DI725 Project Phase III: Parameter-Efficient Fine-Tuning of PaliGemma for Image Captioning

İbrahim Ethem Deveci
Cognitive Sciences
METU Informatics Institute
Ankara, Turkey
ethem.deveci@metu.edu.tr

***Abstract*—This document presents the final report for the *Transformers and Attention-Based Deep Networks* course, centered on improving the image captioning capabilities of the PaliGemma vision-language model (VLM) through QLoRA-based fine-tuning with the RISC dataset. We report on the outcomes of Phase III, which builds upon prior results from the baseline and Phase II. In this phase, we applied two distinct hyperparameter configurations, each trained on the full training and validation sets. To systematically evaluate the effectiveness of the fine-tuned models, we conducted a series of 12 experiments on the complete test set: two models tested across three prompt types and two inference configurations. This phase specifically investigates the efficacy of low-rank adaptation techniques in enhancing model performance under constrained fine-tuning conditions.**

GitHub: GitHub repository of the project phase III.

***Index Terms*—parameter-efficient fine-tuning, quantized LoRA, image captioning, PaliGemma**

I. Introduction

The main question this project aimed to answer was whether parameter-efficient fine-tuning methods, when paired with evaluation criteria that reflect the functional demands of image captioning, can enhance the performance of VLMs such as PaliGemma beyond baseline levels, while maintaining minimal computational overhead. Given the large scale of PaliGemma, full fine-tuning is computationally expensive in practical environments. Therefore, the project focused on exploring parameter-efficient fine-tuning strategies [1], [2] as a scalable alternative for adapting large VLMs to domain-specific objectives.

In Phase II, Quantized Low-Rank Adaptation (QLoRA) [2] was introduced as a parameter-efficient fine-tuning method applied exclusively to the language model component, while freezing the vision tower and multi-modal projector parameters. This approach yielded significant improvements, demonstrating QLoRA's potential in adapting large-scale vision-language models under hardware constraints.

Building upon these results, Phase III explores the efficacy of QLoRA fine-tuning under two distinct hyperparameter configurations. Each configuration was trained on the full training and validation sets to maximize data utilization. To comprehensively evaluate the fine-tuned models, we conducted twelve experiments by testing two models across three prompt types and two inference configurations using the complete test set.

II. Dataset

A. RISC Dataset Overview

The RISC dataset consists of 44,521 remote sensing images (satellite imagery), each with a fixed resolution of 224×224 pixels. Every image is annotated with five distinct captions, resulting in a total of 222,605 captions that describe the visual content of the images. The dataset is split into training (35,614 images), validation (4,453 images), and test (4,454 images) sets.

B. Exploratory Data Analysis

Exploratory data analysis revealed that there are no missing image files. Caption lengths vary significantly, ranging from 4 to 50 tokens, with a mean length of approximately 11.10 tokens per caption. In terms of character length, captions range from 20 to 268 characters, with an average of 61.74.

However, the dataset presents several quality issues that must be addressed. Notably, 14,632 captions are exact duplicates, which can introduce redundancy. More critically, inconsistencies and contradictions exist among captions corresponding to the same image. These include mismatches in object counts and syntactically ill phrases. Such discrepancies can hinder model training and obscure evaluation validity.

A similarity analysis of the captions assigned to each image reveals considerable inconsistency, with a mean BLEU-4 score of 0.21, METEOR of 0.40, and cosine similarity of 0.33 across caption sets. These low scores indicate some divergence among the five captions per image, reinforcing the need for filtering strategies during preprocessing to enhance dataset coherence and training efficacy.

C. Caption Selection and Alignment

To optimize the training process and improve the alignment between visual and textual representations, the same approach as in Phase II was employed for Phase III. From the five available captions per image, a single caption was selected based on its semantic similarity to the corresponding visual content.

This selection utilized CLIPScore [3], which measures the cosine similarity between image embeddings and text embeddings using ViT-B/32 model. For each image, the caption with the highest CLIPScore was retained. In instances where

multiple captions had identical top scores, one was chosen at random.

By reducing each image's annotations to a single, high-quality caption, this approach mitigated noise introduced by duplicated or semantically inconsistent captions and enforced a tighter alignment between visual input and textual supervision. Selecting the caption with the highest CLIPScore enabled both a reduction in computational overhead and an improvement in training efficiency. This strategy rests on the assumption that higher CLIPScores indicate stronger correspondence between image and text, thereby yielding a more informative learning signal.

III. Modeling

PaliGemma is a VLM that employs a unified encoder-decoder architecture designed for multi-modal tasks [4]. The model accepts image inputs together with natural language prompts and generates textual outputs in an autoregressive manner. This flexible framework supports various applications including image captioning, classification, and visual question answering. The architecture consists of three primary components: a SigLIP image encoder specialized for visual feature extraction, a decoder-only language model based on the Gemma architecture, and a linear projection layer that aligns image embeddings with the language model's token embedding space.

For fine-tuning, we utilize a custom dataset class built on PyTorch's `Dataset` interface, designed to handle the RISC dataset images and captions. The `RISCDataset` class loads each image-caption pair, applies preprocessing through the `PaliGemmaProcessor`, and prepares the model-ready inputs. Each training sample consists of a natural language prompt ("Describe this image in detail:"), the associated image, and a caption. These elements are tokenized and formatted into input sequences with appropriate attention masks. Tokens not intended to contribute to the loss are masked using the label value `-100`, ensuring that only relevant parts of the sequence influence model updates.

Fine-tuning is performed using parameter-efficient QLoRA. We configure the model to load in 4-bit precision with nf4 quantization scheme and set computation to the `bfloat16` datatype for efficiency. Two different LoRA hyperparameter configurations were evaluated, both targeting the key projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) within the language model. The configurations varied in rank, alpha, and dropout values to assess their impact on fine-tuning performance.

The pretrained PaliGemma model is loaded with quantization and device mapping specified. The vision tower and multi-modal projector parameters are frozen to reduce training overhead, while LoRA modules are applied exclusively to the language model components. This approach allows fine-tuning to be performed efficiently without updating the entire network, which is critical given the model's size and computational constraints. The hyperparameter choices and LoRA configurations are detailed in Table I.

Hyperparameter	Configuration 1	Configuration 2
`train_batch_size`	1	1
`eval_batch_size`	1	1
`grad_acc_steps`	4	4
`warmup_steps`	500	1000
`learning_rate`	1×10^{-5}	1×10^{-6}
`weight_decay`	1×10^{-6}	1×10^{-6}
`adam_beta2`	0.999	0.999
`optimizer`	adamw_8bit	adamw_torch
`num_workers`	4	2
`LoRA rank (r)`	4	64
`LoRA alpha`	16	256
`LoRA dropout`	0.1	0

TABLE I
TRAINING AND LORA HYPERPARAMETER CONFIGURATIONS USED FOR PHASE III FINE-TUNING.

IV. Evaluation

To assess the performance of the QLoRA fine-tuned PaliGemma models, we conducted a total of 12 experiments. These experiments were structured by evaluating two distinct models across three prompt types and two inference configurations.

Performance was quantitatively measured using the same metrics employed in Phase II to ensure consistency and comparability: BLEU-1, BLEU-2, METEOR, and CLIPScore. These metrics collectively evaluate the linguistic quality and semantic alignment of the generated captions. CLIPScore values are reported in their raw form without any normalization or adjustment. All comparisons across models and configurations rely directly on these scores.

The results of the two fine-tuned models on these metrics were compared against the baseline model and the two prior models obtained from Phase II.

Overall, the Phase III models demonstrate consistent improvements over the Phase II models across all evaluation metrics. These gains reflect enhanced n-gram precision, better semantic alignment with reference captions, and stronger consistency between generated captions and the corresponding image content. Detailed quantitative results are presented in Section V.

This thorough evaluation framework validates the effectiveness of low-rank adaptation under constrained fine-tuning conditions and guides further refinements in model configuration and inference strategies.

V. Results

This section presents the quantitative outcomes of the Phase III experiments. We report evaluation scores in 12 distinct configurations derived from two QLoRA-fine-tuned models, three prompt types, and two inference settings. The results are compared with the baseline model and the two QLoRA models developed during Phase II.

A. Performance Overview and Baseline Comparison

Table II presents the evaluation scores for the baseline model and the two Phase II models fine-tuned using QLoRA with distinct hyperparameter configurations. The first configuration employed a higher rank and alpha (r = 128, alpha =

512), while the second used more compact settings (r = 64, alpha = 256).

The results clearly indicate that both fine-tuned models surpass the baseline across most metrics. Notably, the second QLoRA configuration achieves substantial gains. These findings confirm that QLoRA can significantly enhance the performance of large-scale VLMs without modifying the vision or projection components.

Metric	Baseline	Conf. 1	Conf. 2
BLEU-1	0.0424	0.0372	0.2135
BLEU-2	0.0208	0.0197	0.0764
METEOR	0.0427	0.1088	0.2769
CLIPScore	0.2925	0.2833	0.2667

TABLE II
COMPARISON OF BASELINE AND PHASE II QLORA MODELS ACROSS EVALUATION METRICS.

B. Experimental Results Across Prompts and Inference Modes

Table IV presents a detailed breakdown of the twelve experimental runs. Each model configuration was evaluated using three distinct prompt formulations to systematically assess the impact of prompt variation on inference quality.

- **A**: "Describe this image in detail:"
- **B**: "Describe this image:"
- **C**: "What do you see in this image:"

Furthermore, two decoding strategies were employed for generation, summarized in Table III. Here, greedy decoding selects the most probable token at each step, while beam search with sampling introduces controlled randomness and diversity.

Parameter	Greedy	Beam + Sampling
num_beams	1	4
do_sample	False	True
top_p	–	0.9
top_k	–	50
temperature	–	0.8
early_stopping	–	True
length_penalty	–	1.0
no_repeat_ngram_size	–	2
max_new_tokens	16	16

TABLE III
DECODING STRATEGIES USED DURING INFERENCE.

Conf.	Prompt	Decoding	BLEU-1	BLEU-2	METEOR	CLIPScore
1	A	G	0.3754	0.2314	0.3549	0.2848
1	A	B+S	0.3890	0.2492	0.3670	0.2848
1	B	G	0.3580	0.2057	0.3232	0.2831
1	B	B+S	0.3662	0.2143	0.3411	0.2868
1	C	G	0.0030	0.0016	0.0023	0.2151
1	C	B+S	0.0555	0.0283	0.0446	0.2303
2	A	G	0.3598	0.2071	0.3288	0.2817
2	A	B+S	0.3647	0.2104	0.3331	0.2844
2	B	G	0.2946	0.1413	0.2496	0.2792
2	B	B+S	0.3053	0.1515	0.2728	0.2872
2	C	G	0.0318	0.0169	0.0302	0.2428
2	C	B+S	0.0805	0.0393	0.0678	0.2477

TABLE IV
EVALUATION RESULTS FOR 12 EXPERIMENTAL SETUPS ACROSS CONFIGURATION, PROMPT TYPE (A/B/C), AND DECODING (G: GREEDY, B+S: BEAM+SAMPLE).

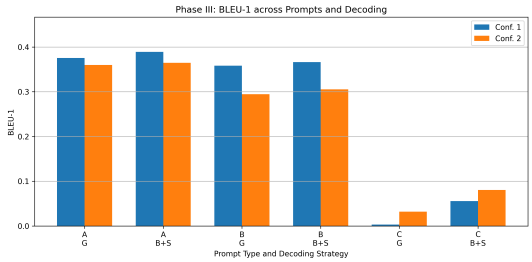

Fig. 1. Phase III BLEU-1 scores across prompt types and decoding strategies for configurations 1 and 2.

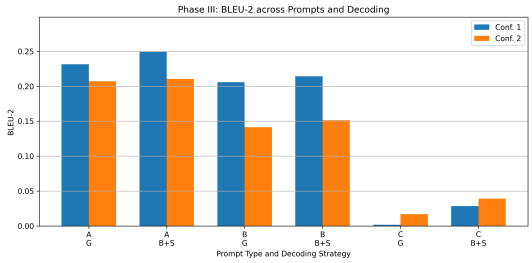

Fig. 2. Phase III BLEU-2 scores across prompt types and decoding strategies for configurations 1 and 2.

VI. DISCUSSION

An important observation from Table IV and the Figures, where each showing models performances across individual evaluation metrics, is that Configuration 1, despite using significantly fewer trainable parameters, consistently outperformed Configuration 2. This outcome indicates that more aggressive parameterization does not inherently lead to better performance. On the contrary, it may be detrimental if not properly regularized or aligned with the model's learning dynamics.

Phase III experiment results also shows how prompt formulation and decoding strategies interact with model architecture. Prompt A ("Describe this image in detail:") consistently yields the highest BLEU and METEOR scores across configurations, likely due to its exact match with the prompt used during fine-tuning. This alignment may have reinforced the model's ability to respond optimally to this specific instruction.

The decoding method also impacts quality. Beam search with sampling generally leads to improved scores compared to greedy decoding, likely due to the added diversity and flexibility it introduces. While differences are modest, they align with expectations that stochastic decoding methods allow models to escape repetitive or generic outputs.

However, a notable pattern emerges in the Prompt C condition ("What do you see in this image:"), where BLEU and METEOR scores sharply decline, often approaching zero, indicating that the generated captions are frequently irrelevant or vacuous. Despite this poor linguistic quality, CLIPScore values remain consistently elevated, even for predictions that are not genuine descriptions, such as generic or interrogative

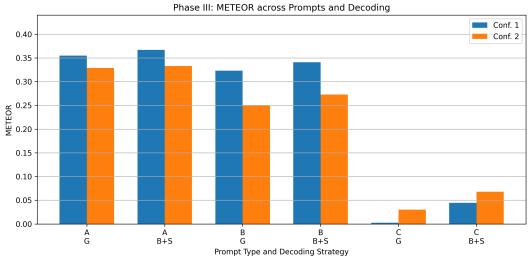

Fig. 3. Phase III METEOR scores across prompt types and decoding strategies for configurations 1 and 2.

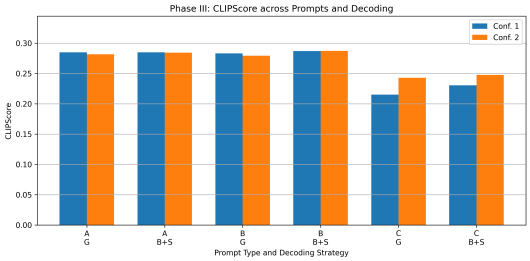

Fig. 4. Phase III CLIPScore values across prompt types and decoding strategies for configurations 1 and 2.

phrases like "What do I see here?" which nonetheless receive relatively high CLIPScore ratings.

This inconsistency highlights a significant limitation of CLIPScore as a standalone measure of caption quality. The metric appears to reward superficial or loosely associated image-related terms that align with the visual embedding, without ensuring true semantic or contextual relevance. As such, CLIPScore should be interpreted with caution and used only as a complementary metric in conjunction with established linguistic evaluations like BLEU and METEOR, and ideally supplemented by human judgment.

The following examples illustrate this phenomenon, presenting the reference caption, the generated prediction, and their corresponding unnormalized and normalized (between 0-100) CLIPScore values:

- **Reference:** There are a few bunkers and a small river and a few paths and many trees on a large green lawn on the golf course.
  Prediction: picture of a golf course
  CLIPScore (unnormalized / normalized): 0.325 / 0.787
- **Reference:** The rectangular farmland is next to the house.
  Prediction: image of a forest of trees
  CLIPScore (unnormalized / normalized): 0.250 / 0.573
- **Reference:** It is a pretty long straight bridge connects the two banks.
  Prediction: picture of a bridge over a river
  CLIPScore (unnormalized / normalized): 0.299 / 0.712
- **Reference:** The cross-line overpass is over the roads in the forest and an open area is next to the overpass.
  Prediction: picture of a road and a bridge
  CLIPScore (unnormalized / normalized): 0.308 / 0.736
- **Reference:** Many buildings are in an industrial area.
  Prediction: What are you looking at?
  CLIPScore (unnormalized / normalized): 0.203 / 0.439
- **Reference:** The church is surrounded by some buildings.
  Prediction: What do I see here?
  CLIPScore (unnormalized / normalized): 0.233 / 0.522

A. Limitations

a) Dataset Caption Usage: While the dataset was filtered to select a single caption per image based on CLIPScore to reduce noise, this approach may discard valuable linguistic diversity. Training with the full set of captions could potentially improve generalization, albeit at the cost of introducing more noise and requiring more sophisticated handling methods.

b) Evaluation Metric Interpretation: The observed divergence between CLIPScore and n-gram metrics highlights the challenges in selecting reliable evaluation criteria. While CLIPScore is designed to capture semantic alignment with the image, it often yielded inflated values for outputs that are linguistically or contextually poor. A notable example arises under the Prompt C condition ("What do you see in this image:"), where BLEU and METEOR scores sharply decline due to irrelevant or vacuous predictions. Despite this, CLIPScore remains consistently elevated, even for generic or interrogative predictions such as "What do I see here?" Consequently, CLIPScore should be interpreted with caution and used only as a complementary signal alongside traditional n-gram-based metrics and, where feasible, human evaluation.

c) Validation Metric Integration: Incorporating evaluation metrics directly into the validation process for model checkpoint selection would enhance performance reliability. However, attempts to compute these metrics during validation caused GPU memory overloads, necessitating reliance on the evaluation loss alone for model selection. Future work should explore more memory-efficient validation schemes or metric approximations to address this limitation.

VII. Conclusion

This work demonstrates the feasibility and effectiveness of parameter-efficient fine-tuning of large-scale VLMs like PaliGemma using QLoRA, particularly for the task of image captioning in resource-constrained environments. Phase III systematically explored the impact of two distinct hyperparameter configurations across varied prompt and decoding strategies. The results confirm that carefully selected QLoRA configurations yield substantial improvements over both the baseline and prior Phase II models without requiring updates to the vision encoder or projector.

Despite these gains, certain limitations remain. First, the relatively modest CLIPScore improvements suggest that while the linguistic quality of the captions has improved, deeper semantic alignment with the visual input could benefit from targeted enhancements to cross-modal representations. Second, the performance degradation observed under prompt type C points to sensitivity in prompt formulation, indicating the need for prompt optimization or prompt tuning. Third, the inference

quality varies across decoding strategies, highlighting the importance of selecting generation parameters aligned with task objectives.

Future work should address these challenges by (i) experimenting with prompt tuning or instruction-based pretraining to increase robustness across prompt types, (ii) incorporating lightweight adaptation layers into the vision or projection modules to enhance cross-modal alignment, (iii) exploring larger-scale ablations across LoRA parameters and decoding settings to more precisely map the performance landscape, and (iv) leveraging full reference captions during training to introduce greater linguistic diversity. Additionally, evaluating on more diverse or domain-specific remote sensing datasets could further validate the generalizability of these methods.

Ultimately, this project illustrates a scalable path forward for adapting VLMs to specialized captioning tasks without the computational burden of full fine-tuning.

REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: efficient finetuning of quantized llms," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, (Red Hook, NY, USA), Curran Associates Inc., 2023.

[3] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022.

[4] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, "Paligemma: A versatile 3b vlm for transfer," 2024.