VLRIMM - Vision-Language Retrieval for Identical Materials Morphology

Kevin Zhang[1], Sartaaj Khan[2], Mohammad Taha[1], Thomas Pruyn[2]
University of Toronto
[1]Department of Materials Science & Engineering
[2]Department of Chemical Engineering & Applied Chemistry

Abstract

VLRIMM is a multi-modal RAG pipeline designed to bridge the gap between Scanning Electron Microscopy (SEM) micrographs and scientific literature. By leveraging DINOv3 for morphological feature extraction and OpenAI text-embeddings for semantic grounding, we enable researchers to use a raw micrograph as a query to retrieve information like synthesis protocols and characterization data from identical micrographs. VLRIMM offers a modular, updateable framework for real-time SEM micrograph analysis.

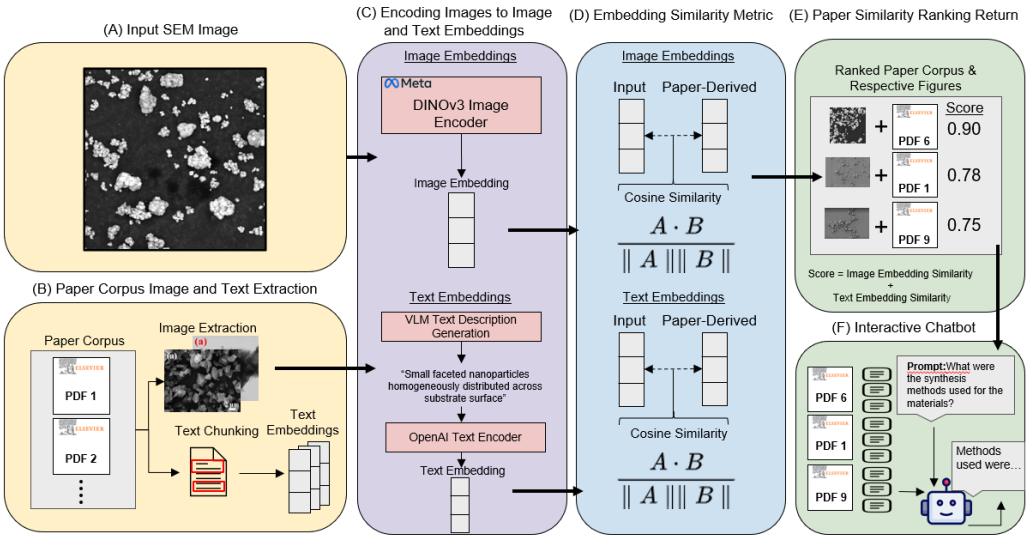


Figure 1. Overview of the VLRIMM workflow.

Motivation

As the "AI for Science" movement accelerates, the volume of published scientific literature and high-resolution characterization data are ever increasing. While Large Language Models (LLMs) provide impressive reasoning capabilities, they suffer from two critical flaws in a laboratory setting: knowledge cutoff (inability to access the latest research) and hallucinations regarding domain-specific information. In materials science, the "ground truth" of a sample's morphology is often captured via SEM. However, papers often lack precise description of SEM morphology and description can be challenging and biased.

VLRIMM addresses this by implementing a multi-modal Retrieval-Augmented Generation (RAG) pipeline. By utilizing DINOv3 [1], a SOTA Vision Foundation Model (VFM) and OpenAI text embeddings, we allow researchers to use a micrograph as a direct query to search for scientific papers with similar micrograph and text description for the sample.

Methodology

Data Scraping & Extraction: For this hackathon, 199 recently published papers were downloaded manually from ScienceDirect using keywords such as "metal powder SEM" and "particles". The download

process can be automated in the future using a data scraper so the paper database can be updated daily. The automated data extraction process from a PDF is as follows: 1) Docling is used to extract all figures, tables, and text from the PDF. 2) Figures are filtered by resolution to remove any icons and logos. 3) Figures are sliced to extract the individual subplots using OpenCV's edge detection while another filter checks the amount of white space and colored pixels to filter out plots, diagrams, and photos. This extracts all SEM micrographs from paper with ~95% accuracy. From this process, approximately 2600 SEM micrographs are extracted.

Image Embeddings: DINOv3-vits16plus is then executed, which is the small-plus variant in the DINOv3 family. This produces a global image embedding of latent dimension 384, which is used as the representation for the micrograph. In testing, DINOv3 accurately captures distribution, particle morphology, porosity, roughness etc. in the micrograph while omitting style artifacts like charging, contrast difference, and markup/labels on the micrograph. Cosine similarity is used to compare between the input micrograph and micrographs from all papers. No significant difference was observed for the top-20 closest micrographs when using different similarity metrics.

Text Embeddings: A VLM was used to generate "img2text" captions for micrographs and embedded them alongside original paper text chunks using OpenAI's text-embedding-3-large. This allows the system to compare user queries, VLM-generated labels, and paper content in a unified semantic space. This also enables easy modification if the user wishes to add additional text information as input.

Similarity Metric & Score: Cosine similarity is used to measure how similar image embeddings and text embeddings are between the input SEM micrograph and the images from the literature where larger values indicate more similar images. The overall similarity score is calculated using a linear combination of the cosine similarity between the image embeddings pair and text embeddings pair.

RAG & ChatBot: During inference, the most similar micrographs are retrieved by comparing image and word similarity as described above. The source papers containing these micrographs are retrieved. Their chunked text embeddings are compared with the user's questions, which then determines the most relevant part of the paper with respect to the user's prompt. The LLM chatbot with a specifically engineered prompt will receive this information and provide a grounded answer to the user's question, including inline citations to the retrieved paper chunks.

Conclusion

VLRIMM demonstrates a transformative shift in how researchers interact with the ever-growing body of scientific literature. By bridging the gap between raw visual data and textual knowledge, we have moved beyond keyword-based search and enhanced LLM knowledge by providing accurate, up-to-date images and text from peer-reviewed scientific papers. We envision that this system can be attached to the live-view of SEM or used as a real-time analysis software that allows researchers to receive immediate, referenced answers. VLRIMM sets the foundation for a universal scientific indexing system where visual and text data is as searchable, queryable, and actionable information.

Reference

[1] Siméoni, O., “DINOv3”, arXiv e-prints, Art. no. arXiv:2508.10104, 2025. doi:10.48550/arXiv.2508.10104.

Appendix

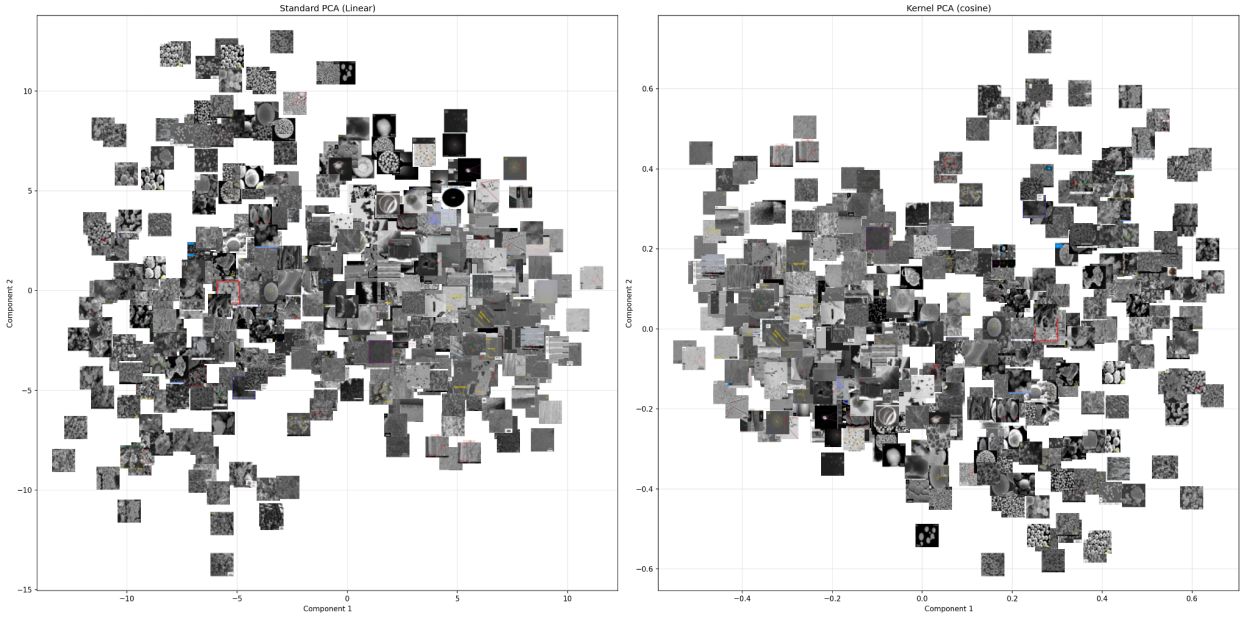


Figure A1. A random subset of the images from the papers plotted in the latent space of DINOv3 using PCA (left) and Kernel-PCA (right). Similar micrographs are clustered together these two reduced dimensional spaces showcasing a structured, well-organized latent space encoding important microstructural information.

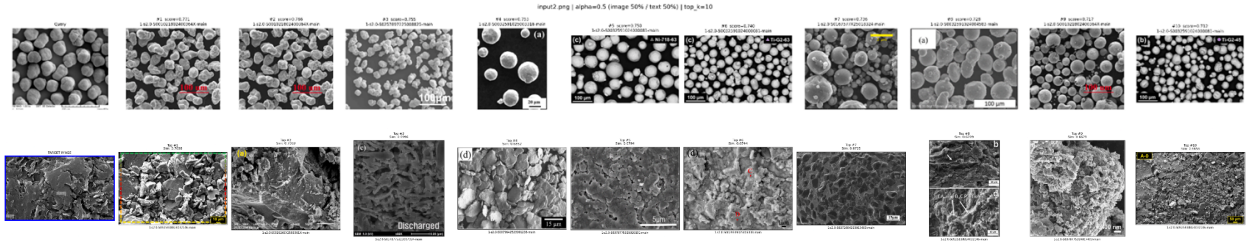
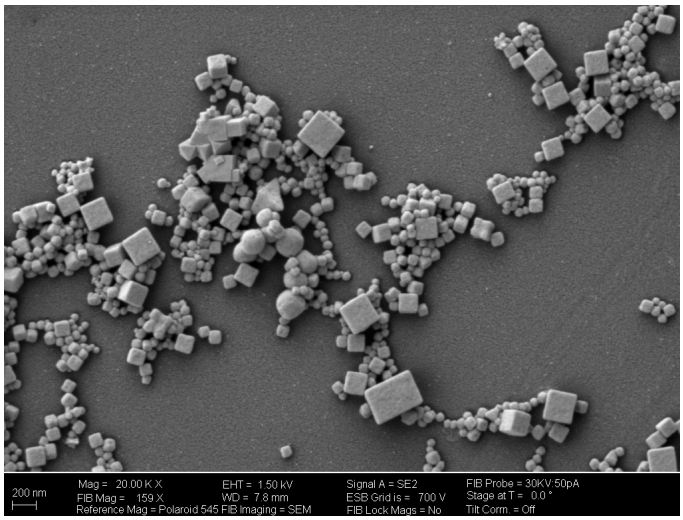
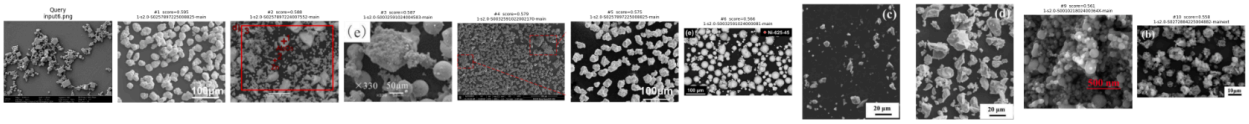


Figure A2. Left most: The micrograph used as a query. Right 10: The top 10 closest micrograph in the latent space of DINOv3, using cosine similarity as the distance metric. We observe micrographs close together in the latent space of DINOv3 shares very similar morphology.





Answer

Based on the image description and the provided evidence snippets, the material synthesis likely involves a process that can produce both cubic and spherical particles, potentially through powder atomization or spray conversion techniques.

1. **Powder Atomization**: The presence of cubic and spherical particles could suggest a powder atomization process. For instance, titanium powders can exhibit different morphologies depending on the atomization conditions, and the presence of voids or hollow particles can occur during this process [1-s2.0-S0032591025003316-main].
2. **Spray Conversion**: Another possibility is spray conversion, where precursor solutions are atomized to form fine droplets that solidify into spherical particles. This process can be influenced by factors such as temperature and the presence of additives like Cr, which can refine particle size and prevent agglomeration [1-s2.0-S0272884225004882-mainext].
3. **Key Parameters**:
   - **Temperature**: Processes like sintering or calcination are mentioned at temperatures ranging from 300°C to 1400°C, depending on the material system and desired microstructure [1-s2.0-S0032591025003316-main], [1-s2.0-S0272884225004882-mainext].
   - **Atmosphere**: The presence of air or inert gases like N2 during processing can affect the oxidation state and phase composition [1-s2.0-S001021802400364X-main].
   - **Time**: Holding times during processes like sintering or reduction are typically around 60 minutes [1-s2.0-S0272884225004882-mainext].
4. **Uncertainty Notes**: Without specific details on the material composition or exact synthesis method, there is uncertainty in pinpointing the precise synthesis route. The presence of both cubic and spherical particles suggests a complex process that may involve multiple steps or varying conditions.

In conclusion, the synthesis likely involves atomization or spray conversion, with key parameters including high temperatures and controlled atmospheres. Further details would be needed to confirm the exact process.

Retrieved papers

- 1-s2.0-S0257897225008825-main
- 1-s2.0-S0257897224007552-main
- 1-s2.0-S0032591024004583-main
- 1-s2.0-S0032591022002170-main
- 1-s2.0-S0032591024000081-main
- 1-s2.0-S0032591025003316-main
- 1-s2.0-S001021802400364X-main
- 1-s2.0-S0272884225004882-mainext

Retrieved evidence chunks

1-s2.0-S0032591025003316-main

chunk 26 | sim=0.586

3.4. Microstructure evolution and mechanical properties

the presence of voids in the central region of the initial Ti powder, indicating the possible existence of hollow particles generated during the powder atomization process. However, the core of the

Fig. 6 displays the microstructure of S1, S2, and S4 at a sintering temperature of 1300 ◦ C. A conspicuous characteristic is the emergence of an α 2 + γ lamellar structure at this temperature. The

chunk 27 | sim=0.575

*Figure A3. An example of the VLRIMM pipeline. Top: The input micrograph. Center: The input micrograph with the top 10 most similar micrographs on the right. Bottom: The answer, referenced paper, and referenced text evidence (only the first chunk is shown, a total of 12 chunks were retrieved) to the user query: "**Based on this image, describe potential synthesis details of this material.**"*