

Multi-Modal Instruction-Tuning Small-Scale Language-and-Vision Assistant for Semiconductor Electron Micrograph Analysis

Sagar Srinivas Sakhinana^{1*}, Geethan Sannidhi^{2†}, Venkataramana Runkana¹

¹TCS Research

²IIIT Pune

sagar.sakhinana@tcs.com, geethansannidhi20@cse.iiitp.ac.in, venkat.runkana@tcs.com

Abstract

We present a novel framework for analyzing and interpreting electron microscopy images in semiconductor manufacturing using vision-language instruction tuning. The framework employs a unique teacher-student approach, leveraging pre-trained multimodal large language models such as GPT-4 to generate instruction-following data for zero-shot visual question answering (VQA) and classification tasks, customizing smaller multimodal models (SMMs) for microscopy image analysis, resulting in an instruction-tuned language-and-vision assistant. Our framework merges knowledge engineering with machine learning to integrate domain-specific expertise from larger to smaller multimodal models within this specialized field, greatly reducing the need for extensive human labeling. Our study presents a secure, cost-effective, and customizable approach for analyzing microscopy images, addressing the challenges of adopting proprietary models in semiconductor manufacturing.

Introduction

Recent advances in AI, such as Large Multimodal Models (LMMs) like OpenAI’s GPT-4 Turbo with Vision (OpenAI 2023), and open-source, small-scale multimodal models (SMMs), such as LLaVA (Liu et al. 2023) and MiniGPT-4 (Zhu et al. 2023), enhance semiconductor manufacturing by analyzing high-resolution electron micrographs. While proprietary LMMs face adoption challenges due to data privacy concerns, SMMs offer cost-effective customization but may lack reasoning and generalization capabilities of proprietary counterparts. Acquiring high-quality training data for SMMs is challenging due to limited and expensive datasets, requiring expert knowledge and annotation tools. The diversity in image characteristics poses challenges for a one-model-fits-all approach across electron micrograph datasets. In our study, we introduce a novel method that utilizes GPT-4 Turbo with Vision, an advanced multimodal large language model, as a robust “teacher” for generating instruction-following data. Specifically, we create question-answer pairs related to nanomaterial image analysis. Using this dataset, we develop the

Multimodal Vision Assistant for Electron Micrograph Analysis (MVaEMa), an end-to-end trained smaller multimodal model (SMM) that is efficient yet powerful. MVaEMa is fine-tuned using the machine-generated dataset, which comprises a comprehensive collection of vision-language corpora for domain-specific customization. Each labeled pair consists of a query image, a related text instruction, and the most accurate response. We utilize vision-language instruction tuning to enhance MVaEMa’s zero-shot capabilities for tasks like visual question answering (VQA) on nanomaterial image analysis. This approach adheres to auto-regressive training, eliminating the need for high-quality, human-annotated image-text pairs for domain-specific adaptation. Training smaller models through vision-language instruction tuning using larger multimodal models is a promising approach, leveraging the knowledge and capabilities of the larger models. This method involves transferring knowledge from the larger model (the teacher) to the smaller model (the student) to enhance performance, enabling better understanding of visual concepts and accurate text generation based on visual content. This method improves grounded language generation and visual reasoning through the distillation of knowledge from teacher models, which is accomplished by aligning the student model’s predictions with those of the teacher model. Furthermore, enterprises can fine-tune the proposed pretrained model, MVaEMa on their proprietary data within their infrastructure, thus ensuring privacy, reducing costs, increasing customization, and enhancing security. Overall, it presents a viable solution potentially democratizing access to their capabilities and accelerating their adoption for various multimodal tasks, aligning with the increasing need for personalized, private AI solutions. We present the architecture of the proposed framework, MVaEMa, in Figure 1 for the zero-shot visual question-answering task. The proposed framework is a small-scale, autoregressive, unified vision-language model that employs an encoder-decoder architecture to process and integrate both text and image modalities. The multimodal input consists of the query microscopic image and the corresponding natural language question (task instruction), with the goal of providing an accurate answer based on the image content. The multimodal model comprises the following components: (a) The instruction-aware **image encoder** uses a self-attention mechanism with a larger global receptive field to analyze visual inputs, capturing salient

*Designed, programmed the software, and drafted manuscript.

†Conducted experiments and analyzed visual results

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

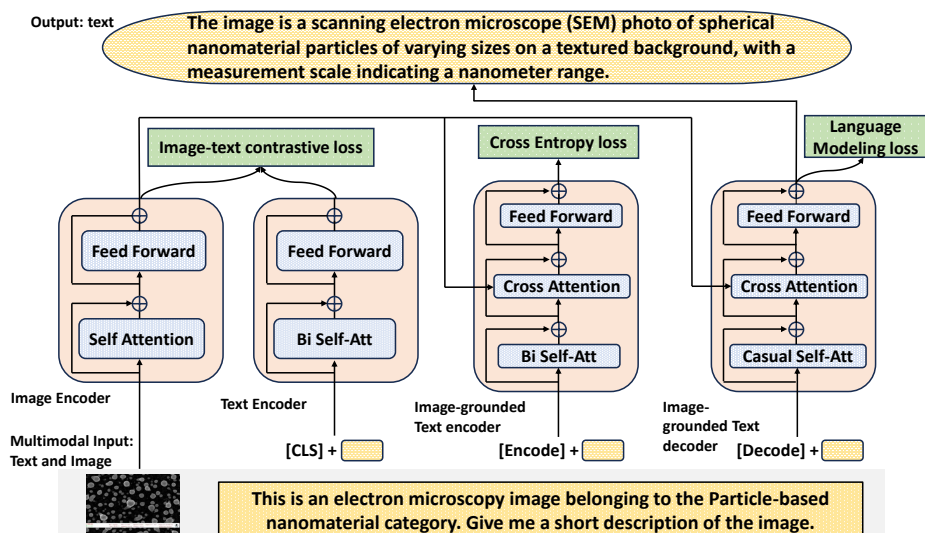
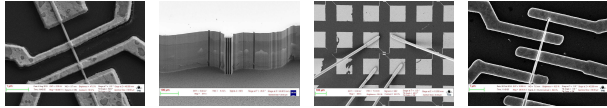


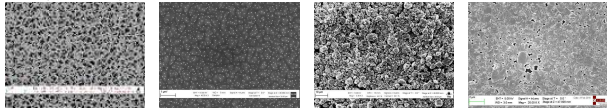
Figure 1: The architecture and objectives of MVaEMa, our proposed multimodal deep learning framework for VQA task in nanomaterial image analysis, are presented. It's a small-scale architecture combining text and image data, trained with vision-language instruction tuning from GPT-4 Turbo with Vision. Optimization uses various loss functions to align multimodal representations for answering image-related questions, showcasing its ability to understand complex intermodal relationships.

information, long-range dependencies and the overall scene composition. This allows the multimodal model to understand the global context of an image in a holistic and flexible manner, highlighting important regions and their contextual relationships while computing expressive image embeddings. (b) **The text encoder** is crucial for understanding and interpreting the query text, ensuring that it can be effectively combined with visual information for cross-modal analysis to provide accurate and relevant answers. The text encoder employs a bidirectional self-attention mechanism to encode linguistic inputs, preserving semantic and learning contextual relationships. We use a $\langle cls \rangle$ token to represent the entire sequence, providing a rich, contextualized representation of the query text essential for integrating with visual information to generate precise descriptions. The $\langle cls \rangle$ token embedding helps the multimodal model focus on relevant parts of the image and guides the answer generation process based on the question's context. The unimodal encoders (i.e., both text and image encoders) compute respective monolithic embeddings, which are jointly trained with a image-text contrastive loss to align the vision and language embeddings. (c) **The image-grounded text encoder** employs an additional cross-attention mechanism to align specific textual information with relevant visual features, computing contextually relevant multimodal representations. We utilize binary cross-entropy loss in image-text matching to assess a multimodal model's ability to correctly match images with text, aiming to minimize the discrepancy between positive and negative image-text pairs. This process results in precise, context-aware textual descriptions that accurately reflect the visual information. (d) **The image-grounded text decoder** utilizes the rich, multimodal representations to generate a syntactically and semantically coherent, contextually relevant textual description corresponding to the visual input. The

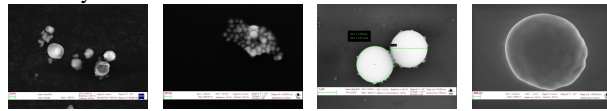
decoder replaces the bi-directional self-attention layers with causal self-attention and employs the same cross-attention layers and feed-forward networks as the image-grounded text encoder for text generation. It is trained with a language modeling loss to produce an output description that accurately reflects the image's content and context, thereby bridging the gap between visual perception and language generation by grounding the output in the image's visual content. The multimodal framework is optimized using a combination of image-text contrastive, cross-entropy, and language modeling loss functions, ensuring alignment between modalities and linguistic accuracy. This sophisticated approach enables the framework to answer questions about images with a high degree of precision and relevance. We train our multimodal framework using a specific type of instruction-following data: VQA task-based image-instruction-answer pairs. Based on this machine-generated data, we design a multimodal prompt to customize the MVaEMa framework, where the objective is to analyze the query image and provide an accurate answer based on the visual content and the specific question. As depicted in Figure 1, we adopt a symbolic approach with prompting mechanism, wherein the prompt (i.e., caption + natural language instruction), the caption explicitly mentions the microscopy image belongs to the predefined nanomaterial category (ground-truth). This description serves as a symbolic representation that the language encoder recognizes, and it decodes this sequence to understand the visual information from an SEM image of nanomaterials. Consequently, it integrates the image information with linguistic context within the multimodal model's processing framework. Nanoimage-based VQA tasks, while advantageous, remain a significant challenge. Figure 2 illustrates the challenges in VQA tasks, which are largely attributed to high intra-class dissimilarity, high inter-class similarity, and the existence of visual patterns



(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 2: The figure shows the challenges in VQA task on electron micrographs in the SEM dataset (Aversa et al. 2018).

at multiple scales, or spatial heterogeneity. The overarching goal of this research is to develop a vision-language instruction tuning framework, utilizing pretrained LMMs such as GPT-4 for training SMMs and address the challenges in VQA tasks for enterprise adoption. The main contributions of our work are as follows:

- The focus of our study is the development of small multimodal models (SMMs), *MV_aEM_a*, using visual instruction tuning. We employ GPT-4, a large, pre-trained multimodal teacher model, to generate diverse instruction-following data that better aligns with human intent. This includes the generation of detailed, context-rich question-answer pairs that explore different facets of microscopic images of nanomaterials. We utilize the high-quality, machine-generated data to provide customized instructions for training SMMs tailored to analyze electron microscopy images of nanomaterials. This teacher-student strategy enables zero-shot learning capabilities in the student models, allowing them to answer visually grounded questions without needing additional human labeling effort. Our approach facilitates knowledge distillation from proprietary LMMs to customized SMMs, improving the performance of the SMMs to be comparable to that of the LMMs on nanomaterial image analysis tasks. The pretrained SMMs can further be fine-tuned by enterprises with their in-house or proprietary data, without having to share sensitive data.
- We present a multimodal machine learning framework designed to process and integrate text and image data for the VQA task. It employs an image encoder with self-attention mechanism to extract salient information from images, as well as a text encoder with bidirectional self-attention to capture contextual language. The unimodal embeddings are then integrated in an image-grounded text

encoder that uses cross-attention mechanism to align text representations with visual cues. This is followed by a text decoder that generates descriptive output capturing the content and context of the image, guided by various loss functions to optimize the learning process. The ultimate goal is to produce text that accurately describes or explains images to assist with interpreting microscopy images.

Proposed Method

Instruction-tuned teacher LMM: We utilize a teacher-student strategy, employing an off-the-shelf, pre-trained large multimodal model to train small-scale multimodal model through instruction tuning on zero-shot VQA tasks. This approach accelerates the student model’s learning, resulting in more accurate, relevant, and appropriate responses for tasks involving visual and linguistic information. In this work, we leverage state-of-the-art instruction-tuned foundational LMMs, such as GPT-4 (OpenAI 2023), which offers efficient and cost-effective text generation with a large context window. By utilizing this general-purpose, large-scale pre-trained vision-language model, we create instruction-following data comprising question-answer pairs by exploring various aspects, such as the microscopic image’s structure and patterns, for customizing SMMs for nanomaterial image interpretation and analysis tasks. This significantly enhances their ability to autonomously handle new queries without relying on human-crafted instructions and aligns them more closely with human intentions. The GPT-4 API is accessible through Multimodal Modeling as a Service (MMaaS), an on-demand service hosted on cloud servers that accepts multimodal inputs, including both images and text, to produce outputs. This approach is similar to how Language Modeling as a Service (LMaaS) (Sun et al. 2022) provides access to Large Language Models (LLMs) for language processing tasks. We generate context-augmented multimodal chain-of-thought (CoT) prompts, that consist of image captions stating the nanomaterial category, along with natural language questions as task-specific instructions, which guide GPT-4 to examine the query nanomaterial image as visual input and generate the answer to produce detailed textual descriptions in response to the natural language question. This process creates instruction-following data for training SMMs to perform VQA task, with GPT-4V leveraging its domain-specific knowledge to provide contextual descriptions based on the visual inputs and image caption, along with the query text serving as labeled data for training the SMMs.

Multimodal Instruction-Following Data: Using GPT-4 to generate domain-specific visual instruction tuning dataset is an effective way to train SMMs for VQA tasks related to nanomaterial images. This approach addresses the scarcity of vision-language instruction-following data and enhances SMMs domain-specific adaptation and alignment abilities, allowing them to perform comparably to proprietary LMMs without requiring excessive computational costs. Transfer learning is also used to improve generalization of SMMs, and the benefits of this approach include: (a) enhancing SMMs reasoning abilities for complex visual questions, (b) improving zero-shot learning for new questions on unseen nanoim-

ages, (c) facilitating knowledge distillation from larger models to transfer insights about nanomaterial structures and patterns, and (d) generating diverse question-answer pairs to enrich training data and expand the smaller models capabilities. Our method employs zero-shot CoT prompting to guide GPT-4 in automatic generation of a novel instruction-following dataset (question-answer pairs) for training SMMs and involves natural language questions that analyze nanomaterials' size, distribution, morphology, and structure in microscopic images. Our approach effectively links natural language instructions (query text) with visual representations (query image), thereby enhancing SMMs' responsiveness to complex visual queries and aiding in understanding the visual representations of concept-based questions and answers. The customized CoT prompt format is as follows:

Prompt 1: **Basics** - This image depicts a nanomaterial. Identify the specific type of nanomaterial depicted in the image.? Additionally, find image scale: real-world length per unit measurement?. **Prompt 2: **Morphology and Structure**** - Describe the overall shape and morphology of the nanomaterials?. Identify any visible layers, phases, or distinct domains?. Assess consistency in size and shape, or note any variability?. **Prompt 3: **Size and Distribution**** - Estimate size/size range of nanostructures?. - Describe distribution - evenly spaced, clustered, or random?. - Comment on any aggregation or bundling visible?. elements/compounds?. **Prompt 4: **Surface Characteristics**** - Describe surface textures - smooth, rough, distinct textures?. - Comment on any visible imperfections like defects, pores, or impurities?. **Prompt 5: **Composition and Elements**** - Note any visible evidence of compositional variations (color, brightness, contrast differences)?. - Identify any labels or markers pointing to specific **Prompt 6: **Interactions and Boundaries**** - Describe visual interactions: touching, fused, or separate?. - Can you distinguish boundaries between structures/phases? Or do they blend without defined borders?. **Prompt 7: **External Environment**** - Note any visible signs of interaction between nanomaterials and surroundings (solvents, polymers, etc.)? - Identify and describe any non-nanomaterial structures/objects present?. **Prompt 8: **Image Technique and Modifications**** - Identify imaging technique used (SEM, TEM, etc.)? - Note any visible post-processing or modifications like false coloring or 3D rendering?. **Prompt 9: **Functional Features**** - Identify any visible functional elements or regions with distinct properties?. - Note if the image shows any dynamic processes, or if it is primarily static?. **Prompt 10: **Context and Application**** - Identify intended use/application of nanomaterials. - Are they experimental samples or theoretical/simulation-based representations?

Model Architecture: Figure 1 illustrates an encoder-decoder architecture designed to comprehend visual and textual inputs and generate coherent responses for complex Visual Question Answering (VQA) tasks. It utilizes a visual transformer for image encoding, dividing the input image into patches and converting them into embeddings. A $\langle cls \rangle$ token encapsulates the global image via self-attention. The text encoder follows the BERT architecture, also starting with a $\langle cls \rangle$ token for summarizing sentences. Unimodal encoders interpret textual questions and analyze visual inputs to generate contextually appropriate responses. The image-grounded text encoder integrates both visual and textual data through cross-attention, understanding image content and query text semantics for accurate answer generation. A $\langle Encode \rangle$ token facilitates multimodal integration, representing the fused image-text representation. The image-grounded text decoder utilizes causal attention for generative decoding, marked by a $\langle Decode \rangle$ token at the start and an end-of-sequence ($\langle EOS \rangle$) token at the end, guiding the auto-regressive decoding mechanism. Our proposed multimodal learning method has three main objectives: understanding-based goals focus on minimizing image-text contrastive and matching losses to comprehend visual and textual content. Generation-based objectives aim to minimize language modeling loss for accurate answer generation. We employ joint optimization, training on all objectives simultaneously to excel in natural instruction-following and visual reasoning for microscopic image-based VQA tasks. The image-text contrastive (ITC) loss minimizes the distance between matching pairs while maximizing it for non-matching pairs, aligning representations in a shared embedding space, based on noise-contrastive estimation principles is expressed as:

$$L_{ITC} = \frac{1}{2}(L_{I2T} + L_{T2I})$$

$$= -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{e^{\text{sim}(v_i, t_i)/\tau}}{\sum_{j=1}^N e^{\text{sim}(v_i, t_j)/\tau}} + \log \frac{e^{\text{sim}(v_i, t_i)/\tau}}{\sum_{j=1}^N e^{\text{sim}(v_j, t_i)/\tau}} \right]$$

Where N is the number of image-text pairs in the batch. v_i and t_i are the embeddings of the image and text, respectively, in the i -th pair. Here, $\text{sim}(v_i, t_i)$ is the similarity score between the i -th image embedding v_i and text embedding t_i , often calculated using the dot product. τ is the temperature parameter that scales the similarity measure. L_{I2T} represents the loss for aligning images to texts (image-to-text contrastive loss) and L_{T2I} is the loss for aligning texts to images (text-to-image contrastive loss). The total ITC loss is the average of these two losses across all image-text pairs in the batch. The loss function drives both the unimodal encoders (visual and text transformers) to align matching image-text pair representations and distinguish non-matching representations, fostering a cross-modal semantic understanding. (b) **The image-text matching (ITM) loss**, using binary cross-entropy loss in multimodal learning, is designed to encourage the image-grounded text encoder to correctly identify whether an image and text representation form a matching pair or not. The parameters of the image-text encoder are updated to minimize this loss, thereby improving the alignment of image-text multimodal representations in the shared embedding space. It penalizes the encoder for incorrect predictions, guiding it to

learn better representations for image-text matching pairs. Let y_i denote the ground truth label for the i -th image-text pair in a batch, where $y_i = 1$ if the image and text match (are relevant to each other), and $y_i = 0$ otherwise. Let p_i be the predicted probability of pairs being positive (matched) that the i -th image and text match. The probability p_i is computed from the output linear layer of the image-grounded text encoder by applying a sigmoid function. The binary cross-entropy loss for the ITM task over a batch of size N can be formulated as follows:

$$L_{\text{ITM}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

(c) **Language modeling loss (LM)** is particularly used for the VQA task, which focuses on generating coherent and contextually relevant text when presented with an image and a question related to that image. The image-grounded text decoder minimizes the LM loss by generating textual descriptions that accurately describe the visual content in images. Specifically, it learns to accurately predict each word in a sentence based on the preceding words and the contextual visual information provided by the corresponding image. The autoregressive decoder aims to maximize the likelihood of the correct words in the text sequence, by refining the model’s ability to understand and answer questions about images. This involves minimizing the negative log-likelihood of the ground truth words under the predicted probabilities of the image-grounded text decoder, thereby leading to improved text generation that aligns with the image.

$$L_{\text{LM}} = -\sum_i^N \log P(w_i | w_{<i}, I, Q)$$

Where L_{LM} represents the language modeling loss, N is the number of words in the text, w_i represents the i -th word in the text, $w_{<i}$ represents all words before the i -th word, I is the image corresponding to the text, and $P(w_i | w_{<i}, I, q)$ is the probability of the i -th word given the preceding words and the image, as predicted by the model. q refers to the question that the generated text aims to answer when conditioned on both the image I and the previous words $w_{<i}$ in the sequence. During inference time, the decoder generates accurate text descriptions for a given image using the knowledge it has acquired during training.

Experiments And Results

Datasets: Our study used the SEM dataset (Aversa et al. 2018) to automate VQA task for nanomaterial image interpretation and analysis. This dataset contains 21,283 electron micrographs across 10 categories, including *particles*, *nanowires*, and *patterned surfaces*. Figure 3 displays the different nanomaterial categories in the SEM dataset. Initial findings (Modarres et al. 2017) on the image classification task were based on a subset, while our research utilized the complete dataset for both the zero-shot VQA and image classification tasks. In our work, to ensure a rigorous comparison with popular baseline models, we employed k-fold cross-validation, as no predefined splits were provided by the dataset curator.

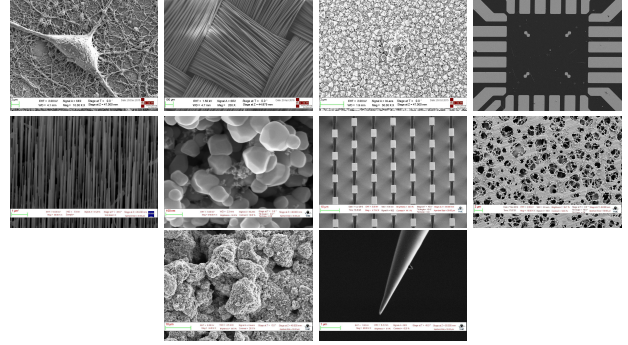


Figure 3: The figure displays nanomaterials from the SEM dataset. From left to right in the first, second, and third rows, we have: *biological*, *fibers*, *films*, *MEMS*; *nanowires*, *particles*, *patterned surface*, *porous sponges*; and *powder*, *tips*.

Experimental Setup: The SEM dataset (Aversa et al. 2018) consists of electron micrographs with dimensions of $1024 \times 768 \times 3$ pixels. We downscale these to $224 \times 224 \times 3$ pixels and normalize the micrographs by adjusting the mean and covariance to 0.5 across channels, resulting in values within $[-1, 1]$. We then tokenize the downscaled and normalized micrographs into non-overlapping 32 pixel patches. The patch and position embedding dimensions are set to 64. We use 10-fold cross-validation and train for 50 epochs with an initial learning rate of $1e^{-3}$ and batch size of 48. For the self, cross-modal and casual attention layers, the number of heads is 4 and key/query/value dimensionality is 16. We employ early stopping on the validation set to prevent overfitting and a learning rate scheduler that halves the learning rate if validation loss stagnates for 5 epochs. We also use the Adam optimization algorithm (Kingma and Ba 2014) to update the framework’s trainable parameters. We assess the performance of MVaEMa in instruction-following and visual reasoning capabilities using the SEM dataset on nanoimage analysis tasks. In our work, we utilize GPT-4 to obtain a multimodal instruction-following dataset (question-answer pairs) on the SEM dataset. We set the temperature to 0.25 to control randomness in text generation and top-p sampling to 0.1 to narrow down word choices for more deterministic output. Additionally, we set the maximum number of output tokens to 3500. We implement the framework in pytorch (Paszke et al. 2019) and pretrained on $4 \times V100$ GPUs.

Due to the potentially high computational cost of using prompting with large multimodal models, we conducted each experiment twice and reported the averaged results.

VQA Results: In VQA tasks, text quality is assessed using metrics like BLEU, METEOR, and ROUGE. BLEU-N evaluates machine-generated text similarity to reference texts based on overlapping n-word phrases, prioritizing precision. METEOR combines unigram precision and recall, incorporating linguistic concepts like stemming and synonym matching for paraphrase handling. ROUGE-N computes overlapping n-grams between candidate and reference texts to assess an-

Method	BLEU-2 (↑)	BLEU-4 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	METEOR (↑)
InstructBLIP	0.570±0.063	0.457±0.078	0.745±0.032	0.648±0.011	0.705±0.042	0.738±0.048
LLaVA	0.620±0.070	0.512±0.085	0.760±0.032	0.668±0.011	0.723±0.042	0.753±0.046
MiniGPT-4	0.680±0.075	0.572±0.090	0.790±0.033	0.698±0.012	0.753±0.043	0.783±0.047
MVaEMa	0.780 ±0.085	0.709 ±0.105	0.860 ±0.036	0.765 ±0.014	0.822 ±0.050	0.853 ±0.055

Table 1: The table presents the experimental results comparing the performance of the MVaEMa framework on the VQA task against the baseline models.

Algorithms		Top-1	Top-5
ConvNets	AlexNet	0.528	0.827
	DenseNet	0.569	0.929
	ResNet	0.485	0.897
	VGG	0.538	0.808
	GoogleNet	0.609	0.969
	SqueezeNet	0.404	0.698
VSL	Barlowtwins	0.148	0.410
	SimCLR	0.130	0.379
	byol	0.143	0.453
	moco	0.169	0.472
	nnclr	0.158	0.563
	simsiam	0.188	0.535
Vision Transformers (ViTs)	CCT	0.570	0.981
	CVT	0.577	0.930
	ConViT	0.609	0.957
	ConvViT	0.319	0.921
	CrossViT	0.442	0.915
	PVTC	0.596	0.964
	SwinT	0.707	0.993
	VanillaViT	0.655	0.970
	Visformer	0.398	0.856
	ATS	0.540	0.973
	CaiT	0.657	0.989
	DeepViT	0.546	0.988
	Dino	0.049	0.437
	Distillation	0.533	0.955
	LeViT	0.624	0.970
	MA	0.202	0.491
	NesT	0.660	0.985
	PatchMerger	0.578	0.975
PiT	0.555	0.979	
RegionViT	0.606	0.948	
SMIM	0.171	0.646	
T2TViT	0.749	0.992	
MVaEMa	0.947	0.988	

Table 2: The table compares our method to baseline algorithms on nanomaterial image classification task. For more information on the baseline algorithms, refer to our previous paper((Sakhinana, Geethan, and Runkana 2023))

answer completeness in VQA, with variants like ROUGE-L

Algorithms		Top-1	Top-5
GCL	GBT	0.547	0.706
	GRACE	0.598	0.750
	BGRL	0.556	0.696
	InfoGraph	0.526	0.702
Graph Neural Networks	APNP	0.632	0.786
	AGNN	0.538	0.894
	ARMA	0.582	0.987
	DNA	0.622	0.916
	GAT	0.491	0.985
	GGConv	0.563	0.992
	GraphConv	0.658	0.996
	GCN2Conv	0.732	0.998
	ChebConv	0.504	0.951
	GraphConv	0.509	0.993
	GraphUNet	0.657	0.978
	MPNN	0.603	0.999
	RGGConv	0.618	0.961
	SuperGAT	0.598	0.985
MVaEMa	0.947	0.988	

Table 3: The table presents a performance comparison of supervised-learning GNNs, self-supervised GCL algorithms, and our novel method for nanomaterial classification task. To learn more about the baseline algorithms, please refer to our previous paper ((Sakhinana, Geethan, and Runkana 2023)).

measuring longest common subsequence matches. These metrics focus on various aspects of text generation, including similarity, linguistic quality, and coherence. Compared to other multimodal models like InstructBLIP((Dai et al. 2023)), LLaVA((Liu et al. 2023)), and MiniGPT-4((Zhu et al. 2023)), MVaEMa excels in seamlessly integrating fine-grained visual details with coherent reasoning for long-form responses, a feature lacking in other models. We argue that preference for long or short responses in VQA tasks should consider question requirements, user needs, and context, aiming for a balance between providing sufficient information and maintaining clarity and conciseness. Table 1 reports the experimental results on the VQA task in comparison to the baselines. Unlike LLaVA and MiniGPT-4, which generate lengthy and less relevant responses, the MVaEMa framework adjusts response length adaptively for optimal relevance. These advantages stem from diverse instruction tuning data and effective architectural design. To compare with our algorithm, we fine-tuned

the baselines on nanoimage analysis tasks and evaluated their performance.

Category	Multi-class metrics		
	Precision	Recall	F1 Score
Biological	0.959	0.975	0.965
Tips	0.937	0.949	0.946
Fibres	0.983	0.992	0.990
Porous Sponge	0.957	0.969	0.953
Films Coated Surface	0.967	0.963	0.971
Patterned surface	0.975	0.971	0.970
Nanowires	0.967	0.974	0.977
Particles	0.963	0.965	0.957
MEMS	0.967	0.960	0.951
Powder	0.969	0.956	0.945

Table 4: Effectiveness of our proposed framework in terms of precision, recall, and F1-score for accurately classifying nanomaterials of different categories.

Image Classification Results: We assessed our proposed framework against common computer vision baselines such as ConvNets, ViTs(al. 2022b,a), and self-supervised vision contrastive learning (VCL)(et al. 2020) algorithms on the zero-shot image classification task. The multimodal prompt (query image and text) didn’t include the image caption. Additionally, we compared the framework’s performance to supervised graph neural networks (GNNs(Rozemberczki et al. 2021; Fey and Lenssen 2019)) and graph contrastive learning (GCL(Zhu et al. 2021)) algorithms, measuring Top-N accuracy for $N = 1, 5$. Tables 2 and 3 show experimental results comparing our framework with baseline algorithms. Under consistent settings, our framework outperformed the baselines, demonstrating a 26.44% relative improvement in Top-1 accuracy compared to the next-best model, T2TViT((Yuan et al. 2021)). We conducted extra experiments to evaluate our framework’s ability to categorize electron micrographs across diverse nanomaterial categories. Using a multi-metric approach, we employed a confusion matrix to calculate precision, recall, and F1-score metrics. This provided insights into how effectively our framework categorized micrographs based on varied structures, patterns, and complexity. The results in Table 4 show that our framework could generalize across various nanomaterial categories, including those with complex patterns.

Ablation Study: To validate the effectiveness of the methods in our framework, we conducted ablation studies by systematically disabling certain methods to create ablated variants and were evaluated using the SEM dataset (Aversa

et al. 2018), with our original framework as the baseline for comparison on both VQA and image classification tasks. The ablation study enables us to verify the efficacy of our methods, substantiate their neural network designs, and justify their inclusion in the framework. A substantial performance decrease in the ablated variants compared to the baseline highlights the importance of the omitted methods. We evaluate the ablated variants performance on metrics such as precision and recall for image classification tasks, or other relevant measures for VQA task. The ablated variants that exclude the image-text contrastive loss(ITC), binary cross entropy loss (CTC), and the self-attention(SA), cross attention(CA), causal self-attention(CSA) mechanisms are denoted as proposed framework “w/o ITC”, “w/o CTC”, “w/o SA”, “w/o CA”, and “w/o CSA” respectively. The abbreviation “w/o” stands for “without”. Across all ablated variants, we observe a consistent decline in performance metrics compared to the baseline. These results clearly validate the crucial contribution of each omitted method through our ablation studies. Tables 5 and 6 shows the ablation study results on the VQA and classification tasks, respectively.

Method	BLEU-4	ROUGE-L	METEOR
w/o ITC	0.579	0.670	0.696
w/o CTC	0.569	0.670	0.696
w/o SA	0.682	0.775	0.794
w/o CA	0.652	0.755	0.769
w/o CSA	0.649	0.740	0.761
MVaEMa	0.709	0.822	0.853

Table 5: Ablation study results on the VQA task.

Algorithms	Avg-Precision	Avg-Recall	Avg-F1 Score
w/o ITC	0.752	0.731	0.717
w/o CTC	0.746	0.773	0.759
w/o SA	0.927	0.912	0.895
w/o CA	0.867	0.872	0.860
w/o CSA	0.843	0.866	0.885
MVaEMa	0.964	0.967	0.962

Table 6: Ablation study results on image classification task.

Conclusion

Our research introduces a small-scale language-and-vision assistant for electron micrograph analysis, trained on a novel dataset generated by GPT-4 Turbo with Vision. This framework excels in visual question answering tasks, particularly in nanomaterial image analysis, while allowing for secure enterprise applications through fine-tuning with proprietary data.



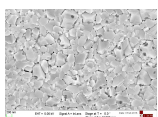
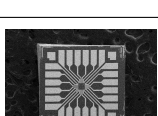
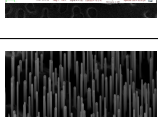
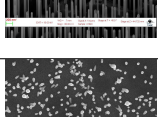
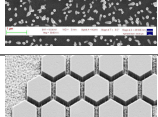
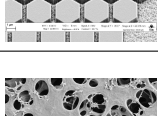
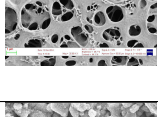
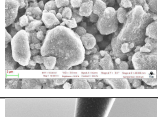
Image	Ground Truth	Answers	BLUE-2	ROGUE-L	METEOR
	The nanomaterials in the image have a dendritic, branching structure with a central node and multiple filament-like extensions.	The nanomaterials in the image possess a dendritic, branching structure with a central node and several filament-like extensions.	0.824	0.895	0.944
	The nanomaterials depicted resemble tightly woven, twisted cables or fibrous strands, densely packed and intertwined.	The nanomaterials depicted appear as tightly woven, twisted cables or fibrous strands, densely packed and interlaced.	0.772	0.839	0.859
	The nanomaterials have a polygonal, plate-like morphology with irregular edges, giving them a shattered glass or cracked ice appearance.	The nanomaterials have polygonal, plate-like morphology with irregular edges, giving them a shattered glass or cracked ice appearance.	0.918	0.974	0.952
	The image depicts a square microfabricated device with uniform linear patterns on a granular semiconductor or nanoparticle substrate.	The image shows a square microfabricated device with uniform linear patterns on a granular semiconductor or nanoparticle substrate.	0.913	0.944	0.999
	The nanomaterials in the image exhibit a needle- or rod-like morphology, standing vertically and densely packed, similar to a bed of nails.	The nanomaterials in the image display a needle- or rod-like morphology, standing vertically and densely packed, akin to a bed of nails.	0.858	0.913	0.954
	The nanomaterials shown are elliptical or rod-shaped with smooth surfaces, scattered randomly across the surface.	The nanomaterials displayed are elliptical or rod-shaped with smooth surfaces, dispersed randomly across the surface.	0.787	0.875	0.861
	The nanomaterials have a hexagonal, honeycomb-like structure, organized in a highly ordered, tessellated pattern.	The nanomaterials display a hexagonal, honeycomb-like structure, organized in a highly ordered, tessellated pattern.	0.886	0.933	0.927
	The nanomaterials exhibit a foam-like structure with a network of interconnected pores of various sizes and irregular shapes, creating a porous, sponge-like morphology.	The nanomaterials display a foam-like structure with a network of interconnected pores of various sizes and irregular shapes, forming a porous, sponge-like morphology.	0.820	0.920	0.913
	The nanomaterials are irregularly shaped, resembling clumped aggregates with a rough, textured surface.	The nanomaterials appear irregularly shaped, resembling clumped aggregates with a rough, textured surface.	0.877	0.920	0.920
	The nanomaterial appears as a sharply pointed, conical structure with a smooth surface, tapering to a fine tip.	The nanomaterial is seen as a sharply pointed, conical structure with a smooth surface, tapering to a fine tip.	0.863	0.920	0.938

Table 7: The table shows illustrative microscopic images, ground-truth and model-generated answers for the question to describe the overall shape and morphology of the nanomaterials. In addition, we report the BLUE-2, ROGUE-L, METEOR scores.

References

- al., N. S. 2022a. VFormer: A modular PyTorch library for vision transformers. *GitHub*. Note: <https://github.com/SforAiDl/vformer>.
- al., P. W. 2022b. Vision Transformer - Pytorch. *GitHub*. Note: <https://github.com/lucidrains/vit-pytorch>.
- Aversa, R.; Modarres, M. H.; Cozzini, S.; Ciancio, R.; and Chiusole, A. 2018. The first annotated set of scanning electron microscopy images for nanoscience. *Scientific data*, 5(1): 1–10.
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-

purpose Vision-Language Models with Instruction Tuning. *arXiv* 2023. *arXiv preprint arXiv:2305.06500*.

et al., I. S. 2020. Lightly. *GitHub*. *Note: https://github.com/lightly-ai/lightly*.

Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Modarres, M. H.; Aversa, R.; Cozzini, S.; Ciancio, R.; Leto, A.; and Brandino, G. P. 2017. Neural network for nanoscience scanning electron microscope image recognition. *Scientific reports*, 7(1): 1–12.

OpenAI. 2023. GPT-4V(ision) System Card.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rozemberczki, B.; Scherer, P.; He, Y.; Panagopoulos, G.; Riedel, A.; Astefanoaei, M.; Kiss, O.; Beres, F.; ; Lopez, G.; Collignon, N.; and Sarkar, R. 2021. PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 4564–4573.

Sakhinana, S.; Geethan, S.; and Runkana, V. 2023. Hierarchical Network Fusion for Multi-Modal Electron Micrograph Representation Learning with Foundational Large Language Models. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Sun, T.; Shao, Y.; Qian, H.; Huang, X.; and Qiu, X. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, 20841–20855. PMLR.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 558–567.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, Y.; Xu, Y.; Liu, Q.; and Wu, S. 2021. An Empirical Study of Graph Contrastive Learning. *arXiv.org*.