

Onboard Vision-Language Inference on an 8 GB Jetson Orin Nano

Accuracy, Cost, and Policy Behavior

Nomos Research

hello@nomos.run <https://nomos.run/research/onboard-vlm/>

NOMOS-TR-2026-04, version 1.1 25 August 2026
Preprint / technical report. Not peer reviewed.

Abstract

A public conference presentation reported that a multi-billion-parameter vision-language model (VLM) could classify overhead imagery on an NVIDIA Jetson in seconds and within an embedded power envelope. We repeat those measurements on the smaller hardware available to us: a physical 8 GB Jetson Orin Nano in its stock 15 W mode. This is not an exact model or flight-platform reproduction. We compare two pinned SmolVLM2 packages on a deterministic, balanced subset of 100 public AID aerial images: a 500M model in BF16 and a 2.2B model loaded in 4-bit NF4. The 500M package classifies 46/100 images correctly at 2.213 s median end-to-end latency, 8.146 W mean whole-board input power, and 18.282 J mean gross energy. The 2.2B NF4 package classifies 100/100 correctly at 4.905 s, 10.983 W, and 53.897 J. The result is a 54 percentage-point gain at $2.22\times$ latency and $2.95\times$ energy. A separate four-image, three-task, three-repeat matrix tests policy behavior: the larger model recognizes all scenes but follows an explicit HIGH/ROUTINE policy in only 6/12 direct requests and obeys an exact label-only contract in 0/100 broad classification requests. Mapping its parsed scene label to priority in deterministic code instead yields 100/100 correct priorities. The unquantized 2.2B BF16 model loads with only 377 MiB host memory available, then the host reboots before one request is committed; the observed reset reason is a watchdog, not proof of root cause. These results support VLMs as bounded perception and semantic-compression components. They do not support delegating mission policy or command authority to a free-form model.

100/100	6/12	100/100
2.2B scene labels correct	direct priorities correct	priorities after code

1 Question and boundary

Compact VLMs can turn an image into a class, priority, or short description. That makes them attractive for spacecraft payloads: a semantic message may be much smaller than the image that produced it, and local interpretation can occur outside a ground-contact window. Gemma 3n and SmolVLM are explicitly designed for resource-constrained multimodal inference [1, 2]. The engineering question is no longer only whether a transformer can execute on embedded hardware. It is whether a useful model fits, what it costs, and where deterministic software still has to take over.

We separate three claims that are often collapsed:

1. *Physical feasibility*: the model loads and completes every declared request while telemetry remains available.
2. *Perception*: the output contains the correct scene label or directly visible description.
3. *Policy behavior*: the output obeys a disclosed mission rule and an exact machine-readable response contract.

Only the first two are plausible roles for a VLM inside a safety-bounded payload. Policy remains deterministic in the architecture tested here. No output enters a spacecraft command path.

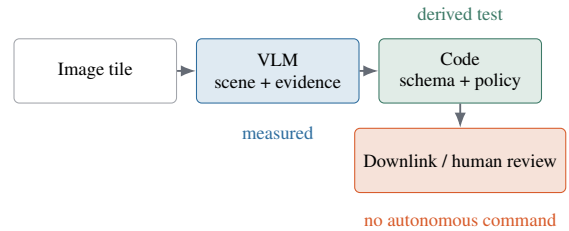


Figure 1: The tested boundary. A VLM produces a bounded perception result. Deterministic code validates schema and applies policy before anything is presented to an operator or another authorized system.

2 Experimental design

2.1 Hardware and software

All measured inference runs execute on the same physical Jetson Orin Nano development board with 7.4 GiB reported usable memory. The board remains in the stock MODE_15W power profile; its governor is observed, not replaced with fixed clocks. NVIDIA documents power profiles as bounds and clock configurations rather than measurements of workload draw [3]. The container uses NVIDIA’s Jetson-specific PyTorch 2.7 build on JetPack 36.4.4 and CUDA 12.9.

The two candidate packages are Apache-2.0 SmolVLM2 checkpoints [4, 5]. The smaller package is 500M-BF16 at pinned revision 7b375e1. The larger is 2.2B-NF4 at revision 482adb5. The latter loads the pinned BF16 checkpoint into bitsandbytes 0.50.1 NF4 weights. NF4 is a 4-bit number format introduced with QLoRA for normally distributed

model weights [6]. Because generic ARM64 CUDA wheels did not target the Jetson L4T runtime and SM87 device, bit-sandbytes was built from pinned source for CUDA 12.9 and SM87, consistent with the project’s Jetson guidance [7].

This comparison intentionally asks a deployment question, not a clean scaling-law question. Model size and weight representation both change. Differences cannot be attributed to parameter count or NF4 alone.

2.2 Images and frozen selection

AID contains 10,000 aerial images across 30 scene classes [8]. We use four classes that match the presentation-shaped task: airport, beach, dense residential, and farmland. Before inference, 25 images per class are ranked by SHA-256 using the fixed seed `nomos-v1m-edge-2026-08-25`. The resulting 100 identities and content hashes are frozen in each input manifest. The first selected image per class forms the smaller presentation matrix.

The images are public aerial RGB tiles collected from Google Earth, not raw spacecraft L0 products. No image is redistributed because the mirror’s license field is unknown. The models are not fine-tuned on AID. We do not test for training-set contamination.

2.3 Tasks and repetitions

The broad sweep presents all 100 images once with greedy decoding and asks for exactly one of four scene labels. One pass measures zero-shot coverage and correctness; it does not measure repeatability.

The presentation matrix uses four images, three tasks, and three repetitions, producing 36 measured requests per model:

1. classify as AIRPORT, BEACH, DENSE_RESIDENTIAL, or FARMLAND;
2. apply HIGH to airport or dense residential and ROUTINE to beach or farmland;
3. describe only visible features in one sentence of at most 30 words.

Classification and priority are scored twice: a lenient parser extracts one unambiguous allowed label, while strict contract scoring requires only the requested label. Description scoring checks a predeclared lexical cue for the true class and the word limit. It is not a complete semantic or hallucination evaluation.

2.4 Physical measurements

Each request starts before image loading and chat-template preprocessing and ends after synchronized generation. We record end-to-end latency, time to first generated token, output tokens, unified-memory views, temperatures, and 50 ms INA3221 VDD_IN samples. Reported power is uncalibrated whole-module input power, not GPU-only draw. Gross request energy integrates those samples. A five-second model-resident idle baseline is recorded separately. Every trial is fsynced before the next begins; no outlier is discarded.

Both measured containers receive a 6,500 MiB memory limit with swap disabled at the cgroup boundary. After model load, a safety gate requires at least 1 GiB host memory available before the first image request.

3 Results

3.1 A larger quantized model changes the answer

Table 1 and Figure 2 report the complete broad sweep. 500M-BF16 returns a contract-compliant label in every request but only 46 are correct. 2.2B-NF4 gets every parsed label correct, but appends a period every time and therefore passes zero strict contracts. The 54-point accuracy gain costs $2.22\times$ median latency, $1.35\times$ mean board power, and $2.95\times$ mean gross energy.

Table 1: One zero-shot classification request on each of 100 frozen images. Power and energy are gross whole-module measurements during each request.

Metric	500M-BF16	2.2B-NF4
Parsed scene correct	46/100	100/100
Exact label contract	100/100	0/100
Median latency (s)	2.213	4.905
Mean board power (W)	8.146	10.983
Mean gross energy (J)	18.282	53.897
Median first token (s)	1.353	4.031
Min. host available (GiB)	3.19	2.77
Telemetry sample errors	0	0

The class breakdown matters more than the aggregate. 500M-BF16 identifies all 25 airports, no beaches, 13 dense-residential scenes, and eight farmlands. It calls 51 non-airport images “airport.” 2.2B-NF4 returns 25/25 in every class. Figure 3 makes this failure mode visible. The selected sample is balanced, so accuracy is also macro recall; it is not representative of an operational class distribution.

3.2 Direct policy fails despite correct scene labels

The presentation matrix is smaller but repeated. All 12 case-task output groups are byte-identical across their three repetitions for both models. Determinism therefore means repeatability under this fixed greedy protocol, not correctness.

500M-BF16 parses the correct scene in 3/12 repeated classification requests, applies the direct priority correctly in 3/12, and includes a predeclared visible class cue in 9/12 descriptions. 2.2B-NF4 parses the correct scene in 12/12 and includes a class cue in 12/12 descriptions, but returns HIGH for every direct priority request. Only six of those 12 answers are correct. Its periods also make every exact classification and priority response noncompliant.

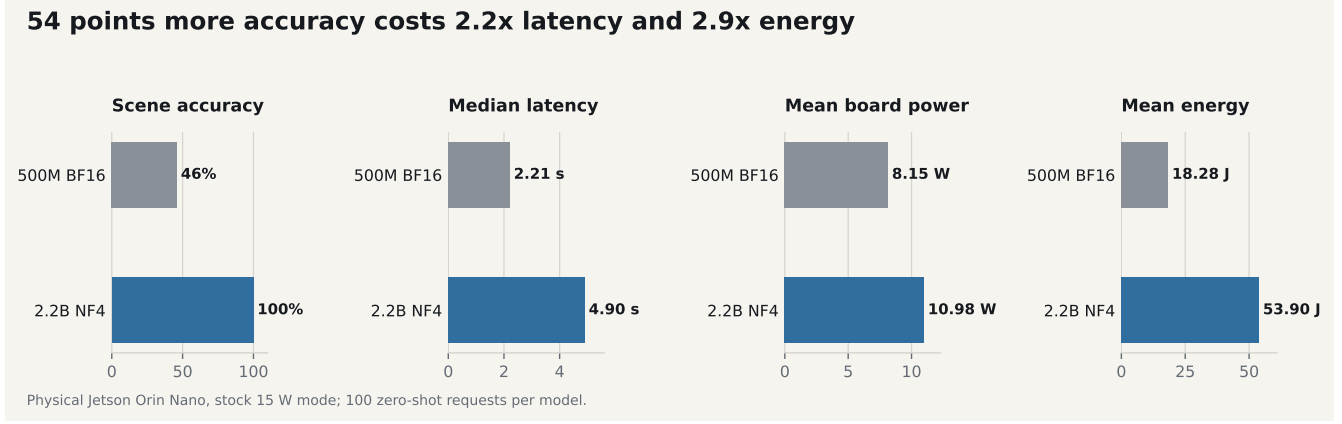


Figure 2: The broad 100-image classification result and its physical cost. Accuracy is parsed label accuracy. Latency includes preprocessing and generation. Power and energy are whole-module measurements.

We take each parsed scene label from the 100-image sweep and apply the same HIGH/ROUTINE mapping in ordinary code. 500M-BF16 then yields 55/100 correct priorities because some scene confusions remain inside the same priority class. 2.2B-NF4 yields 100/100. No second model call or free-form reasoning is required.

3.3 The unquantized model completes no requests

The unquantized 2.2B BF16 checkpoint provides a useful failed baseline. It loads in 112.13 s, with 4.50 GB CUDA allocated, 4.74 GB CUDA reserved, and only 377 MiB host memory available. The host then reboots before the first image request can be committed. Linux reports BCCPLEXWDT as the hardware reset source. That observation localizes the event between model load and the first durable trial; it does not prove memory exhaustion as root cause.

NF4 makes the same checkpoint operational under the declared gate. The quantized load takes 127.22 s, leaves 3.48 GiB host memory available immediately afterward, and completes both the 36-request matrix and 100-request sweep. The load penalty is operationally important: 500M-BF16 loads in 4.73 s. A persistent service can amortize 127 s, but a cold-started or frequently restarted payload may not.

Across the 36-request matrix, maximum measured junction temperature is 55.0°C for 500M-BF16 and 59.9°C for 2.2B-NF4. These are short bench runs in one ambient condition, not a spacecraft thermal validation.

4 What this reproduces

The result reproduces the *measurement shape* of the motivating presentation: image plus prompt enters an onboard-style container; the model produces a classification, priority, or description; and latency, power, memory, repeatability, and task behavior are recorded on physical Jetson hardware. It also extends the four-image demonstration to 100 deterministic images and records a model-fit failure.

It does not reproduce the reported Gemma 3n model, 64 GB Jetson AGX platform, imagery, 30 W setting, software container, or in-orbit environment. Direct numerical comparison would be misleading. Google documents Gemma 3n’s elastic MatFormer design and effective-parameter variants [1]; SmolVLM2 uses a different architecture and training corpus [2]. The available board exposes only the stock 15 W mode without supervised host control during this study.

The AID source material is also not raw L0 imagery. Its labels make it suitable for a transparent ground calibration, not for claims about sensor ingestion, tiling, radiometric correction, cloud screening, or an operator’s actual mission distribution.

5 Implications for onboard observability

This experiment evaluates imagery rather than telemetry diagnosis. It tests one design choice that carries over to observability: which work belongs to a model and which belongs to code.

First, model scale can cross a usefulness threshold. The smaller package physically succeeds and answers quickly, yet its class prior makes the semantic result unreliable. “It runs onboard” is therefore an incomplete gate.

Second, free-form generation is a poor place to enforce machine policy. The larger package sees every selected scene but fails half the direct priority decisions and all exact response contracts. A parser and deterministic mapping turn the same perception output into 100 correct policy labels on this sample. For observability, the analogous design is: let a model summarize bounded evidence, but let code enforce schemas, escalation rules, resource limits, and allowed actions.

Third, quantization is not merely an optimization. On this device it separates a runnable 2.2B service from an unsafe fit boundary. The tradeoff includes a long load path and nearly triple request energy. A real payload needs an explicit resident-model, restart, and degradation policy.

Finally, semantic compression is plausible but unproven

The small model collapsed three classes toward ‘airport’

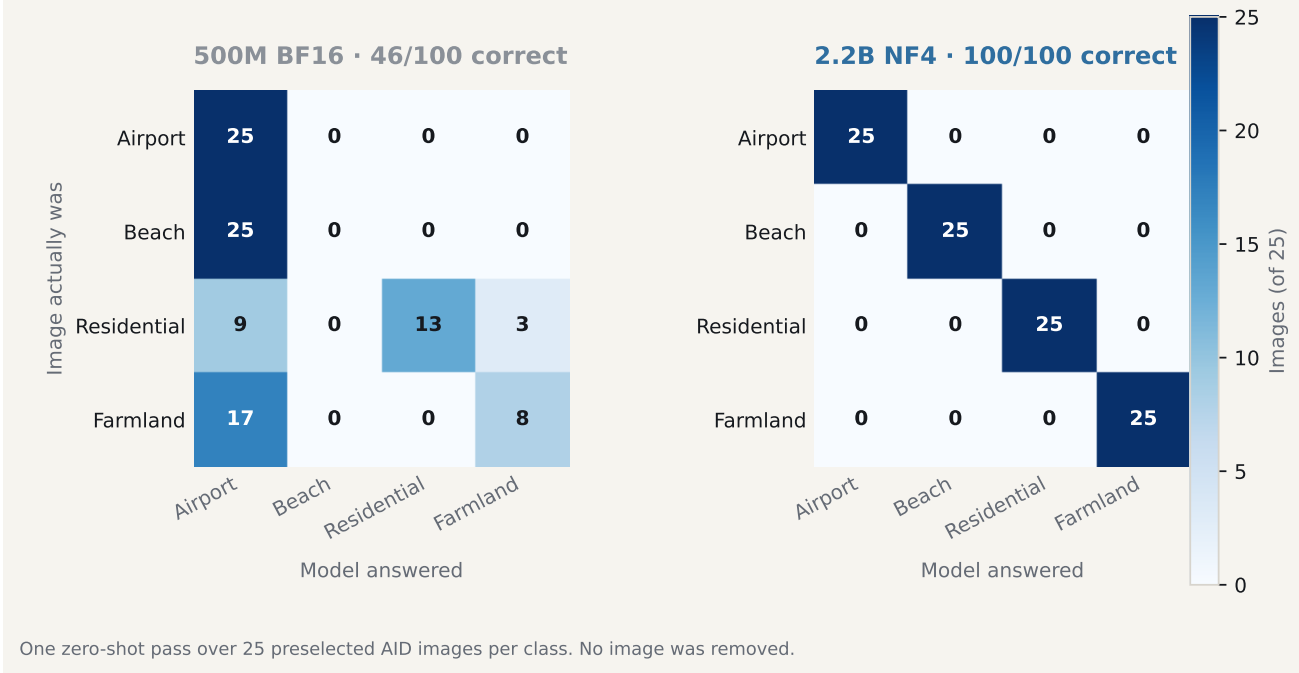


Figure 3: Confusion matrices for the 100-image sweep. The small model’s 46% aggregate hides a strong answer prior. The larger model separates all four selected classes in this calibration set.

here. A short class or description can be much smaller than an image. We did not measure downlink bytes, image value, false-negative cost, or whether the descriptions preserve the evidence an analyst needs. Those require operator-defined tasks and representative private data.

6 Threats to validity

The 100 images cover only four hand-selected AID classes. Selection within each class is deterministic, but the class set is presentation-shaped rather than representative. A perfect result therefore must not be read as general remote-sensing accuracy. Common classes or images related to AID may appear in model training; contamination was not assessed. One broad repetition does not test stochastic stability, while the repeated matrix covers only one image per class.

The model packages differ in parameter count and quantization. We cannot isolate either effect. NF4 is created at load time from the pinned checkpoint rather than served from a pre-quantized artifact. Model-load time therefore includes conversion work. The stock governor, short cadence, ambient bench condition, and uncalibrated VDD_{IN} sensor limit hardware generalization.

The description metric only checks a lexical class cue and word count. It does not detect every unsupported claim. Priority is a toy disclosed mapping, not a mission rule. No radiation, fault containment, real-time scheduling, sensor front end, flight

software, command authentication, digital twin, or human-in-the-loop interface is evaluated.

7 Reproducibility

The public capsule includes both compact 36-request summaries, both 100-request summaries, the cross-model comparison manifests, the sealed BF16 interruption record, exact model and container identifiers, source revision, prompts, class selection seed, measurements, and SHA-256 checksums. Raw trial files remain on the physical bench and are bound by hashes in each summary. AID imagery and model weights are excluded.

No failed request or statistical outlier is removed. The corrupt sparse model download encountered during staging was rejected after upstream LFS SHA-256 verification and is not an input. The BF16 watchdog event is preserved as a result rather than retried under a weakened gate.

8 Conclusion

An 8 GB Jetson Orin Nano can run a useful 2.2B VLM in NF4. On 100 preselected aerial images, it reaches 100/100 parsed scene accuracy at 4.905 s median latency and 10.983 W mean whole-board power. A 500M BF16 package is 2.22× faster and uses about one-third the request energy, but reaches only 46/100.

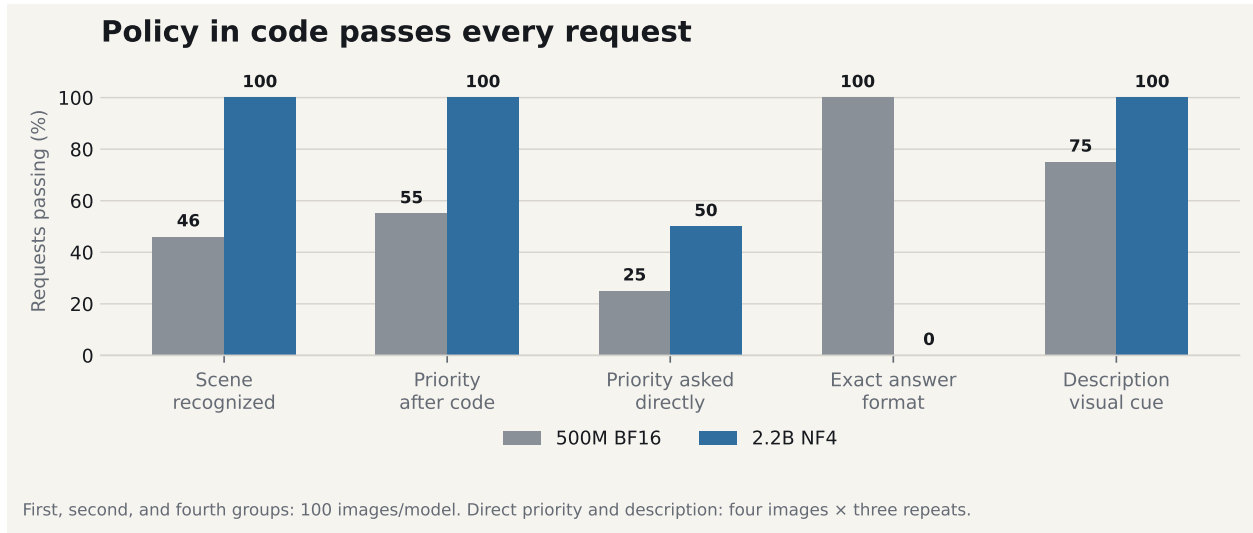


Figure 4: Perception, deterministic policy, direct model policy, schema, and description checks. The groups have different denominators as labelled: broad checks use 100 images per model; direct priority and description use four images repeated three times.

Perfect recognition did not produce reliable direct policy behavior or exact machine output. In the tested composition, the VLM performs perception while deterministic software interprets that observation under mission rules. Command authority remains with a human or another authorized system.

References

- [1] Google, “Gemma 3n model overview.” Google AI for Developers, 2025.
- [2] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. Ben Allal, A. Lozhkov, N. Tazi, V. Srivastav, J. Lochner, H. Larcher, M. Morlon, L. Tunstall, L. von Werra, and T. Wolf, “SmolVLM: Redefining small and efficient multimodal models,” *arXiv preprint arXiv:2504.05299*, 2025.
- [3] NVIDIA, “Jetson orin power management and performance.” Jetson Linux Developer Guide, release 36.4.4, 2025.
- [4] Hugging Face, “SmolVLM2-500M-Video-Instruct model card.” Hugging Face model repository, 2025. Revision 7b375e1b73b11138ff12fe22c8f2822d8fe03467 accessed 25 August 2026.
- [5] Hugging Face, “SmolVLM2-2.2B-Instruct model card.” Hugging Face model repository, 2025. Revision 482adb537c021c86670beed01cd58990d01e72e4 accessed 25 August 2026.
- [6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient finetuning of quantized LLMs,” *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [7] Hugging Face, “bitsandbytes installation guide.” Hugging Face documentation, 2026.
- [8] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.