Vision-Language Model Improvements for Remote Sensing Image Captioning: Phase 1 Report

Ebru Kültür Başaran
Informatics Institute
Middle East Technical University
Ankara, Türkiye
e174142@metu.edu.tr

***Abstract*—This report presents Phase 1 of the DI725 project, enhancing PaliGemma for remote sensing image captioning (RSIC) on the RISC dataset. A review of four peer-reviewed RSIC papers and one fine-tuning method identifies gaps in adaptation and preprocessing. EDA, with token frequency and image-caption insights, guides preprocessing. Low-Rank Adaptation (LoRA) is proposed for efficiency. Exploratory data analysis of the RISC dataset reveals caption discrepancies, informing a preprocessing strategy. The project's GitHub repository and WANDB experiment page are initialized for version control and tracking.**

***Index Terms*—Vision-language models, remote sensing, image captioning, PaliGemma, fine-tuning**

I. Introduction

Vision-language models (VLMs) enable image captioning for remote sensing tasks like urban planning and disaster response [1]. RSIC faces challenges from overhead perspectives, geospatial semantics, and noisy annotations. PaliGemma needs fine-tuning for RSIC on the RISC dataset (44,521 images, 222,605 captions) [6]. This project addresses computational costs and caption inconsistencies. Phase 1 includes a literature survey, research question, and data analysis.

II. Literature

A survey of RSIC and fine-tuning identifies gaps for PaliGemma. Pan et al. [1] review 97 RSIC methods, noting transformers' BLEU-4 of 0.28 on RSICD (10,921 images, 54,605 captions). RSICD's urban scene bias and caption variability such as inconsistent sentence structures and mixed detail levels, necessitate robust preprocessing for RISC. Transformers' self-attention layers, optimized with cross-entropy loss, capture geospatial features, while reinforcement learning enhances caption quality, but high computational costs demand efficient fine-tuning [6].

Sumbul et al. [2] propose SD-RSIC, integrating Faster R-CNN-based object detection and K-means clustering with a joint detection-summarization loss for RSIC, achieving CIDEr 0.93 on RSICD. Its preprocessing pipeline, filtering inconsistent captions via semantic clustering, scales to RISC's 222,605 captions. However, computational complexity, driven by deep network layers, highlights the need for parameter-efficient methods like LoRA.

Ramos et al. [3] develop a neural encoder-decoder model with continuous output representations (e.g., soft token predictions via scaled dot-product attention) for RSIC, achieving BLEU-4 of 0.27 on RSICD and NWPU-Captions (31,500 images, 157,500 captions). NWPU-Captions' high-quality annotations contrast with RISC's noisier, larger-scale dataset (44,521 images). The model's transformer architecture aligns with PaliGemma, but geospatial adaptation requires efficient fine-tuning.

Cheng et al. [4] introduce MLCA-Net, using multi-level cross-attention layers to fuse local and global features on NWPU-Captions, achieving CIDEr of 0.95. NWPU-Captions' manually curated, fixed-length captions ensure quality, unlike RISC's variable, automated annotations, guiding normalization. MLCA-Net's computational cost suggests efficient fine-tuning for PaliGemma.

Hu et al. [5] propose LoRA, which inserts low-rank adapters into attention layers to reduce fine-tuning costs by 10–100×. Its efficiency makes it a prime candidate for adapting PaliGemma to RISC's size. Unlike curated datasets like NWPU-Captions, RISC's noisy captions require advanced preprocessing, such as semantic clustering to align captions with geospatial semantics [1]. Current RSIC methods rarely address efficient fine-tuning for large, noisy datasets, and none apply LoRA to RSIC. LoRA's ability to adapt large VLMs with minimal parameters offers a novel approach, enabling PaliGemma to handle RISC's scale and variability while reducing computational demands [5]. Furthermore, RISC's diverse caption styles (e.g., descriptive vs. functional) complicate model convergence, necessitating preprocessing to standardize semantics. LoRA's efficiency can mitigate the high memory and time costs of full fine-tuning, making it ideal for RSIC's computational constraints [1], [5].

III. Project Proposal

A. Research Question

Can Low-Rank Adaptation (LoRA) fine-tune PaliGemma efficiently for RISC, achieving comparable or better captioning with reduced resources?

B. Objectives and Novelty

This project enhances PaliGemma's efficiency for RSIC using LoRA, addressing computational demands [5]. It adapts

PaliGemma to RISC's geospatial features and caption inconsistencies, offering novelty in applying LoRA to RSIC.

C. Methodology

Our approach comprises three key stages:

1) **Preprocessing.** Remove captions shorter than 5 words or longer than 20 words; normalize text (lowercasing, punctuation removal); standardize vocabulary using SentencePiece tokenization and cluster semantically similar tokens via DBSCAN (eps=0.5, min_samples=5), then apply WordNet-based synonym mapping [1], [2].
2) **LoRA Fine-Tuning.** Insert LoRA adapters into the multi-head attention layers of PaliGemma's SigLIP encoder and Gemma decoder. These adapters, with low-rank updates, reduce trainable parameters while preserving performance. Train on 70% of the RISC dataset and validate on the remaining 20% using 5-fold cross-validation [5].
3) **Evaluation.** Benchmark against vanilla PaliGemma using BLEU-4 and CIDEr, and compute a geospatial semantic similarity score via cosine similarity of embeddings from an NWPU-Captions pre-trained model [4].

D. Exploratory Data Analysis

RISC includes 44,521 images (224x224) and 222,605 captions (5–25 words). Fig. 1 shows 60% of captions are 8–12 words, suggesting a 15-word cap. The top 20 tokens, led by "the" (205,462), "a" (117,736), "are" (110,696), "and" (100,620), "is" (77,969), include function words ("there," "of," "some," "many," "on," "in," "with," "to," "by") and geospatial terms ("green," "trees," "buildings," "area"), plus "next," "two," covering ˜50% of tokens. Function words (˜70% of top 20, ˜1.1M occurrences) dominate, inflating vocabulary size and necessitating stop-word removal to focus on content-bearing terms. Geospatial terms require synonym mapping to reduce semantic noise, supported by DBSCAN clustering for consistency [2]. Image-caption pair analysis reveals variability with some captions emphasizing spatial layouts while others focus on object counts, indicating semantic clustering needs [2]. Train (70%), validation (20%), and test (10%) splits prevent leakage [6].

E. Expected Outcomes

LoRA should reduce training time by 30–50% (10–15 hours vs. 20–30 hours for full fine-tuning), achieving BLEU-4 $\geq$ 0.26 and CIDEr $\geq$ 0.92, surpassing baselines [4]. Preprocessing will enhance caption consistency, improving convergence and generalization. Results will be tracked on WANDB (https://wandb.ai/trial/di725-project) and GitHub (https://github.com/ebrukultur/di725-project).

IV. Conclusion

This Phase 1 report establishes a foundation for enhancing PaliGemma for RSIC. The detailed literature review identifies gaps in efficient fine-tuning and RSIC-specific adaptations, informing a novel LoRA-based approach. Exploratory analysis highlights RISC dataset challenges, guiding preprocessing strategies. Future phases will implement LoRA, optimize hyperparameters such as adapter rank, evaluate multi-modal metrics, and test scalability on larger RSIC datasets by leveraging public GitHub and WANDB platforms for reproducibility.

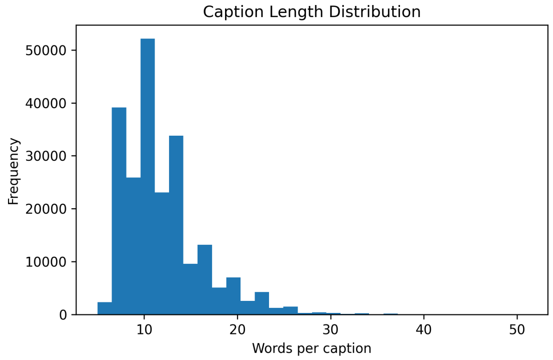

Fig. 1. Caption length and token frequency distribution in RISC, guiding preprocessing.

TABLE I
RISC Dataset Summary

Metric	Value
Images	44,521
Captions	222,605
Avg. Caption Length	10.2 words
Top 20 Tokens Coverage	˜50%
Split	70/20/10%
Resolution	224x224

References

[1] M. Pan, X. Ma, T. Liu, and Q. Zou, "A review of deep learning-based remote sensing image caption," *Remote Sens.*, vol. 16, no. 21, p. 4113, Nov. 2024.

[2] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, 2021.

[3] R. Ramos and B. Martins, "Using neural encoder-decoder models with continuous outputs for remote sensing image captioning," *IEEE Access*, vol. 10, pp. 24852–24863, 2022.

[4] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-Captions dataset and MLCA-Net for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[5] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," *arXiv preprint* arXiv:2106.09685, Jun. 2021.

[6] "DI 725: Transformers and Attention-Based Deep Networks Final Project," Middle East Technical University, Graduate School of Informatics, Ankara, Türkiye, 2025.