MVaEMa: A Vision-Language Assistant with Distilled Domain Knowledge

Empowering Specialized Models via Instruction Tuning

Sakhinana Sagar Srinivas[1] Geethan Sannidhi[2] Venkataramana Runkana[1]

[1]Tata Research Development and Design Centre (TRDDC)
[2]Indian Institute of Information Technology (IIIT), Pune

Presented at the AAAI 2024 Spring Symposium Series
AAAI-MAKE: Empowering LLMs with Domain and Commonsense Knowledge
Stanford University, March 25-27, 2024

Agenda

The Domain: High-Stakes Analysis in Semiconductor Manufacturing

The Critical Role of Nanoscale Imagery

- In advanced semiconductor fabrication, quality control relies on interpreting complex Scanning Electron Microscope (SEM) images.
- This analysis requires deep, specialized **domain knowledge** of materials science and morphology.

The Knowledge Bottleneck

- Expert analysis is precise but slow and expensive, creating a bottleneck in rapid manufacturing cycles.
- Our goal is to automate this process by empowering an AI model with the necessary expert-level domain knowledge.

How do we safely provide an AI with this specialized knowledge?

Approach 1: Large Language Models (GPT-4V)

- ✓ **Strength:** Possess vast, general-world knowledge.
- • **Weakness:** Using them requires uploading proprietary chip designs to third-party APIs—an unacceptable IP security risk. They are also costly for large-scale, continuous use.

Approach 2: Small, Open-Source Models

- ✓ **Strength:** Can be hosted in-house, guaranteeing data security and privacy.
- • **Weakness:** They lack the required domain knowledge and there is no large, expert-annotated dataset to teach them. This is the classic knowledge acquisition bottleneck.

Our work addresses this fundamental challenge: **How do we distill domain knowledge into a secure, specialized model?**

Our Approach: Distilling Knowledge via a Teacher-Student Framework

We introduce **MVaEMa**(**M**ultimodal **V**ision **A**ssistant for **E**lectron **M**icrograph **A**nalysis), a framework designed to bridge the knowledge-security gap.

The Strategy: Guided Knowledge Distillation

We use a two-step process to safely transfer expertise:

1. **The "Teacher" (GPT-4V):** We leverage a large model *offline* as a "domain expert" to generate a rich, instruction-following dataset from unlabeled images.

2. **The "Student" (MVaEMa):** We then use this high-quality, synthetic data to train our smaller, secure, in-house model. MVaEMa effectively learns the distilled domain knowledge from the teacher.

The Result

We create a high-performance model that is:

- **Knowledgeable:** Exhibits deep understanding of a specialized domain.
- **Secure:** Deployed entirely in-house, protecting sensitive IP.
- **Efficient:** Avoids the high costs and latency of constant API calls.

Step 1: Capturing Domain Knowledge via Instruction Tuning. Guiding the Teacher to "Think" like an Expert

Expert-Guided, Zero-Shot Chain-of-Thought (CoT) Prompting

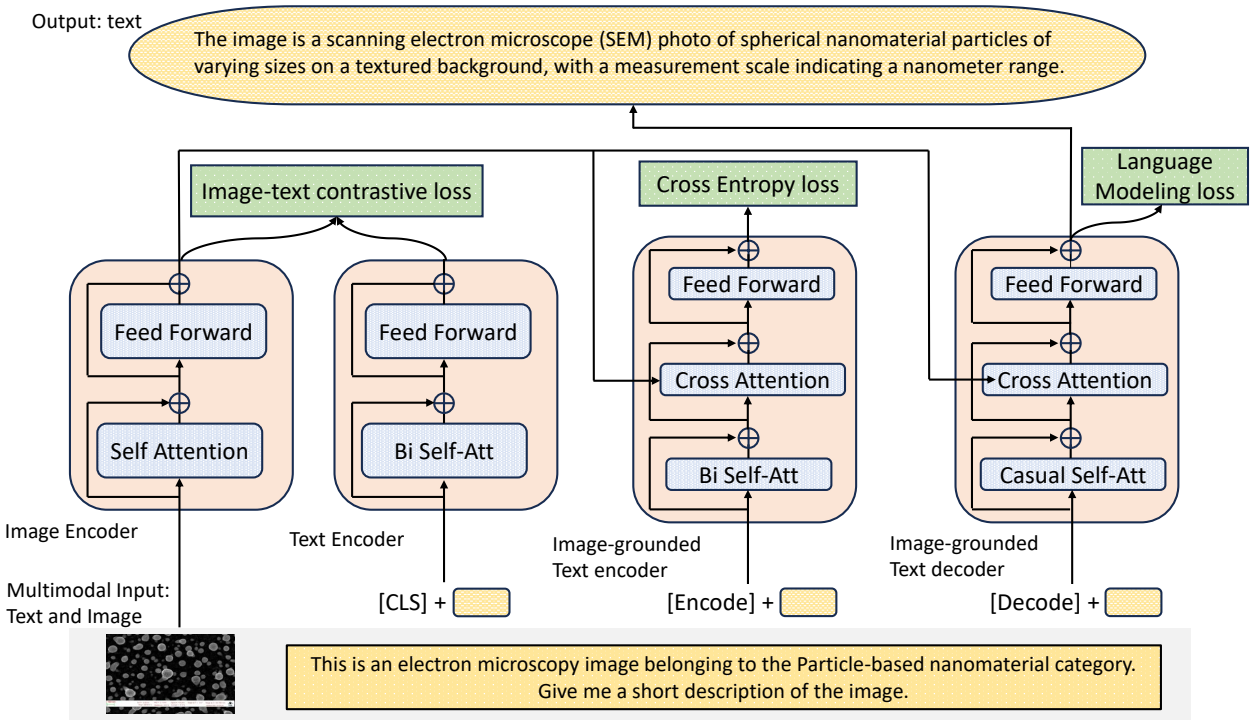
To distill high-quality knowledge, we guide the teacher model's reasoning process using a structured, multi-faceted prompting strategy.

- **Symbolic Anchoring:** Before the question, we provide a "symbolic anchor" by stating the ground-truth category (e.g., *"This is an electron microscopy image belonging to the Particle-based nanomaterial category."*). This primes the model with high-level context.
- **Comprehensive Knowledge Facets:** Using zero-shot CoT, we prompt for a wide range of expert knowledge, covering key areas like *morphology*, *particle distribution*, and *surface defects*.

Outcome: A High-Fidelity, Knowledge-Rich "Textbook"

This process yields a dataset of {Image, Contextualized Prompt, Detailed Answer} triplets. We further ensure the quality of this "textbook" for our student model. We ensure deterministic, fact-based outputs to minimize creative hallucinations.

Step 2: The MVaEMa Architecture for Knowledge Fusion



Our architecture is designed to effectively fuse and reason with multimodal knowledge.

- **Unimodal Encoders:** Process the raw text and image data using self-attention.
- **Image-Grounded Encoder (Fusion):** The core of our knowledge fusion. It uses Cross-Attention to let textual concepts query the image, finding relevant visual evidence.
- **Image-Grounded Decoder (Generation):** Articulates the final answer. It uses Cross-Attention again to ensure the generated text remains factually grounded in the visual evidence throughout the generation process.

Step 3: Embedding Knowledge through Multi-Task Learning

We train MVaEMa on three synergistic objectives to ensure deep and robust learning.

1. Contrastive Loss (ITC)

Teaches: Conceptual Association. Aligns the concepts of "nanowire" in text and image form within a shared embedding space.

2. Matching Loss (ITM)

Teaches: Factual Verification. Trains the model to verify if a textual description is a factually correct match for a given image, sharpening its fine-grained understanding.

3. Language Modeling Loss (LM)

Teaches: Articulate Expression. Enables the model to generate fluent, coherent, and contextually appropriate sentences that accurately express its internal, knowledge-based understanding.

Synergy

These three losses work together: ITC provides a coarse alignment, ITM refines it with fact-checking, and LM uses this grounded understanding to generate accurate text.

Experimental Setup

Dataset

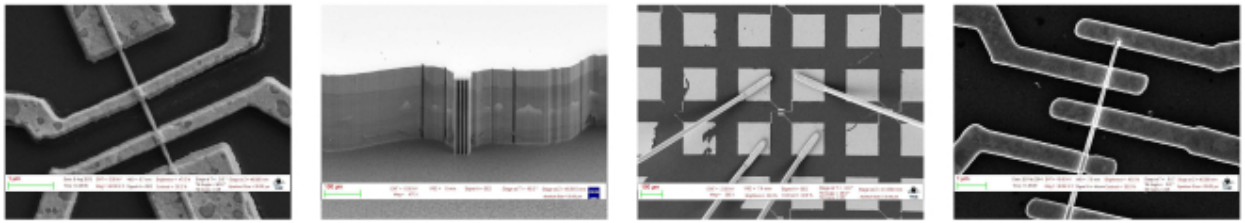
We use the public **SEM Dataset for Nanoscience** (Aversa et al., 2018), containing over 21,000 images across 10 challenging categories (e.g., particles, films, nanowires, MEMS devices).

Evaluation Tasks

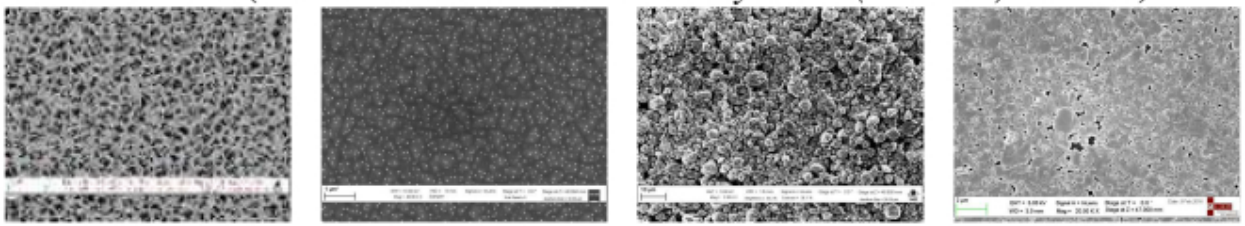
We designed our evaluation to test for deep knowledge, not just surface-level pattern matching:

1. **Visual Question Answering (VQA):** Can the model articulate domain-specific knowledge by describing morphology?
2. **Zero-Shot Image Classification:** Can the model generalize its knowledge to classify images into categories it was never explicitly taught?

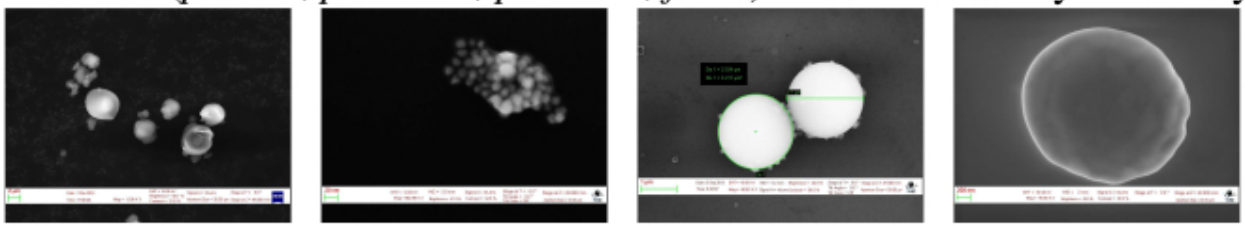
A Look at the Dataset's Complexity



(a) High intra-class dissimilarity(variance) in electron micrographs of a nanomaterial (*micro-electromechanical systems(MEMS)* device).



(b) High inter-class similarity: Electron micrographs of different nanomaterials (*porous, particles, powders, films*) show noteworthy similarity.



(c) Multi-spatial scales of patterns: Nanoparticle electron micrographs exhibit multi-scale spatial heterogeneity.

Figure: Sample images showing the visual diversity of the nanomaterial dataset. Key visual challenges: High inter-class similarity (e.g., particles vs. powders) and significant intra-class variance (e.g., diverse MEMS structures).

Result 1: Articulating Domain-Specific Knowledge (VQA)

Quantitative Comparison

We compared MVaEMa against leading multimodal baselines.

Method	**BLEU-4** (↑)	**ROUGE-1** (↑)	**ROUGE-2** (↑)	**ROUGE-L** (↑)	**METEOR** (↑)
InstructBLIP	0.457	0.745	0.648	0.705	0.738
LLaVA	0.512	0.760	0.668	0.723	0.753
MiniGPT-4	0.572	0.790	0.698	0.753	0.783
MVaEMa (Ours)	**0.709**	**0.860**	**0.765**	**0.822**	**0.853**

Interpretation

Our model's generated descriptions are significantly more aligned with expert-level ground truth. The high METEOR and ROUGE scores indicate superior factual and semantic accuracy, proving the effectiveness of our knowledge distillation approach.

Result 2: Demonstrating Knowledge Generalization (Classification)

Zero-Shot Classification Performance

Algorithm Type	**Top-1 Acc. (↑)**
GoogleNet (Supervised CNN)	0.609
SwinT (Supervised ViT)	0.707
T2TViT (Supervised ViT)	0.749
MVaEMa (Ours, Zero-Shot)	**0.947**

Interpretation

This is a key finding. By training our model to *describe* (a knowledge-intensive task), it develops a conceptual understanding so deep that it outperforms models explicitly trained to *classify*. This shows true generalization of the embedded domain knowledge.

Why does our framework work so well?

The Question: Are all the components necessary?

An ablation study validates our architecture by removing key **modules** to quantify their individual contributions.

Variant (Component Removed)	VQA Performance Drop (METEOR)	Classification Drop (F1-Score)
- Image-Text Contrastive (ITC) Loss	-18.4%	-25.5%
- Image-Text Matching (ITM) Loss	-18.4%	-21.1%
- Cross-Attention in Decoder	-9.9%	-10.6%
- Self-Attention in Encoder	-6.9%	-7.0%

Interpretation

The largest performance drops occur when the core multimodal fusion losses (ITC and ITM) are removed. This confirms that **effectively aligning and verifying the connection between text and image is the most critical factor** for grounding the model's knowledge.

Conclusion and Impact for AAAI-MAKE

Summary of Contributions

- We proposed a practical and effective framework for **distilling and injecting domain knowledge** from large, generalist models into small, secure, specialist models.
- We demonstrated that this distilled knowledge empowers our model, MVaEMa, to achieve state-of-the-art performance in a complex scientific domain.
- Our results show that training for descriptive articulation leads to a more profound and generalizable understanding than training for simple classification.

This work provides a blueprint for building knowledge-intensive AI systems that are both powerful and secure, directly addressing a core challenge for the enterprise adoption of LLMs.

Thank You Questions?

Scan for the full paper