Ali Yiğit Başaran - METU - DI

Efficient Domain Adaptation for Remote Sensing Image Captioning via LoRA Fine-Tuning of PaliGemma

***Abstract*– Remote sensing image captioning (RSIC) poses unique challenges due to the complexity and scale of satellite imagery. This study investigates the effectiveness of Low-Rank Adaptation (LoRA) as a parameter-efficient fine-tuning method for PaliGemma, a large vision–language model, on the RISC dataset. Two LoRA configurations are evaluated against a pretrained baseline using BLEU, METEOR, and ROUGE-L metrics. Results demonstrate that LoRA significantly enhances captioning performance while updating less than 0.1% of the model's parameters. These findings highlight LoRA's potential as a scalable and resource-efficient approach for domain-specific adaptation in RSIC tasks.**

***Keywords*– Remote sensing, image captioning, PaliGemma, parameter-efficient fine-tuning, LoRA, vision–language models, PEFT.**

I. Introduction

Remote sensing image captioning (RSIC) involves generating natural language descriptions for satellite images—a task vital for applications such as urban planning and environmental monitoring. While large vision–language models (VLMs) like PaliGemma show promise in this domain, fully fine-tuning such models is often impractical due to high computational and memory demands.

To address this, parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) offer a scalable alternative by updating only a small subset of model parameters. Although LoRA has shown success in domains like medical VQA, its application to RSIC—especially with general-purpose models like PaliGemma—remains largely unexplored.

This project evaluates whether LoRA can match or surpass the captioning performance of a pretrained baseline on the RISC dataset while significantly reducing training cost. By comparing two LoRA configurations, we aim to identify performance–efficiency trade-offs and explore LoRA's potential for lightweight domain adaptation in remote sensing tasks.

II. Literature Review

Lin et al. introduced **RS-MoE**, a Mixture-of-Experts (MoE) based vision–language model tailored for **remote sensing image captioning (RSIC)** and **VQA** tasks. The model divides the captioning process into sub-components—theme, object, and spatial reasoning—using an instruction router that dynamically assigns each part to a dedicated lightweight LLM [1]. This modular structure is trained in two stages: first by fine-tuning the vision encoder and a single LLM, then by training the full MoE block initialized from the first stage. To enable efficient learning, **LoRA** is applied to attention and feedforward layers across both stages, significantly reducing the number of trainable parameters [1]. Experiments on five RSIC and two RSVQA datasets show that **RS-MoE-1B** performs on par with or better than larger models like **BLIP2-13B** and **MiniGPT-13B** in BLEU, METEOR, ROUGE-L, and CIDEr scores, including zero-shot generalization on unseen datasets [1]. Ablation studies confirm that the instruction router and multi-expert setup enhance caption quality, while the LoRA-based fine-tuning preserves performance with fewer parameters. These results highlight how **parameter-efficient modular fine-tuning strategies**—especially using LoRA—can rival full fine-tuning in RSIC tasks and reinforce the relevance of this approach for adapting general-purpose VLMs like PaliGemma.

Liu et al. proposed **VQA-Adapter**, a parameter-efficient transfer learning approach designed for **medical visual question answering (Med-VQA)**. Built on the CLIP model, this method freezes the visual encoder and introduces a lightweight adapter module between it and the classifier, enabling the model to adapt to downstream tasks by updating only **2.38% of the parameters** [2]. To further enhance generalization, the authors incorporate a **multi-stage label smoothing strategy**, which gradually reduces confidence in target labels across training epochs to prevent overfitting [2]. The model was evaluated on the **VQA-RAD** and **SLAKE** datasets and outperformed seven strong baselines, including fully fine-tuned CLIP (PubMedCLIP), MMBERT, and MEVF models [2]. VQA-Adapter achieved **75.8% accuracy** on VQA-RAD and **81.0%** on SLAKE, setting new state-of-the-art results in both open- and closed-ended VQA tasks. Ablation studies confirmed the effectiveness of both the adapter module and the dynamic label smoothing. These findings demonstrate that large pretrained vision–language models can be efficiently adapted to domain-specific applications through **lightweight tuning strategies**, supporting the broader hypothesis that **LoRA-based methods** may offer similar benefits in **remote sensing image captioning** without the need for full fine-tuning.

Gao et al. presented the **LoRA dataset**, a large-scale benchmark specifically designed to evaluate the **logical reasoning abilities** of VQA models through multimodal information [3]. It contains **200,000 structured questions** grounded in a food-and-kitchen ontology and formal description logic (SROIQ). The dataset covers various logical forms including negation, conjunction, disjunction, and rule-based reasoning, offering three levels of increasing complexity. Each image is linked with questions requiring **multi-step logical inference**, and answers range from binary to compositional formats. Several VLMs—including **VisualBERT**, **MiniGPT4**, and **InstructBLIP**—were benchmarked on the dataset. While simple logical questions were handled moderately well, all models performed poorly on higher-order reasoning tasks; InstructBLIP achieved only **30.5% accuracy** on the most complex cases [3]. The authors observed that models tended to rely on textual priors instead of combining visual and logical information, producing inconsistent outputs. Though this study focuses on a dataset rather than a fine-tuning method, it reinforces the idea that **targeted adaptation strategies like LoRA-based PEFT** may improve general-purpose VLMs on logic-sensitive tasks such as **remote sensing image captioning**, where spatial reasoning and structured interpretation are also required.

Punneshetty et al. proposed a **LoRA-based parameter-efficient fine-tuning** strategy for the large-scale **Idefic 9B** visual language model, targeting performance improvement in **medical VQA** tasks [4]. Their method applies **LoRA** to attention layers while freezing other model components, and integrates **4-bit quantization** to reduce training cost. The model treats VQA as a generative task, producing natural language answers rather than selecting from predefined options. It achieved **72.02 BLEU** and **77% accuracy** on **SLAKE** and **VQA-RAD**, demonstrating strong domain adaptability with minimal parameter updates [4]. The study shows that LoRA can deliver high performance across diverse medical modalities with significantly reduced training overhead. Adapter weights can also be merged back into the base model for simplified deployment. While limitations such as hallucinations and prompt sensitivity are acknowledged, the results confirm that **partial fine-tuning** using LoRA is a viable alternative to full fine-tuning for large VLMs. This aligns with the goal of this project to evaluate LoRA's potential in **remote sensing image captioning**, particularly when using models like **PaliGemma**.

According to recent studies, there is limited investigation into whether LoRA can be applied to **general-purpose VLMs**, such as **PaliGemma,** in the context of fine-tuning with **RSIC**. Moreover, prior works have not directly focused on a comparative analysis between LoRA-based and full fine-tuning strategies on **RSIC dataset**. This project aims to address this gap by empirically evaluating whether LoRA can offer similar or better captioning performance to full fine-tuning, while significantly reducing training cost and memory usage.

III. Project Proposal

1. Research Question:

Do LoRA-based parameter-efficient fine-tuning methods perform comparably to raw pretrained big VLMs (e.g. Paligemma) in remote sensing image captioning tasks in terms of performance and/or resource efficiency?

2. Project Objectives:

This project aims to evaluate the feasibility and effectiveness of Low-Rank Adaptation (LoRA) as a parameter-efficient fine-tuning strategy when applied to PaliGemma, a general-purpose encoder–decoder transformer-based vision–language model (VLM), in the context of remote sensing image captioning (RSIC). The primary objectives are:

12 transformer encoder l

- To **compare the captioning quality** of LoRA-based fine-tuning and pretrained model using established automatic evaluation metrics.
- To **measure resource efficiency**, including the number of trainable parameters, training time.
- To identify **performance–efficiency trade-offs** and determine whether LoRA can offer a scalable, cost-effective alternative for adapting large VLMs to domain-specific tasks such as RSIC.

Importance and Motivation:
Recent advances in parameter-efficient transfer learning (PEFT) techniques, particularly LoRA, have demonstrated remarkable results in medical vision–language tasks by achieving similar performance to full fine-tuning while updating only a small subset of model parameters. However, their application to general-purpose models like PaliGemma remains largely unexplored in remote sensing domains. Remote sensing image captioning presents unique challenges:

- High-resolution, information-dense satellite imagery
- Significant data volume (e.g., 44,521 images and 222,605 captions in RISC)
- The need for scalable, memory-efficient model adaptation

Given the growing size of VLMs and the limited computational resources available in many real-world settings, **developing lightweight yet effective fine-tuning strategies is both practically useful and theoretically significant**. If successful, this study could contribute to a more sustainable and accessible use of large transformer models in earth observation tasks.

Planned Methodology:
We will experiment with two distinct fine-tuning approaches on the **RISC dataset**, which provides 5 human-annotated captions for each of 44,521 satellite images at 224×224 resolution:

- **Pretrained (Baseline):**
  - All parameters of the PaliGemma model (including the SigLIP image encoder and Gemma decoder) are pretrained.
  - No parameter freezing will be applied at pretraining.
  - Serves as the reference point for performance and resource usage.
- **LoRA-based Fine-Tuning (Experimental Setup):**
  - The backbone ViT encoder and the Gemma decoder are **frozen** except for selected attention and feedforward layers.
  - **LoRA modules** will be injected into those layers to enable efficient adaptation.
  - Only a small percentage (typically <2%) of the total parameters will be updated.

The model will be finetuned, and compared with the pretrained baseline by following evaluation criteria:

- **Captioning Performance Metrics:**
  BLEU, METEOR, ROUGEL(under the test set)

- **Efficiency Metrics:**
  Number of trainable parameters, training time in hours

To ensure fair comparison, early stopping or fixed epoch schedules will be employed consistently.

Expected Outcomes:
We expect that LoRA-based fine-tuning will considerably improve captioning performance relative to the pretrained model, while significantly reducing the number of trainable parameters and training time. A small performance gap, if any, may be acceptable in exchange for large gains in efficiency. This result would confirm that parameter-efficient adaptation is viable for transformer-based VLMs in remote sensing applications, contributing to both sustainable AI practices and domain adaptation research. The experiment will also shed light on the limitations of LoRA in multi-modal encoder–decoder settings and help establish guidelines for applying PEFT to large general-purpose models like PaliGemma.

IV. Dataset

This project utilizes the Remote Image Sensing and Captioning (RISC) dataset, which consists of 44,521 satellite images at a fixed resolution of 224×224 pixels, each associated with five human-written captions, totaling 222,605 textual descriptions. The dataset is structured into three splits: **train (35,614 images–80%)**, **validation (4454 images–10%)**, and **test (4454 images–10%)**.

All images are successfully located and used (no missing or unused entries), and each row contains exactly five non-empty captions. The dataset includes samples from three different sources: **NWPU** (31,500 images), **RSICD** (10,921), and **UCM** (2,100). This source diversity reflects varying linguistic styles and domain-specific captioning tendencies, which is relevant for evaluating model generalization across satellite domains.

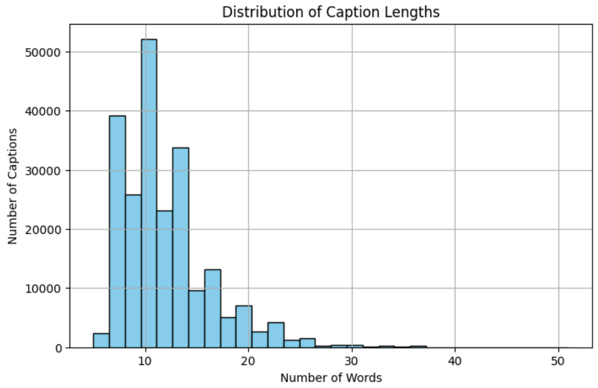

Figure 1: Distribution of Caption Lengths On All Images

As shown in **Figure 1**, the majority of captions in the RISC dataset exhibit a **narrow and consistent length distribution**. The histogram reveals that most captions contain between **8 and 14 words**, with a peak around **10–11 words**. The average caption length is approximately **12 words**, with very few outliers exceeding 30 words.

This regularity in caption length is favorable for transformer-based models such as PaliGemma, which often rely on fixed-length positional encodings and are sensitive to sequence length during training and inference. The low variance also suggests that the captioning style across the dataset is relatively standardized and likely follows concise, structured phrasing.

Such consistency simplifies batching and reduces the risk of padding inefficiency in training. However, it may also reflect **formulaic language or template-style captioning**, which could limit semantic richness or linguistic diversity in model outputs—especially when combined with observed n-gram repetition patterns.

An important quality indicator in image captioning datasets is **caption duplication**. For each image, five captions are expected to be **semantically diverse**. However, our analysis revealed that:

- **~75%** of images have five completely unique captions,
- But **~11%** have all five captions identical,
- Remaining images have 2 to 4 duplicates per set.

The duplication level varies significantly across sources: **98.9% of NWPU samples are unique**, while **RSICD contains 44.65% fully duplicated caption sets**, raising concerns about caption diversity within that domain. Additionally, the validation set shows a relatively high rate of partial duplication (**~25% have 2+ identical captions**), which could affect generalization estimates.

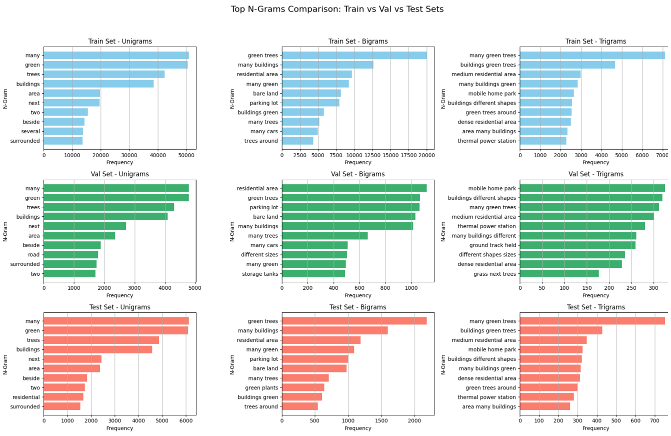

Figure 2: Top N-Grams (1-2-3 Grams) Per Train, Val & Test Sets

Figure 2 illustrates the most frequent unigrams, bigrams, and trigrams across the train, validation, and test splits of the RISC dataset. A high degree of lexical overlap is evident—phrases like *"green trees"*, *"residential area"*, and *"many green trees"* dominate all splits. This repetition suggests a templated captioning style, which may help training convergence but poses risks of overfitting and inflated performance metrics such as BLEU. Since many test captions reuse phrases from the training set, models may score well without genuine semantic understanding. Nonetheless, the presence of domain-specific phrases like *"thermal power station"* and *"mobile home park"* indicates that meaningful content is still present and can guide effective learning if used properly. Overall, Figure 2 highlights the importance of considering linguistic redundancy when evaluating model generalization and caption diversity.

V. Modeling

This section describes the complete methodology adopted to fine-tune the PaliGemma model for the task of image captioning in the context of satellite imagery. The experimentation process involved preprocessing and filtering of captions, implementing parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA), configuring training and validation pipelines, and evaluating model performance using standard generation metrics.

1. Caption Selection and Dataset Preparation

The Remote Image Sensing and Captioning (RISC) dataset contains 44,521 satellite images, each annotated with five descriptive captions. These captions often exhibit lexical overlap, paraphrasing, or varying specificity, which can introduce noise during training and increase the amount of the training and evaluation time considerably. To address these, we implemented a caption filtering strategy aimed at selecting the most semantically representative caption per image.

Each image's five captions were compared against each other using three automatic evaluation metrics — BLEU, METEOR, and ROUGE-L — commonly used in natural language generation tasks. For each candidate caption, we computed its average score by treating the other four captions as references. The caption with the highest combined score was selected as the filtered caption. This process yielded a new CSV file, ***filtered_captions_fixed.csv***, containing only one caption per image.

The dataset was then split into three subsets: train, validation, and test, based on the predefined split column in the CSV. Unused metadata such as source and split columns were discarded after stratification. This resulted in a high-quality, deterministic dataset with reduced ambiguity for training the captioning model.

2. Base Model and LoRA Configurations

The base architecture selected for this project is pretrained **google/paligemma-3b-pt-224**. This model has demonstrated competitive generalization capabilities across various vision-language tasks but has not been specifically trained or adapted for satellite imagery captioning, and open to fine-tuning essential for domain adaptation.

To explore the trade-off between fine-tuning efficiency and model performance, we experimented with two distinct LoRA configurations, each tailored to different objectives in the project pipeline.

The **first configuration**, referred to as *LoRA-2*, was designed as a lightweight adaptation setup to obtain **preliminary results over pretrained baseline model.** This configuration prioritized minimal parameter overhead and faster training convergence. It targeted only the query and output projection submodules within the attention mechanism—components that directly influence the model's generative behavior. This selective targeting ensured that the model could begin adapting to the captioning task without incurring significant computational costs. The configuration was intentionally conservative, making it ideal for early experimentation and direct comparison with the fine-tuned baseline. The **second configuration**, named *LoRA-1*, was constructed for the **benchmarking phase**. Its design emphasized expressiveness and coverage, aiming to improve upon both the pretrained baseline and the preliminary LoRA-2 results. This setup extended LoRA's influence to all four projection components of the attention mechanism—query, key, value, and output—allowing the model to adjust every core stage of the attention computation. This comprehensive adaptation strategy was expected to yield more substantial performance gains, particularly in generating semantically rich and syntactically accurate captions for the satellite image domain. Both configurations utilized dropout for regularization and a scaling mechanism to control the influence of the learned low-rank matrices. While LoRA-2 favored simplicity and stability, LoRA-1 adopted a more aggressive approach to leverage the model's full potential in the benchmarking experiments. Configuration parameters are given in Table-1.

	Rank (r)	**Alpha (α)**	**Dropout**	**Target Modules**
Lora-1	4	8	0.2	q_proj, o_proj, v_proj, k_proj
Lora-2	2	4	0.1	q_proj, o_proj,

Table-1: LoRA Finetune Configuration Parameters

3. Collation & Tokenization

All images were loaded by converting them to RGB. During training and validation, each image was paired with the selected filtered caption. The prompt "caption the image" was prepended to every input text to guide the model's decoder during autoregressive generation.Two distinct collate functions were defined:

Training/Validation Collate: Takes a batch of images and captions, combines them with prompts, and feeds the output to the processor to produce input tensors (pixel values, tokenized text, attention masks).

Test Collate: Takes only images and prompts, omitting target captions to allow the model to freely generate predictions during evaluation.

Tokenization: The **PaliGemmaProcessor** from the **Hugging Face transformers library** serves as a unified interface for preparing both visual and textual inputs for the model. On the textual side, it automatically tokenizes the input prompt and, during training, the corresponding target caption using the model's tokenizer, converting them into sequences of token IDs. It also inserts special <image> tokens when images are present, ensuring correct alignment between vision and language modalities. The processor dynamically creates attention masks to distinguish between meaningful tokens and padding, and it handles padding and truncation to ensure that all input sequences in a batch have uniform length. Additionally, the processor returns all input components as PyTorch tensors, and in our setup, these tensors were cast to bfloat16 precision to reduce memory consumption and speed up computation on supported GPUs without compromising training stability.

VI. Evaluation

1. Training, Validation and Testing Procedures

1.1. Training and Validation Procedures

The model was finetuned with LoRA for **3 full epoch**s using a **batch size of 1**, chosen due to GPU memory limitations associated with the 3B-parameter model. The **AdamW optimizer** was used with a **learning rate of 1e-5** and **weight decay of 1e-3**. A **linear warmup scheduler** was incorporated, reserving the first **10%** of total training steps for gradual learning rate increases, followed by linear decay.

Training involved the following key points:

a. Gradient accumulation was disabled, relying on single-step updates.
b. **Training loss** was logged at regular intervals (per 100 steps) using Weights & Biases (W&B).
c. **Validation loss** was logged at end of each epoch using Weights & Biases (W&B).
d. The learning rate scheduler was updated at the end of each epoch based on the validation loss.

1.2. Testing Procedure

After training, the model was evaluated on the held-out test set under three scenarios:

- **Pretrained Baseline**: Pretrained Paligemma PT 224 model that is provided by Google.
- **LoRA-1 Fine-Tuned:** Finetuning with LoRA-1 Configuration
- **LoRA-2 Fine-Tuned:** Finetuning with LoRA-2 Configuration

For each image in the test set, the model was prompted with ***"caption the image"*** and expected to generate a caption without having access to the ground-truth. Captions were decoded and cleaned to remove prompt for evaluation.

Model performance was evaluated using:

- **BLEU:** Measures n-gram precision and penalizes overly short generations.
- **METEOR:** Accounts for synonymy, stemming, and alignment.
- **ROUGE-L:** Focuses on the longest common subsequence, indicating fluency and recall.

Each metric was computed using the Hugging Face evaluate library. The results were logged to W&B for visual comparison and historical tracking.

VII. Results

The preliminary results, benchmarking, and final experiment results are combined and presented in this section.

Table 2 shows the BLEU, METEOR, Rouge-L, fine-tuning duration (in hours), and number of parameters for the Paligemma models fine-tuned with the pretrained baseline and the LoRA-1 and LoRA-2 configurations specified in the Modeling section.

Figures 3 and 4 show the training and loss curves, respectively, of the Paligemma models fine-tuned with the LoRA-1 and LoRA-2 configurations

Model	BLEU	Meteor	RougeL	Ft Hour	Num. Trainable Params
Baseline Pretrained	0.0044	0.082	0.150	-	3032242416
FT w/LoRA-2	0.113	0.196	0.335	25	577024
FT w/LoRA-1	0.197	0.266	0.487	30	2343936

Table-2: Bleu, Meteor and Rouge-L Metrics & FT Hours - Num. Tr. Param

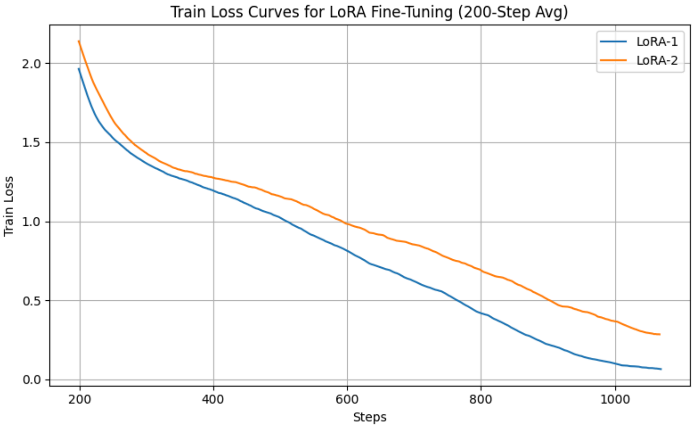

Figure-3: Train Loss of Finetuning Paligemma with LoRA-1/2 Configurations

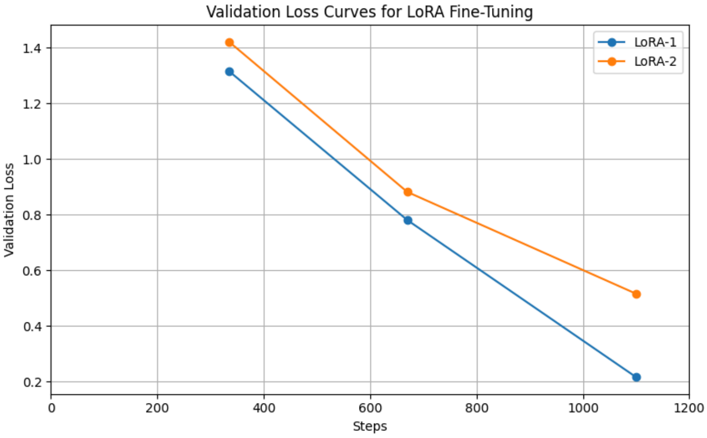

Figure-4: Validation Loss of Finetuning Paligemma with LoRA-1/2 Configurations

VIII. Discussion

1. Preliminary Results

The preliminary results phase focused on evaluating whether the proposed fine-tuning strategy could yield meaningful improvements over the pretrained baseline. Using the **LoRA-2 configuration**, which applies a lightweight adaptation to only the query and output projections of the attention mechanism, we conducted fine-tuning over three epochs on the RISC dataset. The results, presented in Table 2, demonstrate a clear and consistent performance gain across all evaluation metrics.

The baseline model, which had not been fine-tuned on the satellite domain, performed poorly across all metrics. The BLEU score of 0.0044 indicates extremely low n-gram overlap between generated captions and the references. This is expected, as the pretrained PaliGemma model has not been exposed to the specialized visual-textual patterns of remote sensing imagery. Similarly, the METEOR score of 0.0820 and ROUGE-L score of 0.1509 reflect a lack of semantic coherence and limited phrase-level alignment.

After fine-tuning with **LoRA-2**, we observed a dramatic increase in all three metrics:

- **BLEU** improved from 0.0044 to 0.113, suggesting that the model learned to generate captions with significantly more accurate token sequences.
- **METEOR** rose to 0.196, indicating that the model captured more semantically aligned outputs, including better word choices and ordering.

- **ROUGE-L** nearly doubled, increasing from 0.150 to 0.335, reflecting improved structural consistency and overlap with reference captions at the phrase level.

These improvements are especially noteworthy considering the low-rank nature of LoRA-2, which introduces a minimal number of trainable parameters. The model was able to adapt to the new domain efficiently without full-scale parameter updates. This validates both the effectiveness of the LoRA method and the overall soundness of the training pipeline, including preprocessing, prompt engineering, and evaluation.

The number of trainable parameters in the fine-tuned model using the **LoRA-2 configuration** was approximately **577K**, compared to the full model size of over **3 billion parameters**. This means that less than **0.02%** of the model's parameters were updated during training. Despite this extremely small proportion, the model achieved substantial performance gains over the baseline, demonstrating the **effectiveness of parameter-efficient fine-tuning**. This also highlights one of the key advantages of LoRA: it enables domain adaptation with minimal computational cost while preserving the integrity and efficiency of the original pretrained model.

The training loss curve of the fine-tuned model using the LoRA-2 configuration, shown in Figure 3, demonstrates a smooth and consistently decreasing pattern, indicating stable optimization throughout training. While the loss did not converge fully to zero, it approached a low plateau toward the end of the third epoch at the 0.6's. The lack of full convergence is expected given the limited training schedule, small batch size, and the parameter-efficient nature of LoRA-2, which restricts the number of tunable weights. Rather than being a sign of underfitting, this outcome likely reflects a healthy balance between learning and generalization, where the model captures essential patterns without memorizing the data.

The validation loss curve for the LoRA-2 configuration, depicted in orange in Figure 4, shows a steady and meaningful decrease throughout the training process. Starting from a high loss value of approximately 1.42, the model progressively reduces its loss to around 0.53 by the end of the training steps. This downward trend indicates that even with a minimal set of trainable parameters, LoRA-2 enables the model to effectively adapt to the satellite image captioning task. The consistent decline without oscillations or spikes suggests that the training remained stable. Compared to the pretrained PaliGemma model—which did not undergo any task-specific fine-tuning and thus lacks a validation loss trajectory—LoRA-2 demonstrates a clear benefit in learning domain-relevant features and improving model generalization.

After validating the effectiveness of the **LoRA-2** configuration in the preliminary results phase, we conducted a comparative benchmarking study to assess whether a more expressive adaptation strategy could yield further improvements. This involved training a second configuration, **LoRA-1**, which increased the rank and coverage of the adapted attention layers compared to **LoRA-2**. The two configurations, shown in Table 1, were selected to explore the trade-off between training cost and performance in a parameter-efficient fine-tuning context.

2. Benchmarking

After validating the effectiveness of the **LoRA-2** configuration in the preliminary results phase, we conducted a comparative benchmarking study to assess whether a more expressive adaptation strategy could yield further improvements. This involved training a second configuration, **LoRA-1**, which increased the rank and coverage of the adapted attention layers compared to **LoRA-2**. The two configurations, shown in Table 1, were selected to explore the trade-off between training cost and performance in a parameter-efficient fine-tuning context.

The results in Table 2 reveal that **LoRA-1 consistently outperforms both LoRA-2 and the baseline** across all metrics. LoRA-1 improves the BLEU score by approximately **74%** over LoRA-2, and achieves a **44-fold increase** over the pretrained baseline. The METEOR and ROUGE-L scores also see large improvements, reflecting better alignment and fluency in the generated captions. These gains come at a modest cost: although LoRA-1 introduces approximately **four times more trainable parameters** than LoRA-2, it still updates only **~0.077% of the total model parameters**, making it significantly more efficient than full fine-tuning.

The increase in **training time from 25 to 30 hours** is also reasonable given the improved performance and demonstrates the scalability of LoRA-based tuning. LoRA-1 thus proves to be a compelling balance between expressiveness and efficiency, suitable for more advanced fine-tuning stages.

To assess the training stability and learning dynamics of both configurations, we tracked training loss over time. **Figure 3** shows that both LoRA-1 and LoRA-2 exhibit **smooth and stable convergence**, without signs of divergence or oscillation. Importantly, LoRA-1 consistently achieves **lower training loss** at every step, confirming that its higher rank and broader adaptation capacity enable better fitting of the training data.

In the benchmarking phase, a comparative analysis of validation loss trajectories between LoRA-1 (blue line) and LoRA-2 (orange line) reveals the superior learning dynamics of the more expressive LoRA-1 configuration. While both configurations begin with similarly high loss values (around 1.3–1.4), LoRA-1 exhibits a sharper and more consistent descent, ultimately reaching a much lower final validation loss of approximately 0.22. In contrast, LoRA-2, although still improving, plateaus at a higher loss value of around 0.53. This significant gap underscores LoRA-1's enhanced capacity to capture domain-specific patterns, likely due to its broader adaptation of attention submodules and increased parameter rank. The graph thus validates that expanding LoRA coverage and depth yields tangible performance gains without compromising training stability.

The benchmarking results confirm that **LoRA-1 provides a clear performance advantage** over both the lightweight LoRA-2 configuration and the original pretrained baseline. While LoRA-2 remains useful for quick adaptation with minimal resource use, LoRA-1 demonstrates that **a broader and deeper adaptation of the attention mechanism** is more effective for domain-specific caption generation. These findings validate the importance of architectural tuning in vision-language modeling and provide a robust foundation for future work.

3. Final Result

The final results of this study provide clear evidence of the effectiveness and efficiency of LoRA-based fine-tuning for remote sensing image captioning using the PaliGemma model. Among all tested configurations, LoRA-1 emerges as the best-performing setup, achieving the highest scores across all evaluation metrics: BLEU (0.197), METEOR (0.266), and ROUGE-L (0.487). These values represent a substantial improvement over both the pretrained baseline, which showed minimal task-specific performance (BLEU: 0.0044), and the lightweight LoRA-2 configuration (BLEU: 0.113). This performance leap highlights the advantage of LoRA-1's broader adaptation coverage, which targets all four projection components (q, k, v, o) in the attention mechanism. In contrast, LoRA-2, while effective, was limited to the query and output projections, which constrained its expressive power.

The validation loss curves further corroborate these findings: LoRA-1 not only achieved a lower final validation loss but also exhibited a steeper and more stable convergence trajectory compared to LoRA-2. Despite requiring slightly more training time (30 hours vs. 25), LoRA-1 updated only ~0.077% of the total model parameters, maintaining a high degree of efficiency while delivering strong performance. The LoRA-2 configuration, which trained with just 577K parameters (~0.019%), proved effective for early experimentation and rapid prototyping but ultimately lacked the capacity to match LoRA-1's caption quality.

Overall, these results affirm that parameter-efficient fine-tuning via LoRA can rival full-model adaptation, and that careful tuning of LoRA's rank, dropout, and target modules significantly affects downstream performance. LoRA-1, in particular, strikes an optimal balance between training efficiency and captioning accuracy, making it the most practical and scalable configuration for domain adaptation in vision–language modeling under computational constraints.

IX. Conclusion

This study investigated the feasibility and effectiveness of Low-Rank Adaptation (LoRA) as a parameter-efficient fine-tuning strategy for remote sensing image captioning (RSIC) using the PaliGemma vision–language model. The primary aim was to determine whether LoRA could enable effective domain adaptation on the RISC dataset while dramatically reducing computational overhead compared to full model fine-tuning. Through a series of methodical experiments involving two LoRA configurations—LoRA-1 (expressive) and LoRA-2 (lightweight)—the study systematically evaluated the trade-offs between performance and efficiency.

The experimental results clearly support the hypothesis: both LoRA configurations yielded substantial improvements over the pretrained baseline, which performed poorly due to its lack of exposure to domain-specific patterns in satellite imagery. Among them, LoRA-1 consistently outperformed all alternatives, achieving the highest BLEU, METEOR, and ROUGE-L scores, while still updating less than 0.1% of the model's parameters. LoRA-2, although more limited in adaptation scope, demonstrated solid improvements, especially valuable in settings where resource constraints are critical.

The study's success lies not only in the empirical gains but also in affirming that general-purpose large-scale VLMs like PaliGemma can be effectively adapted to specialized domains such as RSIC with minimal parameter tuning. The validation loss curves further reinforced the robustness and stability of the fine-tuning process, with LoRA-1 achieving smoother and deeper convergence.

In conclusion, this work confirms that LoRA is a viable and scalable solution for efficient domain adaptation in multimodal learning tasks. It opens up promising directions for applying parameter-efficient fine-tuning to other resource-intensive domains, making powerful VLMs more accessible and sustainable for real-world applications. Future work may explore combining LoRA with additional adaptation strategies or extending this approach to multilingual or multimodal satellite datasets to further generalize the findings.

X. Future Work

Future work can build on the promising findings of this study by addressing several open directions. First, a direct comparison with a fully fine-tuned version of PaliGemma would offer a stronger baseline for understanding the exact performance trade-offs of parameter-efficient fine-tuning. While this study confirms that LoRA can significantly boost performance with minimal parameter updates, quantifying its gap against full fine-tuning would help determine its practical limits. Additionally, combining LoRA with other parameter-efficient strategies such as prefix tuning or adapter modules may lead to even better efficiency–performance trade-offs. Exploring these modular PEFT methods could also enable multi-task learning scenarios that go beyond captioning and include related tasks like visual question answering in the remote sensing domain.

Another promising direction is extending the model to multilingual or multimodal captioning tasks. Incorporating non-English descriptions or data types such as night-time or SAR satellite imagery would help evaluate the generalizability of LoRA-based tuning in more diverse and realistic settings. Future work could also involve a systematic search over LoRA's hyperparameters—such as rank, dropout rate, and targeted attention layers—to better tailor the tuning process to specific dataset characteristics or hardware constraints.

Beyond training, further research should investigate inference-time performance, particularly in deployment settings with strict latency or memory budgets. Measuring how LoRA impacts model responsiveness and scalability on edge devices or cloud services would be crucial for real-world applications. Finally, complementing current automatic evaluation metrics with human assessments or semantic diversity-focused metrics like SPICE or CIDEr would provide a more nuanced understanding of the quality and usefulness of generated captions. These directions together can help refine and expand the role of LoRA and other PEFT methods in building practical, high-performing, and sustainable vision–language systems for specialized domains like remote sensing.

XI. References & Links

[1] H. Lin et al., "RS-MoE: A Vision–Language Model With Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering," IEEE Trans. Geosci. Remote Sens., vol. 63, 2025.

[2] J. Liu et al., "Parameter-Efficient Transfer Learning for Medical Visual Question Answering," *IEEE Trans. Emerging Topics Comput. Intell.*, vol. 8, no. 4, pp. 2816–2826, Aug. 2024.

[3] J. Gao, Q. Wu, A. Blair, and M. Pagnucco, "LoRA: A Logical Reasoning Augmented Dataset for Visual Question Answering," *NeurIPS Datasets and Benchmarks Track*, 2023.

[4] S. Punneshetty et al., "Fine-Tuning Idefic 9B With LoRA for Multimodal Medical VQA," *ICKECS 2024*, IEEE.

Github: https://github.com/YigitBasaran/DI725_Project

Wandb:
https://wandb.ai/aliyigitbasaran-/DI725_Project?nw=nwuseraliyigitbasaran