Segmentation-driven statistics-aware workflow for detailed scene description of UAV images using Mistral and LORA powered model

Bhargav Parulekar[1], Anandakumar M Ramiya[2]

[1]Indian Institute of Space Science and Technology, Thiruvananthapuram, India - bhargavparulekar11@gmail.com
[2]Indian Institute of Space Science and Technology, Thiruvananthapuram, India - ramiya@iist.ac.in

Keywords: Remote sensing, Scene description, captioning, LLM, segmentation.

Abstract

In the era of explainable AI, rapid data processing, analysis, and generation have become essential. Over the past few years, many approaches have been developed to process such heavy data and present it in an explainable manner, including in the field of remote sensing. One of such applications is remote sensing scene description. Many established workflows and models exist, but these models either fail to incorporate essential geospatial information or suffer from hallucination. We present a hybrid multimodal captioning methodology that tightly couples semantic segmentation outputs (via a LoRA-adapted Segment Anything Model) with a small, high-quality LLM- Mistral to produce descriptive, interpretable, and data-grounded scene captions. Rather than relying on direct image-to-text pipelines, our approach first extracts structured scene statistics (class proportions), spatial context (quadrant dominance and object localization), and color fingerprints (dominant colors per semantic class). These structured signals are converted into compact, factual prompts that the LLM consumes to generate coherent, informative, and verifiable captions. A comparison with the established Florence-2 model in terms of quantitative description demonstrates a significant improvement, with the Precision Vocabulary Index increasing from 0.077 to 0.232 due to the proposed workflow.

1. Introduction

The field of scene description has seen remarkable progress that can be credited to the success of large multimodal models (LMMs). These models are very good at understanding language and producing smooth, natural descriptions for many kinds of images. But this strength often comes with a trade-off: the text may sound fluent, yet not always be factually accurate. LMMs are trained to generate plausible-sounding responses. This can lead them to "hallucinate" objects that aren't actually in the image (Li et al., 2023; Rohrbach et al., 2018; Maynez et al., 2020) to mistake intricate spatial arrangements of objects, or to be unable to capture important quantitative details (Johnson et al., 2017; Hudson and Manning, 2019). Such issues are further exacerbated by weak alignment between visual tokens and language representations, which limits their ability to perform reliable reasoning over structured spatial information **(**Yuksekgonul et al., 2023; Liu et al., 2023**).** This can hinder them with restricted transparent reasoning (Sharma et al., 2024). This can hinder them with restricted transparent reasoning (Sharma et al., 2024).

Such limitations become even more challenging in the case of domain-specific fields like remote sensing, GIS, and urban analysis. In such areas, general-purpose models often fail to capture fine-grained details or reason accurately about spatial patterns **(**Tuia et al., 2011; Long et al., 2015; Cheng et al., 2017**).** Studies have reported notable performance gaps in agricultural remote sensing (Li et al., 2025) and in tasks involving unfamiliar or fast-changing urban environments (Han et al., 2025). Domain experts in these fields, such as scientists, urban planners, and first responders, are in dire need of outputs that are quantitative, spatially precise, and traceable to pixel-level evidence (Zhang et al., 2025**;** Reichstein et al., 2019; Camps-Valls et al., 2021). A generic caption like "an aerial view of a city with buildings and trees" is insufficient. Domain experts require verifiable statements such as: "The northeastern quadrant contains 4.5 hectares of new high-density residential development, replacing 3.1 hectares of former grassland."

Popular image-to-text models struggle to produce this level of detail because they cannot reliably count, measure, or spatially analyze image components in a verifiable manner (Gajbhiye et al., 2022). To address this gap between fluent but vague LMMs and precise yet unstructured segmentation data, we propose a "segmentation-first" hybrid architecture that explicitly separates visual perception from narrative generation.

Our approach combines three components: (1) a segmentation-first pipeline that produces class-indexed masks from either human labels or a LoRA-adapted Segment Anything Model (SAM), (2) extraction of statistics, spatial summaries, and color attributes, and (3) an LLM prompting layer that converts structured visual analytics into factual narratives. The segmentation front end builds on SAM, which offers a general, promptable backbone. Implementation of LoRA allows SAM to adapt to domain-specific textures and scales with minimal overhead, consistent with recent adaptations of SAM for remote sensing (Shan et al., 2025; Zhang et al., 2024). The resulting masks are uniform in structure with human annotations, which are fed into a shared analysis engine that computes class proportions, quadrant summaries, object counts, and dominant color clusters. These quantitative "segmentation ratios" serve as the factual basis for the LLM prompt, a technique shown to improve accuracy in grounded captioning (Zhang et al., 2025). A compact LLM such as Mistral then transforms these concise evidence-rich prompts into human-readable descriptions.

This design improves factual grounding and interpretability compared to direct image-to-text models, which rely on latent visual features without explicit numeric evidence. The two-stage pipeline segmentation followed by language generation is increasingly used for complex reasoning tasks and offers a more controlled route to trustworthy captions.

Earlier image captioning systems used encoder–decoder mechanisms, where CNN features were decoded using RNNs (Vinayls et al., 2015) or attention-based models (Xu et al., 2015). These established the foundation for neural captioning and the mapping from dense visual features to language tokens. Grounded captioning and template-based approaches such as

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-231-2026 | © Author(s) 2026. CC BY 4.0 License.

Neural Baby Talk (Lu et al., 2018), tied language generation to detected objects and regions, demonstrating how explicit visual anchors improve grounding. Similarly, scene-graph and relation-aware captioning incorporated structured representations of object relationships, although these methods still relied heavily on object detection and often showed mixed gains depending on training conditions. Our approach differs by focusing on LULC semantics, area statistics, and color summarization, and then delegating final language generation to an LLM under controlled prompting. SAM acts as our pixel-level backbone, with LoRA ensuring domain-specific adaptation for aerial imagery.Contemporary multimodal instruction-tuned models like the LLaVA family enable open-domain visual dialogue by connecting a vision encoder to an LLM through instruction tuning. While effective for general question answering and freeform descriptions, these models remain prone to hallucination and cannot reliably report numeric or spatial facts unless specifically trained for them. They map visual embeddings to language but do not provide token-level traceability back to pixels.

In contrast, our method begins with segmentation to obtain explicit symbolic representations (class IDs per pixel), applies quantification routines to compute verifiable statistics, and passes these structured descriptors to the LLM. By propagating the input of LLM in explicit symbolic and numerical evidence rather than relying solely on embeddings. This reduces incorrect statements and makes it possible to trace each generated claim back to a specific pixel mask or computation. This approach produces captions that are essential in practical settings like urban planning and environmental monitoring, where transparency and verifiability are crucial.

2. Methodology

The methodology involved assessing the performance of detailed caption generation using our proposed workflow of SAM with LORA guided segmented masks, along with Mistral, against the established vision-language model Florence-2 for the application targeted for land-use landcover analysis under the task of scene description. The complete methodological flowchart is depicted in Figure 1. The method utilised the usage of UAV dataset of the German city of Vaihingen, which has a spatial resolution of 9cm with 5 classes as impervious surfaces(road), trees, buildings, cars, and low vegetation (grass/shrubs).

2.1 Florence-2 for remote sensing scene description

Florence-2-Large is Microsoft’s 771M-parameter vision-language model built on a unified encoder–decoder architecture (Xiao et al., 2024). It processes an image using a ViT-style visual encoder, transforms visual tokens into a shared multimodal latent space, and then uses a text decoder to generate captions, OCR outputs, or region-aware descriptions. The model is trained on a large mixture of vision tasks (captioning, grounding, OCR, VQA), making it more generalizable and consistent than earlier models. For our application, we opt to proceed with application of More_Detailed_Caption.

2.2 Segmentation guided Remote Sensing scene description

2.2.1 Segmentation Backbone (Segment Anything Model) and LoRA Adaptation

SAM is an effective segmentation foundation model capable of zero-shot mask generation across diverse imageries. It provides flexible prompt interfaces and scales with dataset-centric mask generation strategies (Kirillov et al., 2023). However, SAM’s base behavior is not guaranteed optimal for high-altitude or domain-specific textures encountered in aerial imagery. To adapt SAM efficiently, we insert LoRA modules into selected transformer layers and fine-tune on a modest amount of domain data. LoRA allows task-specific adaptation with a tiny fraction of parameters and cost compared to full model fine-tuning, retaining base model weights while learning low-rank update matrices (Hu et al., 2022). This makes it practical to adapt SAM for domain variance without expensive retraining. Operationally, for each input image, we check for the presence of a human-created mask. If present, we skip SAM inference and treat the human mask as ground truth. If not, the SAM+LoRA module produces a mask (or set of masks) that are post-processed and mapped to a canonical class index scheme. This conditional logic conserves compute and retains high-fidelity human labels when available.

2.2.2 Aggregation of derived feature information

From the predicted segmentation masks, we compute three classes of features:

1) Statistical descriptors - class proportions (percent pixel area per class), scene density metrics (e.g., building cover fraction), and class-wise area estimates when geo-referencing metadata exists.

2) Spatial summaries - quadrant decomposition (top-left, top-right, bottom-left, bottom-right) with per-quadrant dominant class and relative percentages; connected-component counts for discrete object classes (buildings, vehicles) to estimate instance counts; and coarse centroids for cluster localization.

3) Features for chromatic analysis – Dominant color centroids were computed and then mapped to general color names to yield interpretable color descriptors (e.g., “rooftops are predominantly reddish-brown”).

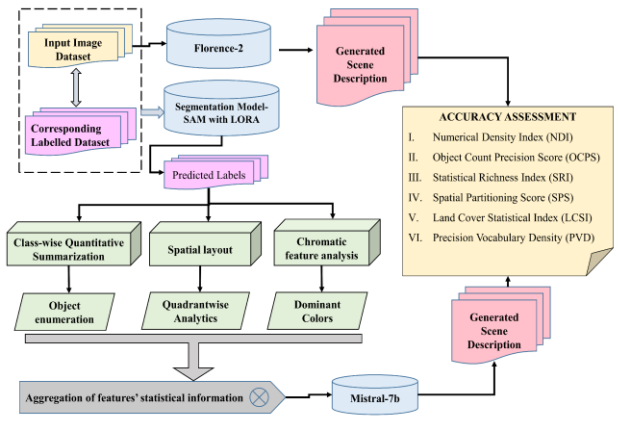

Figure 1. Methodology implemented for the workflow

2.2.3 Propagating information thorugh Mistral

Mistral 7B is a compact yet capable large language model created by Mistral AI (Jiang et al., 2023). It packs roughly 7 billion parameters, which makes it small enough to run on consumer GPUs while still producing strong, coherent outputs. The model uses modern architectural choices like grouped-query attention and an optimized sliding-window approach to keep inference fast and memory-efficient. Because of this balance of speed, accuracy, and openness, it’s become a popular foundation for research and real-world applications.

Rather than sending raw image embeddings, the LLM is given a templated prompt that lists the structured evidence. The prompt contains three concise sections: (A) one-line statistical summary (e.g., “Buildings: 42.8%; Trees: 23.4%; Impervious: 17.9%”), (B) spatial notes (e.g., “Top-right quadrant dominated by buildings, 62% of that quadrant”), and (C) color cues (e.g., “Building rooftops: reddish-brown; Vehicles: white/silver”). The LLM is asked to produce a readable caption of length X that: (i) is factually grounded in the supplied numbers, (ii) avoids inference beyond the supplied evidence, and (iii) optionally suggests a 1-sentence implication for planners (e.g., “high building density suggests urban core”).

Using Mistral provide compact and high-quality choice as the generator offers a cost-effective, controllable language backend with strong English generation quality. As the prompt is explicit and numeric, the LLM’s generative space is constrained. This both improves factuality and reduces hallucination risk compared to LLaVA-style freeform decoders.

2.3 Accuracy metrics and performance evaluation:

Outputs obtained from both the methods were then assessed with following metrics:

1) Numerical Detail Index (NDI): NDI measures how well the caption mentions numerical information, such as counts, quantities, or measurements.

2) Object Count Precision Score (OCPS): measures how accurately and precisely the caption counts objects in the image. High OCPS means the caption correctly identifies the number of key items rather than over- or under-estimating.

OCPS = Count(instances of the pattern[Number] + [Word])

3) Spatial Partitioning Score (SPS) evaluates how well the caption describes spatial arrangement or structure, such as positions, alignments, patterns, or layout of objects

SPS = Count (spatial keywords like "top-left", "bottom-right", "quadrant")

4) Precision Vocabulary Density (PVD): This is a crucial metric for measuring efficiency. A high PVD score indicates the caption is dense with technical terms rather than descriptive filler (for example, phrases like "fiery red leaves," "calm European town").

PVD = Total Count of Technical Terms/Total Number of Words in Caption

5) Land Cover Statistical Index (LCSI): Rather than checking if "trees" are mentioned or if a number is present (like NDI), it specifically checks if a land cover class is associated with a number.

6) Statistical Richness Index (SRI): This is essential for establishing the tone of the caption. The presence of words like "primarily," "constitute," and "distributed" signals an analytical intent.

These features are computed and saved for each scene. Since every number and label is derived from pixel indices, an analyst can always verify claims by inspecting the mask. This transparency becomes crucial for scientific use and regulatory reporting.

3. Results

The result can be visualized in the figure.

3.1 Florence Model

3.1.1 Visual assessment of descriptions:

In the first approach, where input images are fed to Florence-2, depict a detailed description involving land cover classes such as building, roads, trees, cars, and low vegetation as identified as shrubs. The model captures the pattern of arrangement of buildings. However, this model interprets the image to be taken during the fall season, attributing its assumption to trees having red leaves. Here it is creating its own thinking since it lacks the knowledge of an image being a false color composition. Such an implication may be misleading in the applications of remote sensing. The overall tones are described at image levels. The visual results are depicted in Figure 2 and Figure 4 for 2 different regions.

3.1.2 Quantitative Evaluation:

The accuracy metrics evaluation, which had a statistical emphasis, helped understand the behaviour of the model in producing such information from raw imagery. Average values obtained from various metrics are tabulated in the Table. The model was unable to capture details like the numerical density index, OCPS, and SPS, indicating its inability to produce numeric information and quadrant-wise distribution of objects in the scene description. Other parameters, like the statistical richness index and the precision vocabulary density, turned out to be 1 and 0.77, respectively.

3.2 Proposed Model (SAM with LORA + Mistral):

3.2.1 Visual assessment of descriptions:

In this approach, the input relies on the segmented output of the model, which, in our case, produced impressive results after the controlled feed to Mistral-7b. As seen in the figure. The scene generation incorporates quantitative composition of land cover classes, providing an estimate of the overall area occupied by each class in percentages. A detailed quadrant-aware analysis is also available. The chromatic analysis, unlike in the previous case, is extended to the class level, providing finer detailing of objects present in the class. Objects enumeration is described for cars and buildings. The model sticks to the quantitative information and eliminates the cases of hallucination or false assumptions. The visual results are depicted in Figure 3 and Figure 5 for the same regions tested with earlier approach.

3.2.2 Quantitative Evaluation:

As seen in the Table 1, we can see the surge in the metrics tracking numerical count of objects such as numerical density index, spatial partitioning score, and land cover statistical index with an average scores of 14, 10, and 9 respectively. Statistical richness index and precision vocabulary index found it to be 11 and 0.232 respectively.

Averaged Metrics	Florence-2	Ours
Numerical Density Index	0	14
Object Count Precision Score	0	1
Statistical Richness Index	1.182	10
Spatial Partitioning Score	0	10
Land Cover Statistical Index	0	8
Precision Vocabulary Density	0.064	0.237

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-231-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 1. Average Performance metrics comparison between Florence-2 and our proposed method

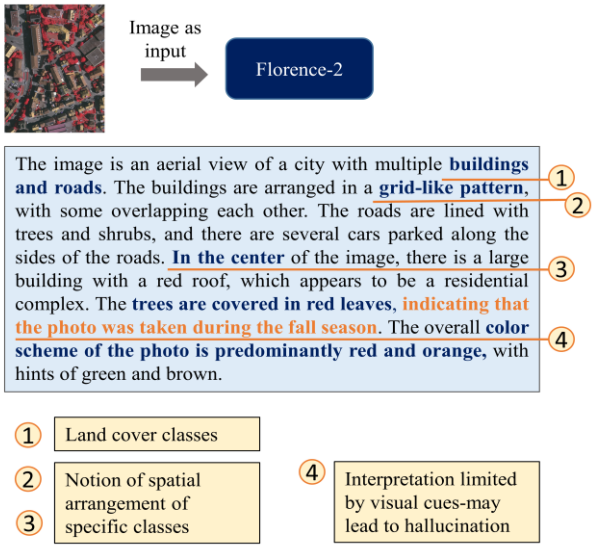

Figure 2. Performance of Florence-2 for remote sensing scene description- Region 1

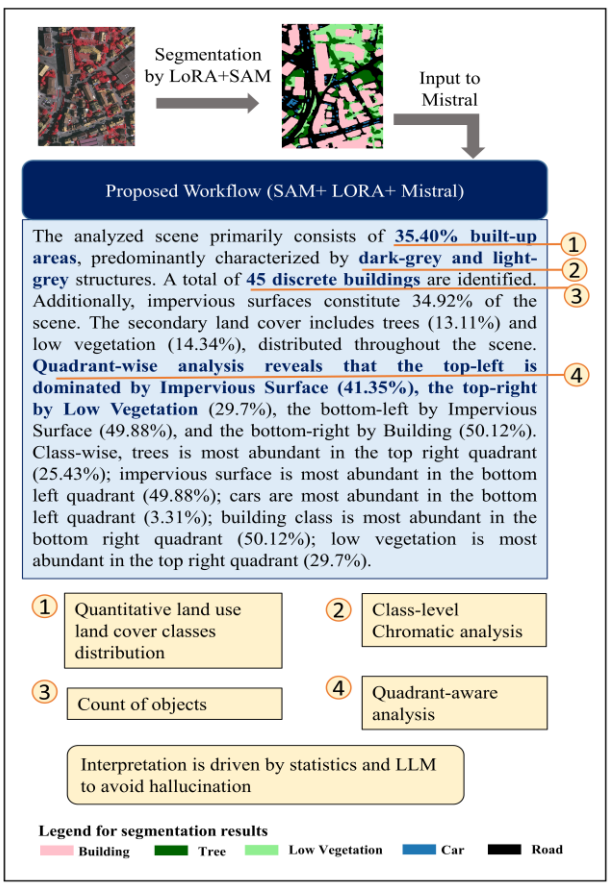

Figure 3. Performance of proposed workflow of segmentation guided Mistral for scene description- Region 1

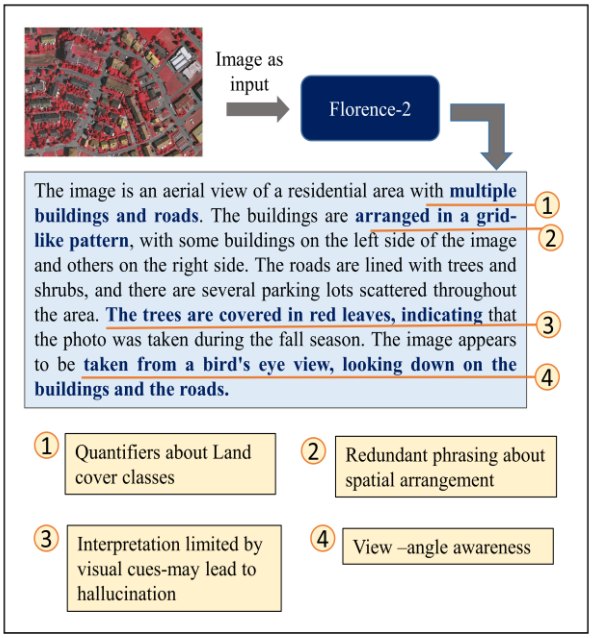

Figure 4. Performance of Florence-2 for remote sensing scene description- Region 2

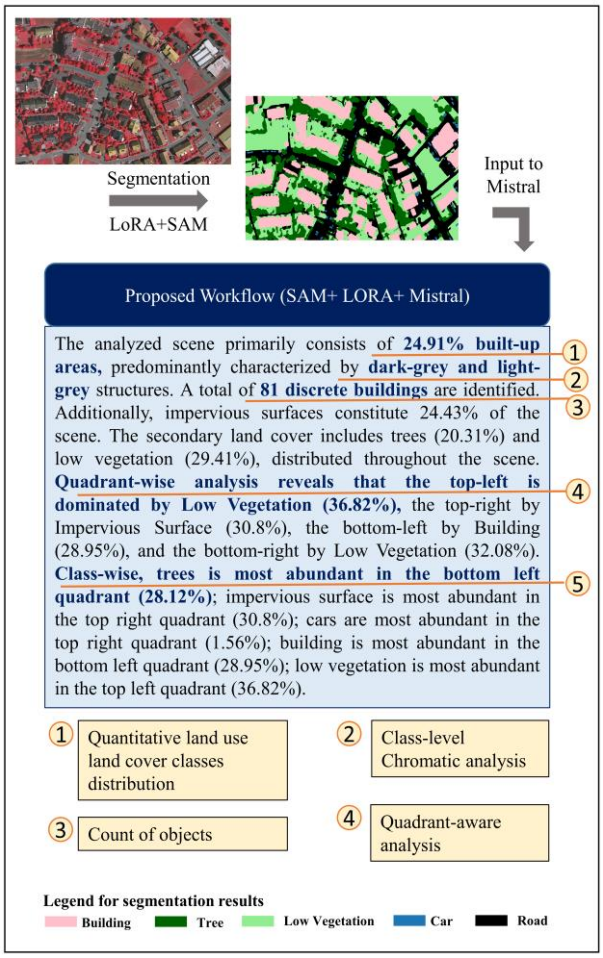

Figure 5. Performance of Florence-2 for remote sensing scene description- Region 2

4. Discussions

The fusion of SAM-LORA with Mistral provided interesting results which extended conventional basic scene description into quantitative results. This process provided following major advantages:
1) Segmentation guided controlled results on the input images of False color compositions
2) Possibility of hallucinations is drastically reduced. This has happened due to following reasons.
i) Symbolic anchoring: Every declarative clause in the caption (e.g., “buildings occupy 42.8% of the image”) is directly traceable to a numeric field in the prompt; the LLM is constrained to rephrase rather than invent.
ii) Limited inference scope: The templated prompt restricts the LLM to avoid assumptions outside the supplied evidence. Because the LLM needs only to perform linguistic surface realization and not latent visual inference, the likelihood of speculative assertions falls.
In contrast, end-to-end visual LLMs map latent multimodal features to language and often learn statistical priors that can dominate when visual cues are ambiguous. By separating perception (segmentation and analytics) from language generation, we achieve a clearer separation of responsibilities and better verifiability.

The following challenges can be faced by the system.
1) Segmentation errors propagation: If the chosen segmentation model miss-segments a class, the caption inherits the error. A provision of reporting uncertainty by various means such as identifying and marking low confidence areas may require human review for critical decisions.
2) LLM outputs is highly sensitive on prompt wording. In our case, we mitigate this with disciplined templates and few-shot examples that bias the LLM toward fact-anchored rephrasing.
3) LoRA adaptation requires a modest domain corpus which can be challenging during some instances. During such unavailability, SAM base masks may be less accurate. In such cases, the methods of active learning can be implemented to curate a small, high-quality fine-tuning set.

5. Conclusion

By building a segmentation-centric front end and transforming the resulting symbolic analytics into controlled LLM prompts, we create captions that are both human-readable and machine-verifiable. This hybrid paradigm significantly narrows the gap between freeform visual language models (which can be eloquent but ungrounded) and strictly analytic reports (which are precise but dry), delivering the best of both worlds for geospatial analysis and automated reporting.

6. References

Camps-Valls, G., Tuia, D., Zhu, X. X., & Reichstein, M. (Eds.). (2021). *Deep learning for the Earth sciences: A comprehensive approach to remote sensing, climate science and geosciences*. John Wiley & Sons.

Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE, 105*(10), 1865–1883.

Gajbhiye, G. O., & Nandedkar, A. V. (2022). Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach. Engineering applications of artificial intelligence, 114, 105076.

Han, J., Ning, Y., Yuan, Z., Ni, H., Liu, F., Lyu, T., & Liu, H. (2025). Large Language Model Powered Intelligent Urban Agents: Concepts, Capabilities, and Applications. arXiv preprint arXiv:2507.00914.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models. In: Proc. Int. Conf. Learn. Represent. (ICLR).

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6700–6709).

Jiang, D., Liu, Y., Liu, S., Zhao, J. E., Zhang, H., Gao, Z., Xiong, H., 2023. From clip to dino: Visual encoders shout in multi-modal large language models. arXiv preprint arXiv:2310.08825.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2901–2910).

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Girshick, R., 2023. Segment anything. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 4015–4026.

Li, Q., Zhang, Y., Mai, Z., Chen, Y., Lou, S., Huang, H., ... & Zheng, J. (2025). Can Large Multimodal Models Understand Agricultural Scenes? Benchmarking with AgroMind. arXiv preprint arXiv:2505.12207.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

Lu, J., Yang, J., Batra, D., Parikh, D., 2018. Neural baby talk. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 7219–7228.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 1906–1919).

Osco, L. P., Wu, Q., De Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., & Junior, J. M. (2023). The segment anything model (sam) for remote sensing applications: From zero to one shot. International Journal of Applied Earth Observation and Geoinformation, 124, 103540.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195–204.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4035–4045).

Shan, Z., Liu, Y., Zhou, L., Yan, C., Wang, H., & Xie, X. (2025). Ros-sam: High-quality interactive segmentation for remote sensing moving object. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 3625-3635).

Sharma, H., & Padha, D. (2024). Domain-specific image captioning: a comprehensive review. International Journal of Multimedia Information Retrieval, 13(2), 20.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., & Muñoz-Marí, J. (2011). A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing, 5*(3), 606–617.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 3156–3164.

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Huang, X., Zhu, X., Yuan, L., 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 4818–4829.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. Int. Conf. Mach. Learn. (ICML), 2048–2057.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it. arXiv preprint arXiv:2210.01936.

Zhang, E., Liu, J., Cao, A., Sun, Z., Zhang, H., Wang, H., & Song, M. (2024). RS-SAM: integrating multi-scale information for enhanced remote sensing image segmentation. In Proceedings of the Asian Conference on Computer Vision (pp. 994-1010).

Zhang, Y., Shen, G., Ning, K., Ren, T., Qiu, X., Wang, M., & Kong, X. (2025). Improving Region Representation Learning from Urban Imagery with Noisy Long-Caption Supervision. arXiv preprint arXiv:2511.07062.

Zhang, Z., Wu, H., Jia, Z., Lin, W., & Zhai, G. (2025). Teaching lmms for image quality scoring and interpreting. arXiv preprint arXiv:2503.09197.