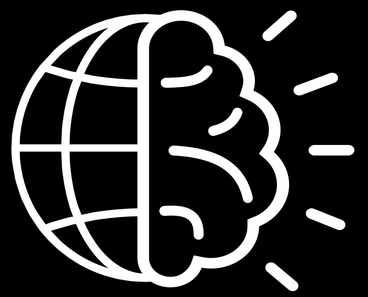
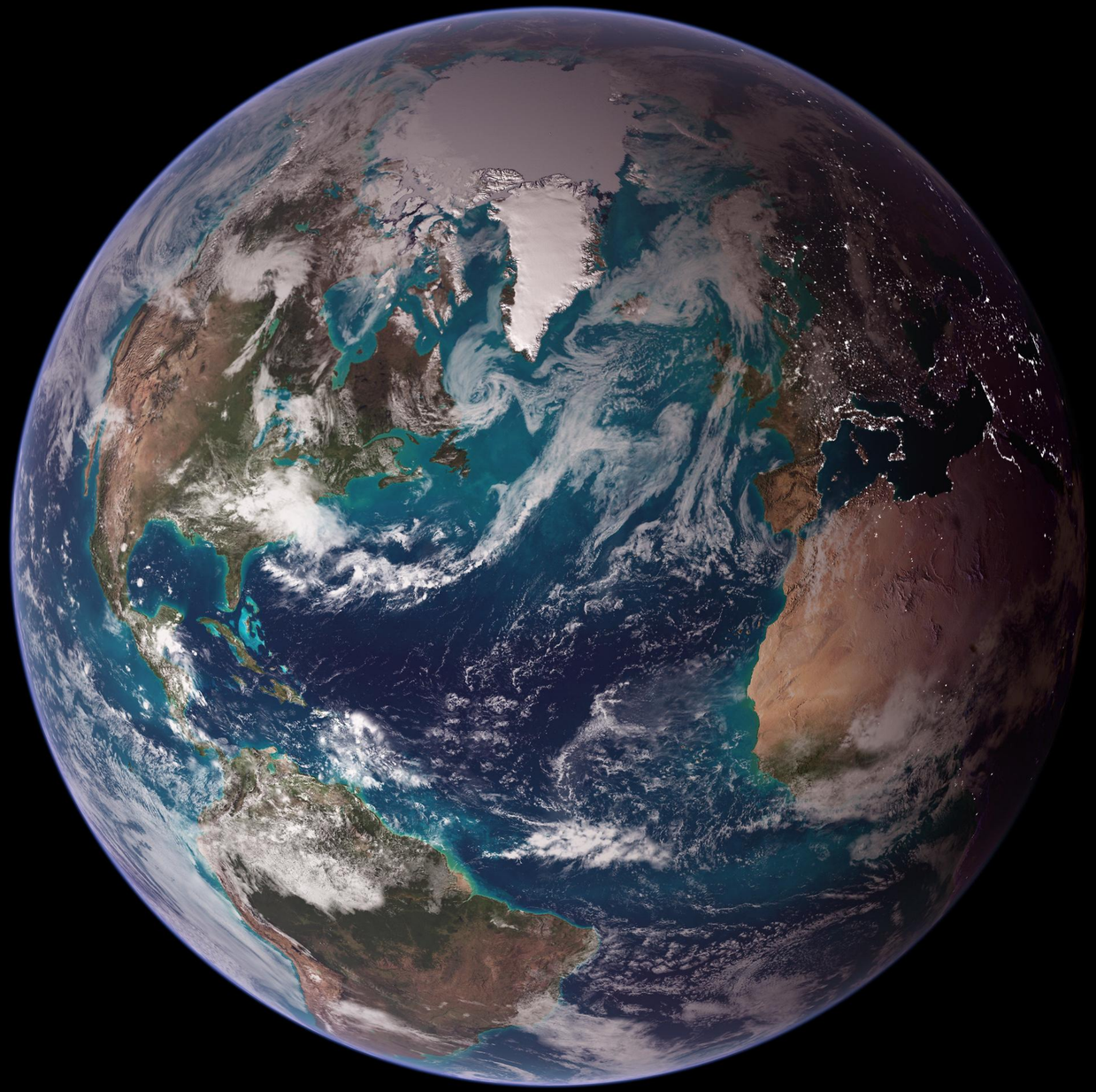
Large-Scale Generative Multimodality for Earth Observation



AI for Good Summit 2025
Gianfranco Basile

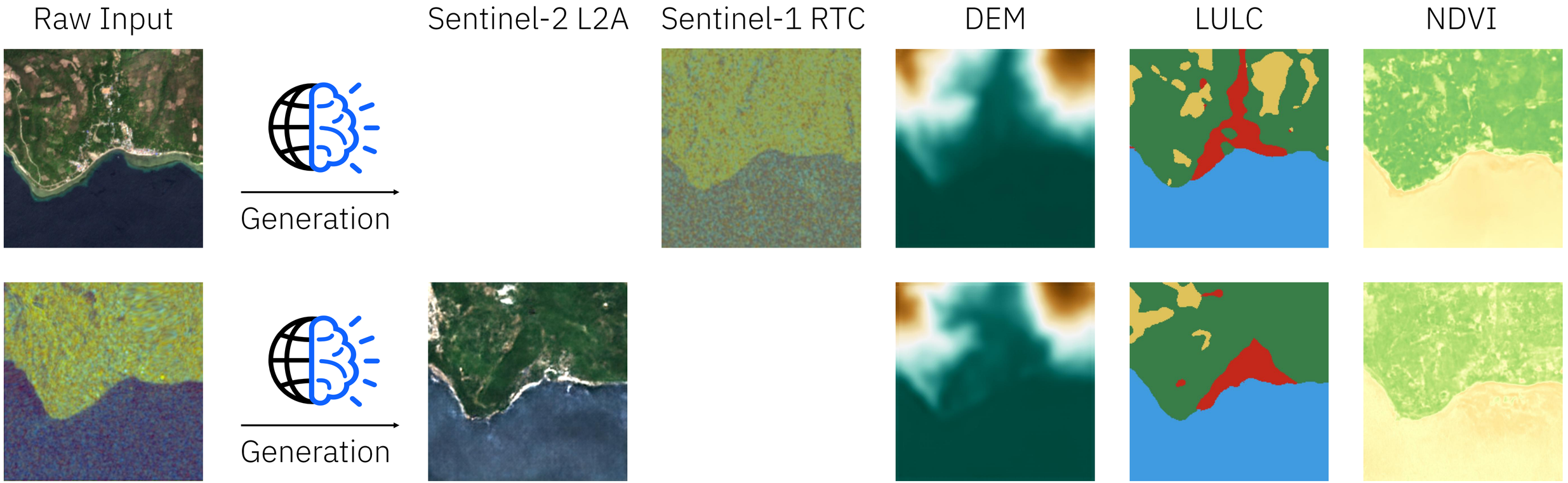
Credit: Johannes Jakubik

IBM **Research**



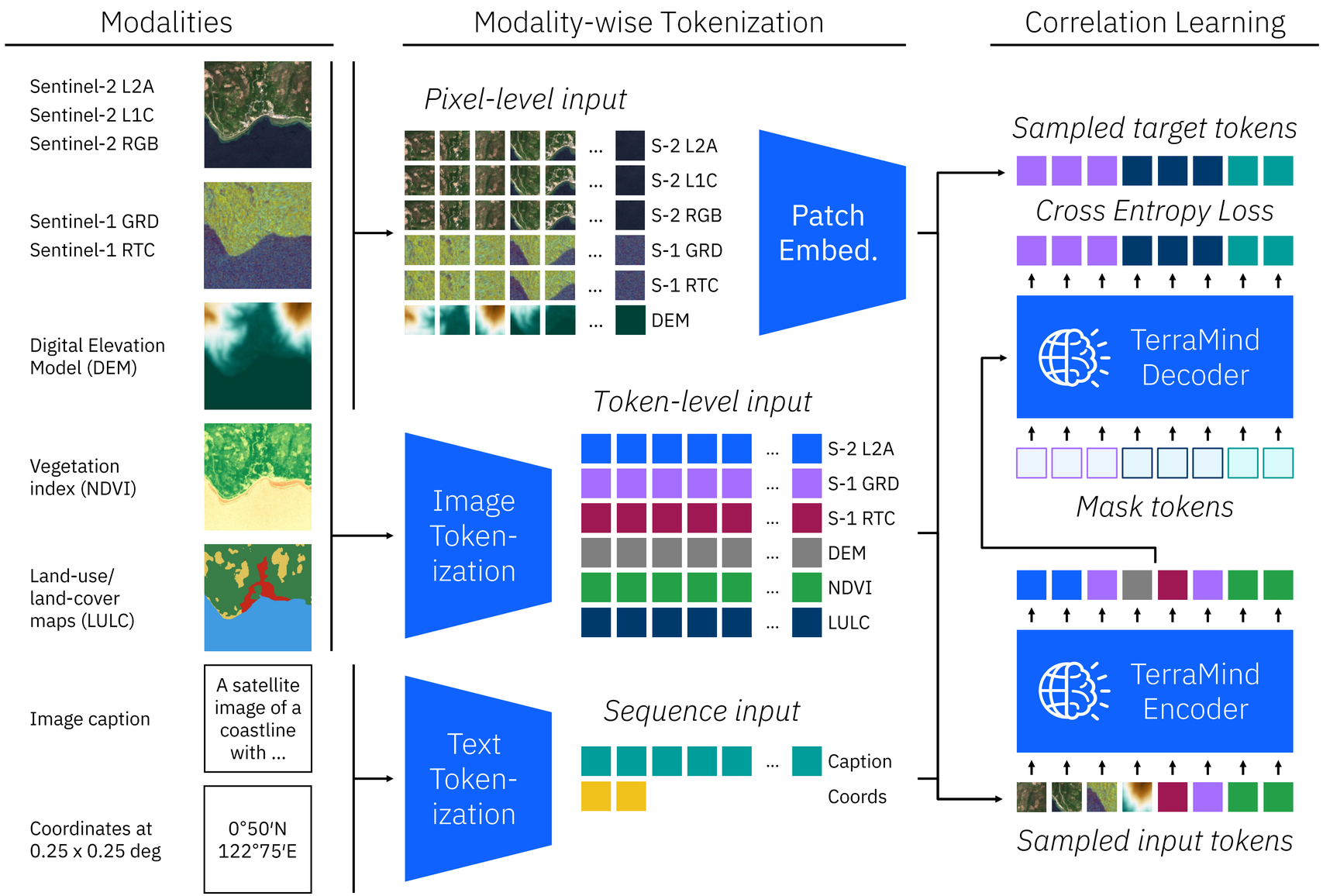
Credit: NASA/Goddard Space Flight Center

TerraMind – our first foundation model with cross-modal understanding



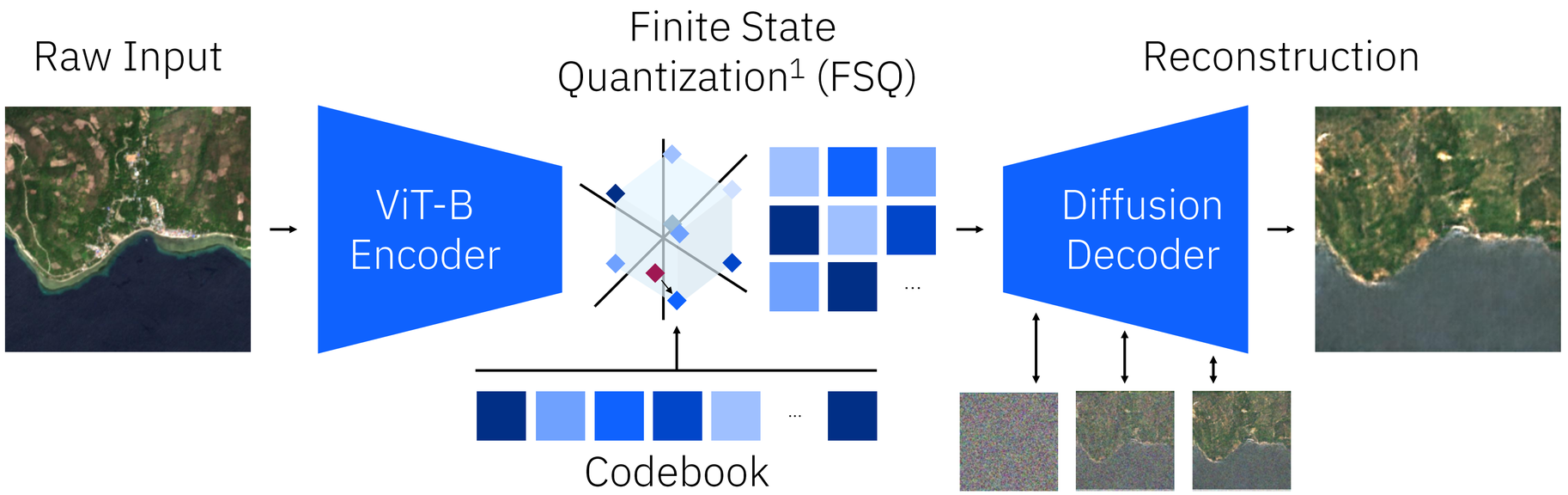
TerraMind

TerraMind is our first any-to-any generative, large-scale multimodal FM for EO and is pre-trained on 500 billion tokens using diverse geospatial data.



Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., ... & Longépé, N. (2025). TerraMind: Large-Scale Generative Multimodality for Earth Observation. *arXiv preprint arXiv:2504.11171*.

Tokens are good pre-training targets

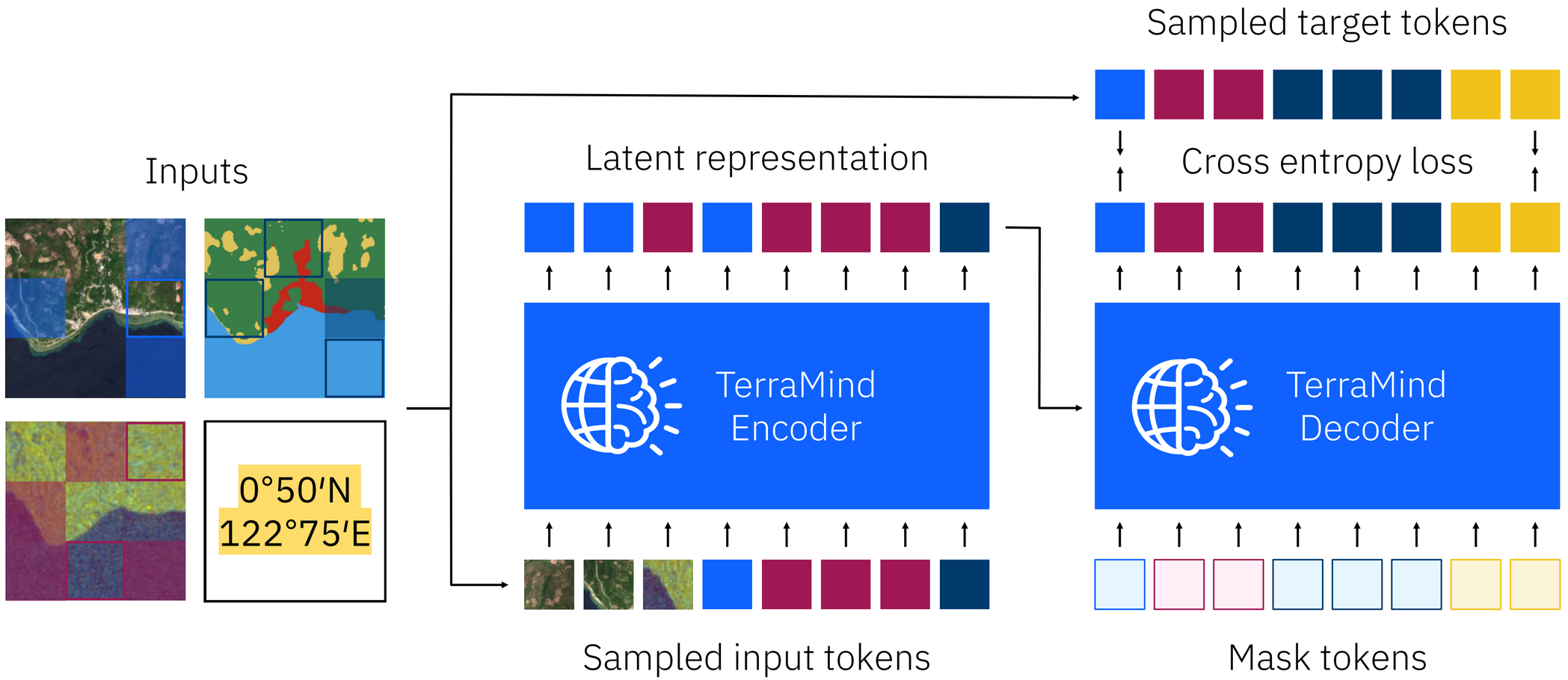


TerraMind uses Vector-quantized Variational Auto-Encoders (VQ-VAE) with diffusion steps for tokenization.

The tokenizer of each modality is trained separate and uses Finite Sate Quantization (FSQ) with a codebook size of 15,360 tokens.

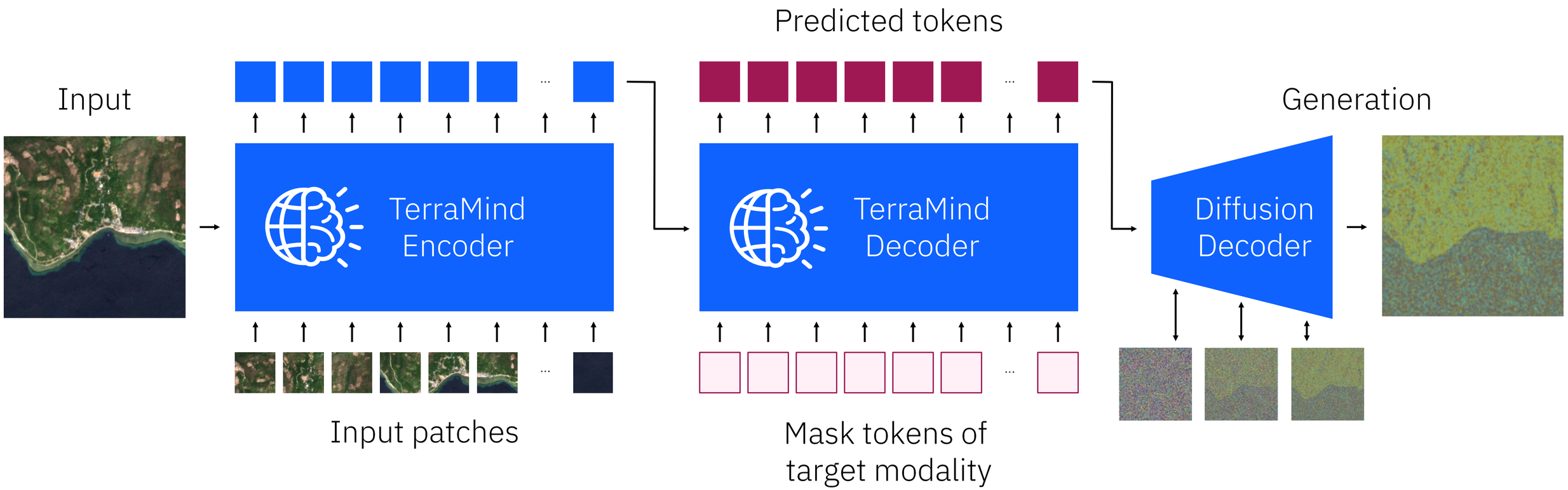
1 Mentzer, F., Minnen, D., Agustsson, E., & Tschannen, M. (2023). Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Fusing the modalities with correlation learning



Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., & Zamir, A. (2023). 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*.

Any-to-any generation

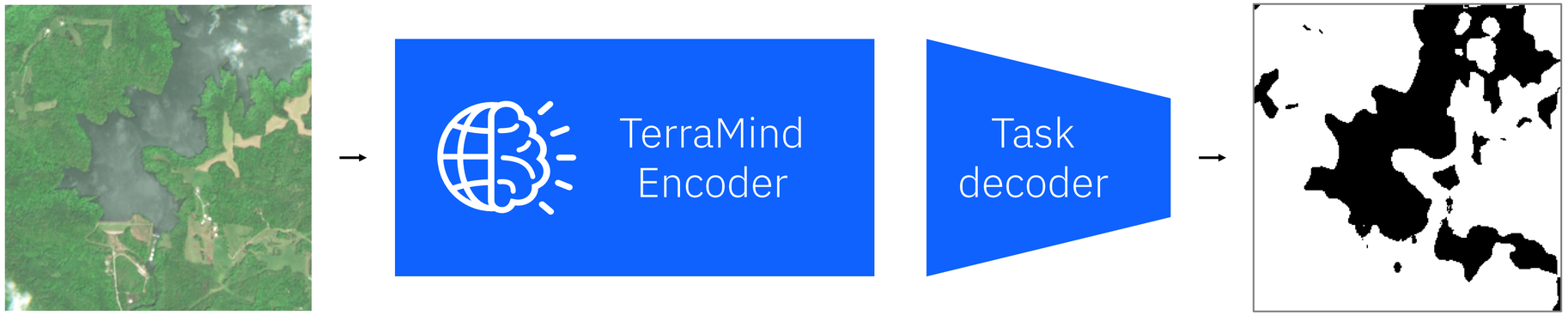


Thinking in Modalities

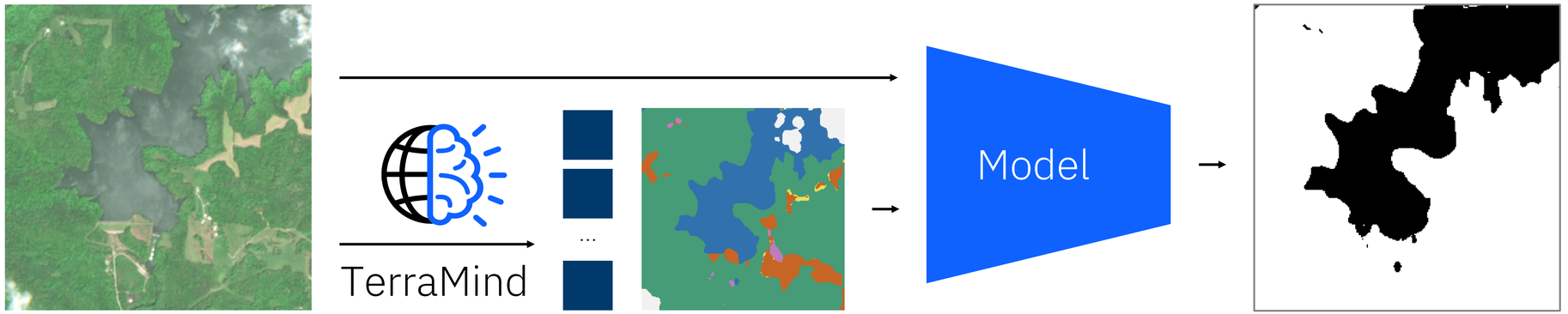
TerraMind enables to enhance fine-tuning by Thinking-in-Modalities (TiM) – generating intermediate artificial data of other modalities.

The raw image and the generated tokens are used as input by the fine-tuned model.

Standard fine-tuning



TiM fine-tuning with intermediate modalities



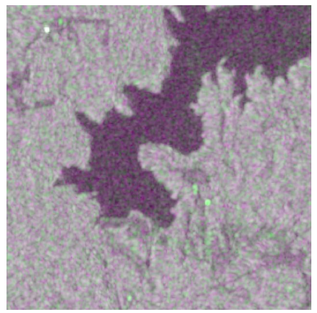
How does TerraMind perform?

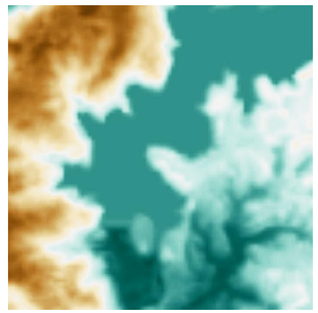
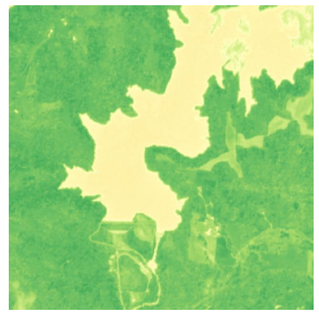
Thinking in Modalities

Comparision between the standard approach with full fine-tuning and Thinking-in-Modalities (TiM) tuning using generated LULC tokens as additional inputs.

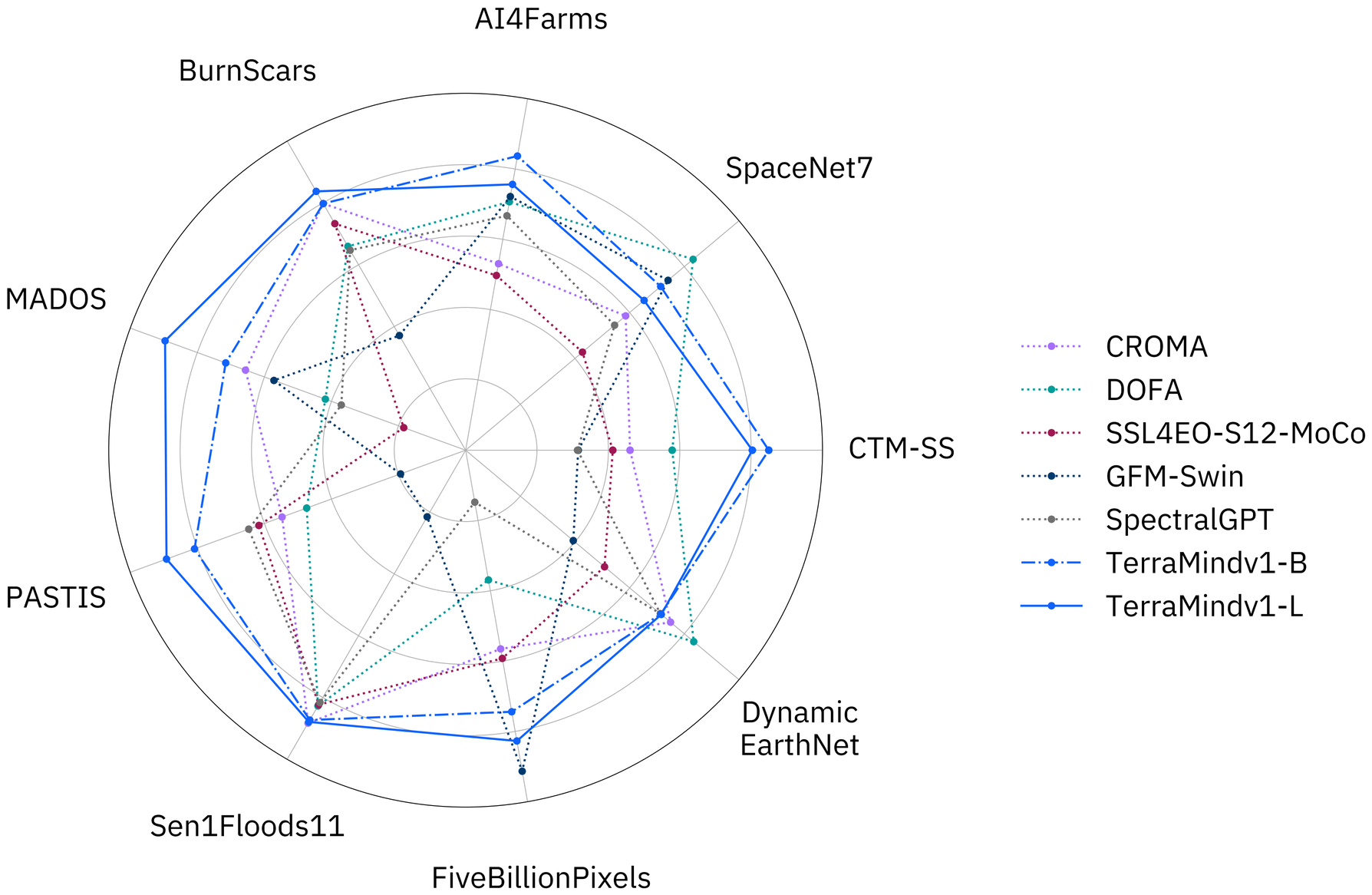
Dataset	Model	Input	IoU_{Water}	mIoU
Sen1Floods11	TerraMind-B	Sentinel-1	68.00	81.06
	TerraMind-B TiM	S-1 + *gen. LULC*	**72.25**	**83.65**
	TerraMind-B	Sentinel-2	82.26	89.70
	TerraMind-B TiM	S-2 + *gen. LULC*	**84.75**	**91.14**
SA Crop Type	TerraMind-B	Sentinel-2	–	41.87
	TerraMind-B TiM	S-2 + *gen. LULC*		**42.74**

Potential TiM modalities for TerraMind include S-2, S-1, LULC, DEM, NDVI, and coordinates.

PANGAEA bench results



PANGAEA bench results for TerraMind and the top 5 EO FMs based on average rank. The mIoU is visualized on a normalized scale.

A little outlook...

Ingesting weather data into TerraMind

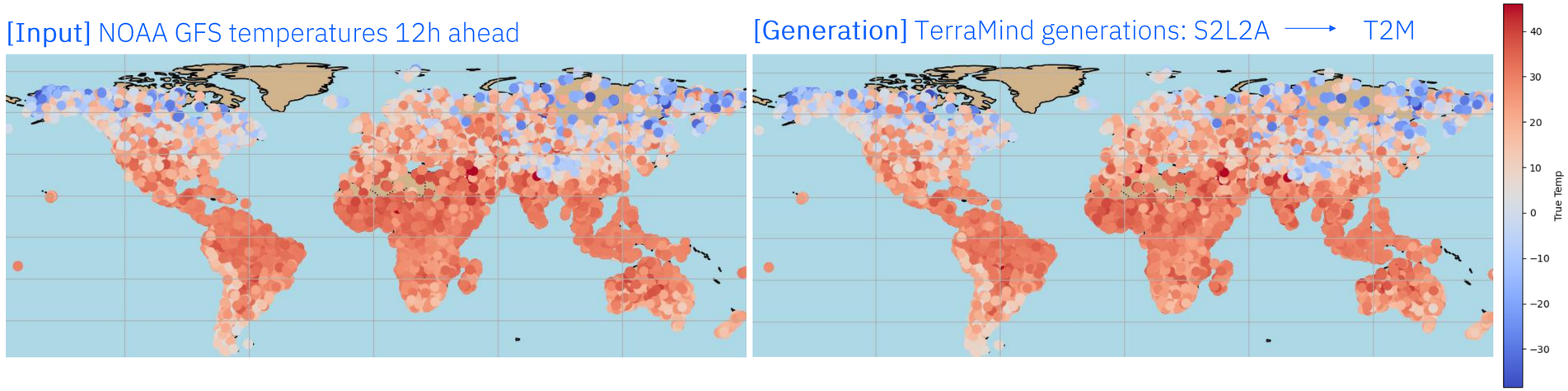
Conditioning generations of temperature profiles on Sentinel-2 images with TerraMind.

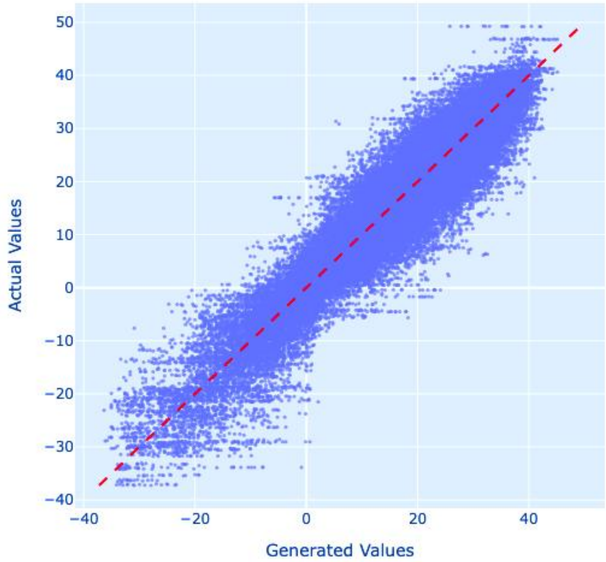
Leveraging timeseries on a daily and hourly basis for correlation learning.

FSQ tokenization of the time series.

[Input] NOAA GFS temperatures 12h ahead

[Generation] TerraMind generations: S2L2A ⟶ T2M

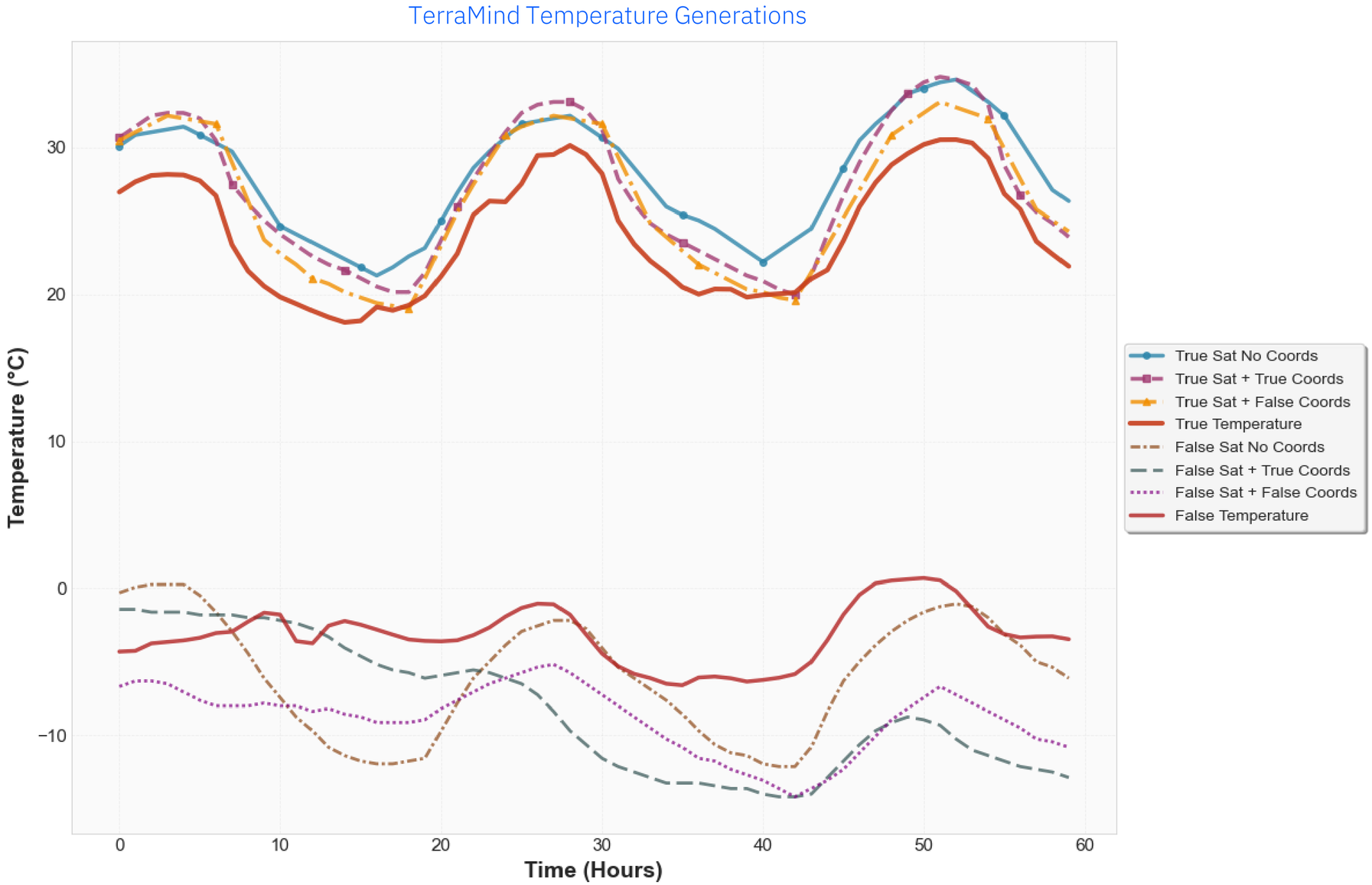




Causality experiments

TerraMind mostly relies on Sentinel-2 images when generating temperature profiles.

Randomized coordinates do not decrease generation accuracy significantly.



Thank you!

Find more information about TerraMind at
https://huggingface.co/ibm-esa-geospatial