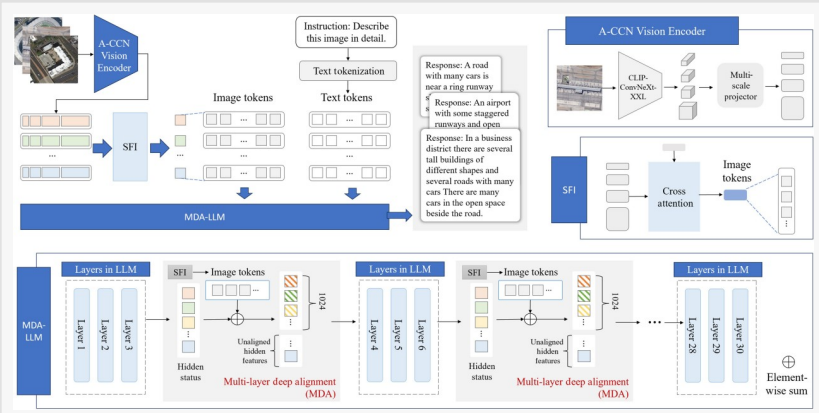
Aquila gives satellites a smarter voice

GA, UNITED STATES, May 6, 2026 /EINPresswire.com/ -- A new artificial intelligence model, Aquila, helps computers read remote sensing images with greater detail and stronger language reasoning. By combining high-resolution image processing with deeper image-text alignment, it can describe scenes more accurately and answer complex visual questions more reliably than earlier systems. The advance could make Earth-observation data more useful for environmental monitoring, urban analysis, agriculture, and disaster response, where subtle spatial details often determine whether an interpretation is merely plausible or truly actionable.



Architecture overview of the proposed Aquila.

Remote sensing is now central to tracking crops, cities, coastlines, ecosystems, and emergency events, yet many AI systems still struggle to understand overhead imagery the way experts do. Earlier remote sensing vision-language models often depend on shallow fusion, loosely connecting image features with language outputs. They also face a scale problem: roads, buildings, harbors, and fields can look very different depending on resolution and ground sampling distance. Even when high-resolution data are available, many models cannot fully preserve fine-grained spatial structure during reasoning. Based on these challenges, deeper research into high-resolution, multi-scale remote sensing vision-language modeling is needed.

Researchers from Wuhan Kotei Informatics Co. Ltd., the Chinese Academy of Surveying and Mapping, Emory University, the University of Science and Technology Beijing, and the China Aero Geophysical Survey and Remote Sensing Center for Natural Resources reported the study in the Journal of Remote Sensing on March 3, 2026. Their model, Aquila, was designed to tackle a persistent bottleneck in Earth-observation AI: how to connect rich visual detail with language-based reasoning without losing the spatial clues that make remote sensing imagery meaningful.

Aquila improves remote sensing image comprehension through two linked innovations. First, it accepts image inputs up to 1,024 × 1,024 pixels, far higher than the 448 × 448 scale supported by many earlier systems. Second, it combines multi-scale image features and repeatedly re-injects

them into the language model, rather than aligning vision and text only once. This strategy produced clear gains: on the challenging FIT_RSFG-Captions benchmark, Aquila outperformed SkySenseGPT by 7.77%, and on FIT_RSFG-VQA it reached 83.87% accuracy, beating SkySenseGPT by 4.11%.

The model is built from three core parts: an Aquila-CLIP ConvNeXt vision encoder, a hierarchical spatial feature integration module, and a multi-layer deep alignment language model based on Llama-3. Instead of relying on a single visual summary, Aquila extracts features from four scales and fuses them with a spatially aware cross-attention design that preserves local structure. This matters in remote sensing, where small objects and spatial layouts often carry the key meaning. In ablation tests, the spatial feature integration module improved captioning by 5.62% and VQA by 6.85% over a concatenation baseline. Adding deep alignment further raised performance by 2.55% in captioning and 4.64% in VQA. Aquila also showed broader grounding ability, reaching an mIoU of 68.33 on the DIOR-RSVG test set.

In the paper, the authors argue that Aquila's gains come from modeling remote sensing imagery the way the domain demands: with high resolution, multi-scale perception, and persistent image-language interaction throughout reasoning. Their results suggest that fine-grained Earth-observation understanding depends not just on bigger models, but on architectures that preserve spatial evidence instead of compressing it away too early.

The team trained Aquila in two stages. First, they aligned image and language features using about 1 million remote sensing image-text pairs while freezing both the vision encoder and language model. Second, they instruction-tuned the system on 1.8 million high-quality pairs using LoRA. Training ran on four NVIDIA A800 GPUs, with images resized to 1,024 × 1,024, no cropping or padding, and benchmarks covering captioning, visual question answering, and grounding tasks.

Aquila points toward a future in which analysts can interact with satellite and aerial imagery through natural language while still retaining expert-level spatial precision. The authors note that the system remains computationally intensive and currently focuses on single-temporal RGB imagery. But its design offers a foundation for broader geo-foundation models that could integrate multi-temporal, multispectral, or SAR data, expanding applications in urban growth tracking, disaster assessment, environmental surveillance, and intelligent geospatial decision-making.

References

DOI
10.34133/remotesensing.1041

Original Source URL
https://doi.org/10.34133/remotesensing.1041

Funding information
This research was supported by the Faculty Startup Fund of Emory College of Arts & Sciences, the Fundamental Research Funds for the Central Universities (FRF-TP-25-008), the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20250171), the National Natural Science Foundation of China (42201440 and 42401500), and the Fundamental Research Funds for Chinese Academy of Surveying and Mapping (AR2410).

Lucy Wang
BioDesign Research
email us here