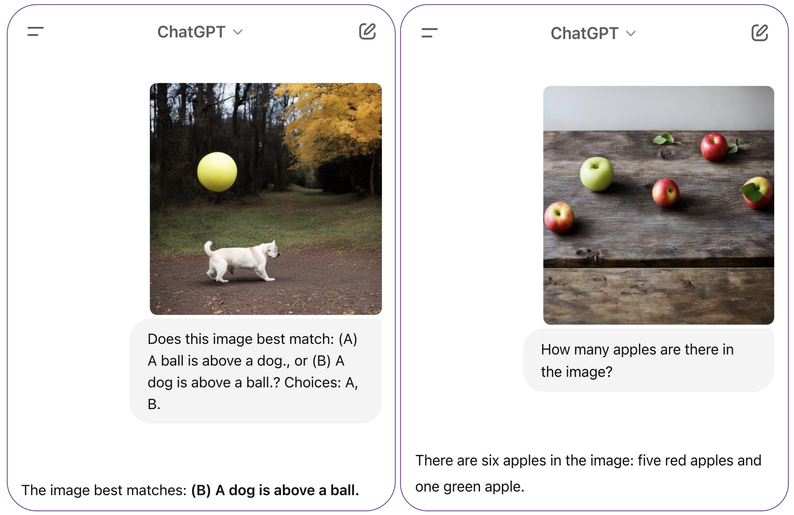
In-Context Learning for Fine-Grained Visual Understanding

Morteza Mahdiani, Alireza Farashah
{morteza.mahdiani, alireza.farashah}@mila.quebec

Mila
Université de Montréal
McGill University

Introduction

RQ1: Does ICL improve visual understanding in MLLMs?
RQ2: How does ICL compare to fine-tuning on VisMin Dataset?
RQ3: What are ICL failure modes for reasoning?



Methodology

Text Query
<**image_0**> Does this image best match: (**A**) A picture of a bridge with an amusement park in the background., or (**B**) A picture of a bridge with a golf course in the background.? Choices: **A**, **B**.
Image Query
<**image_0**> <**image_1**> Which image better aligns with the description: "A picture of a bridge with an amusement park in the background."? Choices: **First**, **Second**.

$$p_\theta(y_{test} \mid x_{test}, \mathcal{C}) = f_\theta\left(x_{test}, \{(x_i, y_i)\}_{i=1}^{8}\right)$$

LlaVa Idefics GPT4o

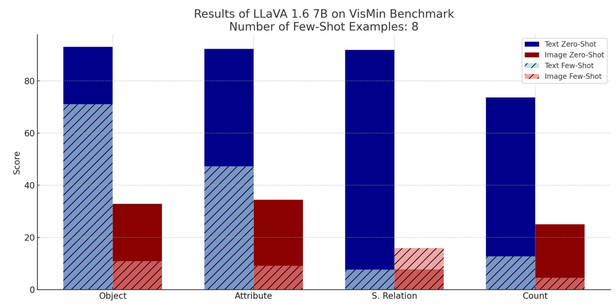
Evaluation

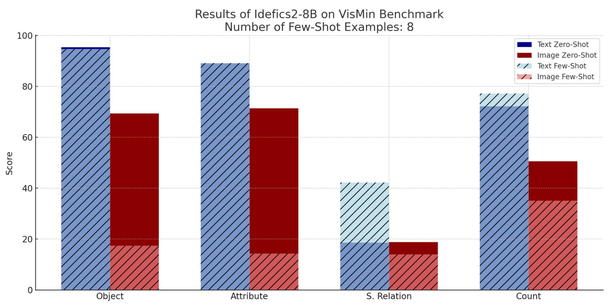
- I_0, I_1 — Image 0 and Image 1
- C_0, C_1 — Caption 0 and Caption 1
- $s(C, I)$ — Predicted Score

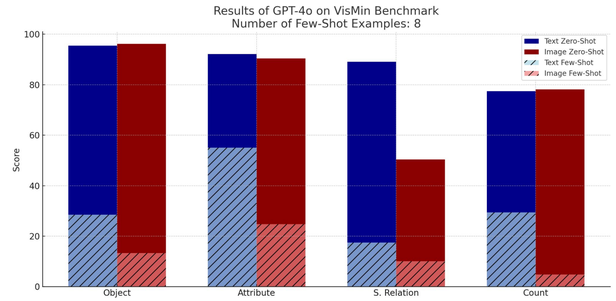
$$f(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \quad \text{and } s(C_1, I_1) > s(C_0, I_1) \\ 0 & \text{otherwise} \end{cases}$$

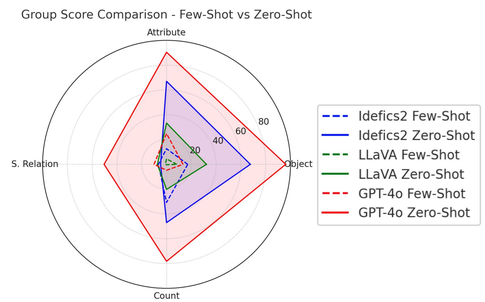
$$g(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \quad \text{and } s(C_1, I_1) > s(C_1, I_0) \\ 0 & \text{otherwise} \end{cases}$$

Experimental Results



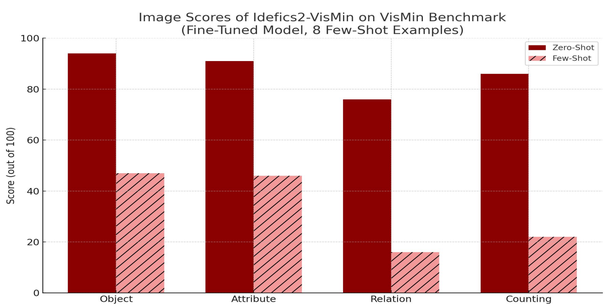


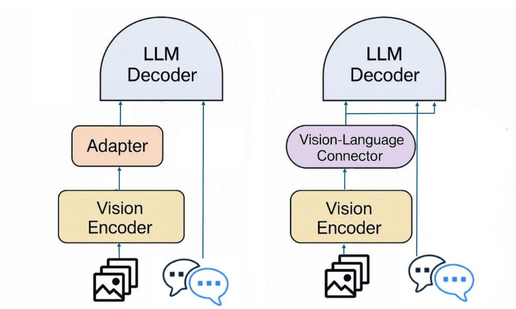


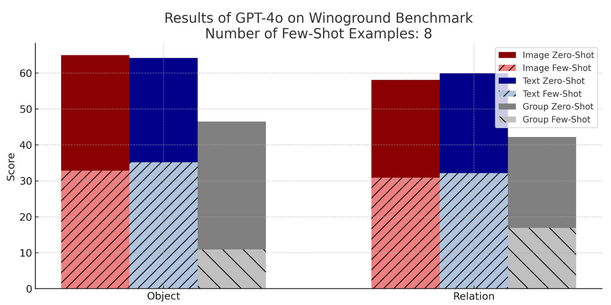


Challenges & Future Direction

- Fine-tuned encoder improves performance
- Few-shot examples can reduce performance
- Reasoning capabilities of VLMs are still limited







Code & Resources