Unsupervised Microstructure Image Analysis and Question Answering Using CLIP and VLMs

Introduction: Materials science relies heavily on microscopy techniques such as scanning electron microscopy (SEM), transmission electron microscopy (TEM), atomic force microscopy (AFM), and optical microscopy to characterize microstructural features. Accurate identification of microscopy imaging techniques and material categories from micrographs requires substantial domain expertise. As microscopy datasets continue to expand across multiple scientific disciplines, scalable and automated classification methods are increasingly necessary to support data organization, interpretation, and reuse.

In this work, we investigate the ability of modern multimodal models to understand materials microscopy images. We first employ CLIP to perform embedding-based zero-shot matching between microscopy images and textual descriptions, enabling unsupervised grouping and weak labeling without task-specific training. These CLIP-generated descriptions are then used as contextual inputs for question answering, where vision language models (VLMs) are prompted to reason over microscopy images through multiple choice and descriptive question formats.

Our objective is to evaluate this two-stage pipeline on a real-world microscopy dataset, analyze the complementary strengths and limitations of similarity based unsupervised labeling and reasoning-based question answering, and demonstrate how they can be combined into a practical, scalable, and computationally lightweight framework for microstructure image understanding.Our goal is to benchmark their performance on a real-world microscopy dataset, analyze their strengths and limitations, and assess how these approaches can be combined into a practical and scalable solution.

Dataset Collection: The dataset used in this study was constructed by web crawling the DoITPoMS Micrograph Library, starting from the publicly accessible micrograph records. Using an automated scraping pipeline, we collected a total of 867 microscopy images along with their associated expert-annotated metadata. The DoITPoMS library is a curated collection of materials micrographs intended for teaching and learning and currently contains over 850 annotated samples. Each image in our dataset includes information about the imaging technique, material category, and descriptive annotations related to the observed microstructure.

The dataset reflects realistic materials science data and exhibits strong class imbalance. Metal and alloy samples dominate the dataset, while ceramics, polymers, composites, and fracture-related images are comparatively underrepresented. This imbalance mirrors the distribution commonly found in public materials databases and presents an additional challenge for automated classification methods.

Experimental Methodology: To examine multimodal understanding, we designed two distinct experimental pipelines. In the first pipeline, we applied CLIP for zero-shot classification across the full dataset of 867 images. Each microscopy image was encoded using a pretrained CLIP image encoder, while textual descriptions of imaging techniques and material categories were encoded using the CLIP text encoder. Cosine similarity between image and text embeddings was used to rank candidate labels. Performance was evaluated using Top-1 and Top-2 accuracy metrics. This approach was chosen for its computational efficiency and ability to scale to large datasets without requiring domain-specific training.

In the second pipeline, we evaluated vision–language models using a multiple-choice question-answering formulation. Due to the higher computational cost of VLMs, we randomly sampled 51 images from the complete dataset. After filtering invalid or incomplete samples, 46 images were used for microscopy technique

classification and 45 images were used for material category classification. Each image was paired with a carefully designed multiple-choice question asking either about the imaging technique or the material category, and the model was required to select a single answer. We evaluated two VLMs, LLaVA-v1.6-Mistral-7B-HF and Qwen/Qwen3-VL-8B-Instruct, using accuracy, balanced accuracy, and macro-averaged precision, recall, and F1-score.

Results: The CLIP-based zero-shot classification achieved strong performance for identifying microscopy techniques. Across the full dataset, CLIP achieved a Top-1 technique accuracy of 0.5356 and a Top-2 accuracy of 0.7981, indicating that the correct technique was frequently among the top-ranked predictions. Material category classification proved more challenging, with CLIP achieving a Top-1 accuracy of 0.3456 and a Top-2 accuracy of 0.5226. These results reflect both visual similarity between certain material classes and the impact of dataset imbalance.

The vision–language model evaluations showed substantial variation across models. Using LLaVA-v1.6-Mistral-7B-HF, microscopy technique classification achieved an accuracy of 0.261 with a balanced accuracy of 0.250, while macro-averaged precision, recall, and F1-score were low, indicating difficulty in distinguishing between imaging modalities. Category classification using the same model achieved a higher accuracy of 0.556, although balanced accuracy remained low due to class imbalance.

In contrast, the Qwen/Qwen3-VL-8B-Instruct model demonstrated significantly improved performance for microscopy technique classification. Qwen achieved an accuracy of 0.913, a balanced accuracy of 0.862, and a macro F1-score of 0.888, indicating strong and consistent discrimination between microscopy modalities. For material category classification, Qwen achieved an accuracy of 0.378 with a balanced accuracy of 0.382 and a macro F1-score of 0.329. While category classification remains challenging, these results indicate improved robustness compared to earlier VLM evaluations.

Discussion

The comparative analysis highlights clear trade-offs between embedding-based and reasoning-based approaches. CLIP provides strong scalability and computational efficiency, making it suitable for large microscopy datasets. Its performance on technique classification suggests that embedding-based models capture meaningful visual semantics related to imaging modalities. However, CLIP's performance degrades for material category classification, particularly for minority classes, and it lacks explicit reasoning or interpretability, underscoring the need for domain-specific adaptation.

Vision–language models offer a more interpretable framework by explicitly reasoning over images through natural language questions. The strong performance of Qwen/Qwen3-VL-8B-Instruct on technique classification demonstrates that newer-generation VLMs can capture fine-grained technical features in microscopy images. However, category classification remains difficult across all models, likely due to subtle visual differences between material classes and limited domain-specific training data. Overall, these results suggest that a hybrid strategy combining scalable embedding-based models with selectively applied reasoning-based VLMs offers a promising direction for practical, automated microscopy analysis.

Acknowledgement
Used Generative AI tools for debugging code and correcting grammatical mistakes and increase readability.