Enhancing Vision-Language Models for Mechanical Tool Understanding

Adithya Kameswara Rao Srecharan Selvam
Carnegie Mellon University
{akameswa, sselvam}@andrew.cmu.edu

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in general visual understanding tasks, yet their application to specialized technical domains remains limited. This course project proposal aims to evaluate VLMs for mechanical tool recognition, usage instruction, and safety guidance—critical capabilities for industrial settings where proper tool handling directly impacts workplace safety and operational efficiency. We propose to build a dataset of mechanical tool images spanning 13 categories (wrenches, hammers, pliers, screwdrivers, bolts, dynamometers, testers, tool boxes, tape measures, ratchets, drills, calipers, and saws) and develop a comprehensive evaluation framework measuring not only conventional accuracy metrics but also safety-critical response characteristics and instruction quality. Our study would evaluate several state-of-the-art VLMs including Qwen2-VL-7B-Instruct, Phi-3-vision-128k-instruct, Llama-3.2-11B-Vision-Instruct, SmolVLM, and PaliGemma on their ability to recognize tools and provide appropriate usage guidance. This proposed research would contribute methodological insights for assessing domain-specific VLM capabilities and explore potential applications for industrial safety and training.

1. Introduction

In industrial and professional settings, accurate identification and understanding of mechanical tools are essential for operational safety and efficiency. Industries ranging from construction and manufacturing to specialized maintenance rely on proper tool usage to prevent accidents and ensure productivity. Improper tool usage can lead to equipment damage, project delays, and reduced operational efficiency, making reliable visual recognition systems valuable for workplace environments.

Contemporary Vision-Language Models (VLMs) have demonstrated remarkable capabilities across diverse domains, from medical diagnostics to artistic creation [6, 7]. These models integrate visual perception with language understanding, enabling them to process images and generate contextually relevant textual responses. However, their application to specialized technical domains, particularly mechanical tool recognition and instruction, remains underexplored. Current VLMs often lack the domain-specific knowledge required to provide precise guidance on specialized tool usage and associated safety protocols.

This proposed course project seeks to address this gap by evaluating VLMs' capabilities in the mechanical tools domain. We would focus on developing a dataset of mechanical tools with annotations to better understand how state-of-the-art VLMs can be evaluated for this specialized domain.

Our project would focus on 13 core tool categories: wrenches, hammers, pliers, screwdrivers, bolts, dynamometers, testers, tool boxes, tape measures, ratchets, drills, calipers, and saws. By evaluating VLMs for this domain, we aim to understand their capabilities in providing contextually appropriate, technically accurate, and safety-conscious guidance when presented with visual tool inputs. Such systems could have significant real-world applications in training programs, workplace safety enforcement, and technical assistance platforms.

The proposed contributions of this project include: (1) comparative analysis of different VLMs' performance on tool recognition tasks, (2) insights into the potential applications of VLMs in industrial safety contexts.

2. Technical Approach

2.1. Dataset

A comprehensive custom dataset would be curated for this study, comprising over 1,000 high-resolution images of mechanical tools. This collection would feature approximately 20 images per tool category at minimum, including wrenches, hammers, pliers, screwdrivers, bolts, dynamometers, testers, tool boxes, tape measures, ratchets, drills, calipers, and saws. Each image would be captured under controlled conditions with varying angles, distances, and lighting scenarios to ensure model robustness against real-world variability.

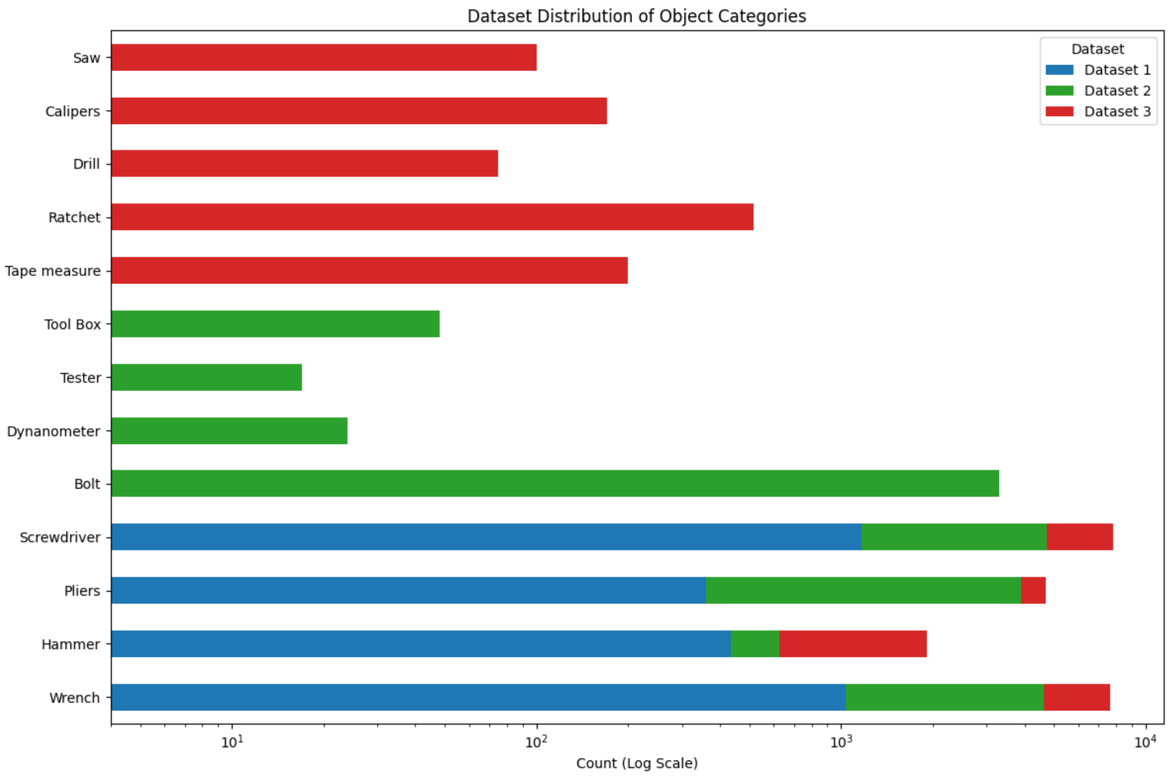

Figure 1. Distribution of mechanical tool categories in our custom dataset, compiled from three distinct sources. The balanced representation across different tool types would ensure comprehensive coverage for industrial applications.

The dataset's distinguishing feature would be its rich annotation scheme. Each image would be paired with detailed textual descriptions that include precise usage instructions, safety protocols, and maintenance guidelines.

2.2. Model Selection

For this study, we propose to evaluate several state-of-the-art Vision-Language Models (VLMs) based on their reported capabilities in multimodal understanding:

- **Qwen2-VL-7B-Instruct** [4]: Developed by Alibaba, a 7B parameter instruction-tuned multimodal model.
- **Phi-3-vision-128k-instruct** [3]: Microsoft's vision-capable model with a 128K token context window.
- **Llama-3.2-11B-Vision-Instruct** [2]: Meta's 11B parameter multimodal model.
- **SmolVLM** [8]: A lightweight VLM optimized for efficient deployment.
- **PaliGemma** [5]: Google's multimodal model built on the Gemma architecture.

These models represent diverse architectural approaches and parameter scales, allowing us to systematically evaluate their performance on mechanical tool understanding tasks.

3. Evaluation Plan

3.1. Performance Metrics

Our planned evaluation framework would employ a set of quantitative and qualitative metrics designed to assess multiple dimensions of model performance:

- **Tool Recognition Accuracy:** Precision, recall, and F1 scores for correctly identifying tools across different viewing angles, lighting conditions, and occlusion levels.
- **Instruction Quality:** A superior model would be used to evaluate the quality of instructions generated by the tested VLMs. This approach would help us assess instruction completeness, correctness, and clarity without requiring extensive human annotation.
- **Safety Guideline Adherence:** A superior model would implement a custom safety-weighted scoring system that evaluates whether critical safety information is included in model responses.

We would analyze how different models perform on these metrics to understand their strengths and limitations for technical tool understanding tasks.

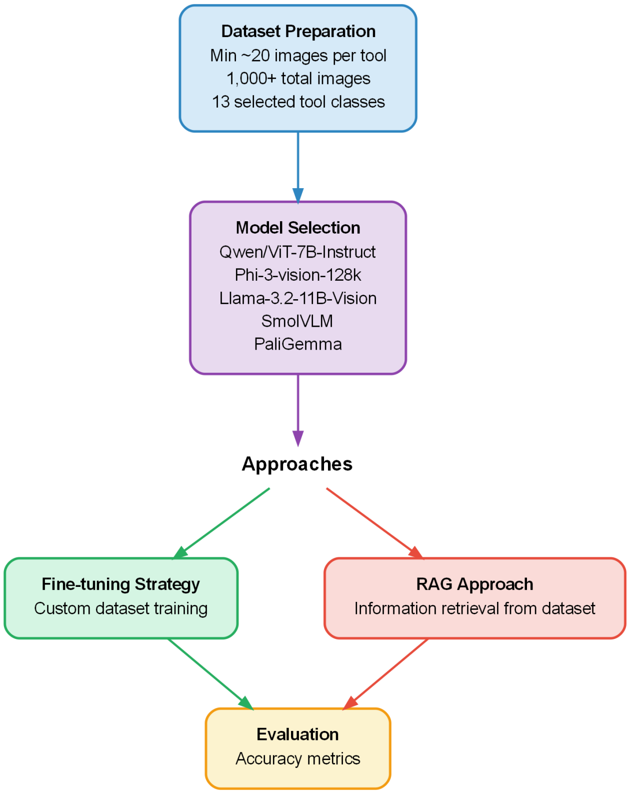

Figure 2. Proposed pipeline showing the data processing, model assessment, and analysis workflow for evaluating VLMs on mechanical tool understanding tasks.

3.2. Design

Our proposed evaluation methodology would follow a structured approach:

1. **Baseline Assessment:** Evaluating each model's performance on our test set without any domain-specific adaptations.
2. **Qualitative Analysis:** Conducting detailed error analysis to identify common failure modes and misunderstandings in tool recognition and instruction generation.
3. **Challenging Scenarios:** Testing models with deliberately difficult scenarios including partially occluded tools, uncommon viewing angles, and ambiguous use cases.

Test prompts would range from simple identification queries (”What tool is this?”) to more complex procedural inquiries (”How should I safely use this tool to accomplish a specific task?”).

3.3. Contributions

This proposed project aims to deliver several key contributions to the field:

- Empirical insights into the current capabilities and limitations of VLMs for specialized domain understanding, particularly for safety-critical applications [1].
- Technical recommendations for improving model performance on domain-specific recognition and instruction generation tasks, with a focus on RAG and fine-tuning strategies as depicted in Figure 2.

Based on preliminary explorations, we anticipate finding significant variation in how well different VLMs handle mechanical tool recognition and safety instruction tasks. The results would provide valuable guidance for future work on enhancing VLMs for industrial applications.

4. Conclusion

This proposed course project aims to advance the application of Vision-Language Models in specialized technical domains by evaluating their capabilities for mechanical tool recognition and instruction generation. Through a systematic assessment of various state-of-the-art VLMs including Qwen2-VL, Phi-3-vision, Llama-3.2-Vision, SmolVLM, and PaliGemma, this research would contribute to our understanding of how multimodal AI systems can support safety-critical industrial applications.

The data collection approach and evaluation methodology outlined in this proposal would help establish benchmarks for assessing VLMs in specialized domains where precision and safety are paramount. By focusing on both recognition accuracy and instruction quality metrics, our evaluation framework would provide a more comprehensive picture of model performance.

While this project would focus specifically on mechanical tools across 13 diverse categories, the methodologies and insights could be applicable to other technical domains where visual recognition combined with specialized instruction is valuable. Future work building on these findings could explore applications in related fields such as medical device operation, construction equipment handling, or specialized manufacturing processes.

As VLMs continue to advance in capabilities and deployment scope, understanding their performance boundaries in specialized domains becomes increasingly important. The insights from this proposed research would help guide future enhancements to make these powerful systems more reliable, accurate, and safety-conscious for real-world industrial applications.

References

[1] Yi Ding, Bolian Li, and Ruqi Zhang. Eta: Evaluating then aligning safety of vision language models at inference time, 2025. 3

[2] Aaron Grattafiori et al. The llama 3 herd of models, 2024. 2

[3] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 2

[4] Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2

[5] Xi Chen et al. Pali-x: On scaling up a multilingual vision and language model, 2023. 2

[6] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study, 2023. 1

[7] Ombretta Strafforello, Derya Soydaner, Michiel Willems, Anne-Sofie Maerten, and Stefanie De Winter. Have large vision-language models mastered art history? *arXiv preprint arXiv:2409.03521*, 2024. 1

[8] Xiang Yin et al. Smolvlm: A small open-source vision-language model. *arXiv preprint arXiv:2311.11696*, 2023. 2