EMBEDDING-ONLY UPLINK FOR ONBOARD RETRIEVAL UNDER SHIFT IN REMOTE SENSING

Sangcheol Sim
Telepix
sim2real@telepix.net

ABSTRACT

Downlink bottlenecks motivate onboard systems that prioritize hazards without transmitting raw pixels. We study a strict setting where a ground station uplinks *only* compact embeddings plus metadata, and an onboard system performs vector search to triage new captures. We ask whether this embedding-only pipeline remains useful under explicit remote-sensing shift: *cross-time* (pre/post-event), *cross-event/location* (different disasters), *cross-site cloud* (15 geographic sites), and *cross-city AOI holdout* (buildings). Using OlmoEarth embeddings on a scaled public multi-task benchmark (27 Sentinel-2 L2A scenes, 15 cloud sites, 5 SpaceNet-2 AOIs; 10 seeds), we find that all effective methods rely on the same uplinked embeddings, but the *optimal decision head is task-dependent*: kNN retrieval is significantly superior for cloud classification (0.92 vs. centroid 0.91; $p<0.01$, Wilcoxon), while class centroids dominate temporal change detection (0.85 vs. retrieval 0.48; $p<0.01$). These results show that embedding-only uplink is the key enabler—once embeddings are onboard, the system can select the best head per task at no additional uplink cost, with all telemetry under 1 KB per query.

1 INTRODUCTION

Operational disaster response faces a mismatch between sensing capacity and limited contact windows for downlink (Denby & Lucia, 2020). Retrieval against a compact onboard memory can prioritize what to transmit; we focus on whether such retrieval remains effective under distribution shift for remote-sensing triage.

We study a strict setting: **ground-to-space uplink is embedding-only (plus metadata)**. Hints are (embedding, metadata) tuples; the onboard system indexes them in a vector database and retrieves the top-k most similar candidates (k typically 1–10) for each new capture. Rather than optimizing a single dataset, we ask whether this pipeline remains useful under **explicit RS shift axes** (Tuia et al., 2016)—*cross-time* (pre vs. post disaster), *cross-event/location* (different disasters and hard negatives), *cross-site cloud* (15 geographic sites), and *cross-city AOI holdout* for buildings—while emitting compact, auditable telemetry as the downlink product.

Related work. ESA's Φ-Sat-1 demonstrated onboard CNNs for EO, reducing downlink by $\sim 90\%$ (Giuffrida et al., 2022); recent surveys cover foundation-model deployment on satellite platforms (Sang et al., 2026). Self-supervised RS encoders—SatMAE (Cong et al., 2022), SpectralGPT (Hong et al., 2024), OlmoEarth (Herzog et al., 2025)—yield capable representations; we evaluate whether they support retrieval under shift, not proposing a new encoder. RAG (Lewis et al., 2020) augments LMs with retrieved context; we adapt retrieval to a satellite setting where "generation" is compact telemetry.

Contributions. We do not propose a new encoder or retrieval algorithm; our contribution is the first systematic evaluation of embedding-only uplink for onboard RS triage under shift. We contribute **(i)** an end-to-end pipeline for embedding-only uplink and onboard retrieval with compact telemetry (Figure 1), **(ii)** a scaled reproducible multi-task benchmark (hazard, change, cloud, buildings; six baselines, 10 seeds, k-sweep, paired significance tests), and **(iii)** the finding that the optimal decision head is *task-structure-dependent*, not difficulty-dependent: kNN retrieval excels for continuous-

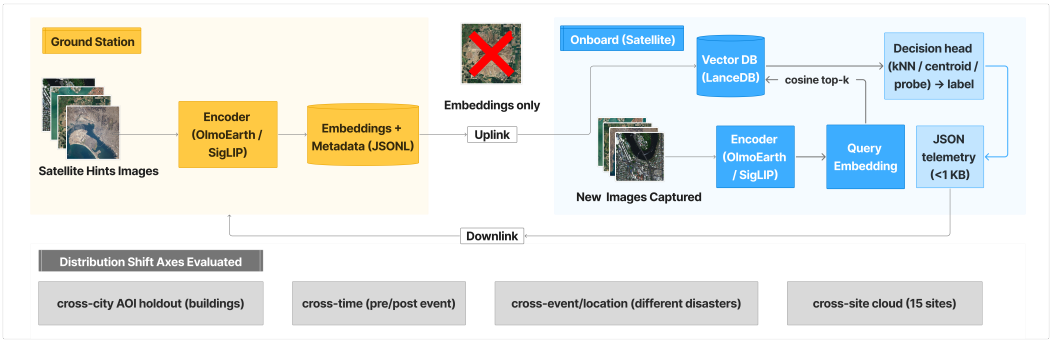

Figure 1: Embedding-only uplink pipeline. The ground station computes hint embeddings and uplinks (embedding, metadata) tuples—no imagery. The onboard system indexes hints in a vector database and retrieves top-k matches for each new capture via cosine similarity, emitting compact JSON telemetry ($\sim$700 B at k=5) as the downlink product.

label tasks while class centroids dominate discrete temporal classification—but all heads rely on the same uplinked embeddings.[1]

2 METHOD

2.1 EMBEDDING-ONLY UPLINK AND ONBOARD INDEXING

On the ground, we compute an embedding per hint image and store it with metadata in JSONL. The onboard system ingests these rows into LanceDB (LanceDB contributors, 2026) and performs cosine nearest-neighbor search at query time. Crucially, uplink carries *no imagery*.

2.2 MULTI-MODAL EMBEDDINGS AND DECISION HEAD

We use task-appropriate embedding backbones for public remote-sensing imagery: OlmoEarth (D=768) for Sentinel-2 L2A 12-band patches (Herzog et al., 2025) and SigLIP (D=1152) for RGB building tiles (Zhai et al., 2023). L2A surface reflectance provides radiometrically consistent inputs; the 12-band spectral signature enables OlmoEarth to capture land-cover semantics without RGB reduction (single sensor: Sentinel-2 at 10 m GSD for hazard/change/cloud; WorldView-3 via SpaceNet-2 for buildings). For a query, we retrieve top-k hints ($k \in \{1, 5, 10\}$) and apply a cosine-similarity-weighted kNN vote (Cover & Hart, 1967) to produce a task label. Patches are 256$\times$256 pixels; embeddings are L2-normalized after OlmoEarth's computed normalization. Cloud labels use STAC `eo:cloud_cover` thresholds (clear$\leq$10%, cloudy$\geq$20%). As non-retrieval baselines, we evaluate (i) a nearest-centroid prototype classifier (Snell et al., 2017) and (ii) a ridge-regression linear probe (Alain & Bengio, 2017) (ℓ_2=10^{-3}), both operating on hint embeddings only.

2.3 COMPACT, AUDITABLE TELEMETRY

The onboard system emits a minimal JSON record per query (task id, label, top-k hints with scores); we report the serialized byte size. Uplink cost is $N_{\text{hints}} \cdot D \cdot b$ bytes per hint-set refresh ($b \in \{4, 2, 1\}$ for FP32/FP16/INT8). Downlink telemetry averages 598–690 B per query (k=5); retrieval latency is $\sim$5 ms (LanceDB cosine, single CPU thread).

[1]Code, data pipeline, and benchmark splits are released open-source: `https://github.com/weirdsim14/orbit-RAG`

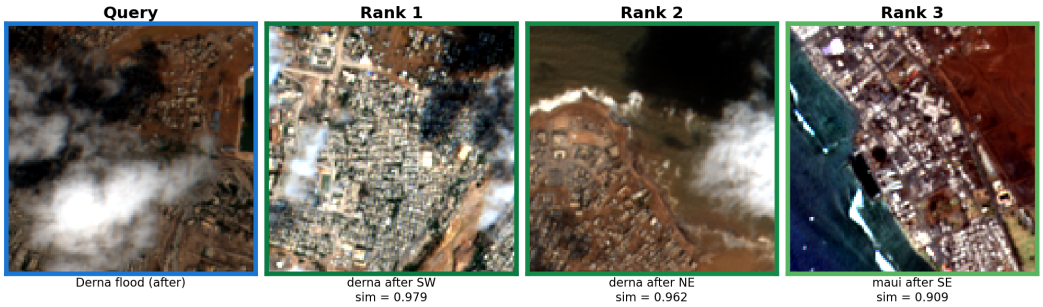

Figure 2: Qualitative retrieval example (hazard task). A Derna flood query retrieves same-event crops (sim>0.96) before a cross-event match (Maui wildfire, sim=0.91), demonstrating embedding-level hazard grouping.

3 BENCHMARK AND METRICS

3.1 MULTI-TASK PUBLIC BENCHMARK

We evaluate four tasks with a unified retrieval interface and **explicit split units** aligned with RS leakage risks (Roberts et al., 2017). The benchmark comprises: 27 Sentinel-2 scenes ($\sim$70 hints, $\sim$30 queries) for hazard; 8 cross-time pairs for change; 15 geographic sites with 5 clear/cloudy scenes each ($\sim$300 hints, $\sim$150 queries) for cloud; and 5 SpaceNet-2 AOIs (up to 30 tiles/AOI, $\sim$40 queries) for buildings. All data sources are public/no-auth (European Space Agency, 2026; Element 84, 2026; Van Etten et al., 2018). We repeat over 10 random seeds; seeds deterministically choose held-out quadrants (Sentinel-2 tasks) and per-AOI tile subsamples (SpaceNet) to report mean$\pm$std.

(1) Hazard retrieval (Sentinel-2 L2A (European Space Agency, 2026; Element 84, 2026; Herzog et al., 2025)): Hazard retrieval tests whether embeddings cluster scenes by disaster type well enough to triage new captures as wildfire, flood, or normal without examining imagery. Each scene is labeled by disaster-type group (wildfire, flood, or normal); hints are multiple quadrant crops per scene; queries use a held-out quadrant (leave-one-crop-out). Retrieval tests whether embeddings group scenes by disaster type: success requires correct-group hints in the top-k (Recall@k) and correct top-1 (Top-1 accuracy). The task is challenging because spectral signatures partially overlap across wildfire, flood, and normal scenes; hard negatives arise when different disasters share visual characteristics. Optional normal-scene queries measure false-positive rate. **(2) Change (cross-time preference)** (Shi et al., 2020): a query from the after scene is correct if similarity to the after group exceeds the before group, testing whether embeddings preserve temporal ordering for triage prioritization. **(3) Cloud classification** (Sentinel-2 L2A): labels from STAC `eo:cloud_cover` threshold; site-holdout split with held-out quadrant crops (anti-leakage). **(4) Buildings presence** (SpaceNet-2 (Van Etten et al., 2018), 0 vs. 1+): hints from non-holdout AOIs, queries from a held-out AOI (cross-city); GeoTIFFs converted to 8-bit RGB with 2–98% percentile stretch before SigLIP embedding.

3.2 METRICS

Per task: **Recall@k** (any of k hints from correct group), **Top-1 accuracy** (nearest hint correct), **time-preference accuracy** (after-group sim $>$ before-group; tests temporal ordering for change triage), **balanced accuracy** (mean per-class recall; handles site-level class imbalance for cloud), **macro-F1** (mean per-class F1; handles label skew for buildings), and **payload size** (serialized JSON bytes per query).

4 RESULTS

All embedding-based methods significantly outperform embedding-free baselines, confirming that the uplinked embeddings are the key enabler. However, the optimal decision head varies by task structure (Table 1; k=5, mean$\pm$std, 10 seeds; significance via paired Wilcoxon signed-rank test).

Table 1: Multi-task results at k=5 (mean±std, 10 seeds, xlarge benchmark). Bold: best non-oracle per row. Significance markers from paired Wilcoxon signed-rank test vs. retrieval: $^{**}p<0.01$, $^{*}p<0.05$.

Task	Metric	Retrieval	Random	Centroid	Lin. probe	No-retr.	Oracle
Hazard (S2)	Recall@5	**1.00**±.00	.15±.09**	.99±.04	**1.00**±.00	.00±.00**	1.00±.00
Hazard (S2)	Top-1	**.91**±.08	.03±.03**	.86±.12	.90±.08	.00±.00**	1.00±.00
Change (S2)	Time-pref	.48±.08	.00±.00**	**.85**±.15**	.24±.19**	.00±.00**	1.00±.00
Cloud (S2)	Bal. acc	**.92**±.04	.50±.05**	.91±.05**	.92±.04	.50±.00**	1.00±.00
Build. (SN2)	Macro-F1	.51±.26	.42±.10	**.70**±.12	.64±.17	.40±.03	1.00±.00

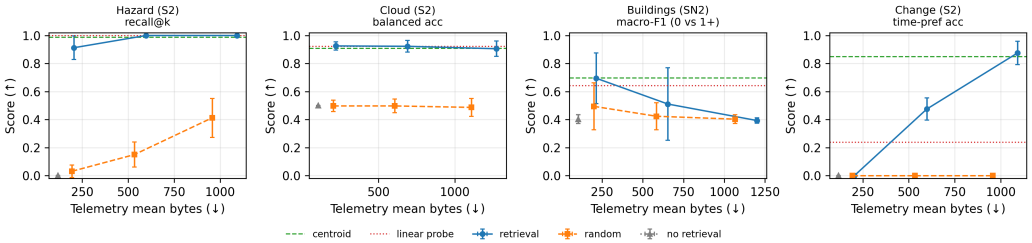

Figure 3: k-sweep: task metric vs. telemetry bytes ($k \in \{1, 5, 10\}$). Dashed horizontal lines show k-independent baselines (centroid, linear probe). Buildings favors small k; change improves with larger k. Error bars: ±1 std over 10 seeds.

Embeddings are the key enabler. Every embedding-based method (retrieval, centroid, linear probe) significantly outperforms random and no-retrieval baselines across all tasks ($p<0.01$), confirming that the uplinked embeddings—not the specific decision head—drive triage performance. All telemetry remains under 700 B per query (k=5).

Optimal head is task-dependent. For *cloud* (15 sites, ∼150 queries), kNN retrieval (0.92) significantly outperforms centroid (0.91; p=0.004). Cloud cover is a continuous spectrum; kNN captures local similarity structure that centroid averaging smooths away. For *change* (8 cross-time pairs), centroid (0.85) significantly outperforms retrieval (0.48; p=0.002). Before/after discrimination is a discrete class-level concept; the centroid captures the mean "post-disaster" pattern more robustly than nearest-neighbor matching. For *hazard*, retrieval and centroid both achieve near-perfect Recall@5 (1.00); Top-1 shows a non-significant advantage for retrieval (0.91 vs. 0.86; p=0.22). Figure 2 illustrates embedding-level hazard grouping qualitatively. For *buildings* under cross-city shift, centroid (0.70) trends above retrieval (0.51; p=0.16), but the difference is not statistically significant, likely due to limited AOI count (5) amplifying variance under cross-city shift.

k-sensitivity reveals task-specific operating points (Figure 3): buildings favors k=1 while change improves with larger k, motivating per-task selection in heterogeneous deployments.

5 Conclusion

Embedding-only uplink supports multi-task RS triage under shift with ∼700 B telemetry (k=5): kNN retrieval excels for continuous-label tasks (cloud, $p<0.01$) while centroids dominate discrete temporal classification (change, $p<0.01$). All heads share the *same* uplinked embeddings, so practitioners select heads per task onboard at no additional uplink cost. Task structure (continuous vs. discrete label) predicts the optimal head, enabling zero-shot head selection without held-out validation data.

Limitations. Single optical sensor (Sentinel-2 L2A, 10 m GSD); cross-sensor, cross-season, noise-robustness, and score calibration remain future work. Near-perfect hazard Recall@5 on 27 scenes warrants larger-scale replication.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Bradley Denby and Brandon Lucia. Orbital edge computing: Nanosatellite constellations as a new class of computer system. In *Proceedings of the 25th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 939–954, 2020.

Element 84. EarthSearch STAC catalog. `https://earth-search.aws.element84.com`, 2026. Accessed 2026-01-28.

European Space Agency. Sentinel-2 MSI level-2a. `https://sentinel.esa.int/web/sentinel/missions/sentinel-2`, 2026. Accessed 2026-01-28.

Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Leonie Buckley, Aubrey Dunne, Chris van Dijk, Marco Esposito, John Hefele, Nathan Vercruyssen, Gianluca Furano, Massimiliano Pastena, and Josef Aschbacher. The Φ-Sat-1 mission: The first on-board deep neural network demonstrator for satellite earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. OlmoEarth: Stable latent image modeling for multimodal earth observation. *arXiv preprint arXiv:2511.13655*, 2025.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024.

LanceDB contributors. LanceDB. `https://lancedb.com`, 2026. Accessed 2026-01-28.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929, 2017.

Hanbo Sang, Limeng Zhang, Tianrui Chen, Weiwei Guo, and Zenghui Zhang. Onboard deployment of remote sensing foundation models: A comprehensive review of architecture, optimization, and hardware. *Remote Sensing*, 18(2):298, 2026.

Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016.

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.