Prompt Engineering for Image Captioning in Remote Sensing using PaliGemma

DI725 Term Project Phase 1 Report

Onat Zeybek Kuşkonmaz
Student Number: 2237634

GitHub: https://github.com/Onatparagus/DI725_TermProject
WANDB: https://wandb.ai/onatzk-metu-middle-east-technical-university/DI725_TermProject

***Abstract*—This project explores prompt engineering and inference benchmarking for satellite image captioning using Google's PaliGemma model. Originally designed as a fine-tuning study, the project shifted to an inference-only approach due to hardware constraints. Through three stages of prompt refinement and an ablation involving decoding parameters and prompt length cues, I analyze refusal rates and output quality. Results show that prompt phrasing significantly affects caption success in zero-shot settings, revealing both opportunities and limitations of frozen vision-language models.**

I. Introduction

Image captioning is a key vision-language task that translates visual information into descriptive natural language. With the release of large pre-trained vision-language models (VLMs) like PaliGemma, opportunities arise to build captioning systems with minimal or no training. However, inference quality can vary significantly depending on prompt formulation and decoding parameters.

II. Literature Review

Text augmentation has proven to be an effective tool in various natural language processing tasks. I reviewed the following relevant works that inform my approach:

A. Back-Translation for Data Augmentation

Sennrich et al. (2016) demonstrated that back-translation could be used to synthesize high-quality data for machine translation tasks. This technique has since been generalized to augment datasets for various NLP tasks. In image captioning, it helps by generating natural sentence variants that preserve semantic meaning, which helps diversify the dataset and improve training quality.

B. Paraphrasing with Pretrained Transformers

Raffel et al. (2020) introduced the T5 model, a versatile text-to-text transformer that includes paraphrasing capabilities. Similarly, Pegasus (Zhang et al., 2020) pre-trains on sentence gaps for summarization and paraphrase generation. These models enable the generation of caption variants, promoting greater linguistic richness in training.

C. Image Captioning Metrics

Zhao et al. (2021) detailed the metrics commonly used to evaluate image captioning tasks such as BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. These methods proved to be more reliable than using human ratings, being less subjective and easier to use on a large scale.

D. Opportunities in Captioning Models

While architectural improvements like LoRA (Hu et al., 2022) allow efficient transformer fine-tuning, simple preprocessing-based enhancements remain underexplored in the context of vision-language models. Text augmentation provides a promising, low-cost opportunity to improve training without increasing computational complexity.

These papers indicate that while complex fine-tuning methods dominate the research, there is a research gap in leveraging simple augmentation strategies for improving captioning diversity, especially in the context of remote sensing.

III. Project Proposal

A. Original Research Question

How can simple text augmentation methods like paraphrasing and back-translation improve the diversity and quality of captions generated by a transformer-based vision-language model on remote sensing datasets?

B. Project Revision

Due to hardware limitations on training or fine-tuning a VLM, I had to change the project proposal to one that is based on inference and prompt engineering. The new research question is:

How does prompt structure affect the quality and reliability of image captions generated by a frozen PaliGemma model in remote sensing applications?

C. Objective and Significance

The primary objective of this revised project is to evaluate the impact of prompt phrasing on the captioning performance of a large vision-language model used in a zero-shot setting. Specifically, I aim to benchmark different prompt sets to generate 5 captions per image like in the dataset, and measure

their effect on output refusal rates, caption diversity, and relevance.

This is significant for two reasons: (1) It demonstrates how much can be gained through careful prompt design without any model fine-tuning, which is crucial in resource-limited environments. (2) It offers insights into using foundation models like PaliGemma for satellite imagery tasks, where pretraining data may not align perfectly with target domains.

D. Dataset and Preliminary Analysis

The RISC dataset consists of 44,521 satellite images, each paired with 5 human-written captions, totaling over 222,000 entries. Captions are short and often exhibit high overlap. For instance, many captions use repeated phrases such as "A building with a red roof." This redundancy limits model expressiveness. My initial analysis confirms an average caption length of 10 words and significant repetition across captions for a single image.

E. Methodology

This project uses the frozen `google/paligemma2-3b-mix-224` model for zero-shot image captioning via prompt engineering. I do not perform any fine-tuning. The model was run on an NVIDIA RTX 2070 Super GPU. Caption generation was handled using HuggingFace Transformers' 'generate()' method. All experiments used the same fixed subset of validation images from the dataset to ensure fair comparison across runs.

I designed three prompt sets (*basic*, *partial*, *descriptive*), each consisting of five distinct prompts. These were iteratively refined over three phases to reduce refusals and improve output relevance. The prompts used can be seen in Table I. Each image was processed with all prompt sets, generating 15 captions per image.

In addition, in an attempt to make the model generate meaningful and descriptive captions, a "word count prompt" was appended to the end of each of the final iteration prompts (e.g., " Use at least 10 words.").

Refusal responses (e.g., "Sorry, as a base VLM I am not trained...") were logged and used to quantify the impact of prompt wording on model behavior. Outputs were saved in structured JSON files, including prompts, generated captions, and reference captions for each image. This enables subsequent evaluation using standard caption quality metrics.

F. Results

The primary evaluation for Phase 2 focused on analyzing how different prompt sets and their refinements affected the image captions and model's willingness to produce responses. Refusal rates were used as the main metric, with a refusal defined as a generated caption consisting of fallback responses such as "Sorry, as a base VLM I am not trained to answer this question.", "unanswerable" or sometimes simply "no".

Each image was captioned 15 times (5 prompts × 3 sets) per iteration, and refusal counts were logged. Figure 1 presents the refusal rates per prompt set across four configurations: the three prompt iterations and an additional Word Count Prompt (WCP) condition appended to Iteration 3.

- **Iteration 1 (Initial Prompts):** 21.0% refusal rate (315/1500). Many prompts were vague or abstract, triggering generic denials.
- **Iteration 2 (Revised Prompts):** 28.9% refusal rate (434/1500). Some edits unintentionally resembled instruction-like phrasing.
- **Iteration 3 (Final Prompts):** Improving over the previous iteration as well as taking what worked better in the first iteration led to a **15.3%** refusal rate (229/1500), confirming that declarative and grounded phrasing improves success.
- **Iteration 3 + WCP:** Appending a word-count cue (e.g., "Use at least 10 words.") caused refusal rates to spike drastically—62.8% refusal rate (942/1500)—showing that the base PaliGemma model does not respond well to explicit instructions.

These results demonstrate that prompt engineering alone can meaningfully improve model reliability, and also highlight the limitations of instruction-free foundation models during inference.

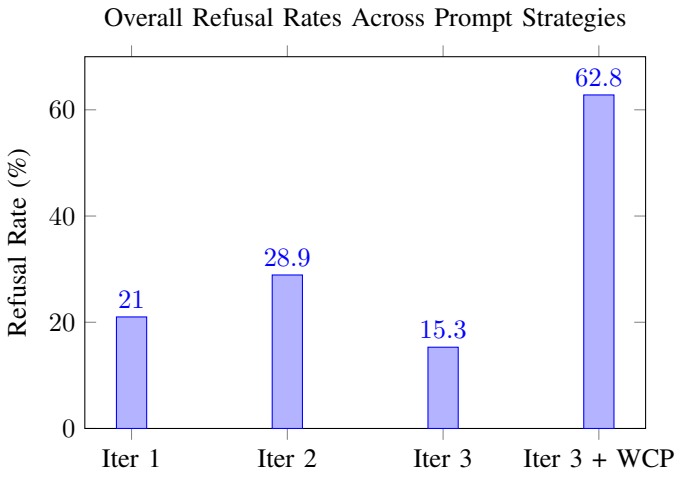

Fig. 1: Overall refusal rate per prompt iteration. Appending a word count instruction caused a large increase in refusals.

G. Phase 3

The remaining work for Phase 3 will focus on benchmarking the quality of generated captions and completing the ablation study. Specifically, I will:

- **Evaluate Caption Quality:** Using standard metrics (BLEU, METEOR, ROUGE-L, CIDEr), I will compare the generated captions from each prompt set and iteration against the five reference captions per image.
- **Analyze Decoding Strategies:** I will conduct ablation experiments on decoding parameters, particularly beam size and maximum token length, to assess their effect on caption diversity and specificity.

TABLE I: Prompt Variations Across Iterations

Prompt Set	Iteration 1	Iteration 2	Iteration 3
Basic	• Write a caption. • What do you see? • Summarize what's visible. • Describe any visible activity. • Describe the visual content of this image.	• Write a caption. • Write a description. • Describe the image. • Caption the image. • Add a caption.	• Write a caption. • Write a description. • Describe the image. • Caption the image. • Add a caption.
Partial	• What subjects are visible? • Describe the type of area shown. • What is the background? • What are the contents of this image? • What are the features of this image?	• Write about the subjects in this image. • Describe the type of area shown. • Describe the background. • What are the contents of this image? • What are the features of this image?	• What subjects are visible? • Describe the type of area shown. • What is the background? • What is shown in this image? • What are the features of this image?
Descriptive	• What does this satellite image show? • Write a detailed description. • What is shown in this aerial view? • Describe all visible elements. • What is happening in this scene?	• Caption this satellite image. • Write a detailed description. • Describe the contents of this image in detail. • Write a detailed caption. • Caption what is happening in this scene?	• Caption this satellite image. • Write a detailed description. • Describe this image in detail. • Write a detailed caption. • Caption what is happening in this scene?

- **Final Report and Visualization:** Results will be compiled with tables and visualizations comparing refusal rates, caption lengths, and metric scores across prompt versions.

The goal of Phase 3 is to consolidate all findings, quantify improvements from prompt engineering, and document effective inference strategies for VLMs in remote sensing applications.

H. Performance Metrics

I will evaluate the generated captions using standard metrics commonly used in image captioning tasks:

- **BLEU** (Bilingual Evaluation Understudy): Measures n-gram precision between candidate and reference captions.
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering): Considers both precision and recall, and includes stemming and synonym matching.
- **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation): Evaluates the longest common subsequence between predicted and reference captions.
- **CIDEr** (Consensus-based Image Description Evaluation): Designed specifically for image captioning; uses TF-IDF weighting to compare candidate captions to multiple references.

These metrics will be used to compare caption quality across different prompt sets and decoding configurations, using the original human-written captions as references.

IV. Conclusion

This project investigated how prompt design affects the performance of a frozen transformer-based vision-language model (PaliGemma) in generating captions for satellite imagery. Due to hardware constraints, I shifted from fine-tuning to an inference-only approach focused on prompt engineering and decoding strategies.

Through multiple iterations of prompt sets, I observed that simple rewording of prompts—making them more grounded and less instructional—significantly reduced refusal rates and improved caption success. Attempts to enforce longer captions via appended word count instructions led to a sharp increase in refusals, highlighting the model's lack of instruction-following capability.

These results show that meaningful performance improvements can be achieved in zero-shot settings through prompt refinement alone, without additional training or domain adaptation. This makes prompt engineering a practical and lightweight alternative to fine-tuning, especially in resource-limited environments. Final benchmarking and ablation studies in Phase 3 will further quantify the impact of prompt style and decoding configuration on caption quality.

References

[1] R. Sennrich et al., "Improving Neural Machine Translation Models with Monolingual Data," ACL, 2016.

[2] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.

[3] J. Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," ICML, 2020.

[4] B. Zhao et al., "A Systematic Survey of Remote Sensing Image Captioning," IEEE Access, 2021.

[5] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.