COMPARATIVE EVALUATION OF VISION-ENABLED LLMS FOR REMOTE SENSING IMAGE CAPTIONING

Robert-Ionuț Vătășoiu, Teodor Costăchioiu, Daniela Faur [1]

Accurate and descriptive captioning of remote sensing images is essential for downstream applications such as land use monitoring, disaster response, and environmental assessment, where timely and interpretable information is critical for decision-making. This paper explores the captioning capabilities of four lightweight, vision-language large language models on high-resolution remote sensing imagery imagery. Specifically, we evaluate Pixtral, Gemma3, LLaVA (7B and 13B) and MiniCPM-V against two well-established annotated datasets: RSICD and UCM captions. Adopting the BLEU, BERT Score, ROUGE, and CIDEr metrics, we compare model outputs to expert-written captions, revealing notable variation in vocabulary diversity, linguistic fluency, and object-counting accuracy. Furthermore, our observations show that the latest models do not invariably outperform preceding versions, underscoring the necessity for more specialized training approaches to address the unique challenges of remote sensing applications.

Keywords: Multi Modal Large Language Models, Captioning, Remote sensing image

1. Introduction

The volume and complexity of Earth Observation (EO) data have increased dramatically, driven by satellite missions such as Sentinel as well as the growing use of unmanned aerial vehicles (UAVs). These systems continuously provide geospatial information with varying resolutions and coverage, contributing to a rich yet challenging data landscape. Hence, there is a growing demand for automatic methods that can generate semantically rich textual descriptions. Remote Sensing Image Captioning (RSIC) addresses this need by bridging the gap between low-level visual content and high-level human-understandable language, thereby enabling more efficient interpretation, retrieval, and decision-making in EO tasks.

At the same time, vision enabled - Large Language Models (LLMs) have evolved rapidly, showing impressive capabilities not only in natural language

[1] Geospatial and Smart Sensors for Environmental Applications Laboratory (GEOSENSE), Campus Research Institute, National University of Science and Technology Politehnica Bucharest, Romania, e-mails: robert.vatasoiu@upb.ro, teodor.costachioiu@upb.ro, daniela.faur@upb.ro

processing but also in multi-modal understanding and knowledge integration. The application of LLMs in the context of EO data is an emerging field with tremendous potential [1], [2]. LLMs can be used to generate natural language descriptions of satellite/aerial observations [3], extract relevant information, or answer context-aware geospatial queries [35]. Moreover, multi-modal models that integrate remote sensing imagery and textual input are paving the way toward semantically enriched remote sensing, offering deeper insights into complex EO data based scenarios.

RSIC has emerged as a key task that bridges computer vision and natural language processing by generating human-readable descriptions of remote sensing imagery. Classical approaches to RSIC are based on encoder-decoder neural network models, combining convolutional, recurrent, and neural attention components. Initially, a convolutional encoder extracts visual feature representations from different regions of the input image. These features are then processed by a recurrent decoder, which generates the caption sequentially, word by word. At each decoding step, an attention mechanism dynamically assigns weights to the image regions based on their relevance to the current word prediction [4, 5, 6]. An encoder-decoder architecture multilevel and contextual attention network (MLCA-Net), is proposed in [7]. A contextual attention module is introduced by MLCA-Net to investigate the latent context concealed in remote sensing images, while a multilevel attention module is utilized to adaptively aggregate image attributes of particular spatial regions and scales. The purpose of the extra step added to a traditional encoder-decoder sequence is to supply supplementary data that could produce more precise descriptions.

The authors of [8] propose a novel joint-training two-stage RSIC method. They used multi-label classification to provide prior information, and designed a differentiable sampling operator to replace the traditional non-differentiable one for indexing classification results. Recent research [9] has explored the use of transformer-based encoder-decoder architectures for remote sensing image captioning. These models have demonstrated improved capability in capturing long-range dependencies and generating semantically coherent descriptions. The method in [10] introduces a retrieval-based topic memory network that integrates shared topic words from multiple reference captions to guide sentence generation, enhancing consistency and reducing ambiguity in remote sensing image captioning. These models, evaluated on datasets like UCM-Caption [11], RSICD [4], have demonstrated the feasibility of RS captioning, but suffered from limited semantic understanding and generalization.

2. **Motivation**

Recent papers in this field, particularly those published in recent years, indicate a notable surge in the application of vision enabled-LLMs to aerial image captioning. This growing interest reflects the advancements in large

multi-modal models. In the context of remote sensing, multi-modal vision-language model have started to be explored for tasks beyond captioning, including Visual Question Answering (VQA) and text-image retrieval [16].While domain-specific models such as RSGPT [12], Earth-GPT [13], GeoChat [14] or BLIP-2 [15] are specifically fine-tuned for geospatial and Earth observation tasks, their complexity, resource demands, and often limited public availability constrain their adoption in practical, lightweight deployments. Moreover, these models are frequently trained on proprietary datasets or with non-transparent pipelines, which hampers reproducibility and wider community benchmarking. In contrast, general-purpose vision-language models such as LLaVA [18] [17], MiniCPM-V [19], and Pixtral [34] are openly available, comparatively lightweight, and easier to adapt across domains via prompt engineering or retrieval-augmented generation. Their architectural modularity—combining open-source visual encoders (e.g., CLIP, ViT) with instruction-tuned LLM backbones like Gemma [22] [23] - makes them attractive candidates for low-resource or edge applications in remote sensing, where rapid deployment, minimal fine-tuning, and cross-task transferability are critical.

There are several compelling reasons to include and evaluate more general-purpose multi-modal vision-language models such as LLaVA [18] [17], MiniCPM-V [19], and Pixtral [34]. These models integrate visual encoders (e.g., ViT, CLIP) with instruction-tuned LLMs like Gemma [22] [23].

While general purpose models are open-source, widely available, and easy to integrate into custom pipelines, many geospatial LLMs are closed, restricted, or API-restricted, which may limit their reproducibility or customization. Moreover, general-purpose multi-modal models are trained on highly diverse and large-scale datasets, which equips them with a remarkable ability to generalize across domains. This cross-domain generalization makes them particularly effective in remote sensing scenarios that involve zero-shot or few-shot learning, where annotated data is limited or unavailable. Some of them offer flexible prompting and multi-modal input.

LLaVA, MiniCPM-V and Pixtral are inherently designed to process both images and text, which is critical when working with annotated remote sensing imagery, semantic segmentation maps, or time series visualizations that require textual interpretation. General-purpose models serve as a valuable baseline for assessing the impact of domain-specific fine-tuning, as seen in models like EarthGPT or GeoChat. Establishing this baseline is crucial for quantifying the performance gains brought by geospatial specialization and for understanding the true contribution of task-specific adaptations.

Lightweight models like MiniCPM-V and LLaVA can be fine-tuned on custom geospatial datasets, enabling tailored applications with lower computational costs than larger, rigid models. Given the rapid release of new LLMs, evaluating models from different periods helps track performance evolution. We therefore selected models released between 2023 and early 2024 to examine

how architectural advances and training scale influence EO data understanding. Exploring commercially available or resource-efficient models is especially relevant for real-world applications where computational limits and scalability matter.

In the context of VQA for aerial imagery, an interesting approach is Co-LLaVA, described in [27], which combines LLaVA-1.5 with a lightweight Contrastive Captioning model. While not a pure captioning paper, it demonstrates a strategy to reduce hallucinations and computational load by having a captioning module assist the LLM in understanding aerial images. This highlights a trend of collaborative models where a captioning sub-model grounds the LLM's descriptions to the image. Co-LLaVA's performance on remote sensing tasks suggests that pairing LLMs with specialized captioning models can enhance caption accuracy for aerial imagery.

The authors of [28] focus on applying existing LVLMs to drone image captioning and analyzing their outputs. The paper evaluates two state-of-the-art vision–LLMs – LLaVA and InstructBLIP – on a new drone data set-AeroCaps and the public VisDrone dataset. While these models can produce detailed aerial captions, the authors report significant hallucination issues. To facilitate research, they release a Labeled Illusions Dataset (LID) marking hallucinated portions of captions. They also show that GPT-4V can be used as an evaluator to quantify hallucination severity in aerial image captions.

This study is motivated by the need to understand which lightweight LLMs are best suited for specific EO data tasks, such as land use classification, environmental monitoring, anomaly detection, or climate-related event interpretation.

The main contributions of this work are summarized as follows:

- **A performance evaluation** of 7B–13B vision–language models, conducted specifically for high-resolution remote sensing image captioning.
- **A comparative analysis** of state-of-the-art V-LLMs, including MiniCPM-V, LLaVA (7B and 13B), Pixtral, and Gemma 3, with respect to caption generation performance.
- **An empirical evaluation** of computational requirements, assessing the feasibility of deploying V-LLMs on consumer-grade hardware in resource-constrained settings.
- **A domain-specific analysis** of model behavior, examining how V-LLMs handle the semantic and structural characteristics of satellite imagery.

3. Methodology

This paper aims to evaluate the performance of vision-enabled LLMs with a relatively low number of parameters, ranging from 7 to 13 billion, to generate descriptive captions for high-resolution remote sensing imagery. In particular, these models can run on affordable consumer hardware, broadening their accessibility for various practical applications. Specifically, our study

examines four models: MiniCPM-V [29], LLaVA [36] (in variants 7B and 13B), Pixtral 12B [34], and Gemma3 [22]. These models are used to generate natural language descriptions for high-resolution remote sensing images, leveraging their integrated visual and linguistic processing capabilities.

Our proposed methodology, detailed in Fig. 1, outlines a pipeline that begins with automatic caption generation, followed by a comparative analysis against expert annotations. This approach facilitates a systematic assessment of the models' abilities to capture salient visual features and articulate them in coherent, contextually relevant language.

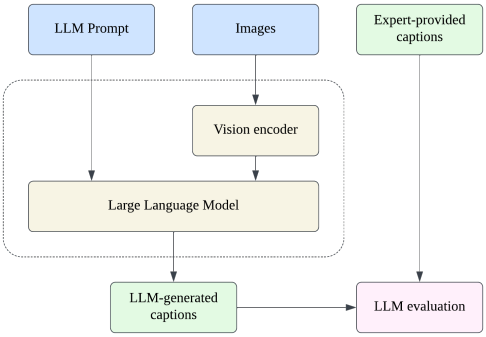


Fig. 1. Framework for LLM captioning evaluation.

To ensure consistency across models, we employed a standardized prompt for caption generation:

> *"You are an AI assistant tasked with captioning an image. Please provide your response strictly in clear, natural, human-readable plain text. Keep the answer informative, one sentence only. Aim to use 15 words."*

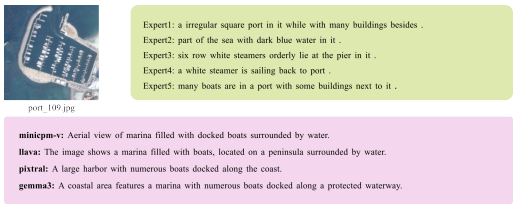
The prompt also included examples to guide the desired tone and structure, such as *"many planes are in an airport"*, *"there is a broad road next to the church"*, or *"several large buildings with parking lots are in a commercial area."*

3.1. **Datasets**

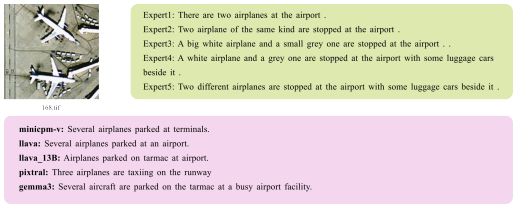
The experimental framework employs two rigorously annotated datasets, RSICD [4] and UCM Captions [11], to ensure a robust evaluation environment.

The RSICD dataset comprises 10921 images, across 30 categories, with an initial total of 24333 textual descriptions. To ensure that each image is paired with five captions, the dataset was extended to 54605 sentences by randomly duplicating existing descriptions for images that originally had fewer than five. An example of image from RSICD dataset can be seen in Fig. 2a.

Similarly, the UCM Captions dataset contains 2100 images organized into 21 classes, including aircraft, beaches, overpasses, and stadiums, with each image also accompanied by five sentences. In instances where fewer than five expert-provided sentences were available, some sentences were repeated to meet the five-sentence criterion. An example of images from UCM captions dataset can be seen in Fig. 2b.



(a) RSICD dataset: marina view



(b) UCM dataset: airport view

Fig. 2. Sample images from the (a) RSICD and (b) UCM datasets.

Compared to the UCM-Captions dataset, RSICD offers a broader coverage, with more categories and a greater number of images per category. However, a notable limitation of RSICD lies in its uneven annotation density. Many images have only one or two associated descriptions - for instance, 4853 out of 10921 images are annotated with just a single sentence. Of the total 54605 captions in the dataset, only 24333 are manually written, while the remaining 30272 were generated by randomly duplicating existing sentences to ensure that each image is linked to five captions.

3.2. Evaluation Metrics

To quantify performance, we use a suite of evaluation metrics commonly adopted in the image captioning domain. This includes: BLEU - Bilingual Evaluation Understudy [30] for assessing n-gram overlaps, BERT Score [31] for evaluating semantic similarity through contextual embeddings, and both ROUGE-1 and ROUGE-L Recall-Oriented Understudy for Gisting Evaluation [32] to measure lexical and sequential overlaps between generated and reference captions. Additionally, the CIDEr metric - Consensus-based Image Description Evaluation [33] captures consensus among multiple human-provided annotations, providing a nuanced measure of caption quality.

BLEU focuses on exact word matching, often missing semantic similarity or object accuracy in complex scenes. ROUGE-L measures recall via longest common subsequences but lacks sensitivity to fine-grained details like object identification. CIDEr evaluates consensus using TF-IDF, yet may fail to capture specific nuances of remote sensing object recognition

Following the quantitative analysis, we provide an in-depth discussion of the results, highlighting each model's strengths and limitations in handling

high-resolution remote sensing imagery and comparing their performance with human experts. These insights deepen our understanding of vision-enabled LLMs and point to future research directions in this emerging field.

4. **Results and discussion**

For all models evaluated, the decoding temperature was set to 0.2, a value empirically selected through preliminary trials. This low temperature ensures deterministic and focused caption generation, reducing randomness and enhancing consistency across outputs, which is particularly important for ensuring consistency and reliability in cross-model evaluation.

TABLE 1. Comparison of Vision-Language Models on the UCM Caption Dataset

Model	BLEU	BLEU-P	BLEU-R	BERT-P	BERT-R	BERT-F1	ROUGE1-F1	ROUGE2-F1	ROUGEL-F1	CIDEr
Pixtral-12B	**0.072**	0.3334	**0.3430**	0.8880	**0.8794**	**0.8836**	**0.2879**	**0.0712**	**0.2434**	**0.0328**
Gemma	0.0294	0.2149	0.3051	0.8690	0.8699	0.8694	0.1985	0.0247	0.1605	0.0093
LLaVA	0.0371	0.3218	0.3044	0.8875	0.8718	0.8793	0.2380	0.0532	0.2078	0.0211
LLaVA-13B	0.0426	**0.3521**	0.2933	**0.8941**	0.8753	0.8845	0.2580	0.0626	0.2278	0.0280
MiniCPM-V	0.0342	0.2783	0.2151	0.8912	0.8741	0.8825	0.1633	0.0404	0.1535	0.0182

The obtained results are discussed from three distinct perspectives: an analysis of individual model performance, an exploration of observed trends and patterns, and a linguistic comparison between expert and model-generated captions.

4.1. **Individual performance analysis of each model**

The results presented in Tables 1 and 2 represent the average scores achieved by each model across the respective datasets, and Fig. 3 shows a comparison for the relative performance between models for the two datasets.

TABLE 2. Comparison of Vision-Language Models on the RSICD Dataset

Model	BLEU	BLEU-P	BLEU-R	BERT-P	BERT-R	BERT-F1	ROUGE1-F1	ROUGE2-F1	ROUGEL-F1	CIDEr
Pixtral	**0.0413**	**0.2919**	0.3109	**0.8909**	**0.8788**	**0.8847**	**0.2517**	**0.0474**	**0.2019**	**0.0172**
Gemma	0.0275	0.2058	0.2970	0.8729	0.8759	0.8743	0.2019	0.0252	0.1574	0.0102
LLaVA	0.0261	0.1888	**0.3359**	0.8659	0.8717	0.8687	0.2150	0.0332	0.1702	0.0132
LLaVA-13B	0.0383	0.3707	0.2535	0.9006	0.8705	0.8851	0.2243	0.0460	0.1954	0.0160
MiniCPM-V	0.0259	0.1927	0.2780	0.8691	0.8742	0.8716	0.1794	0.0212	0.1462	0.0112

Pixtral-12B demonstrates the strongest overall performance across both datasets, consistently achieving high scores in both lexical (BLEU, CIDEr) and semantic metrics (BERT-F1). Particularly notable are the high BERT-F1 scores (0.8836 for UCM and 0.8847 for RSICD), reflecting excellent semantic

alignment between generated and reference descriptions. Pixtral-12B also excels in ROUGE metrics, highlighting its capacity for maintaining lexical and semantic coherence.

Gemma3 shows limited performance across both datasets, consistently achieving the lowest scores, especially in CIDEr and BLEU metrics. This suggests significant difficulties in generating detailed and coherent descriptions. However, moderate BERT-F1 scores (0.8694 for UCM and 0.8743 for RSICD) indicate a reasonable semantic capturing capability, though lacking in lexical precision.

LLaVA displays moderate performance with inconsistencies depending on the dataset analyzed. For the RSICD dataset, it achieves the highest result in BLEU-R (0.3359), suggesting a strong capability for general semantic context rendering, though CIDEr and ROUGE scores remain modest. On the UCM dataset, LLaVA struggles with generating precise lexical descriptions, showing a moderate but superficial approach.

LLaVA-13B, demonstrates significant lexical precision (BLEU-P: 0.3521, BERT-P: 0.8941), indicating strong vocabulary generation quality. However, moderate CIDEr and BLEU scores suggest difficulties in complex semantic descriptions, with a slight tendency toward superficiality. This behavior is consistent on the RSICD dataset, where LLaVA-13B maintains high precision-oriented metrics (BLEU-P: 0.3707, BERT-P: 0.9006) but shows limited gains in recall-sensitive and consensus-based scores such as CIDEr, reinforcing the observation that the model favors accurate lexical choices over rich, detailed semantic coverage.

MiniCPM-V achieves modest results across both datasets, performing poorly in most metrics analyzed. Although it achieves the highest BERT-R score (0.8742) for RSICD, this is insufficiently reflected in its overall performance, with the model generating descriptions lacking detailed lexical coherence and semantic depth.

In conclusion, Pixtral-12B emerges as the most robust and coherent model for both analyzed datasets, demonstrating both versatility and lexical-semantic accuracy. The other models reveal specific areas for improvement, reflecting either insufficient semantic capabilities or lack of coherent lexical details in generated descriptions.

4.2. **Observed trends and patterns**

The results reveal several key trends in vision-language model performance. Larger and more recently fine-tuned models, such as Pixtral-12B, show strong generalization and effectively preserve visual-semantic details, indicating that model scale and alignment quality drive both lexical accuracy and semantic fidelity.

In contrast, models like LLaVA and MiniCPM-V perform moderately well. Although their BERTScore values suggest good semantic alignment, their

lower CIDEr and ROUGE scores reveal weaker handling of complex phrasing and fine-grained details, pointing to limited lexical diversity and contextual precision.

LLaVA-13B and Gemma3 often generate generic, less informative captions, likely due to minimal fine-tuning for remote sensing data—yielding syntactically correct but semantically shallow outputs, as seen in Fig. 3.

Finally, several models display semantic bias: high BERTScore but low CIDEr and ROUGE values suggest that while captions may be broadly relevant, they often lack specificity and fail to fully capture the true visual content.

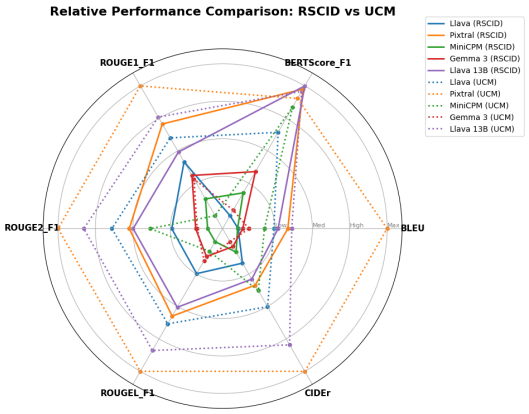


Fig. 3. Relative performance comparison of the models on the two datasets

4.3. Linguistic comparison between expert and model-generated captions

TABLE 3. Number of unique words in captions generated by each model

Remote sensing dataset	Expert annotation	MiniCPM-V	LLaVA	LLaVA-13B	Pixtral	Gemma3
UCM Captions	335	612	845	748	591	603
RSICD	2987	2096	2200	1053	792	799

In addition to the comparative analysis conducted using the aforementioned performance metrics, we also compared the captions provided by the tested LLMs with the annotations included in the datasets. This comparison enabled us to derive several additional observations.

The vocabulary analysis reveals notable differences between LLM-generated captions and expert annotations across both datasets (Table 3). On RSICD, expert captions use a broader vocabulary (2,987 unique words), while LLMs produce more compact ones. Pixtral-12B shows the narrowest range (792 words), suggesting repetitive, generic outputs, whereas LLaVA (2,200) and

MiniCPM-V (2,096) are more diverse but still below expert richness. Interestingly, on UCM, the trend reverses - LLMs exceed experts in vocabulary size: LLaVA (845), MiniCPM-V (612), and Gemma3 (603) versus 335 expert words. This may reflect over-generation or hallucination, adding diversity but reducing domain alignment.

We also noted frequent typographical and grammatical errors in expert annotations, indicating non-native English authorship. Additionally, LLMs tend to use American English, introducing inconsistencies such as 'soccer' vs. 'football' or 'plane' vs. 'airplane', which further lower metric-based scores.

Another notable shortcoming of LLM-generated captions is their inability to accurately count objects within an image. While human experts typically provide precise object counts, LLMs tend to use vague quantifiers such as "many" or "several", reflecting a lack of fine-grained visual grounding required for numerical estimation.

TABLE 4. Percentage of overlapping words - ground truth vs generated captions

Remote sensing dataset	**MiniCPM-V**	**LLaVA**	**LLaVA-13B**	**Pixtral**	**Gemma3**
UCM Captions	22.86%	37.14%	43.05%	42.46%	28.56%
RSICD	13.45%	19.31%	21.73%	26.86%	15.20%

Based on Table 4, the analysis of word overlap between ground truth and generated captions indicates that LLaVA-13B and Pixtral are the top-performing models. LLaVA-13B achieves the highest overlap on the UCM dataset at 43.05%, while Pixtral secures the best result on the RSICD dataset with 26.86%.

The results support the hypothesis that vision-language models vary significantly in the quality of their descriptions, depending on their architecture and fine-tuning. Pixtral clearly outperforms the other models across all major evaluation metrics, demonstrating the ability to generate relevant, coherent, and detailed descriptions. In contrast, the extremely low CIDEr scores of LLaVA-13B and Gemma3 suggest either a lack of effective adaptation to generative tasks or a tendency to produce generic outputs, making them unsuitable for applications that require descriptive precision and nuance.

Furthermore, our results indicate that some newer models, such as Gemma3, under-perform compared to earlier models like LLaVA and MiniCPM-V. This counterintuitive finding suggests that model recency does not necessarily correlate with performance in remote sensing captioning tasks. A potential explanation lies in the differences in pretraining datasets, which may lack sufficient coverage of geospatial or structured visual contexts, thereby limiting the models' ability to generalize effectively to remote sensing imagery.

Acknowledgment

This research was supported by PN-IV-P6-6.3-SOL-2024-2-0251, AI4RISK project - Platform for Fusion and Management of Multi-Source Data Collections Exploitable by Artificial Intelligence Models for Risk Assessment and Predictive Analysis.

REFERENCES

[1] *Y. Bazi et al.*, "RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery," Remote Sensing, vol. **16**, no. 9, p. 1477, 2024.

[2] *H. Lin et al.*, "RS-MoE: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering," IEEE Transactions on Geoscience and Remote Sensing, 2025.

[3] *Y. He and Q. Sun*, "Towards automatic satellite image captions generation using large language models," arXiv preprint arXiv:2310.11392, 2023.

[4] *X. Lu et al.*, "Exploring Models and Data for Remote Sensing Image Caption Generation," IEEE Trans. Geosci. Remote Sens., vol. **56**, no. 4, pp. 2183–2195, 2017.

[5] *S. Zhang et al.*, "Image captioning with visual-semantic aligning attention," Remote Sensing, vol. **9**, no. 6, p. 637, 2017.

[6] *J. Li et al.*, "A multilevel attention model for remote sensing image captioning," Remote Sensing, vol. **12**, no. 10, p. 1604, 2020.

[7] *Q. Cheng et al.*, "NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning," IEEE Trans. Geosci. Remote Sens., vol. **60**, pp. 1–19, 2022.

[8] *X. Ye et al.*, "A Joint-Training Two-Stage Method for Remote Sensing Image Captioning," IEEE Trans. Geosci. Remote Sens., vol. **60**, pp. 1–16, 2022.

[9] *H. Kandala et al.*, "Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning," IEEE Geosci. Remote Sens. Lett., vol. **19**, pp. 1–5, 2022.

[10] *B. Wang et al.*, "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. **13**, pp. 256–270, 2020.

[11] *H. Zhou et al.*, "Self-Learning for Few-Shot Remote Sensing Image Captioning," Remote Sensing, vol. **14**, p. 4606, 2022.

[12] *Y. Zhang et al.*, "RSGPT: A generative transformer model for RSIC," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. **16**, pp. 12345–12356, 2023.

[13] *W. Zhang et al.*, "EarthGPT: A Universal Multimodal LLM for Multisensor Image Comprehension," IEEE Trans. Geosci. Remote Sens., vol. **62**, pp. 1–20, 2024.

[14] *K. Kuckreja et al.*, "GeoChat: Grounded large vision-language model for remote sensing," arXiv preprint arXiv:2311.15826, 2023.

[15] *J. Li et al.*, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv preprint arXiv:2301.12597, 2023.

[16] *X. Li et al.*, "Vision-Language Models in Remote Sensing: Current Progress and Future Trends," IEEE Geosci. Remote Sens. Mag., vol. **12**, no. 2, pp. 32–50, 2024.

[17] *H. Liu et al.*, "Improved Baselines with Visual Instruction Tuning," arXiv preprint arXiv:2310.03744, 2023.

[18] *H. Liu et al.*, "Visual Instruction Tuning," arXiv preprint arXiv:2304.08485, 2023.

[19] *Y. Yao et al.*, "MiniCPM-V: A GPT-4V Level MLLM on Your Phone," arXiv preprint arXiv:2408.01800, 2024. [Online]. Available: https://arxiv.org/abs/2408.01800

[20] *Meta AI*, "Llama 3 Vision: Enabling On-device Multimodal AI at the Edge," 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[21] *H. Touvron et al.*, "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.

[22] *T. Mesnard et al.*, "Gemma: Open Models Based on Gemini Research and Technology," arXiv preprint arXiv:2403.08295, 2024.

[23] *Google DeepMind*, "Gemma: Google's open LLM family," [Online]. Available: https://ai.google.dev/gemma, 2024.

[24] *R. Ramos and B. Martins*, "Using neural encoder-decoder models with continuous outputs for remote sensing image captioning," IEEE Access, vol. **10**, pp. 24852–24863, 2022.

[25] *Z. Chen et al.*, "TypeFormer: Multiscale Transformer With Type Controller for Remote Sensing Image Caption," IEEE Geosci. Remote Sens. Lett., vol. **19**, pp. 1–5, 2022.

[26] *X. Ma et al.*, "Multiscale Methods for Optical Remote-Sensing Image Captioning," IEEE Geosci. Remote Sens. Lett., vol. **18**, no. 11, pp. 2001–2005, 2021.

[27] *F. Liu et al.*, "Co-LLaVA: Efficient remote sensing visual question answering via model collaboration," Remote Sensing, vol. **17**, no. 3, p. 466, 2025.

[28] *D. Basak et al.*, "Aerial Mirage: Unmasking Hallucinations in Large Vision Language Models," in Proc. WACV, pp. 5500–5508, Feb. 2025.

[29] *OpenCSG*, "MiniCPM-V: A lightweight and performant vision-language model," [Online]. Available: https://github.com/OpenCSG/MiniCPM-V, 2024.

[30] *K. Papineni et al.*, "Bleu: a Method for Automatic Evaluation of Machine Translation," in Proc. ACL, Philadelphia, PA, USA, 2002, pp. 311–318.

[31] *T. Zhang et al.*, "BERTScore: Evaluating Text Generation with BERT," arXiv preprint arXiv:1904.09675, 2020. [Online]. Available: https://arxiv.org/abs/1904.09675

[32] *C.-Y. Lin*, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out, Barcelona, Spain: ACL, 2004, pp. 74–81.

[33] *R. Vedantam, C. Zitnick, and D. Parikh*, "CIDEr: Consensus-based Image Description Evaluation," arXiv preprint arXiv:1411.5726, 2015. [Online]. Available: https://arxiv.org/abs/1411.5726

[34] *P. Agrawal et al.*, "Pixtral 12B," arXiv preprint arXiv:2410.07073, 2024.

[35] *J. D. Silva, J. Magalhães, D. Tuia, and B. Martins*, "Large language models for captioning and retrieving remote sensing images," arXiv preprint arXiv:2402.06475, 2024.

[36] *H. Liu, C. Li, Q. Wu, and Y. J. Lee*, "Visual instruction tuning," Advances in Neural Information Processing Systems, vol. **36**, pp. 34892–34916, 2023.