measured whole-board energy, leaving both tiers non-dominated. Under decode/preprocess gaps, every telemetry sample remains at 306 MHz; continuous inference promotes the stock governor mostly to 612 MHz. A separate cadence sweep confirms that latency and energy depend on workload history, not only model identity and a nominal frequency. Finally, a post-capture loader rejects a same-size, one-bit-corrupted engine before TensorRT deserialization. These are ground feasibility and measurement-boundary results. They do not establish physical-DVFS superiority, cross-sensor transfer, calibrated power, spacecraft thermal behavior, or flight readiness.

1 Introduction

Adaptive neural inference can switch among model sizes or exits as energy changes [1], and recent edge systems jointly optimize early exits and DVFS [2]. In satellite systems, EarthSight places model and filter choices inside a broader orbit-to-ground scheduler [3]. These systems make a model ladder plausible, but they do not guarantee that arbitrary model variants provide useful choices on a particular sensor and target computer.

A controller needs at least two non-dominated modes. A slower or smaller network must buy a measurable resource advantage while retaining a task metric that matters. Profiling also needs to control workload cadence: an autonomous hardware governor observes utilization history, so decode gaps and burst structure can change clocks, latency, and energy even when the engine is unchanged.

This technical note reports a separate ground-validation study that follows the frozen *Beyond DVFS* compute-HIL paper [4]. It does not revise that paper’s controller matrix. Its contributions are:

1. A deterministic, target-side fine-tuning run for two MobileNetV3 tiers on a public Earth-observation dataset with a spatial holdout.
2. Twelve complete TensorRT inference runs that expose a two-tier accuracy/board-energy frontier and preserve an anomalous repetition.
3. A physical cadence sweep showing that the stock governor makes workload history part of the measurement condition.
4. A narrowly scoped stored-artifact admission check, together with an explicit distinction from in-memory radiation protection.

The contribution is evidence, not a novelty claim for energy-aware model selection. The central unanswered comparison remains joint model/frequency control versus real fixed scheduling and physical DVFS under identical arrivals.

2 Experimental Protocol

2.1 Data and spatial split

EuroSAT provides 27,000 labeled 64×64 RGB Sentinel-2 patches across ten land-use and land-cover classes [5]. We use TorchGeo’s fixed longitude-based split [6]: 16,200 training, 5,400 validation, and 5,400 test images. The three lists are disjoint and cover the archive. Archive and split checksums are verified before training. A content-tree digest binds every image basename, class index, and SHA-256.

This split measures geographic transfer within Sentinel-2. It is not a train-on-one-sensor, test-on-another protocol. EuroSAT’s class accuracy is also not an operator mission utility metric.

2.2 Models and training

MobileNetV3 Small and Large [7] are initialized from torchvision ImageNet weights and their classifiers are replaced with ten-class linear heads. Every parameter is trainable. Training uses random horizontal and vertical flips, ImageNet normalization, AdamW with learning rate 3×10^{-4} and weight decay

Table 1: One deterministic eight-epoch training run per tier. Accuracy is from the selected checkpoint on the fixed 5,400-image test split.

Tier	Params	Best epoch	Train s	Mean W	Train J	Test acc.
Small	1.528M	7	118.22	7.950	939.81	90.315%
Large	4.215M	5	193.72	8.670	1679.60	91.333%

10^{-4} , cosine annealing, batch size 256, automatic mixed precision, and eight epochs. Checkpoint selection maximizes validation balanced accuracy, breaking ties by loss. The seed is 20260824, deterministic PyTorch algorithms are required, and cuBLAS workspace configuration is fixed before process start.

The runs execute inside a pinned NVIDIA PyTorch container on Jetson Linux R36.4.4. Training power is sampled from INA3221 VDD_IN at a requested 100 ms cadence. VDD_IN includes the complete development board; idle power is not subtracted and the sensor is not externally calibrated.

2.3 Target compilation and inference

The selected checkpoints are exported as FP32 ONNX, then compiled on the target with TensorRT 10.3 using FP16 tactics and FP32 I/O. Build times are 165.31 seconds for small and 179.42 seconds for large. Serialized engines are 4.00 and 9.47 MB. TensorRT engines contain target-specific compiled state and are not treated as portable model files [8].

Each engine executes the complete test set three times in two modes. *Full pipeline* decodes each JPEG, normalizes and transposes it, then invokes TensorRT. *Inference only* preloads every normalized tensor before measurement and invokes the engine continuously. Each run has 100 warmup calls. The same target-side runner records per-image latency, predictions, whole-board power, temperatures, and observed GPU frequency. No run is removed.

3 Results

3.1 Training viability

Table 1 shows that both models can be fine-tuned directly on the board in minutes. The large tier requires 64% more training time and 79% more measured board energy than small, while adding 1.019 test-accuracy points in PyTorch. This demonstrates target-side feasibility, not a reason to train in orbit. Ground or cloud training remains easier to monitor, reproduce, and recover unless local data access or adaptation latency is binding.

3.2 Measured inference frontier

Figure 1 shows the continuous-inference frontier. The large tier gains 0.926 accuracy points and 0.615 balanced-accuracy points over small while adding 0.215 ms median engine latency and 26.9% median board energy per image. Small does not dominate large because it is less accurate; large does not

Table 2: Median of three complete 5,400-image runs. Energy is uncalibrated whole-board VDD_IN energy divided by images.

Mode	Tier	Acc.	Engine ms	Total ms	Images/s	mJ/image
Full	Small	90.463%	2.802	3.978	247.82	16.068
Full	Large	91.389%	2.787	4.002	246.37	17.314
Inference	Small	90.463%	1.227	1.228	793.47	6.506
Inference	Large	91.389%	1.442	1.443	676.32	8.258

dominate small because it is slower and consumes more energy. These two modes are therefore eligible choices for a later controller experiment.

TensorRT accuracy differs slightly from PyTorch accuracy, so the deployed graph is evaluated rather than inheriting checkpoint metrics. Small reaches 90.463% and large 91.389%. The compact summary preserves every run hash and a logit checksum.

The second large full-pipeline repetition slows across decode, preprocessing, and inference. Its mean engine latency is 3.262 ms versus 2.723 and 2.787 ms in the other repetitions. Observed GPU frequency and temperature do not explain the system-wide slowdown. It remains in every aggregate. Three repetitions support a pilot median, not a worst-case execution-time or tail-latency claim.

3.3 Governor and cadence coupling

Every full-pipeline EuroSAT telemetry sample records 306 MHz. Continuous inference promotes the stock nvhost_podgov governor mostly to 612 MHz. This explains why the smaller apparent engine times in inference-only mode cannot be attributed only to removing JPEG and preprocessing work: the workload also drives a different hardware state.

A separate YOLOv8n TensorRT cadence pilot isolates this effect with a fixed engine and 1,000 repeated inference calls. With no scheduled idle gap, weighted GPU frequency is 607 MHz, mean engine latency is 9.19 ms, and energy is 72.13 mJ/image. Adding 10 ms between individual calls holds weighted frequency near 310 MHz, raises mean latency to 16.59 ms, and raises energy to 134.77 mJ/image despite reducing mean board power. Longer idle gaps and burst schedules produce the family in Figure 2.

The result is methodological: a profile indexed only by model and a requested or observed clock can be non-stationary if arrival cadence and governor history are omitted. Future policy comparisons must randomize run order, lock eligible frequencies, measure transition overhead, and preserve the exact arrival process.

4 Stored-Artifact Admission

TensorRT plans are executable, target-specific artifacts. After the measurements above were frozen, the maintained loader was changed to require an expected SHA-256 and optional byte count. It reads the engine once, verifies those exact bytes, and passes the same verified buffer to TensorRT. Unit tests

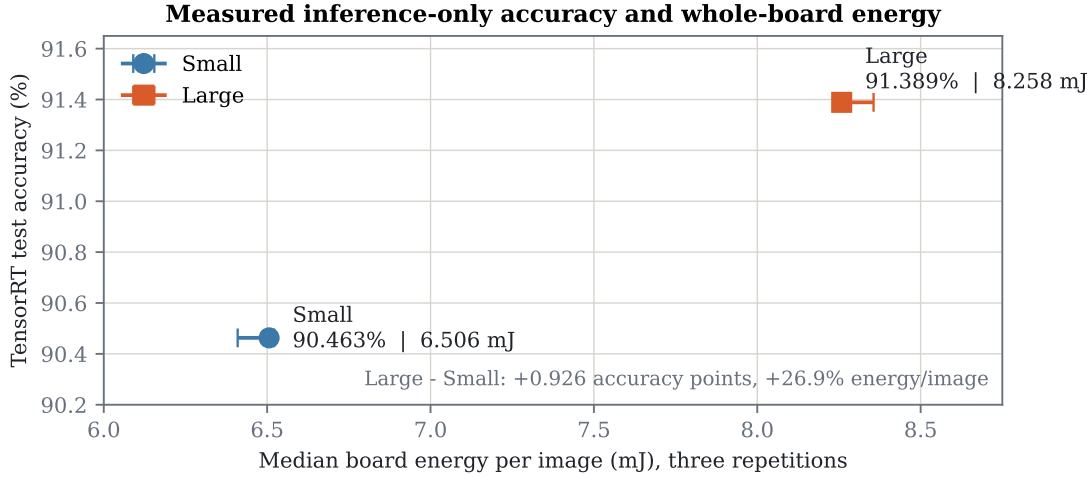


Figure 1: Both TensorRT tiers remain non-dominated on the measured inference-only accuracy/energy plane. Horizontal bars show the minimum and maximum energy/image across three repetitions around the reported median. Accuracy is evaluated over the same fixed 5,400-image test split in every run.

cover valid admission, invalid digest syntax, digest mismatch, and byte-count mismatch.

On the Orin, the approved small engine executes successfully. A same-size copy with one flipped bit is rejected before deserialization. This demonstrates stored-artifact integrity against an approved digest; it does not establish authenticity if the digest and engine are mutable together.

The boundary is also distinct from radiation protection. A bit flip after verified bytes enter host or GPU memory can still silently corrupt inference. SmallSat 2026 results from Goodman et al. measure FP32 weight-bit sensitivity and selective in-place SECDED repair inside cFS [9]. An opaque TensorRT plan does not expose the same weight-buffer interface. Flight treatment may require signed manifests, anti-rollback, ECC-capable memory, redundant inference, output invariants, periodic reload, watchdogs, and a known-good recovery artifact. None is established here.

5 What This Evidence Changes

The frozen *Beyond DVFS* note required a measured non-dominated model set before dynamic selection [4]. The present experiment independently finds a two-tier frontier on a different public Earth-observation task. It also shows that stock-governor state depends strongly on workload shape. Together, the results sharpen rather than expand the controller claim:

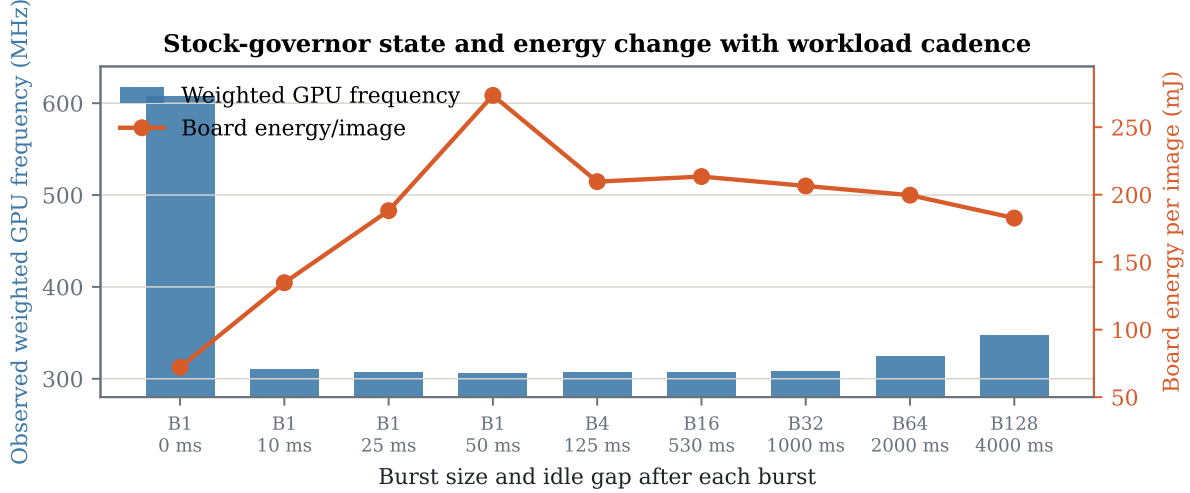
- A small model ladder can expose a real quality/energy choice on Orin-class hardware.
- More tiers are not automatically useful. Every candidate must survive a measured Pareto screen.
- Frequency, latency, and energy profiles need cadence and governor-history controls.
- Fine-tuning viability and inference feasibility do not imply mission value or cross-sensor transfer.

The study does not show that thirty resident variants are better than two. Each additional engine consumes storage and expands compatibility, qualification, update, rollback, and fault-recovery work. Nor does target-side training itself identify a buyer: the economic value depends on sensor data access, label cost, downlink limits, latency requirements, and the operator’s current baseline.

6 Limitations and Next Gate

- **One sensor and one task.** The geographic split remains within Sentinel-2 RGB. No cross-sensor transfer or operator instrument is tested.
- **Pilot replication.** Each tier has one deterministic training run and three inference repetitions. Each YOLO cadence has one repetition.
- **Governor not controlled.** GPU frequency is observed under the stock governor. It is not locked or changed per decision.
- **Board-level telemetry.** INA3221 VDD_IN is not an external calibration-grade meter and includes board activity beyond the GPU.
- **No spacecraft environment.** The development board, container, Python harness, and models are not flight qualified. There is no spacecraft power plant, thermal chamber, radiation environment, attitude history, or fault-management integration.
- **No policy comparison.** This note does not execute fixed schedule, physical adaptive DVFS, or joint control under a common arrival trace.

The next decisive ground experiment is therefore a paired physical comparison. Lock the GPU at each eligible frequency,



One pilot repetition per cadence; COCO is a compute workload, not an EO accuracy test.

Figure 2: Physical stock-governor response to nine workload cadences. Labels give burst size and the scheduled idle gap after each burst. The experiment uses COCO only as a fixed compute workload; it makes no satellite-imagery accuracy claim. Each cadence has one pilot repetition.

randomize repeated runs, include transition overhead, and apply identical arrivals to fixed-model, fixed-schedule, adaptive-DVFS, and joint model/frequency policies. Report completed work, task quality, calibrated energy, peak power, tail latency, and transition count. The comparison still remains synthetic until an operator supplies a trace, objective, and current rule.

7 Reproducibility Statement

The publication capsule contains the compact EuroSAT and cadence summaries, the exact captured training and inference entry points, the summarizers, subsequent integrity loader and tests, figure generator, paper source, and SHA-256 checksums. The recorded inference source is frozen at commit e62ff3c14cef1465c3854afd13950091f055b461; the later artifact-admission change is explicitly separated.

Key identities are:

Artifact	SHA-256
EuroSAT summary	82afead59ed4ea3dde7c1a1f41f085c3e445d0705735d6affaff1a918c25aa15
Cadence summary	800aa20d06d1f1eac8535ca73ab3795e1a8898186e8098c1d65061058a16c1f4
Training source	597aa05e9d8a071b78669b8445e7b677c7afc713d1018d521dc7c49ea5dbf844
Captured runner	ffc27cc8c36d53a7ed7b679f854302b393603936817c53fb1742be2b83262116

Raw data arrays, checkpoints, ONNX exports, and engines remain on the measured system and are bound by path, byte count, and hash in the summaries. The capsule supports result inspection and figure regeneration, not a one-command hardware reproduction.

8 Conclusion

Small-model training and TensorRT deployment on an 8 GB Jetson Orin Nano are straightforward enough to test on the ground. Establishing a useful adaptive runtime is harder. In this study, two MobileNetV3 tiers form a measured accuracy/board-energy frontier, but the stock governor also changes state with workload cadence. That coupling makes experimental protocol part of the product claim.

The result justifies a controlled physical-DVFS comparison. It does not justify a spacecraft, cross-sensor, or operator-value claim. Preserving that boundary is the purpose of this companion note.

References

- [1] Marcello Bullo, Seifallah Jardak, Pietro Carnelli, and Deniz Gündüz. Energy-aware dynamic neural inference. *arXiv preprint arXiv:2411.02471*, 2024.
- [2] Ziyang Zhang, Yang Zhao, Ming-Ching Chang, Changyao Lin, and Jie Liu. E4: Energy-efficient DNN inference for edge video analytics via early-exit and DVFS. *arXiv preprint arXiv:2503.04865*, 2025.
- [3] Ansel Kaplan Erol, Seungjun Lee, and Divya Mahajan. EarthSight: A distributed framework for low-latency satellite intelligence. *Proceedings of Machine Learning and Systems*, 2026. arXiv:2511.10834.
- [4] Nomos Research. Beyond DVFS: Queue-stable model-tier selection for resource-constrained onboard inference. Technical Report Nomos-TR-2026-01, Nomos Research, 2026.

- [5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- [6] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. TorchGeo: Deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2022.
- [7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.
- [8] NVIDIA Corporation. TensorRT engine compatibility, 2026. Accessed 2026-08-24.
- [9] Ryan Goodman, Carson Cramer, Braden Eichmeier, Emily Gerber, and Christopher Roscoe. Bit-flip sensitivity and selective FP32 protection for onboard neural networks in flight software. In *Small Satellite Conference*, 2026.