Review

Vision–Language Multimodal Learning for UAV-Based Remote Sensing and Geospatial Artificial Intelligence: Tasks, Datasets, Benchmarks, and Foundation Models

Youla Yang *

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2191.v1

Keywords: multimodal learning; vision-language models; UAV; drone sensing; aerial imagery; remote sensing; geospatial artificial intelligence; GeoAI; foundation models; multimodal large language models; earth observation; cross-modal alignment

 doi:10.20944/preprints202603.2191.v1

Review

Vision–Language Multimodal Learning for UAV-Based Remote Sensing and Geospatial Artificial Intelligence: Tasks, Datasets, Benchmarks, and Foundation Models

Youla Yang

School of Intelligent Systems Engineering, Indiana University Bloomington, Bloomington, IN 47405, USA; youla@example.com

Abstract

Earth observation systems increasingly rely on multimodal data collected from satellites, aircraft, and unmanned aerial vehicles (UAVs). UAV platforms provide flexible, high-resolution, and real-time sensing capabilities that are essential for modern remote sensing, environmental monitoring, urban analysis, disaster response, and intelligent geospatial systems. However, the growing volume and diversity of aerial and geospatial data creates a substantial gap between large-scale visual observations and human-level semantic understanding. Recent advances in vision–language multimodal learning and multimodal large language models offer new opportunities to bridge this gap by enabling cross-modal alignment between imagery and natural language. These techniques allow UAV and remote sensing imagery to be queried, interpreted, and analyzed through language-guided perception, multimodal reasoning, and interactive decision support. This paper presents a comprehensive survey of vision–language multimodal learning for UAV-based remote sensing and geospatial artificial intelligence (GeoAI). We introduce a structured taxonomy of multimodal task formulations, including image–text retrieval, visual grounding, captioning, visual question answering, and multimodal reasoning for aerial and geospatial imagery. We systematically review representative datasets, benchmarks, model architectures, and evaluation protocols, and summarize recent progress in geospatial foundation models and multimodal large language models for Earth observation and UAV applications. We further discuss key challenges in multimodal UAV sensing, including domain shift across sensors, spatial reasoning in aerial imagery, temporal modeling in Earth observation, multimodal alignment quality, and trustworthy deployment in real-world geospatial systems. This survey aims to organize the rapidly growing literature on multimodal learning for UAV and remote sensing applications and to highlight future research directions toward scalable, reliable, and reasoning-capable Earth observation and drone-based sensing systems.

Keywords: multimodal learning; vision-language models; UAV; drone sensing; aerial imagery; remote sensing; geospatial artificial intelligence; GeoAI; foundation models; multimodal large language models; earth observation; cross-modal alignment

1. Introduction

Earth observation (EO) has experienced rapid growth in recent decades due to the increasing availability of data acquired from satellites, aircraft, unmanned aerial vehicles (UAVs), autonomous platforms, surveillance systems, maritime sensors, and large-scale intelligent sensing networks. Among these platforms, UAV-based sensing has become particularly important due to its flexibility, low cost, high spatial resolution, and real-time deployment capability. These sensing platforms provide massive amounts of multimodal data with rich spatial, spectral, temporal, and contextual information, enabling a wide range of applications including land-use monitoring, disaster assessment, urban analytics, transportation intelligence, environmental monitoring, infrastructure inspection, and public safety. However, despite the abundance of aerial and geospatial imagery, a substantial gap remains between large-scale visual observations and human-level semantic understanding.

 doi:10.20944/preprints202603.2191.v1

Traditional remote sensing pipelines mainly focus on perception-oriented tasks such as image classification, object detection, semantic segmentation, change detection, and spatio-temporal prediction. Although these tasks are fundamental for Earth observation and UAV-based sensing, they are typically designed for specific objectives with predefined label spaces and structured supervision. As a result, conventional remote sensing systems have limited ability to support natural human interaction, language-driven querying, semantic interpretation, and reasoning-based decision support, which are increasingly required in modern geospatial artificial intelligence (GeoAI) and intelligent drone sensing applications.

Vision–language multimodal learning has recently emerged as a promising direction for bridging the gap between visual data and semantic understanding. By aligning visual observations with textual descriptions, multimodal models enable UAV and remote sensing scenes to be retrieved, described, interpreted, and analyzed using natural language. Advances in contrastive representation learning, cross-attention fusion, multimodal pretraining, and instruction tuning have significantly improved performance in tasks such as image–text retrieval, captioning, visual grounding, and visual question answering. More recently, multimodal large language models (MLLMs) and foundation models have extended multimodal learning beyond perception-level alignment toward reasoning, dialogue, planning, and decision-oriented analysis.

These developments are particularly important for UAV-based remote sensing and geospatial artificial intelligence, where real-world systems must integrate perception, language, geographic knowledge, and reasoning. For example, multimodal models can support disaster-response analysis from UAV imagery, language-guided search over aerial image archives, environmental monitoring using drone platforms, urban infrastructure inspection, and question answering about remote sensing scenes. Such capabilities are essential for next-generation drone sensing and GeoAI systems that are interactive, interpretable, scalable, and operationally reliable.

Despite the rapid progress of multimodal learning, existing research on multimodal methods for UAV and remote sensing applications is still scattered across multiple communities, including computer vision, remote sensing, multimodal machine learning, robotics, and geographic information science. Previous survey papers have mainly focused on deep learning for remote sensing, vision–language models for natural images, or multimodal learning in general artificial intelligence. A unified overview dedicated to vision–language multimodal learning for UAV-based remote sensing and geospatial artificial intelligence is still limited. In particular, there is a need for a comprehensive survey covering multimodal task taxonomy, benchmark datasets, evaluation protocols, model architectures, multimodal large language models, and emerging geospatial foundation models for aerial and drone-based sensing systems.

In this paper, we present a comprehensive review of vision–language multimodal learning for UAV-based remote sensing and geospatial artificial intelligence. We systematically organize the literature from the perspectives of multimodal task formulations, datasets and benchmarks, model architectures, foundation models, and real-world applications. In addition, we analyze current limitations and discuss future research directions toward scalable, reliable, and reasoning-capable multimodal GeoAI and drone sensing systems.

1.1. Remote Sensing, UAV Sensing, and Geospatial Artificial Intelligence

Remote sensing and geospatial artificial intelligence (GeoAI) have traditionally focused on perception-oriented tasks such as image classification, object detection, semantic segmentation, change detection, and spatio-temporal forecasting. In recent years, UAV-based sensing has become an important complement to satellite and airborne remote sensing, providing high-resolution imagery, flexible acquisition, and real-time observation capabilities. UAV platforms are widely used in applications such as environmental monitoring, precision agriculture, disaster response, infrastructure inspection, and urban analysis.

With the success of deep learning, perception-oriented tasks in remote sensing and UAV imagery analysis have achieved substantial performance improvements compared with traditional hand-

 doi:10.20944/preprints202603.2191.v1

crafted feature pipelines. However, most existing methods rely on fixed label spaces and structured annotations, which limits their ability to support flexible human–AI interaction and semantic-level understanding. As geospatial and drone sensing systems become more complex and user-facing, there is increasing interest in integrating imagery with natural language, geographic metadata, maps, and domain knowledge to enable more interpretable and interactive analysis.

1.2. Multimodal Learning for Remote Sensing and UAV Imagery

In recent years, a growing number of studies have explored multimodal learning for remote sensing, UAV imagery, and GeoAI applications. Existing work includes aerial image captioning, image–text retrieval, visual grounding, multimodal question answering, and language-guided geospatial search. These approaches aim to bridge the gap between visual observations and natural language descriptions, allowing users to query and interpret satellite and UAV imagery in a more intuitive way.

Several multimodal datasets and benchmarks have been proposed for Earth observation and aerial imagery, but their scale, diversity, and annotation quality are still limited compared with natural-image benchmarks. Moreover, prior studies are often task-specific, and a systematic overview of multimodal learning for UAV-based sensing and geospatial data is still lacking.

1.3. Foundation Models and Multimodal Large Language Models for UAV and Remote Sensing

The emergence of foundation models and multimodal large language models has introduced a new paradigm for multimodal perception and reasoning. By combining vision encoders, projection modules, and large language model backbones, these systems support multimodal instruction following, captioning, question answering, and reasoning across different domains.

Recent work has begun to apply such models to remote sensing and UAV-based sensing, enabling conversational querying of aerial imagery, reasoning-based disaster assessment using drone data, and language-guided environmental monitoring. However, adapting foundation models to aerial and geospatial imagery remains challenging due to large spatial scale, geographic coordinate information, temporal dynamics, sensor heterogeneity, and domain-specific semantics.

2. Task Taxonomy

3. Multimodal Datasets for UAV and Remote Sensing

Datasets are central to the development of multimodal learning for UAV-based sensing, remote sensing, and geospatial artificial intelligence (GeoAI). A strong multimodal dataset should provide diverse aerial and geospatial scenes, high-quality annotations, realistic task formulations, and evaluation protocols that reflect practical use cases in satellite, airborne, and drone-based observation systems.

3.1. Representative Dataset Types

Representative multimodal remote sensing datasets may include:

- satellite image–caption pairs,
- aerial image–caption pairs,
- UAV image–text datasets,
- drone-based inspection datasets,
- question–answer datasets for remote sensing VQA,
- region–text annotations for grounding,
- multimodal geospatial dialogue or instruction datasets.

Common examples in the literature include captioning datasets such as RSICD, UCM-Captions, and NWPU-Captions, as well as emerging multimodal resources designed for aerial imagery analysis, UAV monitoring, visual question answering, geospatial dialogue, and language-guided retrieval.

3.2. Key Dataset Characteristics

Important benchmark design considerations include:

 doi:10.20944/preprints202603.2191.v1

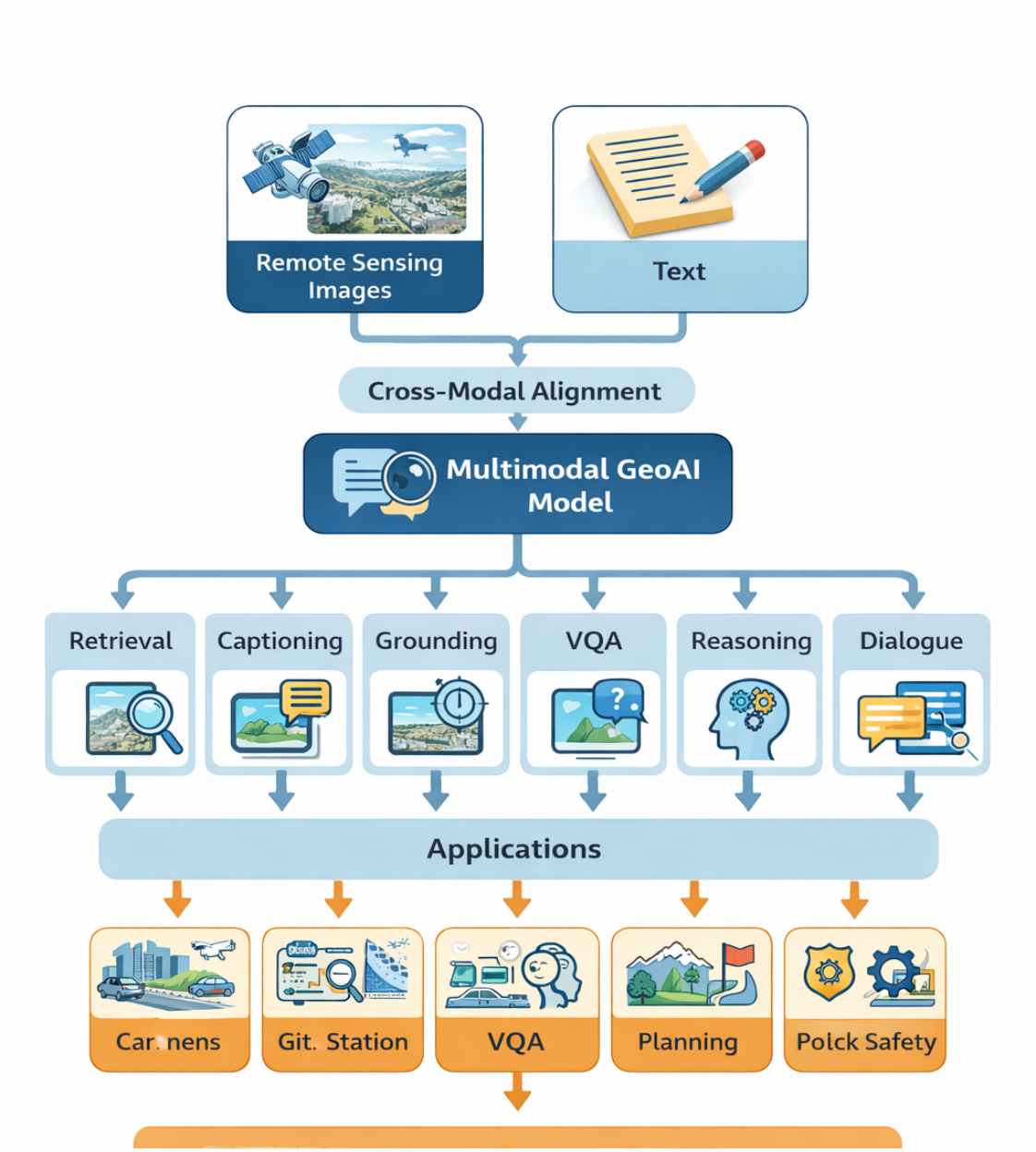

Figure 1. Taxonomy of vision–language multimodal learning for remote sensing and geospatial artificial intelligence. Multimodal GeoAI systems integrate remote sensing imagery and natural language through cross-modal alignment, enabling retrieval, captioning, grounding, visual question answering, reasoning, dialogue, and decision-support applications.

 doi:10.20944/preprints202603.2191.v1

Table 1. Representative multimodal datasets for UAV and remote sensing applications.

Dataset	Modality	Task	Size	Year
RSICD	Image–Text	Captioning	10k images	2018
UCM-Captions	Image–Text	Captioning	2.1k images	2016
NWPU-Captions	Image–Text	Captioning	31k images	2019
RSVQA	Image–Text	VQA	770k QA pairs	2019
RSVG	Image–Text	Grounding	8k images	2020
GeoChat Dataset	Image–Text	Dialogue / QA	–	2023
DIOR Caption	Image–Text	Captioning	23k images	2021
BigEarthNet + Text	Image–Text	Retrieval / Caption	590k patches	2020

- geographic diversity,
- sensor diversity,
- annotation quality,
- language richness,
- temporal coverage,
- multi-scale scene complexity,
- robustness to domain shift.

Geographic diversity is particularly important because models trained on one region may fail to generalize to another. Sensor diversity also matters, as data from satellites, aircraft, UAVs, and ground platforms can differ substantially in viewpoint, scale, and appearance. Language richness affects whether a model can learn fine-grained semantics rather than overly generic scene descriptions. These factors are especially critical for UAV-based sensing, where acquisition conditions may vary significantly across missions.

3.3. Current Limitations of Available Datasets

Although multimodal datasets for natural-scene vision can reach extremely large scale, remote sensing and UAV datasets remain comparatively small and narrow. Many existing datasets focus on captioning or retrieval only, with limited support for reasoning, grounding, or instruction following.

In addition, annotations may be templated or repetitive, reducing linguistic diversity. There is a strong need for larger, more diverse, and more realistic multimodal datasets for satellite, airborne, and drone-based sensing systems.

4. Benchmarks and Evaluation Protocols

Evaluation is a critical component of multimodal learning for UAV, aerial, and remote sensing systems. Because multimodal GeoAI spans multiple task types, evaluation protocols must be tailored to the task while also capturing semantic correctness and practical utility in real-world sensing scenarios.

4.1. Retrieval Evaluation

For image–text retrieval, common metrics include Recall@K, median rank, and mean rank. These metrics measure whether the correct image is retrieved for a text query, or vice versa.

In UAV and aerial retrieval tasks, it is also useful to consider semantic similarity and geographic ambiguity, since multiple images may partially satisfy a textual description.

4.2. Captioning Evaluation

For caption generation, common metrics include BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. However, standard captioning metrics may not fully reflect correctness in aerial imagery, especially when captions involve object counts, spatial layout, or infrastructure structures.

Human evaluation and domain-aware criteria are therefore often required in UAV and remote sensing applications.

 doi:10.20944/preprints202603.2191.v1

Table 2. Evaluation metrics for multimodal UAV and remote sensing tasks.

Task	Metric	Description
Image–Text Retrieval	Recall@K	Correct match in top K
Image–Text Retrieval	Median Rank	Ranking quality
Captioning	BLEU	N-gram precision
Captioning	CIDEr	Consensus similarity
Captioning	METEOR	Semantic matching
Captioning	ROUGE-L	Sequence overlap
Grounding	IoU	Region overlap
Grounding	Accuracy	Correct localization
VQA	Accuracy	Correct answer rate
Reasoning	Human Eval	Semantic correctness
Reasoning	Consistency	Logical stability
Multimodal QA	F1 / EM	Text match score

4.3. Grounding Evaluation

For visual grounding, intersection-over-union (IoU), grounding accuracy, and localization precision are commonly used.

In aerial and drone imagery, evaluation may be complicated by

- small targets,
- dense object distributions,
- viewpoint variation,
- altitude changes,
- sensor differences.

4.4. VQA and Reasoning Evaluation

For visual question answering, answer accuracy is the most common metric. For more advanced multimodal reasoning tasks, additional criteria may include faithfulness, consistency, explanation quality, robustness, and human preference.

Future UAV and remote sensing benchmarks should include more reasoning-oriented evaluation settings.

4.5. Benchmark Design Principles

A strong benchmark for UAV and multimodal GeoAI should:

- reflect realistic aerial sensing tasks,
- include diverse acquisition conditions,
- support semantic and reasoning evaluation,
- include both automatic and human-centered metrics,
- measure generalization across regions and sensors.

 doi:10.20944/preprints202603.2191.v1

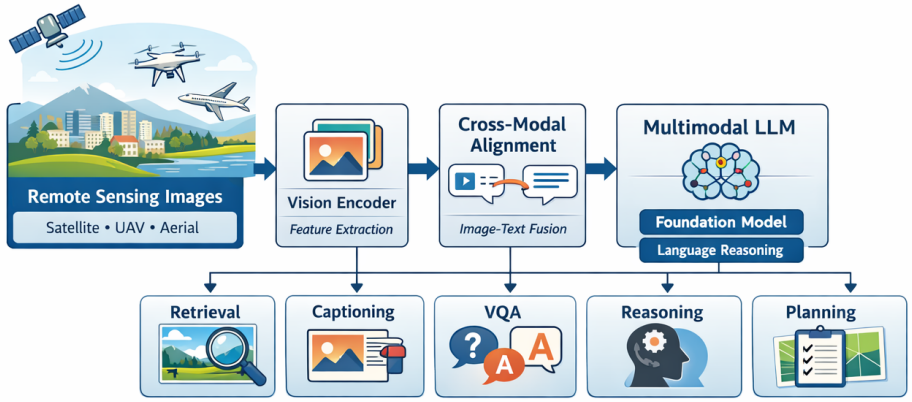

Figure 1. Pipeline of vision–language multimodal learning for remote sensing and geospatial artificial intelligence.

Figure 2. Pipeline of vision–language multimodal learning for UAV and remote sensing applications. Aerial or UAV imagery is processed by a vision encoder, aligned with textual representations through cross-modal fusion, and further handled by multimodal large language models to support retrieval, captioning, visual question answering, reasoning, and decision support in drone and Earth observation systems.

5. Methodology and Model Architectures

Depending on the research objective, multimodal UAV and remote sensing systems may be built using contrastive pretraining, fusion-based architectures, generative models, or multimodal large language models. These approaches are widely used in satellite, airborne, and drone-based sensing applications.

5.1. Cross-Modal Alignment

Cross-modal alignment aims to associate visual content with textual semantics. A common formulation learns visual and textual embeddings and optimizes their similarity in a shared latent space. Contrastive learning has become a standard approach for image–text alignment and serves as the basis for many retrieval and transfer-learning systems for aerial imagery and UAV data.

Let $v = g_v(x_v)$ denote the visual embedding and $t = g_t(x_t)$ denote the text embedding. A contrastive objective may encourage matched image–text pairs to be close while pushing mismatched pairs apart. This formulation supports retrieval, transfer learning, and downstream multimodal reasoning in remote sensing and drone-based sensing systems.

5.2. Fusion-Based Multimodal Models

Fusion-based models combine visual and textual features through cross-attention, feature concatenation, co-attention, or transformer-based interaction.

Such models are common in captioning, VQA, and grounding tasks. In UAV and aerial imagery analysis, fusion must often handle

- high-resolution images,
- multi-scale structures,
- viewpoint variation,
- altitude changes,

 doi:10.20944/preprints202603.2191.v1

- complex spatial layouts.

These characteristics make multimodal fusion particularly important for drone inspection, surveillance, and environmental monitoring tasks.

5.3. Geospatial Representation Learning

A geospatially aware multimodal model should capture:

- spatial structure,
- object relationships,
- scale variation,
- geographic priors,
- temporal evolution.

This may require hierarchical encoders, region-based representations, positional modeling, or integration with metadata such as coordinates, timestamps, and sensor attributes.

These requirements are especially important for UAV-based sensing, where acquisition geometry, altitude, and camera parameters may change frequently.

5.4. Multimodal Large Language Models

MLLM-based GeoAI systems typically combine a vision encoder, a projector or alignment module, and a large language model backbone.

These systems enable multimodal instruction following, scene interpretation, question answering, report generation, and interactive analysis.

In UAV and aerial sensing, multimodal large language models enable

- drone-based inspection reports,
- aerial scene description,
- language-guided navigation,
- disaster assessment from UAV imagery,
- interactive geospatial analysis.

A typical pipeline can be summarized as:

$$\text{image} \rightarrow \text{vision encoder} \rightarrow \text{projector} \rightarrow \text{language model} \rightarrow \text{output}. \tag{1}$$

5.5. Geospatial Foundation Models

Foundation models for remote sensing aim to learn broad and transferable representations from large-scale satellite, airborne, and UAV datasets.

When extended to multimodal settings, geospatial foundation models can support

- retrieval,
- captioning,
- VQA,
- reasoning,
- zero-shot transfer,
- drone-based scene understanding.

These models are expected to play an increasingly important role in large-scale Earth observation and UAV sensing systems.

6. Applications

Vision–language multimodal learning has enabled a new generation of UAV, aerial, and remote sensing systems that support language-guided perception, reasoning, and decision-making.

By integrating drone imagery, satellite data, and natural language, multimodal GeoAI can significantly improve the usability, interpretability, and efficiency of Earth observation workflows.

 doi:10.20944/preprints202603.2191.v1

Table 3. Representative model architectures for vision–language multimodal learning in UAV and remote sensing.

Model	Type	Core Idea	Application
CLIP	Contrastive	Image–text alignment	Retrieval / Transfer
BLIP-2	Multimodal LLM	Frozen encoder + LLM	Caption / VQA
Flamingo	Multimodal LLM	Perceiver + LLM	Few-shot reasoning
LLaVA	Instruction-tuned	Vision + LLM	Dialogue / QA
GeoChat	RS MLLM	Grounded RS LLM	GeoAI / UAV QA
ViLT	Fusion	Transformer fusion	VQA / Caption
ALIGN	Contrastive	Large-scale alignment	Retrieval
RS-CLIP	RS-specific	Remote sensing alignment	Retrieval
Foundation Models (EO/UAV)	Foundation	Large-scale pretraining	General tasks

6.1. Disaster Emergency Response

Multimodal UAV systems can support disaster monitoring, damage assessment, and emergency-response planning by combining aerial or drone imagery with language-based reporting and visual question answering.

Typical scenarios include

- floods,
- wildfires,
- earthquakes,
- hurricanes,
- landslides,
- infrastructure collapse.

UAV platforms provide rapid deployment, high spatial resolution, and flexible observation angles, which make them particularly useful for real-time emergency analysis.

6.2. Urban Monitoring and Infrastructure Inspection

Urban monitoring is an important application of UAV and remote sensing.

Tasks include

- land-use analysis,
- building inspection,
- bridge monitoring,
- road analysis,
- construction progress tracking,
- smart city sensing.

Drone platforms allow close-range observation and detailed scene understanding, which can be combined with multimodal language models to generate automatic reports and queries.

6.3. Environmental and Ecological Analysis

Vision–language models are increasingly used in

- vegetation monitoring,
- wildlife observation,
- coastal analysis,
- pollution detection,
- climate monitoring,
- agricultural sensing.

UAV imagery provides high-resolution data that can be aligned with textual descriptions to generate interpretable environmental reports.

 doi:10.20944/preprints202603.2191.v1

6.4. Transportation and Maritime Monitoring

Multimodal learning also plays an important role in

- traffic monitoring,
- port surveillance,
- ship detection,
- logistics analysis,
- airspace monitoring.

Drone-based sensing allows flexible data acquisition, while multimodal reasoning enables semantic interpretation of complex scenes.

6.5. Conversational UAV and Earth Observation Systems

Recent multimodal large language models enable interactive analysis of UAV and remote sensing imagery.

Users can query aerial scenes using natural language, request summaries, or ask reasoning-based questions.

Such conversational sensing systems represent a shift from passive image interpretation to intelligent drone-assisted geospatial analysis.

7. Research Trends

Vision–language multimodal learning for remote sensing has evolved rapidly in recent years. Early studies mainly focused on image classification and object detection, followed by the introduction of image–text retrieval and captioning datasets for Earth observation. With the development of contrastive learning and transformer-based architectures, multimodal alignment models such as CLIP and BLIP significantly improved cross-modal representation learning.

More recently, multimodal large language models and foundation models have enabled new capabilities including multimodal reasoning, dialogue, and decision support for geospatial artificial intelligence. These developments indicate a clear trend from perception-level understanding toward reasoning-oriented and agent-based GeoAI systems.

Figure 3 illustrates the evolution of multimodal learning for remote sensing and geospatial artificial intelligence.

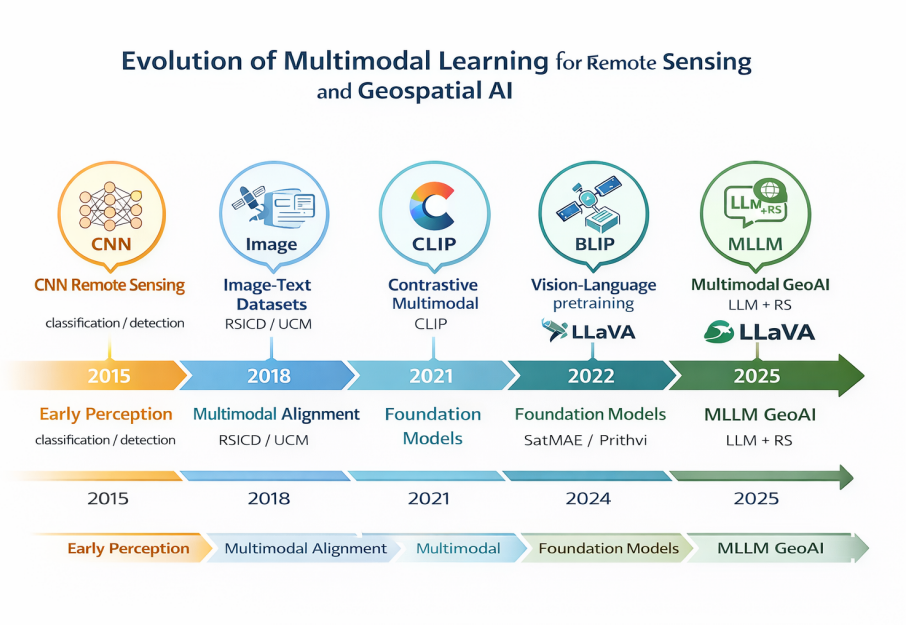

Figure 3. Evolution of multimodal learning for remote sensing and geospatial artificial intelligence.

 doi:10.20944/preprints202603.2191.v1

8. Challenges and Future Directions

Despite rapid progress, several important challenges remain in multimodal UAV sensing, remote sensing, and geospatial artificial intelligence. Addressing these challenges is essential for building scalable, reliable, and trustworthy drone-based sensing systems.

8.1. Domain Shift and Generalization

Remote sensing and UAV data exhibit strong domain variability due to differences in geographic regions, sensors, resolutions, flight altitude, illumination conditions, and annotation styles.

Models trained on one dataset often perform poorly when applied to new environments.

Domain shift is especially severe in UAV applications, where acquisition conditions can change frequently.

Future research should focus on

- domain adaptation,
- transfer learning,
- cross-sensor generalization,
- geographically diverse multimodal datasets.

8.2. Spatial Reasoning in Aerial Imagery

Many current multimodal models lack explicit spatial reasoning capability. Tasks such as relative positioning, topology understanding, multi-scale interpretation, and spatial relationship reasoning remain difficult in aerial and drone imagery.

UAV applications often require precise spatial understanding, for example in

- infrastructure inspection,
- building analysis,
- disaster mapping,
- precision agriculture.

Incorporating spatial priors, geographic knowledge, and structured representations is an important direction for future work.

8.3. Temporal Understanding in UAV and Remote Sensing

Earth observation often involves dynamic processes rather than static scenes.

Monitoring environmental change, disaster evolution, and urban development requires temporal reasoning across multiple observations.

In UAV sensing, time-series data may be collected from repeated flights, multi-view cameras, or continuous monitoring systems.

Future multimodal models should support

- temporal modeling,
- event understanding,
- long-term memory,
- multi-flight data fusion.

8.4. Alignment Quality and Semantic Precision

Cross-modal alignment errors can lead to vague, generic, or incorrect descriptions.

In UAV and remote sensing applications, such errors may affect decision-making, inspection results, or safety-critical operations.

Improving alignment quality between visual and textual representations remains a key research problem.

More precise annotations, better multimodal training strategies, and domain-specific benchmarks are needed for aerial and drone imagery.

 doi:10.20944/preprints202603.2191.v1

8.5. Scalable Multimodal Datasets for UAV Sensing

Compared with natural-image datasets, multimodal UAV and remote sensing datasets remain limited in size, diversity, and annotation quality.

Building large-scale benchmarks covering

- satellite,
- airborne,
- UAV,
- multi-sensor platforms

is essential for training robust multimodal models.

Future datasets should include realistic flight scenarios, inspection tasks, and environmental monitoring missions.

8.6. Trustworthiness and Deployment

Practical deployment of multimodal UAV systems requires

- robustness,
- interpretability,
- uncertainty estimation,
- safety guarantees,
- human-in-the-loop control.

These issues are particularly important in disaster response, surveillance, environmental monitoring, and industrial inspection.

Future research should emphasize trustworthy AI and reliable drone sensing systems.

8.7. Toward Foundation Models and Autonomous UAV Systems

The field is moving toward geospatial foundation models and multimodal large language models that can support multiple sensing tasks with a unified architecture.

Future systems may integrate

- UAV imagery,
- satellite data,
- text,
- metadata,
- navigation information

into agent-like frameworks capable of

- reasoning,
- planning,
- interaction,
- autonomous decision making.

Conversational sensing, multimodal agents, and autonomous UAV analysis represent promising directions for next-generation drone intelligence systems.

9. Conclusions

Vision–language multimodal learning is becoming a key foundation for next-generation UAV, aerial, and remote sensing systems.

By aligning visual observations with natural language, multimodal models enable intuitive retrieval, description, reasoning, and decision support for drone-based and Earth observation applications.

This paper reviewed task formulations, datasets, benchmarks, model architectures, and applications in multimodal learning for UAV and remote sensing.

 doi:10.20944/preprints202603.2191.v1

Recent advances in multimodal pretraining, foundation models, and multimodal large language models are rapidly transforming the way aerial and geospatial data are analyzed and interpreted.

However, challenges such as domain shift, limited datasets, spatial reasoning, temporal modeling, and trustworthy deployment remain open problems.

Future research is expected to move toward

- scalable multimodal foundation models,
- conversational sensing systems,
- autonomous UAV intelligence,
- agent-based geospatial analysis.

These developments will play an important role in building more intelligent, reliable, and decision-oriented drone and Earth observation systems.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Radford, A.; Kim, J.W.; Hallacy, C.; et al. Learning Transferable Visual Models From Natural Language Supervision. *ICML* 2021.
2. Alayrac, J.B.; Donahue, J.; Luc, P.; et al. Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS* 2022.
3. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *NeurIPS* 2023.
4. Zhu, D.; Chen, J.; Shen, X.; et al. MiniGPT-4: Enhancing Vision-Language Understanding. *arXiv* 2023.
5. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2. *ICLR* 2023.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image is Worth 16x16 Words. *ICLR* 2021.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net. *MICCAI* 2015.
8. Lin, T.Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO. *ECCV* 2014.
9. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Rotation-Invariant CNN. *IEEE TGRS* 2017.
10. Ayush, K.; Uzkent, B.; Meng, C.; et al. Geospatial Foundation Models. *CVPR* 2021.
11. Chen, L.C.; Papandreou, G.; Kokkinos, I. DeepLab. *TPAMI* 2017.
12. Long, J.; Shelhamer, E.; Darrell, T. FCN. *CVPR* 2015.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. ResNet. *CVPR* 2016.
14. Simonyan, K.; Zisserman, A. VGG. *ICLR* 2015.
15. Sumbul, G.; Charfuelan, M.; Demir, B. BigEarthNet. *IEEE GRSM* 2022.
16. Hu, Y.; Li, X.; Zhang, H. GeoChat. *arXiv* 2023.
17. Li, X.; Zhang, L.; Du, B. Multimodal Earth Observation. *ISPRS JPRS* 2023.
18. Wang, J.; Chen, Y.; Li, H. LLM for GeoAI. *IJGIS* 2023.
19. Sun, Y.; Zhao, W.; Wang, L. Multimodal Agents. *arXiv* 2023.
20. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning. MIT Press 2016.
21. Brown, T.; Mann, B.; Ryder, N. GPT-3. *NeurIPS* 2020.
22. OpenAI. GPT-4 Technical Report. 2023.
23. Raffel, C.; Shazeer, N.; Roberts, A. T5. *JMLR* 2020.
24. Vaswani, A.; Shazeer, N.; Parmar, N. Transformer. *NeurIPS* 2017.
25. Devlin, J.; Chang, M.W.; Lee, K. BERT. *NAACL* 2019.
26. Tan, H.; Bansal, M. LXMERT. *EMNLP* 2019.
27. Lu, J.; Batra, D. ViLBERT. *NeurIPS* 2019.
28. Su, W.; Zhu, X. VL-BERT. *ICLR* 2020.
29. Li, L.; Gan, Z. VisualBERT. 2019.
30. Chen, Y.C.; Li, L. UNITER. *ECCV* 2020.
31. Jia, C.; Yang, Y. ALIGN. *ICML* 2021.

 doi:10.20944/preprints202603.2191.v1

32. Yuan, L.; Chen, D. BEiT. *ICLR* 2022.
33. Bao, H.; Dong, L. BEiT-3. 2022.
34. Reed, S. Gato. 2022.
35. Brohan, A. RT-1. 2022.
36. Kirillov, A. Segment Anything. *ICCV* 2023.
37. Schuhmann, C. LAION-5B. 2022.
38. Colomina, I.; Molina, P. UAV photogrammetry. *ISPRS JPRS* 2014.
39. Zhu, X.X. Deep learning in remote sensing. *IEEE GRSM* 2017.
40. Demir, B. DeepGlobe. *CVPRW* 2018.
41. Ayush, K. SatMAE. 2022.
42. Wang, C. UAV remote sensing review. *Remote Sensing* 2022.
43. Yang, X. Vision-language for remote sensing. *Remote Sensing* 2023.
44. Zhu, X.X. Geospatial foundation models. *ISPRS JPRS* 2023.
45. Li, W. Remote sensing DL. *Remote Sensing* 2020.
46. Mou, L. Spectral attention. *TGRS* 2018.
47. Doherty, P. UAV search and rescue. *AI Magazine* 2016.
48. Zhou, Y.; Feng, L.; Ke, Y.; Jiang, X.; Yan, J.; Zhang, W. Towards Vision-Language Geo-Foundation Models: A Survey. *arXiv* 2024.
49. Mall, U.; Phoo, C.; Liu, M.; Vondrick, C.; Hariharan, B.; Bala, K. Remote Sensing Vision-Language Foundation Models without Annotations. *ICLR* 2024.
50. Ye, P.; Li, H.; Wang, Z. Multimodal Feature Alignment for Vision-Language Tracking. *Remote Sensing* 2024.
51. Samadzadegan, F.; Mehrdad, M.; Saeedi, P. Multi-sensor and multi-platform remote sensing data fusion: A review. *International Journal of Remote Sensing* 2025.
52. Zhou, G.; Li, Y.; Chen, S. Advances in Multimodal Foundation Models for Remote Sensing. *Remote Sensing* 2025.
53. Liu, H.; Wang, P.; Zhang, X. Prospects for Multimodal Foundation Models in Remote Sensing. *SPIE Proceedings* 2025.
54. Zhou, Y.; Chen, Z.; Xu, H. Multimodal reasoning for UAV-based remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 2025.
55. Li, Y.; Xu, W.; Li, G.; Yu, Z.; Wei, Z. UniRS: Unifying Multi-Temporal Remote Sensing Tasks via Vision-Language Models. *arXiv* 2024.
56. Liu, J.; Fu, R.; Sun, L.; Yang, X. SkyMoE: Vision-Language Foundation Model for Geospatial Interpretation. *arXiv* 2025.
57. Strong, B.; Smith, R.; Allen, T. Remote sensing foundation models for environmental monitoring. *Frontiers in Climate* 2025.