MARVL: Materials science Aware Reasoning dataset for Vision–Language Models

Mohd Zaki[1,*], Indrajeet Mandal[2,*], Prince Sharma[1], Megha Mondal[3], Sudhakar Kumar[3], Vishal Bhaskar[3], Arjun Chand[3], Shivnandi[3]

[1]Hopkins Extreme Materials Institute, Johns Hopkins University, USA.
[2]School of Interdisciplinary Research, Indian Institute of Technology Delhi, India.
[3]Department of Materials Science and Engineering, Indian Institute of Technology Delhi, India.

Email: mzaki4@jh.edu, indrajeet.mandal@sire.iitd.ac.in, psharma47@jh.edu

Abstract

Microscopy image interpretation in materials science relies heavily on experts to distinguish true structural features from artifacts such as tip distortions, noise, or instrument-induced errors. This dependency slows high-throughput research and limits the capabilities of self-driving laboratories that require rapid, reliable imaging feedback. We introduce MARVL (Materials science Aware Reasoning with Vision-Language Models), a multimodal dataset spanning AFM, SEM, and TEM that integrates literature-derived knowledge with experimentally collected images. MARVL includes images with respective captions and descriptions extracted from the peer reviewed research papers. This dataset lays the foundation for training high quality multimodal models for autonomous, high-quality image interpretations in materials science.

1. Introduction

High-resolution microscopy techniques such as Atomic Force Microscopy (AFM), Scanning Electron Microscopy (SEM), and Transmission Electron Microscopy (TEM) are central to materials science research. However, the interpretation of these images remains heavily dependent on domain experts who must manually distinguish genuine surface features from imaging artifacts such as tip-induced distortions (see figure 1), drift effects, noise patterns, or contrast anomalies. This expert-driven interpretation process is slow, subjective, and difficult to scale—especially in high-throughput or automated experimental workflows.

In the context of emerging self-driving laboratories, this challenge becomes a critical bottleneck. Autonomous experimentation pipelines rely on rapid, accurate feedback from imaging outputs, yet current systems lack the capability to identify whether an image contains artifacts that could invalidate downstream analysis. Without reliable automated reasoning, self-driving labs cannot adaptively modify experiments, discard compromised data, or self-correct imaging conditions.

While recent advances in Vision–Language Models (VLMs) demonstrate strong general reasoning and pattern recognition, the materials science domain lacks high-quality, structured datasets that map microscopy images to scientifically grounded explanations. This gap prevents VLMs from achieving expert-level interpretability and from reliably identifying artifacts across different imaging modalities.

To address this, we present MARVL (Materials-Science Aware Reasoning with Vision–Language Models)—a multimodal dataset designed to train and benchmark VLMs on the interpretation of AFM, SEM, and TEM images. MARVL integrates: (1) Automated literature extraction of imaging artifacts, terminology, mechanisms, and descriptions. (2) Experimentally collected microscopy images, labeled for artifact presence, type, and physical origin. (3) Structured, text-based reasoning annotations to support explainable AI in scientific imaging. MARVL aims to enable next-generation VLMs that can reason about nanoscale structures, distinguish real features from artifacts, and support autonomy in materials characterization workflows.

2. Methodology

The proposed pipeline for creating the dataset is inspired by Venugopal et al. [Patterns, Volume 2, Issue 7, 100290], Gupta et al. [npj Comput Mater 8, 102 (2022)] and Ahlawat et al. [arXiv:2412.09560] for extracting the figures, their captions, followed by descriptions from the XMLs. The data is stored in the form of jsonl where each line in the file is a dictionary with keys as *pii, figure id, caption, descriptions, image*. The *pii* is the identifier unique to Elsevier journals, *figure id* is the figure number in the paper, *caption* is the figure caption, *descriptions* is a list containing all paragraphs for the research paper where the given image is referenced.

The GitHub repository contains the codes, sample XML files (for open access papers) and their content for visualization purpose. The interested readers can use create a bigger version of the dataset using their own APIs to download the XMLs and figures followed by execution of our scripts.

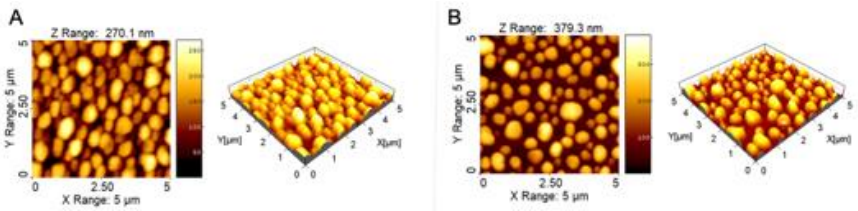
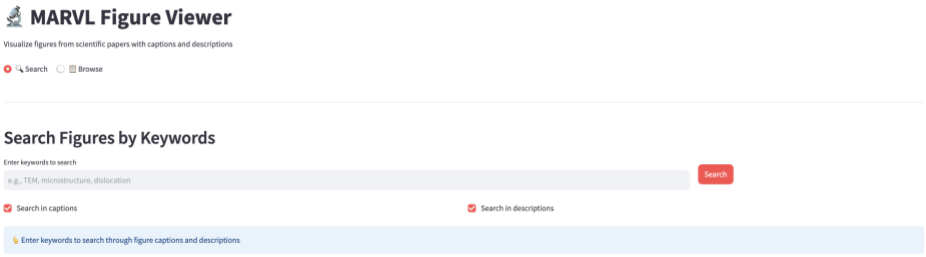


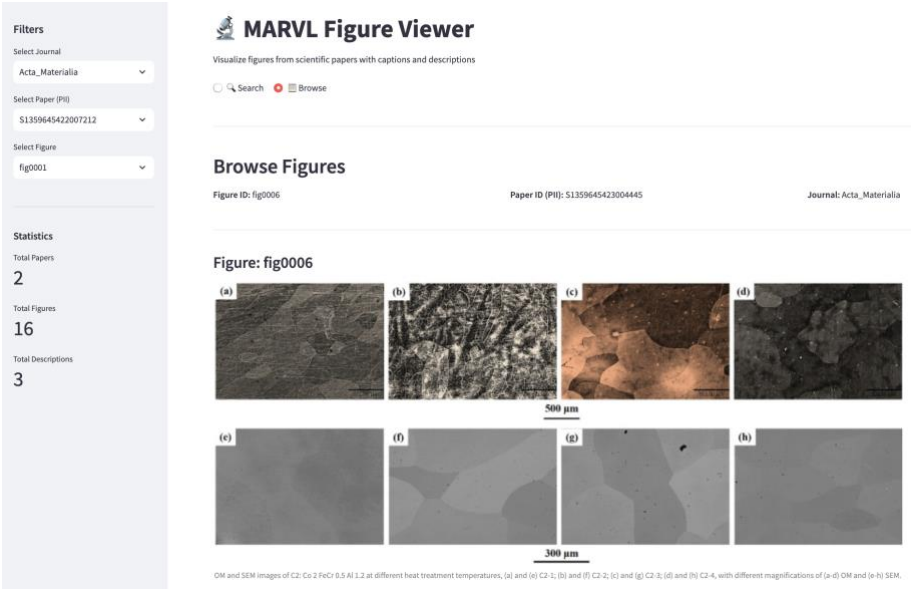
Figure 1. AFM images of gold nanoislands on a glass surface: (A) showing tip-artifact distortion of the gold nanoparticles, and (B) acquired without tip-artifact effects.

3. Results

The dashboard can be used in two modes: search and browse:



The figure below shows the visualization dashboard with a SEM micrographs, retrieved through search mode, and then expanded to view full details including its caption, and extracted descriptions.



3. Conclusion

In this work, we developed MARVL, the first dedicated, multi-modality materials-science reasoning dataset designed for Vision–Language Models. By integrating curated AFM, SEM, and TEM images with literature-derived reasoning annotations, MARVL addresses a critical gap in automated microscopy interpretation. The dataset enables AI systems to distinguish real nanoscale features from imaging artifacts, provides interpretable scientific explanations, and supports the development of autonomous materials research pipelines. Our contribution establishes a foundation for training VLMs capable of expert-level analysis while reducing reliance on manual interpretation. MARVL is a step toward fully self-driving laboratories where imaging feedback loops become reliable, automated, and scalable.