Introduction

Core Capabilities

- Unlike traditional computer vision models, VLMs are not bound by a fixed set of classes or a specific task such as classification or detection.
- Retrained on a vast corpus of text and image / video caption pairs, VLMs can be trained in natural language and used to handle many classic vision tasks plus new AI-powered generative tasks such as summarization and visual question answering.

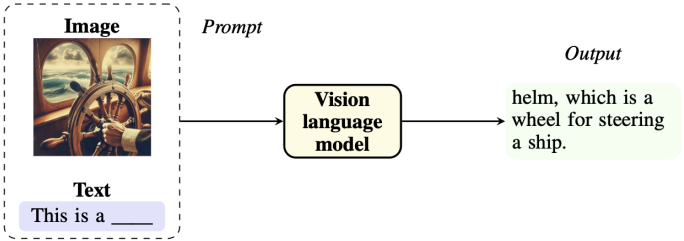

Figure 3: Illustrations of Visual Question Answering