Parameter-Efficient Fine-Tuning and Explainability in Vision-Language Models for Remote Sensing 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—Explainable artificial intelligence, vision language model, parameter-efficient fine-tuning, PaliGemma.

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.

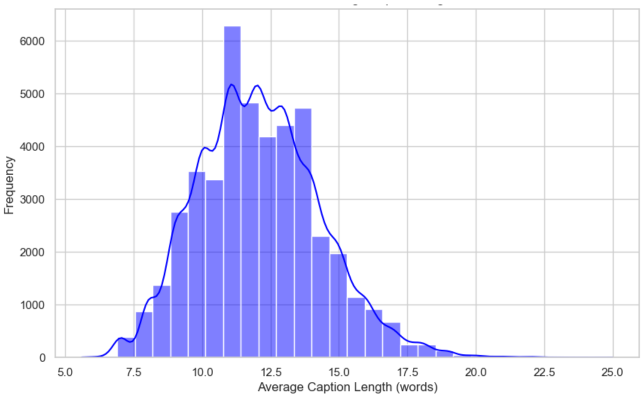

Figure 1 - Distribution of Average Caption Length

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.

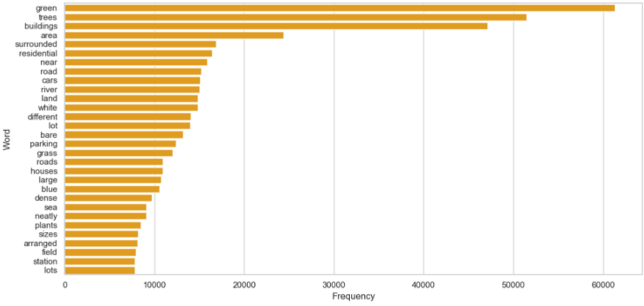

Figure 2 - Top 30 Most Frequent Words in Captions (Stopwords Removed)

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. Modeling

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.

C. *Parameter Efficient Fine Tuning*

To address the challenges of large-scale model adaptation under computational and memory constraints, this study employs Parameter Efficient Fine Tuning (PEFT)—with a particular emphasis on Low-Rank Adaptation (LoRA) and 4-bit quantization—enabling efficient transfer of PaliGemma to the remote sensing image captioning domain. The entire process is carefully orchestrated to maximize adaptation efficiency, minimize training overhead, and ensure reproducibility.

The pipeline begins with the initialization of the PaliGemma-3B-PT-224 model, loaded from pre-trained weights using the Hugging Face Transformers library. Recognizing the value of modularity, we freeze all parameters within the vision tower and multi-modal projector modules. This approach confines learning to the language decoder, which is the primary locus of domain adaptation for caption generation tasks.

The next stage involves injecting LoRA adapters into the decoder's core projection layers. With the rank hyperparameter set to r=8, the LoRA configuration balances expressive adaptation with a minimal parameter footprint. LoRA's architecture allows only the injected low-rank matrices to be updated during training, leaving the vast majority of the model's parameters untouched, thereby reducing both compute and memory demands.

To further boost efficiency, the model is loaded with 4-bit quantization (nf4 mode) using BitsAndBytes, which compresses model weights without sacrificing substantial representational power. This configuration enables effective training and inference on standard GPUs, making advanced VLM adaptation accessible without specialized hardware.

The fine-tuning process leverages a custom collate function and the Hugging Face Trainer API. During training, a fixed prompt is prepended to each image-caption pair, instructing the model to generate scene-specific captions. The training loop uses a batch size of two, a learning rate of 2e-3, bfloat16 precision, and an epoch-based or step-based stopping criterion, with periodic checkpointing for model safety.

During development, we encountered an important technical challenge: the model tended to echo the instructional prompt at the start of each generated caption during inference. This occurred because the prompt was always provided as input, both in training and during generation, and the model learned to reproduce the prompt as the start of its output. To resolve this, we adjusted our inference procedure to remove the prompt from the generation input—passing only the image to the model at inference time. This change aligned the generation regime with standard practices in image captioning, ensuring that the output consists solely of the caption relevant to the input image, free from prompt echo. This fix was critical for fair metric evaluation and for generating captions suitable for downstream human interpretation or further analysis.

At the conclusion of training, only the LoRA adapter weights were saved, alongside the processor configuration, allowing efficient storage and reloading for subsequent experiments or deployments.

V. Evaluation

Evaluation of the proposed parameter-efficient fine-tuned vision-language model is conducted using both quantitative and qualitative methodologies to ensure comprehensive insight into model performance and transparency. Quantitative evaluation is anchored in established image captioning metrics—BLEU (for n-gram overlap), METEOR (for semantic and alignment-based similarity), and BERTScore F1 (for contextual similarity via deep language representations). These metrics collectively capture syntactic, semantic, and contextual aspects of generated captions. Evaluation is performed on a held-out subset of the Remote

Image Sensing and Captioning (RISC) dataset, representing diverse, human-annotated satellite imagery.

To complement these metrics and elucidate the model's internal reasoning, three advanced post-hoc explainability techniques are employed: Grad-CAM, Integrated Gradients, and PixelSHAP. Each method is chosen for its unique perspective on spatial and semantic reasoning.

Grad-CAM generates spatial attention maps by leveraging gradient-weighted activations from intermediate model layers, thereby highlighting regions in the input image that most strongly influence the output. As depicted in Figure 3, the left panel shows the original input, the middle panel displays the Grad-CAM heatmap, and the rightmost panel overlays this heatmap onto the original image. This visualization provides an intuitive, high-level summary of the model's semantic focus, pinpointing salient areas such as aircraft and runways that correspond to objects referenced in the generated caption.

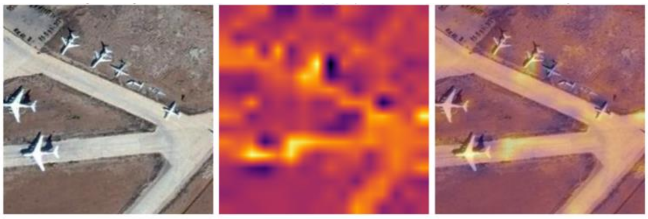

Figure 3 - Original, Grad-CAM Map and Grad-CAM Overlay

Integrated Gradients attribute pixel-wise importance scores by integrating gradients from a reference (baseline) image to the actual input, thus quantifying the contribution of each pixel to the model's prediction. In Figure 4 (left), the resulting saliency map reveals fine-grained structural sensitivity, with the model assigning high attribution to image regions corresponding to critical scene elements demonstrating robust alignment with the captioned content.

PixelSHAP, an adaptation of the SHAP framework for vision tasks, evaluates the marginal contribution of localized image patches via occlusion analysis. By systematically masking image segments and measuring the change in prediction, PixelSHAP constructs a region-wise heatmap (Figure 4, right), where warmer colors indicate areas that are most influential for the generated caption. This model-agnostic approach provides a complementary, interpretable quantification of regional importance, supporting the interpretability of complex vision-language predictions.

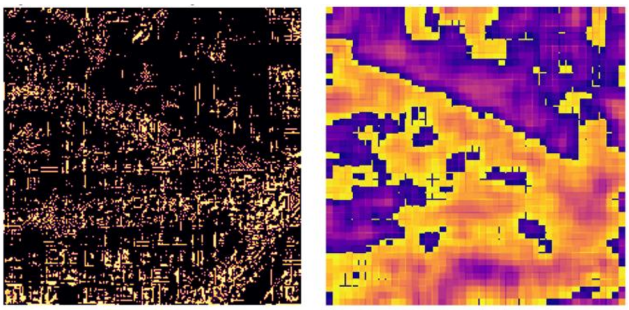

Figure 4 - Integrated Gradient Attribution (Avg over RGB Channels), PixelSHAP-style Occlusion Attribution

Collectively, these visualization techniques enhance transparency and interpretability by elucidating how the fine-tuned model grounds its textual outputs in the underlying visual evidence, thus supporting both qualitative validation and expert trust in model decisions.

VI. Results

The empirical results of both the baseline and PEFT models are summarized in Table I. The baseline, representing the zero-shot performance of PaliGemma on the remote sensing test set, achieved a BLEU score of 0.0037, a METEOR score of 0.1269, and a BERTScore F1 of 0.1224. Following LoRA-based PEFT, substantial gains were observed: BLEU increased to 0.2047, METEOR to 0.3981, and BERTScore F1 to 0.9134. These improvements reflect the effectiveness of LoRA adaptation in aligning the VLM with domain-specific visual and textual distributions.

Model	BLEU	METEOR	BERTScore_F1
Baseline	0.0037	0.1269	0.1224
PEFT (LoRA)	0.2047	0.3981	0.9134

Table 1 - Performance of the LoRA-adapted Model (subset data)

To further contextualize these quantitative gains, we provide a representative example from the held-out test set (Figure X), comparing the model-generated captions with the human-annotated ground truth:

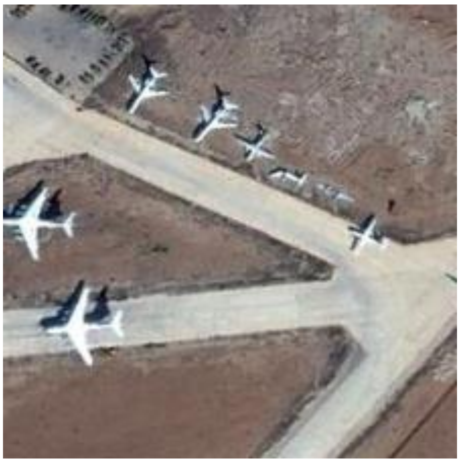

Figure 5 - Test set example 1

Ground Truth: There is a plane on the runway and three planes in the open space next to the runway.

Baseline: The aircraft are parked on the runway.

PEFT: There are several planes parked on the lawn next to the runway.

The baseline model produces a generic caption, missing the precise spatial arrangement and the plurality of planes, which are essential for detailed scene understanding in remote sensing. In contrast, the PEFT-adapted model more accurately reflects the complex spatial relationship, mentioning both "several planes" and their location "next to the runway." While the exact number is not stated, the improvement in specificity demonstrates the benefit of domain-adapted fine-tuning.

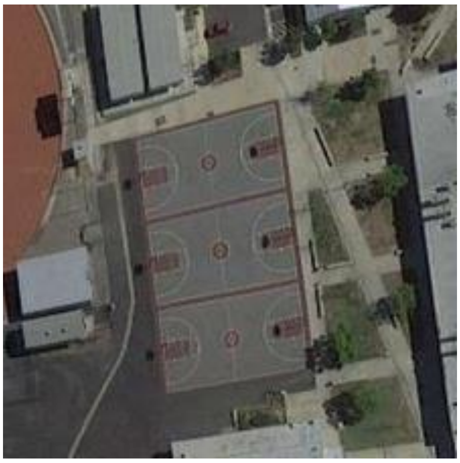

Figure 6 - Test set example 2

Ground Truth: There is a big basketball court in the center of the picture.

Baseline: There are several basketball courts next to buildings.

PEFT: There are three basketball courts in the middle of the road, and there are many buildings and some trees around the basketball court.

Both models recognize the presence of basketball courts, but the baseline response remains vague, lacking detail about the layout and context. The PEFT model provides enhanced spatial reasoning ("three basketball courts in the middle of the road") and incorporates additional contextual information ("buildings and some trees around"), reflecting improved grounding and scene compositionality post-adaptation.

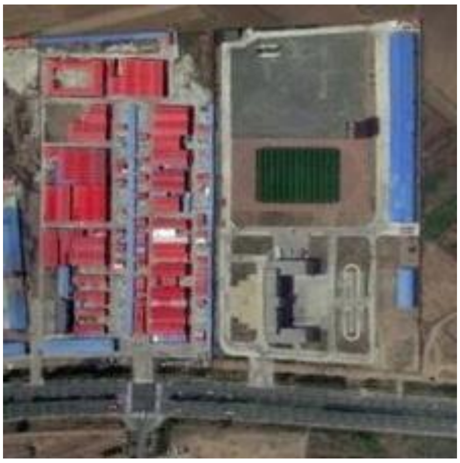

Figure 7 - Test set example 3

Ground Truth: The industrial area has some red workshops of different sizes, other buildings and an open area.

Baseline: There are buildings and a sports field.

PEFT: The playground is next to the road and the buildings are next to the playground.

This scene is visually complex, involving multiple object types and spatial zones. The baseline model generates a terse caption, omitting both color and spatial details. The PEFT model exhibits improved localization ("playground is next to the road," "buildings are next to the playground") but still abstracts over object color and type. This suggests that, while PEFT enables better scene parsing and relative positioning, further improvements in object attribute recognition (such as color) may require additional domain adaptation or more granular supervision.

VII. Discussion

The results affirm that parameter-efficient fine-tuning via LoRA yields significant improvements over the baseline PaliGemma model, bridging the gap between generic VLM capabilities and the domain-specific requirements of remote sensing captioning. The marked increase in all metrics—especially the leap in BERTScore F1—indicates enhanced semantic alignment and contextual understanding post-adaptation. Qualitative inspection of generated captions corroborates these findings, as fine-tuned outputs demonstrate greater specificity and alignment with human-annotated descriptions.

During experimentation, an initial challenge arose from the model's tendency to echo input prompts within generated captions—a well-documented artifact in instruction-tuned transformers when decoding is not properly delimited. This was systematically addressed by omitting the textual prompt during inference and ensuring correct use of the processor interface, thereby enforcing prompt masking and restoring generation fidelity. This correction was crucial in achieving the reported performance gains and underscores the importance of rigorous inference protocol in VLM applications.

The explainability analysis provides further validation of model trustworthiness. Integrated Gradients and PixelSHAP consistently highlight spatial structures that are salient to both human observers and the model, while Grad-CAM overlays visually confirm attention to relevant semantic regions. These findings support the model's capacity for transparent, justifiable predictions, a key criterion for deployment in high-stakes remote sensing domains.

VIII. Conclusion

This study demonstrates the efficacy of parameter-efficient fine-tuning, specifically LoRA, in adapting large vision-language models for domain-specific image captioning in remote sensing. Substantial performance improvements were achieved relative to the pre-trained baseline, with both automatic metrics and qualitative outputs confirming gains in relevance and descriptiveness. The integration of post-hoc explainability techniques further enhanced the transparency and interpretability of model decisions, offering visual evidence of alignment between model attention and annotated scene content.

Future work may extend this approach to additional PEFT strategies and explore more granular XAI methods for sequence-level explanation. The reproducible framework established herein provides a foundation for scalable, interpretable, and trustworthy deployment of VLMs in real-world, high-stakes domains.

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