EARTHEMBEDDINGEXPLORER: A WEB APPLICATION FOR CROSS-MODAL RETRIEVAL OF GLOBAL SATELLITE IMAGES

Yijie Zheng[1,2*] **Weijie Wu**[1,2] **Bingyue Wu**[2,3] **Long Zhao**[1] **Guoqing Li**[1]
Mikolaj Czerkawski[4] **Konstantin Klemmer**[5,6]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
[4]Asterisk Labs
[5]LGND AI, Inc.
[6]University College London

ABSTRACT

While the Earth observation community has witnessed a surge in high-impact foundation models and global Earth embedding datasets, a significant barrier remains in translating these academic assets into freely accessible tools. This tutorial introduces EarthEmbeddingExplorer, an interactive web application designed to bridge this gap, transforming static research artifacts into dynamic, practical workflows for discovery. We will provide a comprehensive hands-on guide to the system, detailing its cloud-native software architecture, demonstrating cross-modal queries (natural language, visual, and geolocation), and showcasing how to derive scientific insights from retrieval results. By democratizing access to precomputed Earth embeddings, this tutorial empowers researchers to seamlessly transition from state-of-the-art models and data archives to real-world application and analysis. The web application is available at `https://modelscope.ai/studios/Major-TOM/EarthEmbeddingExplorer`.

1 INTRODUCTION

Recent foundation models enable reusable representations for search, clustering, and downstream tasks, especially when paired with large embedding datasets such as Major TOM embeddings (Czerkawski et al., 2024). Representative models span different supervision signals and modalities, including language–image alignment (FarSLIP (Li et al., 2025), SigLIP (Zhai et al., 2023)), self-supervised visual features (DINOv2 (Oquab et al., 2024)), and location–image alignment (SatCLIP (Klemmer et al., 2025a)).

Despite this progress, turning "published embeddings" into a practical workflow is still difficult: users often need to download large archives, run embedding pipelines, implement vector search, and build visualization tooling. This gap limits hands-on use beyond expert teams, and motivates standardized, accessible access to Earth embeddings (Klemmer et al., 2025b; Fang et al., 2026).

This tutorial introduces **EarthEmbeddingExplorer**, an interactive web app that operationalizes *precomputed* satellite image embeddings for cross-modal retrieval and qualitative analysis. It supports text-, image-, and location-based queries, global similarity-map visualization, and inspection/export of the top retrieved tiles. In this tutorial, we provide (i) ready-to-use embeddings for four representative models, (ii) a cloud-native deployment on open platforms, and (iii) a step-by-step walkthrough grounded in real-world case studies.

*zhengyijie23@mails.ucas.ac.cn

Model	Arch.	Train	Input bands	Input size	Dim.	Dtype
DINOv2	ViT-L/14	LVD-142M	RGB	224×224	1024	float32
FarSLIP	ViT-B/16	RS5M, MGRS	RGB	224×224	512	float16
SatCLIP	ViT16-L40	S2-100K	Multi-spectral	224×224	256	float16
SigLIP	ViT-SO400M	WebLI	RGB	384×384	1152	float16

Table 1: Embedding models used in EarthEmbeddingExplorer. We report architecture, training datasets, input resolution, embedding dimensionality, and embedding dtype for reproducible comparison.

2 EARTHEMBEDDINGEXPLORER

2.1 EMBEDDING MODELS

EarthEmbeddingExplorer currently includes four complementary embedding models (Table 1) to support different query modes and comparison needs. FarSLIP (Li et al., 2025) and SigLIP (Zhai et al., 2023) enable *text-to-image* retrieval; DINOv2 (Oquab et al., 2024) provides strong image features for *image-to-image* retrieval; and SatCLIP (Klemmer et al., 2025a) enables *location-to-image* retrieval. All models also support image queries, enabling users to contrast semantic alignment (text-supervised) against visual similarity (self-supervised) in a unified interface.

2.2 EMBEDDING DATASETS

We utilize MajorTOM-Core-S2L2A (Francis & Czerkawski, 2024) as the imagery source. The dataset is indexed via a systematic grid of approximately 10×10 km cells, ensuring comprehensive spatial coverage. To maintain global diversity while keeping the tutorial lightweight, we uniformly subsample $1/9$ of the Major TOM grid and crop a central 384×384 pixel patch from each cell. This process yields 248,719 unique patches, representing approximately 1.4% of Earth's land surface (Figure 5). Following the Major TOM Embedding Expansions standard (Czerkawski et al., 2024), we release these as precomputed embeddings stored in GeoParquet shards. This cloud-native format enables high-speed lookups and efficient partial downloads, which are essential for real-time interactive visualization in our web application.

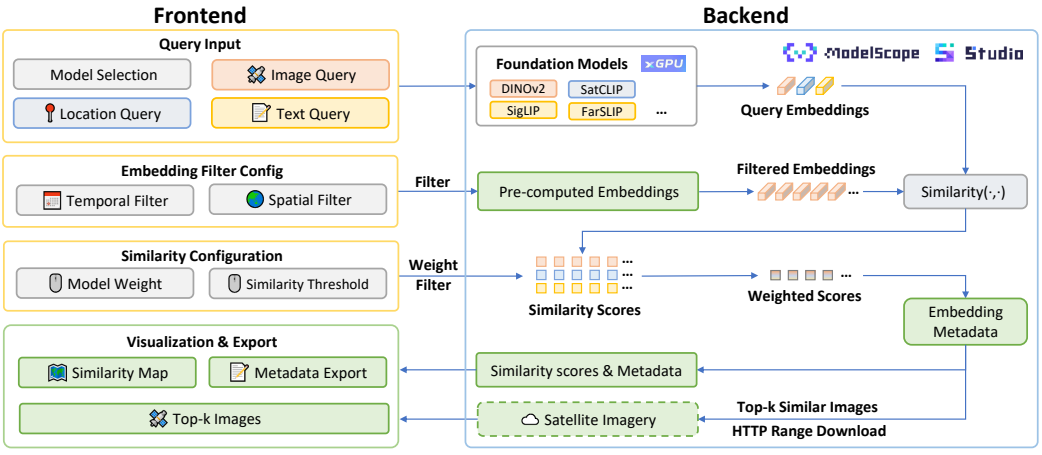


Figure 1: A cloud-native retrieval pipeline based on ModelScope Studio.

2.3 SYSTEM ARCHITECTURE

Figure 1 summarizes the cloud-native design. Queries are embedded with selected models, matched via vector similarity search over precomputed embeddings, and visualized as a similarity map and top-k retrieved images. We offer ModelScope deployments with free GPU runtime, allowing users to run the tutorial without local setup. The frontend is built with Gradio (Abid et al., 2019). As

shown in Figure 2, the left panel configures inputs, while the right panel visualizes similarity maps and retrieved examples.

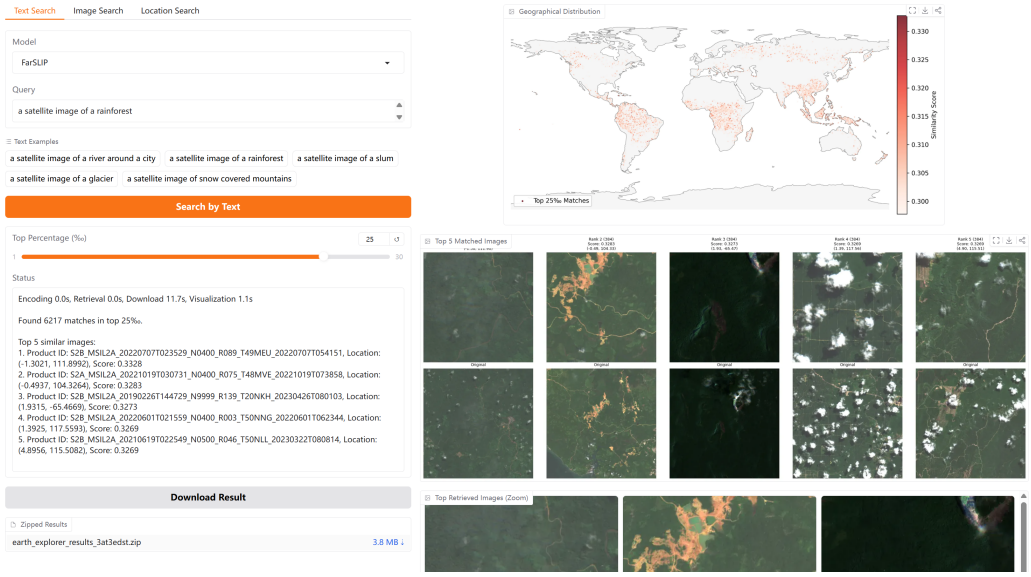
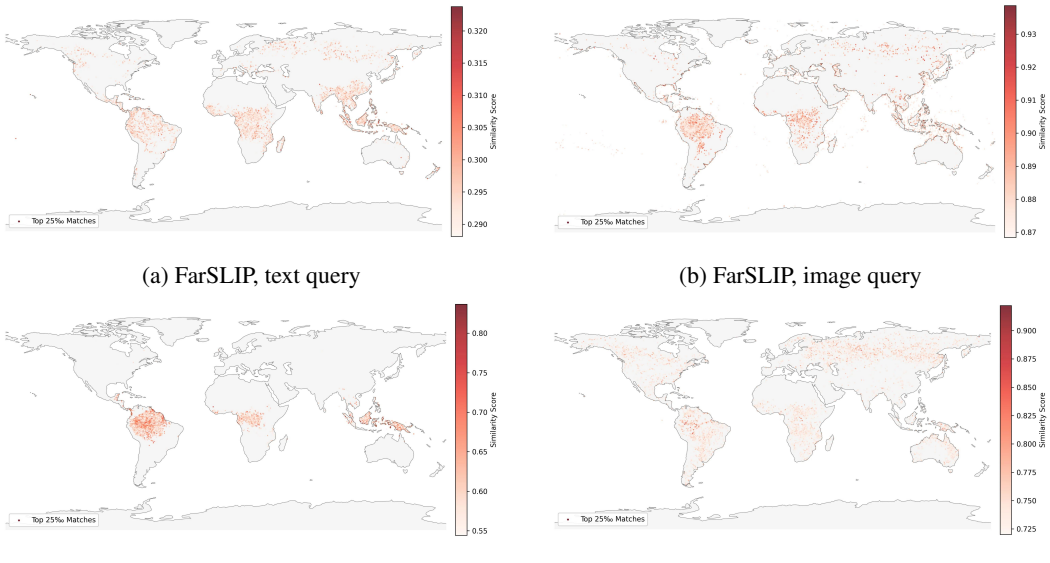


Figure 2: EarthEmbeddingExplorer user interface.



(a) FarSLIP, text query

(b) FarSLIP, image query

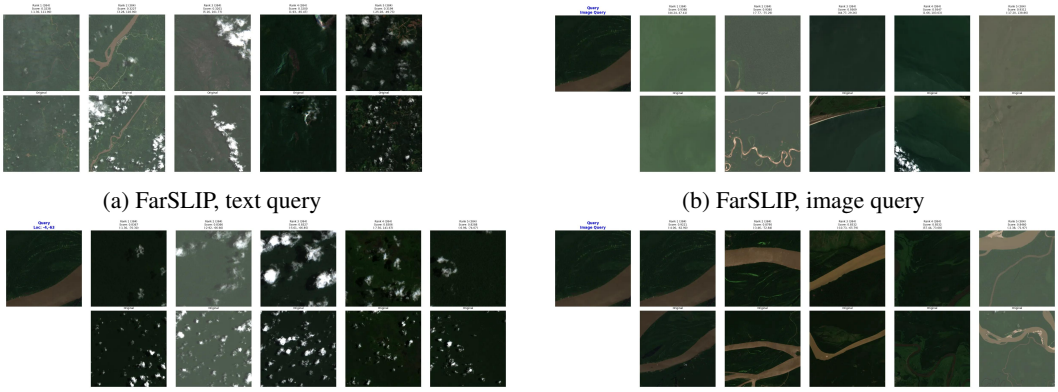
(c) SatCLIP, text query

(d) DINOv2, image query

Figure 3: Geographic distribution of retrieved matches under a top-2.5% threshold for different models and query modalities.

3 Tutorial Walkthrough & Case Study

In practice, users can synthesize results by comparing similarity "hotspots" and top matches across different models or prompts. The following case study demonstrates how query modalities shape retrieval patterns. Additional cross-model comparisons are detailed in the Appendix.



(a) FarSLIP, text query

(b) FarSLIP, image query

(c) SatCLIP, text query

(d) DINOv2, image query

Figure 4: Top-5 retrieved tiles for the same case study in Figure 3.

We demonstrate the workflow with a rainforest retrieval case study. For *text-to-image* search, we use the prompt `a satellite image of a tropical rainforest`. For *image-to-image* search, the query is an image patch centered at (4°S, 63°W) near Rio Purus (an upstream tributary of the Amazon). For *location-to-image* search, we use the same coordinates as the location query.

Figure 3 compares the geographic distribution of high-scoring matches across models and query modalities. With a text query, FarSLIP concentrates matches in humid tropical regions, reflecting semantic alignment with the concept *rainforest*. In contrast, SatCLIP produces a stronger location-consistent prior: high-scoring matches are largely restricted to the tropical belt, including major rainforest regions in the Amazon Basin, the Congo Basin, and Southeast Asia. For image queries, the highest-scoring matches are relatively more geographically dispersed, as similar visual patterns (e.g., rivers, dark vegetation) can occur in multiple climates and continents.

For more detailed inspection, Figure 4 shows the top-5 retrieved patches for each setting. The text-based retrievals are generally semantically consistent with rainforest scenes and often include cloud cover, which is common in these regions. In contrast, DINOv2 (self-supervised) tends to emphasize fine-grained visual cues: four of its top-5 results contain wide rivers, suggesting that river morphology dominates the embedding similarity. FarSLIP image retrieval is closer to its text-based behavior—returning rainforest-like patches that are less dominated by the river pattern—highlighting the difference between semantic alignment and purely visual similarity.

These visualizations also reveal limitations of current foundation models. Even with a well-specified concept prompt (e.g., "tropical rainforest"), FarSLIP can occasionally retrieve patches outside the expected climate zone, suggesting limited geographic/climatic priors in the embedding space. For image-based retrieval, we also observe occasional implausible matches (e.g., ocean tiles).

4 Conclusions and Roadmap

EarthEmbeddingExplorer packages precomputed Earth embeddings into an interactive, reproducible workflow for cross-modal retrieval and rapid qualitative evaluation. It is intended both for model developers (to stress-test representations at global scale) and for geoscience users (to quickly find and export regions of interest from flexible text/image/location queries).

Next steps include: (i) expanding spatial and temporal coverage (more grid cells, timestamps, and sensors), (ii) accelerating retrieval with dedicated vector databases and quantization, and (iii) supporting community contributions of new embedding expansions and models under the Major TOM embedding standard for consistent comparison and reuse. By fostering a community-driven ecosystem, this platform will further bridge the gap from academic publication to practice, accelerating model development and geoscientific research.

ACKNOWLEDGMENT

This work was supported by the National Earth Observation Data Center Research Project. We thank ModelScope for providing free access to high-performance GPU resources for deploying web applications.

REFERENCES

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.

Mikolaj Czerkawski, Marcin Kluczek, and Jędrzej S. Bojanowski. Global and dense embeddings of earth: Major tom floating in the latent space. *arXiv preprint arXiv:2412.05600*, 2024.

Heng Fang, Adam J Stewart, Isaac Corley, Xiao Xiang Zhu, and Hossein Azizpour. Earth embeddings as products: Taxonomy, ecosystem, and standardized access. *arXiv preprint arXiv:2601.13134*, 2026.

Alistair Francis and Mikolaj Czerkawski. Major tom: Expandable datasets for earth observation. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2935–2940, 2024. doi: 10.1109/IGARSS53475.2024.10640760.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025a.

Konstantin Klemmer, Esther Rolf, Marc Rußwurm, Gustau Camps-Valls, Mikolaj Czerkawski, Stefano Ermon, Alistair Francis, Nathan Jacobs, Hannah Rae Kerner, Lester Mackey, Gengchen Mai, Oisin Mac Aodha, Markus Reichstein, Caleb Robinson, David Rolnick, Evan Shelhamer, Vincent Sitzmann, Devis Tuia, and Xiaoxiang Zhu. Earth embeddings: Towards ai-centric representations of our planet. *EarthArXiv*, December 2025b. doi: 10.31223/X5HX9S. URL `https://eartharxiv.org/repository/view/11083/`. Preprint.

Sumin Lee, Sungwon Park, Jeasurk Yang, Jihee Kim, and Meeyoung Cha. Generalizable slum detection from satellite imagery with mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Zhenshi Li, Weikang Yu, Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Pedram Ghamisi, and Xiao Xiang Zhu. Farslip: Discovering effective clip adaptation for fine-grained remote sensing understanding. *arXiv preprint arXiv:2511.14901*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Zhiyuan Xie, Umesh K. Haritashya, Vijayan K. Asari, Michael P. Bishop, Jeffrey S. Kargel, and Theus H. Aspiras. Glaciernet2: A hybrid multi-model learning architecture for alpine glacier mapping. *International Journal of Applied Earth Observation and Geoinformation*, 112:102921, 2022.

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.

A APPENDIX

This appendix provides supplementary details, including the geographical distribution of our sampled grids (Figure 5) and further case studies evaluating model behaviors.

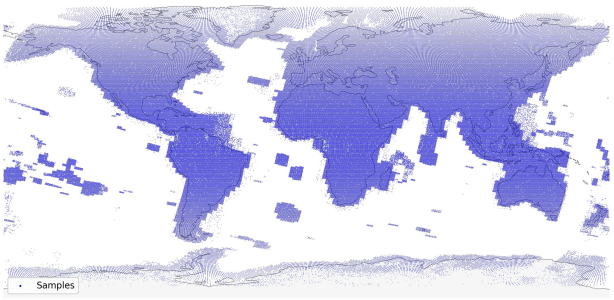
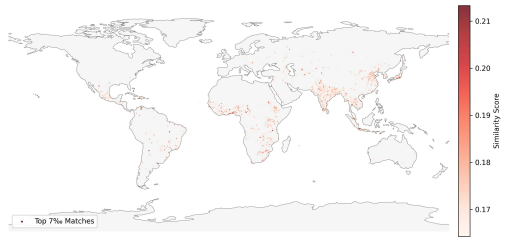


Figure 5: Geographical distribution of sampled grids

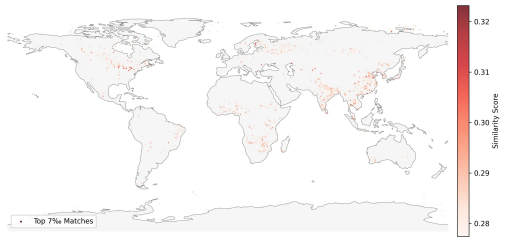
A.1 ADDITIONAL CROSS-MODEL COMPARISON

We further compare vision–language models by contrasting SigLIP and FarSLIP on text-to-image retrieval using prompts from two representative Earth observation applications (Lee et al., 2026; Xie et al., 2022): one socio-economic prompt (`a satellite image of a slum`) and two natural-scene prompts.

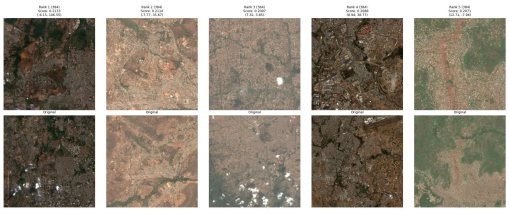
Socio-economic concepts Figure 6 compares similarity maps and top-5 matches for the `slum` prompt. SigLIP produces concentrated high-similarity regions in parts of South Asia, Latin America, and West Africa, suggesting that it captures visual cues that are often associated with informal settlements in satellite imagery. FarSLIP, despite being trained on remote-sensing image–text pairs, yields a more diffuse set of high-similarity responses, including substantial activation outside the regions highlighted by SigLIP. We attribute this model behavior to FarSLIP's pretraining data, which is drawn from several remote-sensing datasets with limited classes and therefore includes few images or labels related to the concept of "slum".



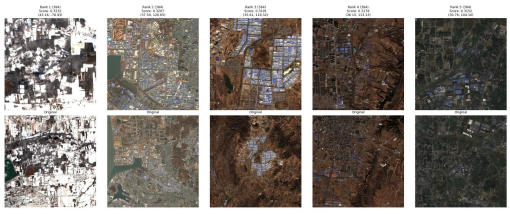
(a) SigLIP, similarity to `slum`



(b) FarSLIP, similarity to `slum`



(c) SigLIP, top-5 matches



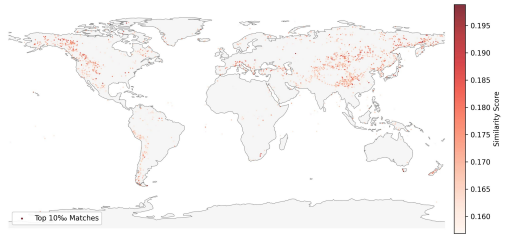
(d) FarSLIP, top-5 matches

Figure 6: Comparison of SigLIP and FarSLIP on socio-economic text-to-image retrieval for `slum`.

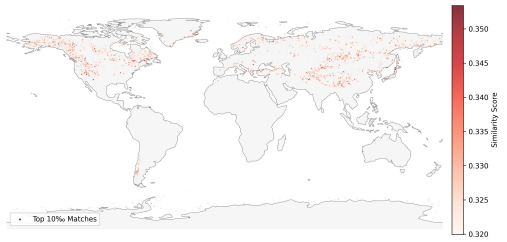
Natural features We next contrast two related cryosphere prompts, `a satellite image of snow covered mountains` and `a satellite image of a glacier`. Figure 7 shows the corresponding similarity maps, and Figure 8 provides the top-5 retrieved tiles.

For `snow covered mountains`, the two models exhibit different geographic concentrations. In this example, FarSLIP places high similarity along major high-elevation belts in Asia (e.g., the Himalayas, Kunlun, and Tianshan ranges), whereas SigLIP shows comparatively stronger responses over the Andes and New Zealand's Southern Alps, reflecting geographic biases in the models.

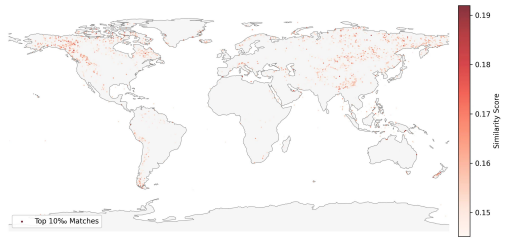
For `glacier`, the global retrieval distribution of the two models also varies substantially. FarSLIP assigns higher similarity to polar regions and the Antarctic margin, while SigLIP omits the Antarctic region; this may be due to a lack of polar data in SigLIP's pretraining corpus.



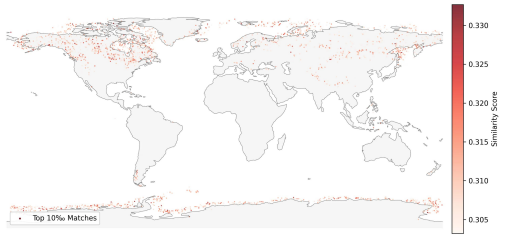
(a) SigLIP, similarity to `snow-covered mountains`



(b) FarSLIP, similarity to `snow-covered mountains`

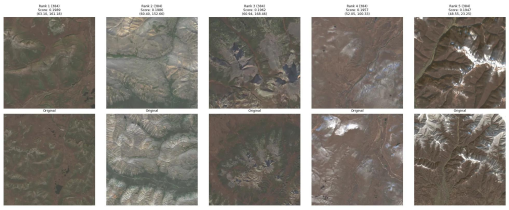


(c) SigLIP, similarity to `glacier`

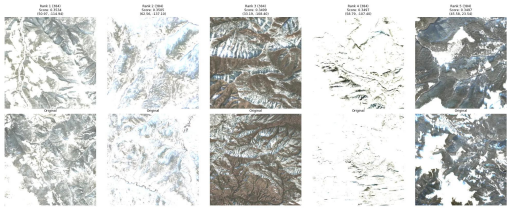


(d) FarSLIP, similarity to `glacier`

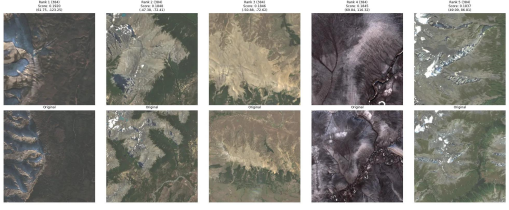
Figure 7: Comparison of SigLIP and FarSLIP on retrieving snow-covered mountains and glaciers.



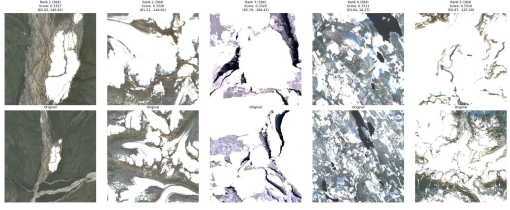
(a) SigLIP, top-5 matches to `snow-covered mountains`



(b) FarSLIP, top-5 matches to `snow-covered mountains`



(c) SigLIP, top-5 matches to `glacier`



(d) FarSLIP, top-5 matches to `glacier`

Figure 8: Top-5 retrieved tiles for the prompts in Figure 7.