Efficient Fine-Tuning and Explainability in VLM for Image Captioning

Esra Şekerci
Department of Information Systems
Middle East Technical University
Ankara, Turkey
esra.sekerci@metu.edu.tr

***Abstract*—This study explores the integration of parameter-efficient fine-tuning (PEFT) strategies and post-hoc explainability techniques in vision-language models for remote sensing image captioning. Building on the PaliGemma foundation we apply and compare three advanced PEFT techniques: Low-Rank Adaptation (LoRA), Prompt Tuning, and Adapter Tuning. The model is trained on the Remote Image Sensing and Captioning (RISC) dataset, comprising over 220,000 annotated captions. Initial benchmarking evaluates captioning quality using BLEU and METEOR scores, while ablation studies investigate the trade-offs between performance and fine-tuning overhead. Complementary to model adaptation, explainability methods such as Integrated Gradients, SHAP, and Grad-CAM are implemented to assess transparency and decision traceability.**

Keywords—Vision-language models, parameter-efficient fine-tuning, PaliGemma, remote sensing, explainable AI.

I. 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 [1]. 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 [2]. 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 [3]. 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 will 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.

II. 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 [4]. Prompt tuning offers an alternative by learning soft task-specific embeddings that steer model behavior without modifying the original weights [5]. 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 [6]. 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 [7] 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 [8] builds on game-theoretic principles, offering additive feature attributions with consistency and local accuracy guarantees; its recent extension, PixelSHAP, applies this framework to vision-language models by perturbing segmented image regions, enabling structured and model-agnostic visual explanations [9]. Attention-based methods, while often used qualitatively, are enhanced with visualization

tools such as Grad-CAM [10] for highlighting the salient regions of an image contributing to the model's prediction.

III. Dataset

A. 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.

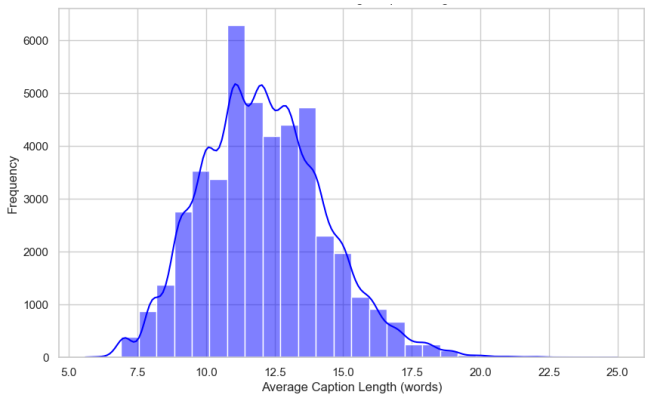

Figure 1 - Distribution of Average Caption Length

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.

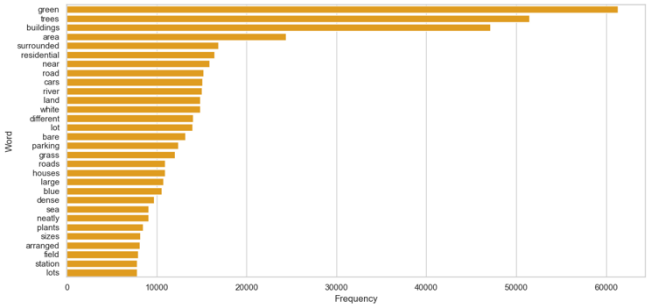

Figure 2 - Top 30 Most Frequent Words in Captions (Stopwords Removed)

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.

B. Data Splits and Prepocessing

To support reproducible experimentation, we follow an 80/10/10 split of the Remote Image Sensing and Captioning (RISC) dataset at the image-level: 80% of images for training, 10% for validation (used exclusively for prompt and hyperparameter tuning), and the remaining 10% held out for final test evaluation. We enforce strict non-overlap between these subsets to prevent any leakage of visual content or captions.

Every raw satellite image is uniformly resized to 224 × 224 pixels and channel-normalized using the ImageNet statistics (mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5]). Captions undergo canonical cleaning by lowercasing, trimming leading/trailing whitespace, and collapsing multiple spaces. We tokenize on whitespace and punctuation, then truncate or pad each token sequence to a fixed maximum length to bound both memory footprint and generation latency. This preprocessing pipeline ensures that all inputs to PaliGemma share a consistent spatial resolution and textual length, which is critical for stable training and fair comparison across prompt-tuning configurations.

IV. Methodology

A. Evaluation Metrics and Benchmarking

To provide a comprehensive and multidimensional assessment of model performance, we employ a suite of quantitative evaluation metrics commonly used in the image captioning and vision-language literature. BLEU [11] is used to evaluate local n-gram overlap between generated and reference captions, capturing syntactic fluency and phrase-level accuracy. However, as BLEU may penalize legitimate paraphrasing, we complement it with METEOR [12], which incorporates synonymy, stemming, and alignment-based matching to provide a more semantically aware evaluation. To further capture deep contextual similarity, we report BERTScore F1 [13], which leverages contextual embeddings from large pre-trained language models, offering robustness to lexical and syntactic variation and improved correlation with human judgment.

For model performance contextualization, we will also cevaluate results on established benchmarks and transfer tasks, as recommended by the PaliGemma model documentation. Benchmarking is crucial for evaluating the generalization and transferability of vision-language models to diverse tasks and datasets beyond the training distribution. The PaliGemma foundation model is evaluated across a variety of academic tasks—such as COCO Captioning, NoCaps, VQAv2, TextVQA, and others—using metrics like CIDEr, BLEU, and accuracy, ensuring a fair and transparent comparison of capabilities with the state-of-the-art. Importantly, none of these benchmarks are included in the pre-training corpus, preventing data leakage and enabling a genuine test of transfer performance. This benchmarking approach is essential for quantifying both the “out-of-the-box” (zero-shot) capabilities and the improvements attained through fine-tuning or parameter-efficient adaptation, providing a reliable reference for interpreting model performance in specialized domains such as remote sensing image captioning.

B. Baseline Model

The baseline for this study is established using the pre-trained PaliGemma vision-language model, which is evaluated on the remote sensing image captioning task

without any task-specific adaptation or parameter-efficient tuning. Since each image in the dataset is paired with five human-authored captions, we first experimented with a semantic similarity approach using CLIP embeddings to identify, for each image, the caption most closely aligned with its visual content. However, qualitative analysis of these selections revealed that this method did not consistently yield the most informative or contextually rich captions.To address this, we systematically reviewed alternative strategies and empirically observed that the longest available caption for each image typically provided the most detailed and comprehensive scene description. Based on this observation, we adopted the longest caption selection strategy for both training and evaluation, ensuring greater consistency and informativeness in the target outputs.

Additionally, we systematically investigated the influence of prompt design and token length settings on model performance—a critical but often underappreciated aspect of vision-language modeling. Drawing on theories from instruction tuning and controlled generation in large language models, we experimented with several prompt formulations (e.g., "<image> Describe the scene," "<image> Explain the layout of the terrain," "<image> Provide a detailed summary," etc.) to determine which most effectively elicited descriptive and relevant captions in the remote sensing context. Similarly, we performed ablation studies varying the maximum token length during generation. This allowed us to analyze the trade-off between brevity and descriptiveness, as longer sequences theoretically enable the capture of more detailed spatial and semantic relationships but may also increase the risk of verbosity or off-topic content.

Ultimately, the baseline model was evaluated using the prompt and token size configuration empirically shown to maximize output informativeness and metric scores on the validation set. Training was conducted using cross-entropy loss and the AdamW optimizer, with early stopping based on validation performance to prevent overfitting. Evaluation employed established image captioning metrics to capture various aspects of output quality, from n-gram precision to semantic consensus. Baseline results demonstrated that, while the pre-trained model produced fluent and structurally correct captions, it frequently lacked the specificity and granularity required for domain-focused remote sensing applications. This systematic baseline assessment highlighted the limitations of direct application of large pre-trained models and underscored the necessity for parameter-efficient adaptation and domain-aligned fine-tuning, which are explored in subsequent stages of this research.

Figure 3 presents a qualitative example from the test set, illustrating the model's generative output alongside the corresponding ground-truth reference caption. The generated caption successfully identifies the presence of aircraft and provides a general description of their location on the runway, but it is less specific than the reference, which details both the number and relative position of the planes. This example typifies the baseline model's behavior: it can capture the main scene elements but often lacks the fine-grained spatial detail required for expert-level remote sensing analysis.

Figure 3 - Test Set Example

Top: Input image.
Generated caption: "The aircraft are parked on the runway."
Reference caption: "There is a plane on the runway and three planes in the open space next to the runway."

Table 1 summarizes the quantitative evaluation of the baseline PaliGemma model on the test set. Using the empirically selected prompt and token length configuration, the model achieves a BLEU score of 0.0837, METEOR of 0.127, and BERTScore F1 of 0.122. These results, consistent with prior studies in zero-shot captioning, reflect the challenge of adapting general-purpose vision-language models to specialized domains without further fine-tuning. The modest metric scores underscore the necessity of parameter-efficient adaptation to bridge the gap between pre-trained representations and the domain-specific requirements of remote sensing captioning.

prompt	max_ tokens	BLEU	METEOR	BERTScore_ F1
<image> explain the layout of the terrain and infrastructur.	30	0.0037	0.1269	0.1224

Table 1 - Baseline Model Performance on the Test Set.

C. Parameter Efficient Fine Tuning

Our parameter-efficient fine-tuning experiments aim to adapt the PaliGemma vision-language model for remote sensing captioning while adhering to strict computational constraints. By sampling a representative subset of the training data, we enable rapid and iterative experimentation that remains theoretically valid for comparative evaluation. LoRA modules are integrated into the language generation layers, with all visual and multimodal encoders frozen, reflecting the principle that adaptation is most critical in the generative head for domain-specific transfer.

The test set remains strictly reserved for final evaluation, ensuring an unbiased measure of generalization. This workflow demonstrates the principled application of parameter-efficient adaptation techniques to large vision-language models under real-world resource constraints. Ongoing work will further scale these experiments to the full dataset in subsequent phases of the project.

prompt	max_ tokens	BLEU	METEOR	BERTScore_ F1
<image> explain the layout of the terrain and infrastructur.	30	0.0053	0.1387	0.1322

Table 2 - Performance of the LoRA-adapted Model (subset data)

Results are based on a debug-sized subset using LoRA and 4-bit quantization. Final evaluation will be conducted on the full dataset in future work.

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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

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

[6] 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

[7] 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

[8] 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

[9] Goldshmidt, R. (2025). *Attention, Please! PixelSHAP Reveals What Vision–Language Models Actually Focus On*. arXiv preprint arXiv:2503.06670. https://arxiv.org/abs/2503.06670

[10] 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

[11] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://doi.org/10.3115/1073083.1073135

[12] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 65–72). https://aclanthology.org/W05-0909

[13] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1904.09675