Efficient Fine-Tuning and Explainability in Vision-Language Models for Image Captioning

Introduction

Vision-Language Models (VLMs) constitute a critical class of multimodal architectures that integrate visual and textual information through joint transformer-based representation learning. These models have demonstrated remarkable success across a range of tasks including image captioning, visual question answering, and multimodal retrieval. With 3 billion parameters, PaliGemma merges the SigLIP-So400m vision encoder and the Gemma-2B decoder to support instruction-tuned multimodal learning at scale. Designed for transferability, it generalizes effectively across more than 40 benchmarks, encompassing both general-purpose and specialized domains such as remote sensing (Beyer et al., 2024). Its architecture exemplifies the growing trend toward open, efficient VLMs optimized for downstream fine-tuning and zero-shot generalization.

PaliGemma builds on two theoretical contributions. First, SigLIP introduces a pairwise sigmoid loss that eliminates the need for global similarity normalization, enabling efficient training at extremely large batch sizes without sacrificing alignment quality (Zhai et al., 2023). Second, Gemma leverages an encoder-decoder architecture adapted from decoder-only LLMs, delivering a superior quality-efficiency trade-off. This design enhances representational depth and improves instruction-following ability while preserving inference-time efficiency (Zhang et al., 2025). Together, these components advance the state of scalable, adaptable, and interpretable VLMs.

Despite these advances, a key research gap persists: the lack of systematic integration between parameter-efficient fine-tuning (PEFT) methods and post-hoc explainability frameworks in vision-language modeling—particularly in high-stakes domains such as remote sensing, where transparency and resource constraints are critical. This study addresses that gap by asking: *How can the integration of advanced PEFT strategies and explainability techniques improve the performance, efficiency, and interpretability of the PaliGemma model in remote sensing image captioning?*

To answer this question, the study wil set out three interrelated objectives. First, it will empirically evaluate the effectiveness of advanced PEFT techniques for adapting PaliGemma to the remote sensing domain while minimizing fine-tuning overhead. Second, it will implement and compare multiple explainability methods to interpret and validate the model's decision-making behavior. Finally, the study will assess the combined impact of these fine-tuning and explainability strategies on key performance indicators within practical multimodal captioning scenarios. Through this integration, the research aims to contribute a reproducible and theoretically grounded framework for advancing scalable, transparent, and domain-adapted VLMs.

Literature Review

This section reviews recent theoretical developments in PEFT and explainable artificial intelligence (XAI), which form the methodological foundation of this study. PEFT has become increasingly important as large pre-trained models are adapted to new tasks with minimal computational overhead. One of the most widely adopted strategies is Low-Rank Adaptation (LoRA), which injects trainable low-rank matrices into frozen attention layers to enable efficient weight adaptation (Hu et al., 2022). Prompt tuning offers an alternative by learning soft task-specific embeddings that steer model behavior without modifying the original weights (Lester et al., 2021). In addition, adapter tuning has emerged as a modular and highly extensible PEFT strategy. Adapter modules are lightweight bottlenecks inserted between transformer layers, allowing models to be incrementally adapted to new tasks while keeping the backbone frozen (Houlsby et al., 2019). Adapters achieve performance near full fine-tuning with orders of magnitude fewer trainable parameters, making them especially attractive for large-scale or multi-task vision-language pipelines.

Complementary to these approaches, XAI techniques offer tools for interpreting the decision-making behavior of complex models. Integrated Gradients (Sundararajan et al., 2017) provide a theoretically grounded method based on path integrals, computing the contribution of each input by comparing gradients along a straight-line path from a baseline to the input. SHAP (Lundberg & Lee, 2017) builds on game-theoretic principles, offering additive feature attributions with consistency and local accuracy guarantees. Attention-based methods, while often used qualitatively, are enhanced with visualization tools such as Grad-CAM (Selvaraju et al., 2017) for highlighting the salient regions of an image contributing to the model's prediction.

Exploratory Data Analysis

The dataset employed in this study comprises 44,521 remote sensing images, each annotated with five human-written captions, resulting in a total of 222,605 caption instances. The dataset has been stratified into training (80%), validation (10%), and test (10%) subsets, supporting standard supervised learning procedures.

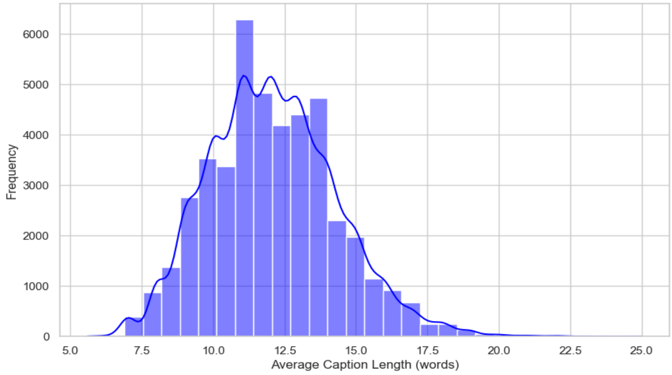

Figure 1- Distribution of Average Caption Lenght

Descriptive analysis reveals that the average caption length is approximately 12.08 words, with a standard deviation of 2.23. The minimum caption length is 5.6 words, and the maximum extends to 25 words. A histogram of average caption lengths (Figure 1) confirms a right-skewed distribution, indicating that the majority of captions fall within a concise and consistent length range—an attribute beneficial for model generalization.

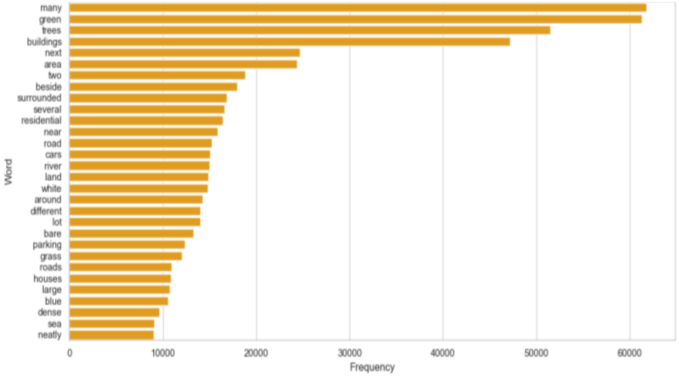

Figure 2 - Top 30 Most Frequent Words in Captions (Stopwords Removed)

A frequency analysis of caption tokens, after removing common English stopwords, highlights domain-relevant keywords such as "buildings", "green", "trees", "area", and "road". These observations suggest that the dataset is rich in spatial and compositional descriptors, offering strong alignment with real-world remote sensing applications. The top 30 most frequent words were visualized in a bar chart (Figure 2) to understand the distributional focus of linguistic elements.

References

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., ... Zhai, X. (2024). *PaliGemma: A versatile 3B VLM for transfer*. arXiv. https://arxiv.org/abs/2407.07726

Houlsby, N., Giurgiu, A., Jastrzębski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). *Parameter-Efficient Transfer Learning for NLP*. In International Conference on Machine Learning (ICML). https://arxiv.org/abs/1902.00751

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2022). *LoRA: Low-rank adaptation of large language models*. arXiv. https://arxiv.org/abs/2106.09685

Lester, B., Al-Rfou, R., & Constant, N. (2021). *The power of scale for parameter-efficient prompt tuning*. arXiv. https://arxiv.org/abs/2104.08691

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://arxiv.org/abs/1705.07874

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319–3328). https://arxiv.org/abs/1703.01365

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). *Sigmoid loss for language-image pre-training*. arXiv. https://doi.org/10.48550/arXiv.2303.15343

Zhang, B., Moiseev, F., Ainslie, J., Suganthan, P., Ma, M., Bhupatiraju, S., Lebron, F., Firat, O., Joulin, A., & Dong, Z. (2025). *Encoder–decoder Gemma: Improving the quality–efficiency trade-off via adaptation*. arXiv. https://arxiv.org/abs/2504.06225