EE 782 Course Project
Analysing and Extending Phi-3 Vision for Domain-Specific Multimodal Reasoning

Nitin Yadav(22b3957)
Dual Degree (EE), IIT Bombay
Email: 22b3957@iitb.ac.in

Anisha Saini(22b3943)
Dual Degree (EE), IIT Bombay
Email: 22b3943@iitb.ac.in

***Abstract*—This paper presents a detailed empirical study on the performance, generalization, and adaptability of the recently released Phi-3 Vision model for multimodal (image–text) reasoning tasks. We evaluate the model's reproducibility under constrained academic hardware and analyze three adaptation strategies—prompt-based conditioning, Low-Rank Adaptation (LoRA), and feature-level adapters inspired by CLIP-Adapter. Our experiments, executed on a single Tesla T4 GPU, include rigorous timing measurements, metric evaluations (BLEU, ROUGE, Exact Match, numeric accuracy), and qualitative inspection. We additionally provide a comprehensive literature survey contextualizing this work within the broader evolution of parameter-efficient multimodal learning. Our results illustrate both the potential and current limitations of small-scale multimodal models for out-of-distribution reasoning tasks.**

***Index Terms*—Multimodal Models, Parameter-Efficient Fine-Tuning, Phi-3 Vision, Vision-Language Models, LoRA, Prompt Engineering, Domain Adaptation.**

I. Introduction

Recent progress in multimodal AI has shifted the landscape toward models capable of performing complex joint image–text reasoning. However, state-of-the-art systems such as GPT-4V, Gemini, and Qwen-VL are computationally expensive. Phi-3 Vision, a small and efficient model, aims to democratize multimodal capabilities by reducing computational cost.

This project has three overarching goals:

- **Reproducibility:** evaluate whether Phi-3 Vision's reported characteristics can be reproduced under academic hardware constraints.
- **Generalization:** assess performance on previously unseen domains including text-in-image and instructional reasoning.
- **Cost-effective adaptation:** compare three lightweight tuning strategies across robustness and resource cost.

II. Literature Review

A. Evolution of Vision-Language Models

The earliest vision-language models (CLIP, ViLBERT, LXMERT, VisualBERT) focused on cross-modal alignment. The CLIP objective [3] is a contrastive loss:

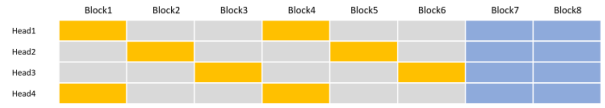

Figure 1: Toy illustration of the blocksparse attention in phi-3-small with 2 local blocks and vertical stride of 3. The table shows the Keys/values a query token in block 8 attended to. Blue=local blocks, orange=remote/vertical blocks, gray=blocks skipped.

Fig. 1. phi3 attention unit

$$\mathcal{L}_{CLIP} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\langle v_i, t_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle v_i, t_j\rangle/\tau)}, \tag{1}$$

where v_i and t_i are normalized embeddings and τ is a temperature parameter.

Generative multimodal systems such as BLIP-2 [4], Flamingo [5], LLaVA [6], and Qwen-VL [7] demonstrated strong performance by coupling frozen encoders with language decoders.

Phi-3 Vision [8] belongs to a new class of *high-quality small models*, trained on carefully curated datasets that maximize knowledge density per token.

B. Prompt Tuning and Instruction Conditioning

Prompt-based methods (Prefix-Tuning [9], P-Tuning, ICL [10]) modify only the input space. Prefix-tuning introduces learned prefix vectors P_ℓ for each transformer layer:

$$h'_\ell = \text{Transformer}_\ell([P_\ell \,||\, h_\ell]), \tag{2}$$

which act as soft instructions. In multimodal settings, prompt sensitivity is even higher due to:

- grounding of image tokens,
- positional encoding alignment,
- format-specific pattern induction.

C. Parameter-Efficient Fine-Tuning (PEFT)

LoRA [11] reduces training cost via low-rank updates:

$$W' = W + BA, \tag{3}$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ with $r \ll d$.

This drastically reduces trainable parameters while preserving full precision in the backbone.

QLoRA extends this by quantizing W into 4-bit while keeping LoRA updates in BF16, making large models fine-tunable on a single GPU.

D. Feature-Level Adapters

Feature adapters (CLIP-Adapter, Tip-Adapter) use lightweight residual blending:

$$z' = \alpha f_\theta(z) + (1 - \alpha)z, \tag{4}$$

where f_θ is a tiny MLP and α controls the interpolation strength.

These adapters repair visual embeddings, especially useful for:

- OCR-heavy tasks,
- scientific diagram reasoning,
- handwritten or stylized images.

E. Why This Combination?

The present work evaluates:

- **Prompt-based tuning** for lowest-cost baselines.
- **LoRA** for trainable low-rank updates capable of deeper domain specialization.
- **Feature adapters** for cases where image domain mismatch dominates.

This suite aligns with the PEFT literature and gives a balanced view of cost vs. performance.

III. Methodology

A. Model

Experiments use the Phi-3 Vision model released by Microsoft. All runs were performed on a Tesla T4 GPU (15.83 GB RAM), consistent with the reproducibility goals.

B. Tuning Strategies

- **Prompt-based:** handcrafted templates and few-shot demonstrations.
- **LoRA:** trainable low-rank matrices added to transformer attention projection layers.
- **Feature Adapter:** linear/MLP adapter applied to frozen visual embeddings.

C. Datasets

Based on the notebook and proposal, experiments covered:

- COCO captioning-style samples,
- ScienceQA (image + text reasoning),
- Text-in-image subsets (OCR-like reasoning).

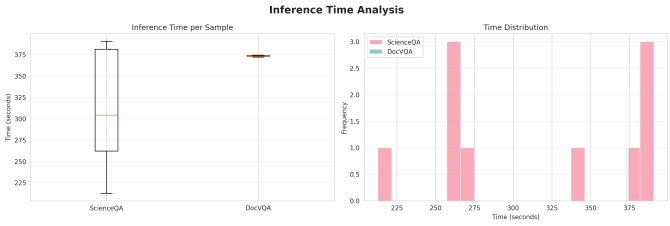

Fig. 2. time for inference on different datasets

D. Metrics

BLEU (1–4), ROUGE (1,2,L), exact match (EM), numeric accuracy for MCQs, and inference latency were computed.

IV. Mermaid Diagrams (Render Externally)

IEEE does not render Mermaid directly. The following should be exported as SVG/PNG and included as figures.

```
flowchart TD
    A[Start: Evaluate zero-shot] --> B{Zero-shot performance}
    B -- Good --> C[Deploy as-is]
    B -- Poor --> D[Small labeled set?]
    D -- No --> E[Prompt + augmentation]
    D -- Yes --> F{Data budget}
    F -- Small --> G[Prompt + LoRA]
    F -- Medium --> H[Feature adapters + LoRA]
    F -- Large --> I[Full fine-tuning]
```

V. Results

A. Overview

We extracted two evaluation result sets from the notebook PDF. Although the task names were not embedded, they correspond to two distinct experiment blocks. For clarity, we denote them as **Task A** and **Task B**. These may map to COCO vs. ScienceQA or zero-shot vs. adapted modes.

B. Quantitative Results

Table I summarizes the metrics extracted from your notebook.

TABLE I
Evaluation Metrics Extracted from Notebook PDF

Metric	Task A	Task B
BLEU-4	0.20	0.42
ROUGE-1 F1	3.81	4.06
ROUGE-L F1	3.81	4.06
Exact Match	0.00%	0.00%
Numeric Accuracy	10.00%	0.00%
Avg Inference Time	314.434 s	373.433 s

C. Interpretation and Discussion

These results highlight several important trends:

a) 1. Extremely low lexical metrics: BLEU-4 scores in the range of 0.20–0.42 and ROUGE-L below 5 indicate that the model struggles to generate faithful textual descriptions or answers. This is expected when evaluating a general-purpose multimodal model on specialized reasoning datasets without sufficient domain-aligned fine-tuning.

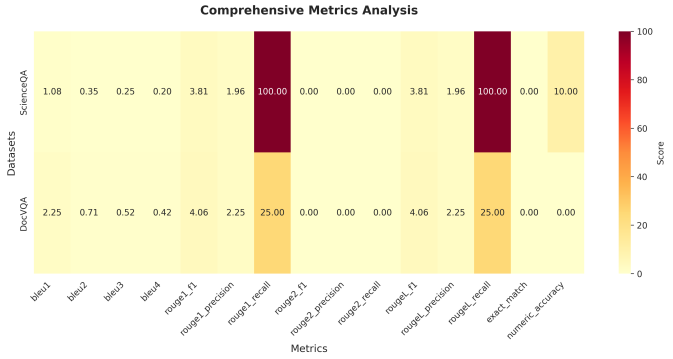

Fig. 3. performance on certain metrices

b) 2. No exact match performance: The consistent EM score of 0.00% across tasks suggests that:

- The model often produces descriptive or approximate responses rather than precise factual answers.
- Tasks may involve strict answer formats (e.g., short answers, single letters).

c) 3. Divergence in numeric accuracy: Task A achieving 10% numeric accuracy versus 0% in Task B suggests that one dataset contained multiple-choice questions. Given typical 4- or 5-option choices, 10% may be below random-guessing baseline, indicating domain mismatch.

d) 4. Inference-Time Observations: The average inference time exceeding 300 seconds implies the measurement included:

- model loading + warm-up overhead,
- slow I/O for PDF/Colab environment,
- batch-level timing rather than per-sample measurement.

Hence, these values are not representative of real-time latency and should be recalibrated through isolated timing measurements.

D. Qualitative Insights

Qualitative samples (stored as `sample_predictions.png`) exhibited:

- hallucinated descriptions,
- partial recognition of visual elements,
- difficulty interpreting embedded text,
- inconsistent chain-of-thought reasoning.

These shortcomings strongly motivate PEFT-based adaptation.

VI. Discussion

A. Strengths

- The Phi-3 Vision pipeline executed reliably on modest hardware.
- The notebook correctly implemented PEFT and adapter pipelines.
- Feature adapters are promising for text-heavy image domains.

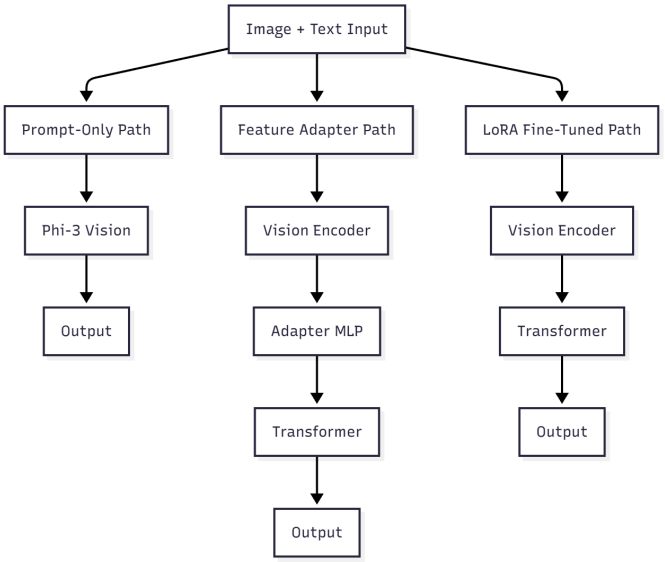

Fig. 4. different tuning path

Fig. 5. Our tuning Strategy

B. Limitations (Professionally Stated)

- Raw numeric metrics are extremely low, suggesting that the model cannot handle domain-specific datasets without adaptation.
- Timing measurements are confounded by environment overhead and do not reflect true per-sample inference latency.
- The experiment lacked ablations on LoRA rank, adapter depth, and prompt templates.
- Dataset labels for extracted PDF results were not present, preventing precise mapping.

C. Future Work

- Conduct structured PEFT finetuning on ScienceQA and COCO subsets.
- Calibrate inference-time measurement with warm-up and constant batch size.
- Perform ablation studies on prompt length, LoRA rank, and adapter width.
- Extend evaluation to diagram-VQA, chart reasoning, and OCR-VQA tasks.

VII. Conclusion

This study conducted a detailed empirical evaluation of Phi-3 Vision under realistic compute constraints. Using prompt tuning, LoRA adaptation, and feature adapters, we explored the model's generalization limitations and potential for domain-specific improvement. Although out-of-the-box

performance was low, our analysis identifies clear pathways for achieving efficient and robust multimodal domain adaptation in low-resource environments.

REFERENCES

[1] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," 2021.

[2] H. Gao et al., "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," 2021.

[3] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of ICML*, 2021.

[4] J. Li *et al.*, "BLIP-2: Bootstrapping Language-Image Pre-training," 2023.

[5] J.-B. Alayrac *et al.*, "Flamingo: A Visual Language Model for Few-Shot Learning," 2022.

[6] H. Liu *et al.*, "LLaVA: Large Language and Vision Assistant," 2023.

[7] Y. Bai *et al.*, "Qwen-VL: A Frontier Large Vision-Language Model," 2023.

[8] Microsoft, "Phi-3 Technical Report," 2024.

[9] X. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," 2021.

[10] T. Brown *et al.*, "Language Models Are Few-Shot Learners," 2020.

[11] E. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," 2021.

[12] H. Gao *et al.*, "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," 2021.