Talk is Cheap: Querying SAR with Text via Ground-Photo Alignment

Mikolaj Czerkawski
Asterisk Labs, London, UK
miko@asterisk.coop

Abstract

Synthetic Aperture Radar (SAR) backscatter does not resemble ground appearance, so optical vision-language models do not generalise to it; the recent SAR-text corpora that have begun to fill the gap (SARChat-2M, SARLang-1M, SAR-TEXT, FSAR-Cap) still require SAR-specific text construction — via templates, detection-label expansion, or automated narrators. GLUE-Link *bypasses that requirement entirely. A single* 1×1 *linear projection maps any frozen satellite encoder into the SigLIP-2 [7] ground-photograph embedding space, supervised only by geographic co-location with LUCAS field photos — no labels, no captions, no SAR-language pairs. With a pure SAR encoder (SSL4EO-S1) the linked representation reaches* 45.4% *zero-shot top-1 on 8-class LUCAS land cover and Spearman* ρ=0.890 *caption agreement, matching the multispectral MS-CLIP (*45.7%*,* ρ=0.549*) and approaching the native optical model from a concurrent companion work [22] (*50.7%*,* ρ=0.912*). Any new satellite encoder — including closed multi-sensor models such as AlphaEarth — can be made text-queryable in one linear training step on existing ground-photo surveys.*

1. Introduction

Earth observation archives contain a massive amount of information about our planet. However, the most popular sensors, such as Sentinel-2, often rely on multi-spectral passive sensing with optical components, which comes at the risk of many observations being obscured by clouds. While potentially useful for weather and climate analyses, cloudy images obscure the information about the surface of the planet. On the other hand, active sensors like Synthetic Aperture Radar (e.g. Sentinel-1) have the all-weather capability at the cost of being generally difficult to interpret or annotate for non-expert humans. This represents a challenging trade-off between optical data, which gets affected by weather but is easier to annotate, and SAR data, which penetrates the clouds, but is less visually intuitive.

This trade-off is the reason why the majority of vision-language solutions in Earth observation (MS-CLIP [8], ChatEarthNet [9], RS5M/GeoRSCLIP [10]) have historically focused on the optical domain. Large-scale SAR image-text corpora only began appearing in 2025 (SARChat-2M [11], SARLang-1M [12], SAR-TEXT [15], FSAR-Cap [16]), and they all construct supervision indirectly — expanding existing detection labels or scene tags into text via templates or automated narrators — rather than collecting natural language descriptions of SAR scenes from human annotators, since the public cannot read SAR backscatter.

This work introduces a bypass bridge, called GLUE-Link. It is based on the existing pioneering work on GLUE (Geometrically-Lifted Unified Earth Embeddings), where densely distributed ground-level images (205 thousand geo-referenced points across Europe with four cardinal directions from the LUCAS survey) are used to connect latent space of existing off-the-shelf vision-language models like SigLIP with satellite image encoders. While the initial GLUE work also prioritises multi-spectral optical sensors (and achieves state of the art performance), this work builds on the realisation that ground-level embedding targets can be linked to any existing geospatial encoder, including those relying on SAR data. Furthermore, training a GLUE-Link adapter can be very cheap once the embeddings from a geospatial encoder are precomputed.

As a result, we introduce a powerful and interoperable paradigm for linking any geospatial AI model with text. GLUE-Link is a one-layer linear projection from any satellite encoder into natural language domain (here: SigLIP-2 embeddings). To demonstrate this, a suite of multiple text-aligned SAR encoders is released, with no need of direct SAR-language supervision. Comparison across five encoder families (Table 1): single-date SAR (SSL4EO-S1, DINOv2-S1, BetaEarth-S1), annual fused (Tessera), and closed multi-sensor (AlphaEarth) suggests the protocol generalises across input modalities and embedding sizes (64–$1024\,$d, including a pure-SAR 64-d encoder). Finally, retrieval agreement with ground-photo SigLIP-2 (Table 2) shows that GLUE-Linked SAR retrieves the same tiles a ground photographer would, 10–$25\times$ above chance.

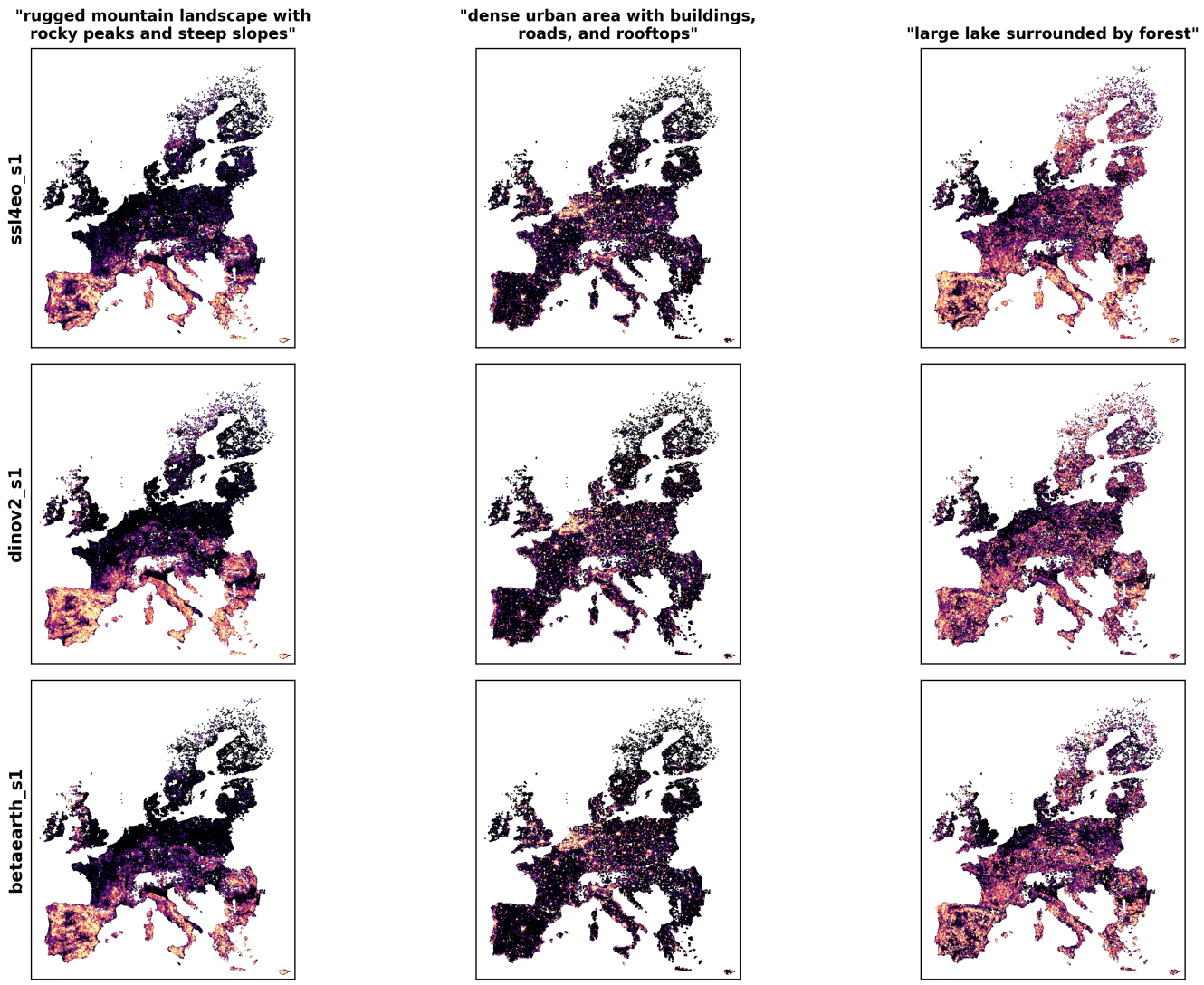

Figure 1. Continental-scale text queries via GLUE-Link. Rows: three pure-SAR encoders (SSL4EO-S1; DINOv2-S1; BetaEarth-S1, a 64-d encoder). Columns: three text queries. Each panel scatters ∼200K LUCAS points across Europe, coloured by cosine similarity between the linked SAR embedding and the SigLIP-2 text embedding (PowerNorm γ=2.5 to emphasise the upper tail). The same SAR representation answers all three queries; no labels, captions, or SAR-text pairs were used in training. Geographic patterns are consistent across encoders despite a 16× range in embedding size. Point coverage follows the LUCAS field survey, which samples EU member states on accessible terrain; non-EU areas (most of Norway) and sparsely-surveyed high-alpine and remote regions therefore appear as gaps in coverage rather than as model failures. The bright regions coincide with well-known references — orogenic belts for mountains, major agglomerations for urban — with quantitative validation against gold ground-photo retrieval in Tables 1–2; the lake-and-forest query is the weakest of the three and over-responds in some non-lake regions.

2. Related work

Caption-supervised satellite VLMs.

Solutions like MS-CLIP [8] (10 Sentinel-2 bands; bands 1 and 10 are omitted), ChatEarthNet [9] (9 Sentinel-2 bands), and RS5M/GeoRSCLIP [10] (RGB, 5M web-aggregated pairs) train on hundreds-of-thousands to millions of image-caption pairs from optical satellites. SAR-specific corpora are addressed separately below; common to both families is that some form of (image, text) pairing must be constructed and curated before training can begin.

Ground-photo–supervised satellite VLMs.

A second family of methods avoids satellite-text annotation entirely by exploiting geographic co-location with crowdsourced or surveyed ground photographs. Sat2Cap [18] (Bing Maps RGB at 0.6 m/px), GRAFT [17] (NAIP RGB at 1 m/px and Sentinel-2 RGB-only B4/B3/B2), SenCLIP [19] (Sentinel-2 RGB), and TimeSenCLIP [20] (Sentinel-2 RGB or multispectral time series) align satellite tiles to image features of nearby photographs taken on the ground; text-queryability is then inherited at inference from the frozen vision-language model whose image space is targeted. All four methods use optical satellite input, and the protocol has not previously been extended to SAR. The concurrent companion work GLUE [22] sits in the same family and serves as the optical reference point in Table 1; the architectural extensions GLUE-Link introduces to make the protocol work for SAR (and for tiny closed-model embeddings) are described in Section 3.

SAR-text corpora and SAR VLMs (concurrent, 2025).

Four large SAR image-text datasets appeared in 2025: SARChat-2M [11] ($\sim$2M dialogue and VQA pairs), SARLang-1M [12] ($>$1M pairs spanning 59 cities), SAR-TEXT [15] (130K pairs produced by the SAR-Narrator auto-generation framework), and FSAR-Cap [16] (14K images and 72K pairs built on the FAIR-CSAR detection benchmark). All four construct text indirectly — by expanding existing detection bounding boxes or scene tags through templates or LLM rewriters — rather than by collecting natural language descriptions of SAR scenes from human annotators, since the public cannot read SAR backscatter.

The released models alongside these corpora are predominantly *generative*. SARChat ships fine-tunes of Qwen2VL, LLaVA-1.5, InternVL2.5 and similar conversational backbones for SAR question-answering; these emit text rather than expose a cosine-comparable image embedding, so they do not slot into a CLIP-style retrieval benchmark without protocol changes. The most direct contrastive counterpart is SARCLIP [13] (IEEE TGRS 2025), a CLIP-style SAR-text foundation model trained on the SARCAP corpus ($\sim$400K image-text pairs derived from SAR detection and land-cover labels via the SARTEX textualisation strategy), released with pretrained weights. A second, concurrent work that also calls itself SARCLIP [14] (and is sometimes referred to as SARVLM, arXiv October 2025) reports stronger numbers on a $\sim$1M aggregated dataset but has announced weights as forthcoming.

SARCLIP's pretrained weights are distributed only via Baidu Netdisk; a request to re-host them on HuggingFace was declined by the authors, who note that the model is tightly coupled to a custom data-loading and preprocessing pipeline and that using it in a different harness "may lead to incorrect or misleading results" [13]. We therefore do not attempt a head-to-head numerical comparison in this workshop submission. The contrast is itself informative: GLUE-Link is encoder-agnostic and protocol-portable by construction, whereas SARCLIP's text-queryability is bound to its training-time pipeline.

More fundamentally, GLUE-Link is complementary to all SAR-text pretraining — including SARCLIP and SARVLM. It sees no SAR-text pair (synthetic, template-derived, or otherwise); supervision is location-aligned *natural ground photographs*, and text-queryability is inherited at inference from a frozen general-purpose VLM (SigLIP-2). SAR-text pretraining and GLUE-Link-style ground alignment could be combined, and neither subsumes the other.

3. Method

The method can take any existing satellite encoder that produces a feature grid $\mathbf{F} \in \mathbb{R}^{H\times W\times d}$. The target is defined as the ground-level photo embeddings from the LUCAS dataset using SigLIP-2 (giant) $\mathbf{g}_v \in \mathbb{R}^{1536}$, where each view sees roughly the 90° wedge in front of it. A soft directional mask $\mathbf{M}_v(i,j) = \max(0, \cos(\theta_{ij} - \theta_v))$ weights the spatial cells by alignment with view direction.

GLUE-Link is a single learnable 1×1 linear layer $\mathbf{W} \in \mathbb{R}^{d\times 1536}$ applied at every spatial location, producing $\mathbf{P} = \mathbf{FW}$. The prediction is pooled as follows: $\hat{\mathbf{g}}_v = \text{norm-mean}_{i,j}(\mathbf{M}_v(i,j)\,\mathbf{P}(i,j))$. The model is trained with a single cosine loss against the matching ground embedding, summed over four views:

$$\mathcal{L} = \frac{1}{4}\sum_v (1 - \cos(\hat{\mathbf{g}}_v, \mathbf{g}_v)).$$

At inference time, pooling is dropped, and the output features $\mathbf{P} \in \mathbb{R}^{H\times W\times 1536}$ can be compared with a text embedding via cosine similarity, which can be represented as alignment maps at the resolution of each encoder's feature grid: 14×14 for SSL4EO-S1 and (after adaptive pooling) BetaEarth-S1, and 50×50 for AlphaEarth and Tessera. BetaEarth-S1's native output is dense at the input pixel grid ($224 \times 224 \times 64$, i.e. 10 m); we adaptive-pool to 14×14 in this work for storage parity with the ViT encoders.

The following encoders are evaluated in this work:

- **SSL4EO-S1** [2]: 2-band (VV, VH) Sentinel-1 specialist model; ViT-L/14 grid $14 \times 14 \times 1024$.
- **DINOv2-S1**: general-purpose DINOv2 [6] applied to a (VV, VH, VV$-$VH) false-colour composite in place of RGB input; output grid $16 \times 16 \times 1024$.
- **BetaEarth-S1** [5]: open SegFormer-B2 encoder with FiLM modality conditioning, used in S1-only mode (VV, VH); 64-d dense output at the native 10 m pixel grid ($224 \times 224 \times 64$), adaptive-pooled to $14 \times 14 \times 64$ in this work for storage parity.
- **Tessera** [3]: annual S1+S2 embedding product; grid $50 \times 50 \times 128$.

- **AlphaEarth** [4]: closed multi-sensor annual model (Google DeepMind); embeddings only; $50 \times 50 \times 64$.

4. Experiments

4.1. Setup

LUCAS 2018 is split by NUTS-2 region: 160,936 train / 21,439 val / 22,634 test points; 22,441 test tiles have all required caches and enter Tables 1–2. Each LUCAS photograph is encoded once with SigLIP-2 (Giant, 1536-d) and cached; the tile-level reference embedding is the L_2-normalised mean of the four cardinal views. Training uses 4 (view, tile) pairs per point (643,744 pairs total), AdamW (lr 10^{-3}, weight decay 10^{-4}, cosine schedule), 20 epochs, batch size 256, on a single A100. The linear head is dwarfed by I/O: end-to-end training takes $\sim$17 min for BetaEarth-S1 (64-d) up to $\sim$2 h for SSL4EO-S1 (1024-d).

Caption-agreement vocabulary. A fixed set of 757 free-form landscape phrases organised into 46 categories spanning natural cover (forests, water, mountains), built environment (residential, transport, energy, heritage), country-specific motifs, and sensory or seasonal descriptors (e.g. *"conifer plantation in regular rows"*, *"ridge and furrow medieval ploughing pattern"*). Each descriptor is encoded once by SigLIP-2 and used as a fixed probe set. The vocabulary was curated to reflect what a person might recognise standing at a location in Europe rather than to map onto any pre-existing classification schema, and is the same set used in the GLUE companion work [22].

Metrics. Val/Test cosine is the mean cosine similarity between the wedge-pooled prediction and the matching ground-photo embedding. Top-1 is zero-shot 8-class LUCAS land cover: because the linked encoder targets ground-photo SigLIP-2, the class prompts are phrased as photographs, not aerial views (*"a photograph showing artificial land and built-up areas"*, etc.); the tile takes the highest-cosine prompt. The two Spearman ρ columns measure different things: in Table 1 the rank is per tile over the 757 descriptors (caption-profile agreement); in Table 2 the rank is per query over the 22,441 tiles, averaged across 100 descriptors sampled with a fixed seed. P@K is the top-K overlap of the same two rankings (random: $K/22{,}441$). MS-CLIP and CLIP rows use the same protocol; GLUE numbers come from [22].

4.2. Zero-shot performance

Pure SAR is text-queryable. SSL4EO-S1 reaches 45.4% zero-shot top-1 and ρ=0.890 caption agreement with no optical input. This matches the 10-band MS-CLIP on classification (45.7%) and exceeds it on ρ by +0.34. DINOv2-S1 (46.1% / 0.892) and BetaEarth-S1 (45.1% / 0.891) land within a percentage point. All three pure-SAR encoders sit within $\sim$5 pp of the optical companion work [22] (50.7%).

Table 1. Zero-shot performance on 22,441 held-out LUCAS test tiles (subset of the 22,634 split with all required caches). **Mod.** is the encoder's input modality (SAR: VV/VH only; SAR+OPT: SAR + optical time series; MULTI: closed multi-sensor annual). Val/Test: cosine between pooled prediction and ground-photo SigLIP-2. Top-1: 8-class LUCAS via zero-shot SigLIP-2 prompts. ρ: Spearman caption agreement over 757 descriptors. †: pure SAR. ‡: GLUE is concurrent companion work [22] (in preparation), reproduced for context.

Encoder	Mod.	Val	Test	Top-1%	ρ
This work — GLUE-Link					
AlphaEarth (64d)	MULTI	0.881	0.881	**54.6**	**0.921**
Tessera (128d)	SAR+OPT	0.881	0.879	**54.6**	0.919
DINOv2-S1† (1024d)	SAR	0.864	0.866	46.1	0.892
BetaEarth-S1† (64d)	SAR	0.864	0.865	45.1	0.891
SSL4EO-S1† (1024d)	SAR	0.860	0.863	45.4	0.890
Reference — native optical					
GLUE‡ [22]	rgb	—	—	50.7	0.912
MS-CLIP [8]	10-band	—	—	45.7	0.549
CLIP [21]	rgb	—	—	36.9	0.444

Table 2. Retrieval agreement with ground-photo SigLIP-2 on the test set. For 100 descriptors sampled from the 757-concept vocabulary, all 22,441 tiles are ranked by cosine similarity twice: once using the linked encoder, once using the gold four-view-averaged ground photo. **P@K**: top-K overlap of the two rankings. ρ: Spearman correlation of the full per-query rankings. Random baseline at P@100: 0.004. †: pure SAR.

Encoder	Mod.	P@100	P@500	ρ
Multi-sensor / fused (with optical)				
AlphaEarth (64d)	MULTI	**0.112**	**0.237**	**0.567**
Tessera (128d)	SAR+OPT	0.104	0.221	0.545
Pure SAR† (no optical input)				
BetaEarth-S1 (64d)	SAR	0.055	0.143	0.381
DINOv2-S1 (1024d)	SAR	0.048	0.141	0.403
SSL4EO-S1 (1024d)	SAR	0.043	0.124	0.378
Random baseline	—	0.004	0.022	0.000

Embedding size does not bound SAR performance. BetaEarth-S1 (64-d) matches the 1024-d ViT-L encoders on every metric in Table 1 despite using 16$\times$ fewer dimensions; the bottleneck for pure-SAR open-vocabulary retrieval is upstream of the linear link.

Multi-sensor / fused encoders set the upper bound. AlphaEarth and Tessera reach 54.6% / $\rho \approx 0.92$, exceeding the optical baseline by +3.9 pp. These rows include optical input and are not apples-to-apples with the pure-SAR rows; they characterise the GLUE-Link envelope.

DINOv2 transfers to SAR. Frozen DINOv2 on (VV, VH, VV$-$VH) RGB matches SAR-specialist SSL4EO-S1 (+0.7 pp, +0.002 ρ); natural-image texture statistics carry over without radar pretraining. A single 1$\times$1 linear head

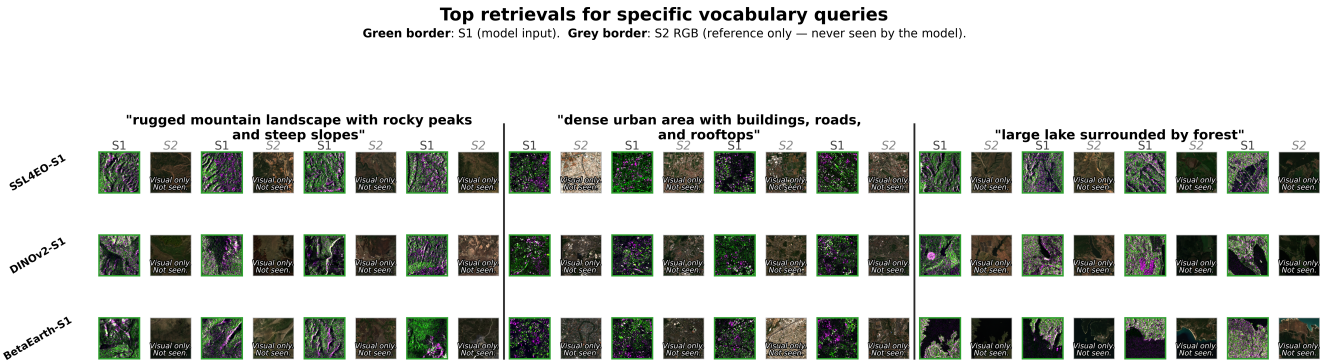

Figure 2. Top-4 retrievals for three text queries on the held-out test set, by SSL4EO-S1, DINOv2-S1, and BetaEarth-S1 linked via GLUE-Link. For each retrieved tile we show the S1 VV/VH composite (**green border**: model input) alongside the corresponding Sentinel-2 RGB (**grey border**: reference only — never seen by the model during training or inference). The S2 panels confirm that all three SAR encoders retrieve tiles whose visual content matches the natural-language query.

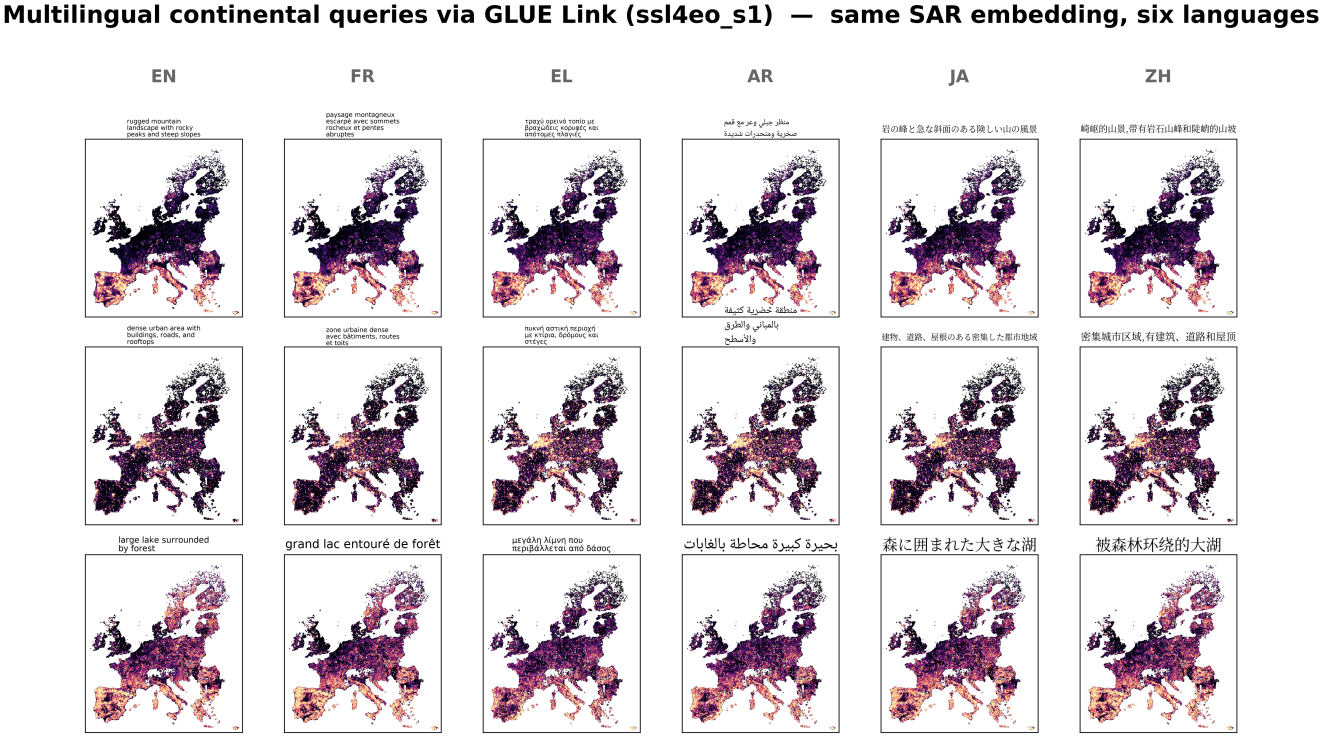

Figure 3. Multilingual continental queries via GLUE-Link (SSL4EO-S1). Rows: three queries. Columns: same query in six languages spanning four scripts (Latin, Greek, Arabic, Han, Kana). The SAR embedding is identical across columns; only the text input varies. Geographic patterns (Alps and Pyrenees for mountains; European agglomerations for urban; Scandinavian/Finnish lake districts for lake) are consistent across languages, demonstrating that GLUE-Link inherits SigLIP-2's multilingual text encoder for free.

suffices throughout.

4.3. Retrieval agreement with ground photos

Table 2 asks whether a satellite-only retrieval surfaces the same tiles a ground photographer would: rank all 22,441 test tiles by cosine similarity once using the linked encoder and once using the four-view-averaged ground-photo SigLIP-2, for 100 descriptors sampled with a fixed seed.

All five encoders are well above chance ($\sim$9–25$\times$ at P@100, $\sim$6–11$\times$ at P@500). AlphaEarth tops both columns (P@100$=$ 0.112, ρ=0.57). Pure-SAR encoders trail by $\sim$2$\times$ on P@K but remain decisively above random. Within the pure-SAR group BetaEarth-S1 wins on P@K at $1/16$ the embedding width, consistent with Table 1; on full-ranking ρ, however, DINOv2-S1 is slightly ahead (0.403 vs

0.381), so BetaEarth-S1's edge is concentrated in the head of the ranking.

Per-tile caption ρ exceeds 0.89 for every encoder (Table 1), but per-query retrieval ρ is much lower: GLUE-Link learns ranking-*consistent* but not rank-*identical* representations. P@K remains many multiples above chance throughout.

4.4. Qualitative analyses

Continental-scale plausibility (Fig. 1). Across the $\sim$200K LUCAS points the rugged-mountain query highlights the Alps, Pyrenees, Carpathians and Scandinavian highlands as continuous bright arcs along known orogenic belts; the dense-urban query picks out major European agglomerations (London, Paris, Madrid, the Ruhr, Po Valley conurbations); the lake-and-forest query concentrates on Scandinavian and Finnish lake districts, although it is the least reliable of the three and produces spurious responses in some non-lake regions. These continental patterns align with well-known geographic references — orogenic belts, major agglomerations, northern lake districts — without any geographic supervision, and are quantitatively validated against the gold ground-photo retrieval in Tables 1–2. Apparent gaps over parts of the Alps, Scandinavia and Scotland reflect LUCAS survey coverage (EU member states, accessible terrain) rather than missing predictions, suggesting that the linked SAR encoders capture genuine geographic semantics rather than spurious tile properties.

Top retrievals (Fig. 2). Inspecting the highest-scoring tiles per query confirms that each encoder retrieves the right scenes for the right reasons. Rugged mountain landscapes give strong layover and shadow patterns in SAR; all three pure-SAR encoders retrieve Alpine and Pyrenean tiles whose S2 reference confirms steep, rocky topography. Dense urban areas produce bright double-bounce returns from building corners, so retrievals concentrate on European city centres. Large lakes surrounded by forest yield a high-contrast SAR signature — smooth water returns near-zero backscatter while the surrounding forest is diffusely bright — and all three encoders surface Scandinavian and other northern-European lake-and-forest tiles.

Multilingual transfer (Fig. 3). Because every linked encoder targets the SigLIP-2 image-text space, any of the 109 languages covered by SigLIP-2 can be used as a query at inference with no additional training. SSL4EO-S1 answering the three queries in six languages spanning four scripts (Latin, Greek, Arabic, Han, Kana) gives essentially the same continental response in every column; multilingual coverage is inherited from the frozen text encoder rather than re-acquired.

5. Conclusion

Geographic co-location of satellite tiles with ground photographs, paired with a single 1×1 linear projection, is enough to make any frozen satellite encoder text-queryable through a frozen general-purpose VLM. No SAR-caption pairs, no labels and no architectural changes are needed; the protocol generalises across modalities (pure SAR, SAR+optical time series, multi-sensor) and embedding sizes (64–1024 dimensions) and inherits multilingual coverage from the SigLIP-2 text encoder. For SAR specifically, this hands all-weather monitoring workflows a natural-language interface at no annotation cost.

Several caveats temper the result. LUCAS is European, so tropical, arid and desert biomes remain unverified. The output resolution follows each encoder's feature grid — around 160 m for the ViT-L caches and our pooled BetaEarth-S1 cache, around 45 m for AlphaEarth and Tessera — even though BetaEarth-S1 natively predicts at the 10 m pixel grid. The cached extraction windows are also not co-extensive across encoders (2240 m for the ViT-style and BetaEarth-S1 caches against 500 m for AlphaEarth and Tessera), a historical artifact that wedge pooling absorbs at training time but that a journal-length follow-up should remove by re-extracting at a common ground extent. Finally, per-point cosine agreement is high (ρ>0.89) yet retrieval P@100 stays modest ($\leq$ 0.11): the protocol learns a ranking-consistent rather than rank-identical representation.

The most natural next steps follow from these caveats: temporal-aware ground supervision to capture seasonal variation, extension to other under-served modalities such as LiDAR or hyperspectral, and training against non-LUCAS ground-photo corpora — GBIF, iNaturalist, or Mapillary — to escape the European footprint.

References

[1] Eurostat. LUCAS land use and coverage area frame survey, 2018.

[2] Y. Wang, C. M. Albrecht, et al. SSL4EO-S12: a large-scale multimodal multitemporal dataset for self-supervised learning in EO. *IEEE GRSM*, 2023.

[3] Z. Feng, C. Atzberger, et al. TESSERA: temporal embeddings of surface spectra for Earth representation and analysis. *CVPR*, 2026. *arXiv:2506.20380*.

[4] C. F. Brown, M. R. Kazmierski, et al. AlphaEarth Foundations: an embedding field model for accurate and efficient global mapping from sparse label data. *arXiv:2507.22291*, 2025.

[5] M. Czerkawski. BetaEarth: emulating closed-source Earth observation foundation models through their public embeddings. *ISPRS Congress*, 2026.

[6] M. Oquab, T. Darcet, et al. DINOv2: learning robust visual features without supervision. *TMLR*, 2024.

[7] M. Tschannen, A. Gritsenko, et al. SigLIP 2: multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025.

[8] C. T. Marimo, B. Blumenstiel, et al. Beyond the visible: multispectral vision-language learning for EO. *arXiv:2503.15969*, 2025.

[9] Z. Yuan, Z. Xiong, L. Mou, X. X. Zhu. ChatEarthNet: a global-scale, high-quality image-text dataset empowering vision-language geo-foundation models. *Earth Syst. Sci. Data*, 17:1245, 2025.

[10] Z. Zhang, T. Zhao, Y. Guo, J. Yin. RS5M and GeoRSCLIP: a large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE TGRS*, 2024.

[11] Z. Ma, X. Xiao, S. Dong, P. Wang, H. Wang, Q. Pan. SARChat-Bench-2M: a multi-task vision-language benchmark for SAR image interpretation. *arXiv:2502.08168*, 2025.

[12] Y. Wei, A. Xiao, Y. Ren, Y. Zhu, H. Chen, J. Xia, N. Yokoya. SARLANG-1M: a benchmark for vision-language modeling in SAR image understanding. *arXiv:2504.03254*, 2025.

[13] P. Wang, Z. Lu, Y. Li, B. Ding, D. Zhang. SARCLIP: the first vision-language foundation model for SAR image. *IEEE Trans. Geosci. Remote Sens.*, 63:5223211, 2025.

[14] Q. Ma, Z. Wang, W. Liu, X. Lu, B. Deng, P. Duan, X. Kang, S. Li. SARVLM: a vision language foundation model for semantic understanding and target recognition in SAR imagery. *arXiv:2510.22665*, 2025.

[15] Y. He, X. Cheng, J. Zhu, et al. SAR-TEXT: a large-scale SAR image-text dataset built with SAR-Narrator and a progressive learning strategy for downstream tasks. *arXiv:2507.18743*, 2025.

[16] J. Zhang, L. Zhang, B. Zou. FSAR-Cap: a fine-grained two-stage annotated dataset for SAR image captioning. *arXiv:2510.16394*, 2025.

[17] U. Mall, C. P. Phoo, M. K. Liu, C. Vondrick, B. Hariharan, K. Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment (GRAFT). *ICLR*, 2024.

[18] A. Dhakal, A. Ahmad, S. Khanal, S. Sastry, H. Kerner, N. Jacobs. Sat2Cap: mapping fine-grained textual descriptions from satellite images. *CVPR EarthVision*, 2024.

[19] P. Jain, D. Ienco, R. Interdonato, T. Berchoux, D. Marcos. SenCLIP: enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting. *WACV*, 2025. arXiv:2412.08536.

[20] P. Jain, D. Marcos, D. Ienco, R. Interdonato, T. Berchoux. TimeSenCLIP: a time series vision-language model for remote sensing. *ISPRS J. Photogr. Remote Sens.*, 236:99–119, 2026. arXiv:2508.11919.

[21] A. Radford, J. W. Kim, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[22] Anonymous. GLUE: multi-lingual querying of Earth from space. Manuscript in preparation, 2026.