Optimized Bi-Directional Satellite Image Interpretation System using VLM+LLM for Industrial Monitoring in Indonesia

Prabu Kresna Putra
Research Center for Data and Information Science
National Research and Innovation Agency, Banding, Indonesia.
prab003@bppt.go.id
ORCID: 0000-0001-9246-8863

Abstract

This research proposal outlines the development of an innovative system for interpreting satellite and drone imagery using advanced Vision-Language Models (VLMs) and Large Language Models (LLMs), specifically tailored for the Indonesian language context. The system is designed to address the inherent challenges in interpreting massive and complex remote sensing data, where traditional methods often prove inadequate. The proposed solution allows users to upload images, select an Area of Interest (AOI), input text-based queries, process and interpret data automatically using multimodal AI models, perform intelligent advanced analysis, and generate information in an easily understandable narrative or descriptive text format. The primary contribution of this research lies in enhancing remote sensing data analysis through sophisticated multimodal integration, emphasizing Indonesian language processing, and the ability to generate human-comprehensible insights. The significance of this system is substantial, offering the potential to significantly improve decision-making processes in critical applications such as disaster management and environmental monitoring in Indonesia.

1. Introduction

Remote sensing technology plays a vital role in global monitoring of the Earth's surface, with diverse applications in Indonesia. This technology is crucial for disaster management, environmental monitoring, urban planning, agriculture, and marine affairs (Martino et al., 2009; Chuvieco, 2020; Sadewa & Supriyadi, 2024). Given Indonesia's vast geography and vulnerability to natural disasters, advanced remote sensing capabilities are integral to national development and resilience. The need for localized and context-aware remote sensing solutions in Indonesia is urgent. Generic global systems may not fully address the specific challenges faced by this archipelago nation. This indicates that the proposed system's focus on the Indonesian language and specific case studies is not merely an academic exercise but a response to a real national need. This means that the system's design must inherently consider cultural, linguistic, and environmental nuances, moving beyond generic VLM/LLM applications to specialized, localized ones. This emphasis on the Indonesian language elevates the significance of this feature from a mere characteristic to a core enabling factor for practical impact.

Despite its numerous benefits, interpreting vast, complex, and high-velocity satellite data remains a significant challenge. Traditional methods, such as manual interpretation, visual inspection, and simple quantitative analysis, prove inefficient and inflexible for effectively handling large datasets. Human limitations in managing extensive study areas necessitate machine assistance in data interpretation (Lillesand et al., 2015). The shift from manual interpretation to machine assistance signifies a fundamental paradigm change from human-centric, subjective analysis to AI-driven, scalable, and objective interpretation. This transformation is not just about efficiency but also

about unlocking levels of detail and speed previously unattainable. AI, particularly VLMs and LLMs, is capable of processing patterns and relationships that are too subtle or too numerous for the human eye, and can standardize interpretations, thereby reducing subjectivity. The "generative" nature of VLMs (Weng et al., 2025) allows for flexible outputs beyond predefined categories, which is a major departure from traditional "discriminative" models. This shift redefines the role of remote sensing professionals from interpreters to system architects and validators, focusing on model training, ethical deployment, and solving complex problems leveraging AI-generated insights.

Advances in machine learning and artificial intelligence have introduced Vision-Language Models (VLMs) such as BERT, GPT, and ViT, which integrate natural language processing (NLP) with computer vision (Li et al., 2022; Zhang et al., 2021). These models promise to enhance remote sensing data interpretation. VLMs frame tasks as generative models, aligning language with visual information, thereby enabling the handling of more challenging problems and greater flexibility compared to discriminative models (Weng et al., 2025). Furthermore, Large Language Models (LLMs) offer unprecedented advancements in language understanding and sophisticated reasoning capabilities (Weng et al., 2025). The shift from "discriminative models" to "generative models" is a fundamental and important trend. It means that the system does not merely classify or detect predefined objects, but can *describe* and *reason* about novel or complex scenarios, providing richer, more nuanced outputs. For example, a discriminative model might classify "forest" or "urban area." A generative VLM could describe "a dense, healthy forest canopy with signs of recent logging on the eastern edge, bordering a newly developed residential area with red-roofed houses." This level of detail and contextualization is precisely what "narrative or descriptive text" refers to. This generative capability enables a more human-like interaction with remote sensing data, transforming raw data into actionable intelligence that can be directly consumed by decision-makers without extensive GIS or remote sensing expertise.

This research aims to develop an advanced model for interpreting remote sensing and drone images, providing detailed textual descriptions, and to create a system that generates visual representations of regions based on textual input, thereby enhancing decision-making. The system will allow users to upload images, select an AOI, input text-based queries, process data with VLMs/LLMs, determine advanced analysis, and generate narrative output. The "bi-directional" nature of the proposed system (image-to-text and text-to-image/visual representation) is a key differentiator, indicating a more sophisticated interaction model than common unidirectional interpretation systems. This hints at a movement towards conversational AI for remote sensing, where users can iteratively refine their queries and analyses. This capability enables "visual question answering" (VQA) where users ask questions about images, and "text-to-image generation" for scenario planning or visualizing hypothetical changes based on textual descriptions (e.g., "show me what this area would look like with a new road here"). This iterative feedback is crucial for complex decision-making. This bi-directional capability transforms remote sensing tools from passive analysis platforms into active intelligent assistants that can engage in complex reasoning and scenario exploration with users, significantly expanding their accessibility and utility.

2. Literature Review

2.1. Vision-Language Models (VLMs): Architecture and Evolution

Vision-Language Models (VLMs) are a subset of artificial intelligence models that combine data from both visual and textual modalities to perform various tasks. These models integrate computer vision techniques with natural language processing (NLP) methods (Zhang et al., 2021). Early examples of VLMs include BERT, developed by Google for NLP and integrating visual information for tasks like image annotation (Devlin et al., 2018); ViT, one of the first models to propose using transformers for image processing (Fu, 2022); LXMERT, specifically designed for vision-language tasks and trained on image-text pairs for image explanation and visual question answering (Tan & Bansal, 2019); UNITER, which integrates information from images and text using multilevel transformers (Chen et al., 2020); and Llama, also developed by Google Research to understand the relationship between text and images (Touvron et al., 2023). Recent advancements have led to models like LLaVA and GPT-4, which have driven significant progress in the remote sensing domain (Weng et al., 2025).

The evolution from early VLMs like BERT (primarily NLP adapted for vision) to specialized models like LXMERT and then to large general models like Llama, LLaVA, and GPT-4 reflects a trend towards increasingly unified and powerful multimodal understanding. This implies that the proposed system can leverage highly general pre-trained models, reducing the need for extensive domain-specific training from scratch, while still requiring fine-tuning for remote sensing specifications (Weng et al., 2025). This shift towards larger foundation models means they are "pre-loaded" with vast general knowledge, which can be transferred to remote sensing tasks through fine-tuning. This differs from earlier models that might have required more specialized architectural designs for each task. This makes the development process more efficient and the models more generalizable. This trend suggests that future remote sensing AI will increasingly rely on adapting general-purpose foundation models rather than building highly specialized models from scratch, accelerating innovation and reducing development costs.

2.2. The Role of Large Language Models (LLMs) in a Multimodal Context

Large Language Models (LLMs), exemplified by ChatGPT and LLaMA, have demonstrated unprecedented advancements in language understanding, reasoning, and generation (Weng et al., 2025). These models are increasingly viewed as "world models" capable of simulating and predicting real-world scenarios

. Multimodal LLMs (MLLMs) integrate LLMs to handle various data types such as text, images, and spatial information, enabling multi-task learning through few-shot or zero-shot techniques (Yuan et al, 2025). The concept of LLMs as "world models" has profound implications for remote sensing. It suggests that LLMs can go beyond mere description to infer causal relationships, predict changes, and simulate scenarios based on visual and textual input. This elevates the system from a descriptive tool to a predictive and analytical one, which is crucial for applications like disaster management. For instance, in disaster management, an LLM as a world model could not only describe flood damage but also, with historical data and current conditions, predict areas most vulnerable to future flooding or simulate the impact of mitigation strategies. This moves beyond static analysis to dynamic and predictive intelligence. This capacity transforms remote sensing from a reactive monitoring tool into a proactive planning and forecasting instrument, enabling more effective policy-making and resource allocation in critical domains.

2.3. Principles of Multimodal Data Integration (MDI) in GeoAI

Multimodal Data Integration (MDI) is defined as the process of combining information from various sources or formats (text, images, audio, sensor data) to obtain more comprehensive information about a particular case study or problem. This integration allows for richer analysis

and understanding that would not be possible by examining each modality separately (Bellandi et al., 2022). The MDI framework involves Data Collection, Preprocessing, Feature Extraction, Alignment and Fusion, Modeling and Analysis, and Evaluation and Interpretation. OmniGeo is a Multimodal LLM (MLLM) tailored for geospatial artificial intelligence (GeoAI), capable of processing and analyzing heterogeneous data sources, including satellite imagery, geospatial metadata, and textual descriptions (Yuan et al, 2025). The emphasis on "Alignment and Fusion" within the MDI framework and the discussion of methods to "bridge the gap between modalities" (Tao et al, 2025) reveal a key technical challenge: ensuring that information from different modalities (e.g., pixels and words) are meaningfully connected. This implies that the success of the proposed system relies not only on powerful VLMs/LLMs but also on robust techniques for aligning cross-modal representations. This involves designing or fine-tuning projection layers or attention mechanisms that can map high-dimensional image features to the semantic space of language, and vice versa, without losing critical information or introducing noise. The "semantic gap" is a recognized challenge. The robustness of the proposed system's interpretation and generation capabilities will directly correlate with the sophistication and effectiveness of its multimodal alignment mechanisms, making this a critical area of focus in the modeling phase.

2.4. Recent Advancements in VLM/LLM for Remote Sensing

VLMs have shown significant progress in various remote sensing tasks, including geophysical classification, object detection, and scene understanding (Weng et al., 2025; Park et al, 2025). These models offer advantages in handling long-tail distributions and multi-task learning within a single framework. Datasets are also evolving from manual to combined to automatically annotated, with large-scale datasets like SkyEye-968K, MMRS-1M, RS5M, and SkyScript emerging (Weng et al., 2025). The trend towards "automatically annotated datasets" is a game-changer for scaling VLM/LLM applications in remote sensing. This indicates a shift from labor-intensive, expert-driven data labeling to AI-assisted data generation, which can rapidly expand available training data for specific tasks and languages, directly supporting the feasibility of fine-tuning for Indonesian. Automated annotation, for example, using LLMs with external data sources like maps (Anderson et al., 2025), can create large-scale, domain-specific datasets (such as fMoW-mm (Anderson et al., 2025)) that would otherwise be prohibitively expensive to produce manually. This directly addresses the data scarcity challenge often faced in specialized language or low-resource domains. This methodological shift in data creation is crucial for the scalability and adaptation of VLM/LLM solutions, democratizing access to powerful AI for a wide range of geospatial applications and linguistic contextsb (Weng et al., 2025).

Table 1 Key Vision-Language Models and Their Relevance to Remote Sensing Tasks

Model	Description	Relevance to Remote Sensing (RS) Tasks	Key Features / Strengths	Source
BERT	Bidirectional Encoder Representations from Transformers. Developed by Google for NLP.	Image annotation, visual language understanding.	Integrates visual information with language understanding.	(Devlin et al., 2018)
ViT	Vision Transformer. One of the first models to propose using transformers for image	RS image processing, classification.	Converts images into pixel sequences and applies transformers, eliminating the need for	(Fu, 2022)

	processing.		feature preprocessing like convolution.	
LXMERT	Learning Cross-Modality Encoder Representations from Transformers.	Image explanation, visual question answering (VQA).	Combines visual and text representations in one large transformer model, trained on image-text pairs.	(Tan & Bansal, 2019)
UNITER	UNiversal Image-TExt Representation Learning.	Understanding image-text relationships in vision-language tasks.	Integrates information from images and text using multilevel transformers.	(Chen et al., 2020)
Llama	Language Model for Multimodal Access. Developed by Google Research.	Multimodal data understanding, text and image integration.	Integrates NLP and image processing techniques for better utilization of multimodal data.	(Touvron et al., 2023)
CLIP	Contrastive Language-Image Pre-training.	Zero-shot image understanding, cross-modal tasks.	Drives significant advancements in various cross-modal tasks and zero-shot image understanding.	(Weng et al., 2025)
LLaVA	Large Language and Vision Assistant.	Enhanced performance, diverse capabilities, conversational interaction in RS data analysis.	High-performing VLM that emerged after the release of GPT-4.	(Weng et al., 2025)
GPT-4	Generative Pre-trained Transformer 4.	Advanced reasoning, language understanding, text generation.	Highly sophisticated large language model, driving advancements in RS VLMs.	(Weng et al., 2025)
OmniGeo	Multimodal LLM (MLLM) tailored for GeoAI.	Processes satellite imagery, geospatial metadata, textual descriptions.	Integrates geospatial information from various modalities, handles multiple heterogeneous geospatial tasks simultaneously.	(Yuan et al, 2025)
LANGO	LANGuage-guided Object detection.	Object detection in aerial images, addressing scene and instance variations.	Incorporates language-guided learning to address variations like illumination	(Park et al, 2025)

3. Proposed Research Methodology

This research methodology is divided into four main stages: data collection, preprocessing, modeling, and evaluation. This flow may change according to the research focus, scope, and limitations of the research problem.

3.1. System Architecture

The proposed system will operate as an interactive platform allowing users to upload remote

sensing (satellite or drone) imagery and specify an Area of Interest (AOI). Subsequently, users can input text-based questions or commands in Indonesian. The system will process these inputs using integrated VLMs and LLMs, perform the specified analysis, and generate outputs in an easily understandable narrative or descriptive text format. This architecture is designed to support a bi-directional workflow, enabling both image-to-text interpretation and visual generation from text.

3.2. Data Collection

The research will commence with data collection, which will then be stored in a database. The data includes images obtained from satellite imagery and drones, as well as text data to be used as a corpus. Image data forms the basis for the image interpreter model, while text data helps understand user commands and provides textual output. Additionally, datasets from previous research can be combined to enhance the overall knowledge base.

3.3. Data Preprocessing

Utilizing BRIN's High-Performance Computing (HPC) resources, the collected image data will undergo preprocessing to produce standardized data with higher informational value. This includes techniques such as cloud data removal. Similarly, text data will be processed to clean, standardize, and extract useful features. This involves data cleaning, word simplification, normalization, and other spatial and text data analysis methods.

3.4. VLM and LLM Modeling

The next stage involves fine-tuning several popular vision-language models. Fine-tuning adapts a pre-trained model to specific tasks or data by continuing to train the model with smaller or task-specific datasets. This research focuses on using the Indonesian language, where tasks involve interpreting images based on text commands and providing textual responses. The model will interpret the selected image and generate answers based on existing repository knowledge. The task model will also be tailored to specific case studies, such as disaster management or health applications.

3.5. Advanced Analytical Capabilities

The proposed system will support a range of advanced analytical capabilities driven by VLMs and LLMs:

- **Object Detection and Semantic Classification:** VLMs will be employed to identify objects within images, such as buildings, vehicles, or land cover types, and semantically classify scenes. Recent advancements in object detection, such as the LANGO framework, demonstrate how language-guided learning can significantly improve performance by addressing external variations (e.g., weather conditions, illumination, viewpoint and scale changes) often found in aerial and satellite imagery (Park et al, 2025). This enables the system to find foreground candidate regions and effectively suppress clutter from the background, even for small objects.

- **Quantitative Estimation and Visual Question Answering (VQA):** The system will be capable of performing quantitative estimations, such as counting objects or measuring areas, and supporting visual question answering (VQA) where users can ask specific questions about image content. While LLMs have demonstrated advanced reasoning capabilities, recent research on datasets like SensorQA indicates that they still face challenges in precise quantitative question answering for sensor data, especially those involving time-related queries or complex numerical data (Reichman et al., 2023). This suggests that the development of quantitative estimation capabilities will require specific attention to model design and fine-tuning strategies to overcome these limitations, possibly by integrating specialized modules for accurate numerical information extraction.
- **Multitemporal Change Analysis:** Modern change detection algorithms will be applied to compare multi-temporal remote sensing images to identify areas of change. This enhances the ability to monitor and respond to dynamic environmental conditions, such as deforestation, urbanization, or disaster impacts.

3.6. Narrative and Descriptive Generation

One of the primary goals of this system is to generate information in an easily understandable narrative or descriptive text format for users. LLMs offer a promising alternative for generating more descriptive and context-rich captions, moving beyond traditional rule-based methods that often lack the depth to capture complex wide-area scenes (Anderson et al., 2025). However, LLM-generated text for remote sensing data often remains generic and, importantly, is prone to "hallucinations" (incorrect or non-existent information) (Anderson et al., 2025). To address this, methods will be implemented to measure and mitigate hallucinations in LLM-generated text. This may include integrating external data sources such as maps and metadata to provide rich context, which has been shown to reduce hallucination rates and improve caption specificity (Anderson et al., 2025).

Table 2 Recent Datasets for VLM/LLM in Remote Sensing

Dataset	Year	Image Source	Image Size	Primary Task	Notes	Source
HallusionBench	2023	-	-	VQA	Manual dataset for VQA.	(Weng et al., 2025)
RSICap	2023	DOTA	512	IC (Image Captioning)	Manual dataset.	(Weng et al., 2025)
CRSVQA	2023	AID	600	VQA	Manual dataset.	(Weng et al., 2025)

SATIN	2023	Million-AID, WHU-RS19, SAT-4, AID	≈ 775K	SC (Scene Classification)	Combined dataset.	(Weng et al., 2025)
GeoPile	2023	NAIP, RSD46-WHU, MLRSNet, RESISC45, PatternNet	600K	-	Combined dataset.	(Weng et al., 2025)
SatlasPretrain	2023	UCM, BigEarthNet, AID, Million-AID, RESISC45, FMoW, DOTA, iSAID	512	-	Combined dataset.	(Weng et al., 2025)
RSVGD	2023	DIOR	800	VG (Visual Grounding)	Combined dataset.	(Weng et al., 2025)
RefsegRS	2024	SkyScapes	512	RRSIS (Referring Remote Sensing Image Segmentation)	Combined dataset.	(Weng et al., 2025)
SkyEye-968K	2024	RSICD, RSITMD, RSIVQA, RSVG	968K	-	Combined dataset.	(Weng et al., 2025)
MMRS-1M	2024	AID, RSIVQA, Syndney-Captions	1M	-	Combined dataset.	(Weng et al., 2025)
RS5M	2024	LAION2B-en, LAION400M, LAIONCOCO	5M	-	Automatically annotated dataset, LLM-generated captions.	
SkyScript	2024	Google Earth Engine, OpenStreetMap	2.6M	-	Automatically annotated dataset, LLM-	(Weng et al., 2025)

					generated captions.	
fMoW-mm	2025	fMoW	-	ATR (Automatic Target Recognition)	New multimodal dataset with satellite imagery, maps, metadata, and text annotations, addresses hallucination.	(Anderson et al., 2025)
ChatEarthNet	2024	Sentinel-2, WorldCover	256	-	Automatically annotated dataset, LLM-generated captions, primarily land cover descriptions.	

References

1. Anderson, M., Cha, M., Freeman, W. T., Perron, J. T., Maidel, N., & Cahoy, K. (2025). *Measuring and mitigating hallucinations in vision-language dataset generation for remote sensing* [Preprint]. arXiv. https://arxiv.org/abs/2501.14905
2. Bellandi, V., Ceravolo, P., Maghool, S., & Siccardi, S. (2022). Toward a general framework for multimodal big data analysis. *Big Data, 10*(5), 408–424.
3. Chen, Y.-C., Li, L., Yu, L., Lu, J., Li, X., & Wang, X. (2020). UNITER: Universal image-text representation learning. In *European Conference on Computer Vision* (pp. 104–120). Springer.
4. Chuvieco, E. (2020). *Fundamentals of satellite remote sensing: An environmental approach*. CRC Press.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding* [Preprint]. arXiv. https://arxiv.org/abs/1810.04805
6. Fu, Z. (2022). *Vision Transformer: ViT and its derivatives* [Preprint]. arXiv. https://arxiv.org/abs/2205.11239
7. Park, S., Kim, H., Park, B., & Ro, Y. M. (2025). *Language-guided learning for object detection tackling multiple variations in aerial images* [Preprint]. arXiv. https://arxiv.org/abs/2505.23193

8. Li, F., Zhang, P., Gan, Z., Li, X., Wang, X., & Liu, J. (2022). *Vision-language intelligence: Tasks, representation learning, and large models* [Preprint]. arXiv. https://arxiv.org/abs/2203.01922
9. Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.
10. Martino, L., Ulivieri, C., Jahjah, M., & Loret, E. (2009). Remote sensing and GIS techniques for natural disaster monitoring. In *Space Technologies for the Benefit of Human Society and Earth* (pp. 331–382).
11. Yuan, L., Mo, F., Huang, K., Wang, W., Zhai, W., Zhu, X., Li, Y., Xu, J., & Nie, J.-Y. (2025). *OmniGeo: Towards a multimodal large language models for geospatial artificial intelligence* [Preprint]. arXiv. https://arxiv.org/abs/2503.16326
12. Reichman, B., Yu, X., Hu, L., Truxal, J., Jain, A., Chandrupatla, R., Šimunić Rosing, T., & Heck, L. (2025). *SensorQA: A human-created dataset and benchmark for QA interactions between humans and long-term time-series sensor data* [Preprint]. arXiv. https://arxiv.org/abs/2501.04974
13. Sadewa, A. H., & Supriyadi, A. A. (2024). The use of remote sensing in monitoring shoreline change: Implications for maritime area security. *Remote Sensing Technology in Defense and Environment, 1*(1), 28–35.
14. Tan, H., & Bansal, M. (2019). *LXMERT: Learning cross-modality encoder representations from transformers* [Preprint]. arXiv. https://arxiv.org/abs/1908.07490
15. Tao, L., Zhang, H., Jing, H., Liu, Y., Yan, D., Wei, G., & Xue, X. (2025). *Advancements in visual language models for remote sensing: Datasets, capabilities, and enhancement techniques* [Preprint]. arXiv. https://arxiv.org/abs/2410.17283
16. Touvron, H., Lavril, T., Izacard, G., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models* [Preprint]. arXiv. https://arxiv.org/abs/2302.13971
17. Weng, X., Pang, C., & Xia, G.-S. (2025). Vision-language modeling meets remote sensing: Models, datasets and perspectives. *IEEE Geoscience and Remote Sensing Magazine*. (Early Access)
18. Zhang, P., Li, F., Hu, X., Wang, X., & Liu, J. (2021). VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5579–5588).