Preliminary Investigations of a Multi-Faceted Robust and Synergistic Approach in Semiconductor Electron Micrograph Analysis: Integrating Vision Transformers with Large Language and Multimodal Models

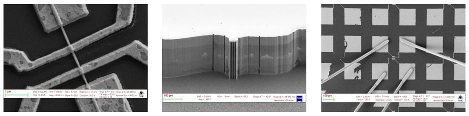
Anonymous submission

Abstract

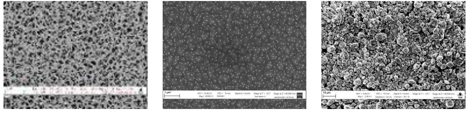
Characterizing materials using electron micrographs is crucial in areas such as semiconductors and quantum materials. Traditional classification methods falter due to the intricate structures of these micrographs. This study introduces an innovative architecture that leverages the generative capabilities of zero-shot prompting in Large Language Models (LLMs) such as GPT-4(language only), the predictive ability of few-shot (in-context) learning in Large Multimodal Models (LMMs) such as GPT-4(V)ision, and fuses knowledge across image-based and linguistic insights for accurate nanomaterial category prediction. This comprehensive approach aims to provide a robust solution for the automated nanomaterial identification task in semiconductor manufacturing, blending performance, efficiency, and interpretability. Our method surpasses conventional approaches, offering precise nanomaterial identification and facilitating high-throughput screening.

Introduction

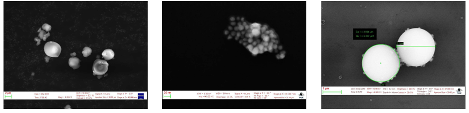
Semiconductors have been the backbone of technological advancements in modern electronics, driving growth and innovation in computing and communication systems, among others. The semiconductor process comprises three main stages: (a) design and development, during which fabless firms create chip blueprints, specifying the architecture, functions, and specifications of the miniaturized chips; (b) fabrication, where specialized foundries manufacture chips by etching integrated circuits onto silicon wafers using intricate technologies; and (c) testing and assembly, during which chips undergo rigorous testing and are subsequently assembled into protective packages for integration into electronic devices. This collective effort results in the production of high-quality semiconductor components suitable for a wide range of applications. The state-of-the-art imaging and analysis methods(Holt and Joy 2013) are indispensable in semiconductor manufacturing for the development of next-generation miniaturized chips, especially those sized at 7 nm or smaller. The pursuit of miniaturized chips below 7 nm technologies introduces a level of complexity and precision that significantly increases the risk of errors in the manufacturing process. These errors can compromise the consistency of high-quality chips and amplify the variability in chip performance, posing a substantial challenge for manufacturers aiming to produce reliable and advanced chips at this scale. The semiconductor industry utilizes various advanced electron beam tools, including scanning and transmission electron microscopy, to create images or micrographs of semiconductor materials, structures, and devices at the micro and nanoscale with high resolution and detail. These tools contribute to quality control, process monitoring, failure analysis, and materials characterization in the semiconductor industry. Automated labeling of electron micrographs, though advantageous, poses a considerable challenge due to the level of detail, complexity of patterns, and information density involved. These challenges arise primarily from the high inter-category similarity (similar-looking or indistinguishable) between different nanomaterials, high intra-category dissimilarity within nanomaterials (distinct or differing appearances), and the presence of intricate visual patterns in nanomaterials across various scales (spatial heterogeneity). The complexities of automated nanomaterial identification tasks are illustrated in Figure 1.



(a) High intra-class dissimilarity: *MEMS* devices exhibit a high degree of heterogeneity.



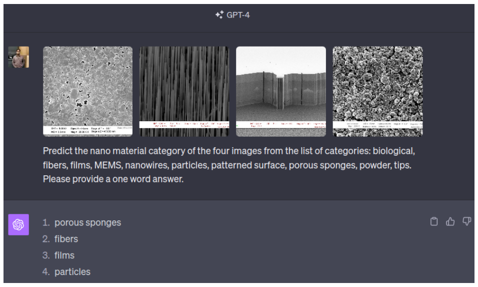
(b) High inter-class similarity: Different nanomaterial categories (*listed from left to right as porous sponges, particles and powders*), are similar-looking or indistinguishable.



(c) Multi-spatial scale (spatial heterogeneity) patterns in electron micrographs of *nanoparticles* are evident.

Figure 1: The figure provides a visual representation of the challenges of classifying electron micrographs in the SEM dataset(Aversa et al. 2018).

Recently, unimodal Large Language Models (LLMs) such as GPT-4 (language-only)(OpenAI 2023a) which are pretrained autoregressive large-scale models on extensive, diverse text corpora in unsupervised learning settings following a fundamental paradigm "prompt and predict" approach, have significantly transformed natural language processing(NLP), achieving improved performance across a wide range of NLP tasks, demonstrating better logical reasoning abilities, and generating human-like text. Zero-shot Chain of Thought(Zero-Shot CoT)(Wei et al. 2022) and Few-Shot (In-Context) learning(Few-Shot ICL)(Brown et al. 2020) are prompt engineering strategies for designing and crafting tailored prompts for utilizing general-purpose LLMs in specialized language-based tasks or associated new, unseen problem-solving scenarios, thereby eliminating the need for traditional task-specific fine-tuning. Zero-Shot CoT relies on customized instructions without requiring explicit task-specific demonstrations(input-output pairs), requiring the language model to generalize from the implicit knowledge acquired during training to generate the output for the downstream task. Conversely, Few-shot ICL incorporates a few guiding demonstrations to learn from analogy along with the task-centric instructions to guide LLMs to generate the output simply by conditioning on the prompts. In recent times, OpenAI's GPT-4 with Vision (GPT-4V)(OpenAI 2023b), which possesses the ability to process and understand images, represents a significant advancement in the domain of large multimodal models (LMMs). It is more versatile than GPT-4, as it has broken the text-only barrier of previous language models, introducing visual understanding and analysis as a new dimension. GPT-4V is designed to accept multiple modalities, including both images and text as input, and generate text outputs. GPT-4V incorporates visual processing capabilities, enabling it to analyze image inputs provided by the user in conjunction with text, thereby facilitating visual question answering. Despite its advanced capabilities, when tested on SEM images(Aversa et al. 2018) for nanomaterial categorization, GPT-4V incorrectly classified them, highlighting the limitations of LMMs. Figure 2 illustrates these limitations of GPT-4V.



Predict the nano material category of the four images from the list of categories: biological, fibers, films, MEMS, nanowires, particles, patterned surface, porous sponges, powder, tips. Please provide a one word answer.

1. porous sponges
2. fibers
3. films
4. particles

Figure 2: The electron micrographs shown above were provided as input to GPT-4V(Yang et al. 2023) for nanomaterial categorization to determine how the multimodal model classifies nanomaterials in SEM images across different structural categories from a predefined list. However, the LMMs predictions were incorrect, with the actual nanomaterial categories being films, nanowires, MEMS and powder. It highlights the inherent limitations of visual processing capabilities of even advanced LMMs such as GPT-4V, reminding users to approach predictions with a degree of skepticism.

Despite advances in the use of language-only LLMs such as GPT-4, LMMs like GPT-4V and other behemoths across various scientific disciplines, the synergistic integration of foundational LLMs and LMMs with computer vision algorithms in semiconductor research, particularly for the automated electron micrograph identification task, remains an underexplored approach in the evolution of intelligent semiconductor manufacturing processes. In this study, we introduce an automated nanomaterial identification framework, which is built upon cross-modal electron micrograph representation learning, referred to as `CM-EMRL` for brevity. Figure 3 illustrates the framework. The objective is to utilize the complementary strengths of LMMs, LLMs, and small-scale language models(LMs) to establish a more robust and accurate predictive framework. Closed-source LLMs like GPT-4(OpenAI 2023a), while proficient in language understanding, have a black-box nature, and lack interpretability for downstream applications since they typically do not provide direct access to logits or token embeddings. In addition, their jack-of-all-trades approach makes them incredibly resource-intensive for repurposing and often poorly aligned with specialized tasks. On the other hand, open-source smaller LMs like BERT(Devlin et al. 2018) following "pretraining and fine-tuning" approach, while cost-effective for repurposing through fine-tuning to align with specialized tasks and be interpretable, may fall short in terms of reasoning and generalization, often yielding less coherent and contextually relevant responses compared to LLMs. LMMs such as GPT-4V (OpenAI 2023b) are more potent and versatile than language-only LLMs, as they incorporate multi-sensory capabilities that combine visual and language understanding. This enables users to instruct the multimodal model to analyze image inputs alongside textual information. Consequently, it offers the ability to perform complex tasks that require an understanding of both text and visual inputs, producing output that is contextually relevant to the combined data. LMMs excel in multimodal processing, with their remarkable capabilities to analyze images, identify objects, transcribe text, and decipher data, but they grapple with challenges related to interpretability, bias, unpredictability, and resource consumption. Navigating these challenges among LLMs, LMMs, and small-scale LMs demands a fine balance between performance, efficiency, and interpretability. Our study introduces a novel approach to the automatic nanomaterial identification task, combining the strengths of LMMs, LLMs, and small-scale LMs. The main contributions of this work can be summarized as follows:

- **Utilizing Vision Transformers (ViT) for Holistic Image Representation:** An input image is divided into patches treated as tokens, converted into 1D vectors, and enhanced with positional embeddings for location context. A classification token is added to achieve a global image representation. This token sequence is processed by a transformer architecture, specifically ViT(Dosovitskiy et al. 2020), to identify relationships between different image regions. The output corresponding to the classification token provides a comprehensive image representation. This appraoch employs the transformers to encapsulate the entire image visual context by treating the classification token's latent representation as an image-level embedding.
- **Zero-shot Chain-of-Thought(CoT) LLMs Prompting and Cross-Modal Alignment:**Our study leverages powerful LLMs through Language Model as a Service (LMaaS), using their text-based inference APIs without accessing

the parameters and gradients of the LLMs. Compared to fine-tuning task-specific LLMs, LMaaS efficiently deploys a single versatile LLM for various tasks using task-specific prompts. It optimizes tasks without backpropagation, ensuring low costs irrespective of the model size. The open-ended natural language CoT prompts guide LLMs to generate comprehensive textual descriptions of nanomaterials, covering their structure, synthesis methods, properties, and applications. After obtaining these technical descriptions, a smaller pretrained language model (LM) is used to summarize the LLM-generated content and compute high-level text embeddings by aligning them through supervised fine-tuning for the downstream nanomaterial identification task. Furthermore, in a cross-modal alignment scenario, a scaled dot-product attention mechanism matches image embeddings with their corresponding text-level embeddings. This mechanism calculates similarity scores to determine the best text match for a given image, ensuring robust alignment between different modalities. In brief, we utilize CoT LLM prompting to extract domain-specific knowledge and obtain image-aligned (nanomaterial-specific) text-level embeddings.

- **In-Context Learning in LMMs via Few-Shot Prompting for Nanomaterial Identification:** Utilizing few-shot prompting can quickly adapt LMMs to perform new tasks, like nanomaterial identification, without extensive fine-tuning. By providing LMMs with a limited set of image-label pairs, these models can predict the category of unfamiliar or unseen nanomaterial images. Two strategies, random and similarity-driven sampling, influence the selection of these pairs. Instead of updating model parameters through supervised learning, this approach leverages the multimodal model's inherent knowledge, generating prediction embeddings by solely conditioned on the prompt.
- **Unified Attention Layer:** We utilize a hierarchical multi-head attention mechanism to compute a cross-modal embedding from image-level, text-level, and prediction embeddings. This robust framework effectively integrates diverse information across these modalities, producing a holistic representation that can improve nanomaterial identification tasks.

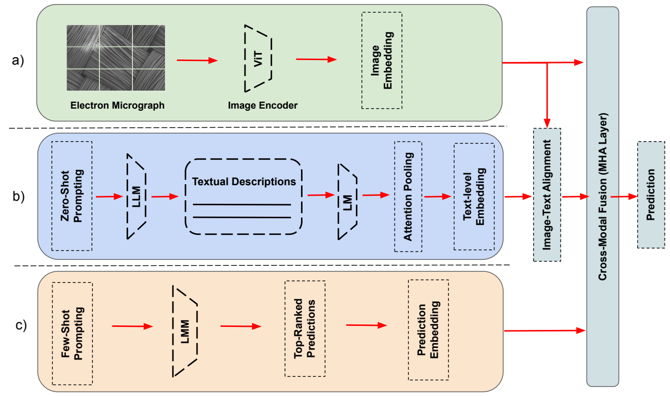


Figure 3: Our framework includes three methods: (a) Image Encoder (ViT), (b) Zero-Shot CoT prompting with LLMs, (c) Few-shot prompting with LMMs, and (d) an output layer modeled with the multi-head attention (MHA) mechanism for integrating cross-domain embeddings and facilitating label prediction.

Problem Statment

Our study focuses on the classification of electron micrographs, an inductive learning challenge where the goal is to categorize previously unobserved micrographs by leveraging a labeled dataset, represented as $\mathcal{D}_L = (\mathcal{I}_L, \mathcal{Y}_L)$. We train a multi-modal encoder, defined by the non-linear transformation $g_\gamma : \mathcal{I} \rightarrow \mathcal{Y}$ on the labeled dataset in the context of supervised machine learning to predict the labels ($\mathcal{Y}_U$) for unlabeled micrographs ($\mathcal{I}_U$). Within this context, γ denotes the trainable parameters, with the primary aim being to reduce the loss $\mathcal{L}_{\mathcal{I}}$, which is framed as:

$$\min_{\gamma} \mathcal{L}_{\mathcal{I}}\left(\mathcal{I}_i, \gamma\right) = \sum_{(\mathcal{I}_i, y_i) \in \mathcal{D}_L} \ell\big(g_\gamma(\mathcal{I}_i), y_i\big) \qquad (1)$$

where $y_i^{\text{pred}} = g_\gamma(\mathcal{I}_i)$ represents the predictions from the multi-modal encoder, $\ell(\cdot, \cdot)$ signifies the cross-entropy loss.

Proposed Method

Electron Micrograph Encoder: Let's consider an input image $\mathbf{I}$, which is represented as a 3D tensor with dimensions $H \times W \times C$, where H represents the image's height in pixels, W represents its width in pixels, and C represents the number of channels of each pixel within the image. We divide the image into smaller, non-overlapping regions or patches to obtain a sequence of tokens. Each patch is treated as a token and has a fixed size with spatial dimensions of $P \times P \times C$, where P denotes the patch size. The total number of patches is given by $n = \left(\frac{HW}{P^2}\right)$. We then linearly encode each patch, each of which has an overall size of P^2C, to flatten it into a 1D vector represented as $\mathbf{I}' \in \mathbb{R}^{n \times d}$, where d is the patch embedding dimension. To provide the model with spatial information, we add positional embeddings to each patch. These positional embeddings are learnable vectors representing the position of each patch within the image grid. They help us understand the relative positions of different patches, and we add the position embeddings element-wise to the patch embeddings. In addition, we append a classification token <*cls*> to the patch sequence. This token aggregates information from all patches and provides an embedding of the entire patch sequence, creating a global representation of the entire input image. We input this augmented token sequence into ViT(Dosovitskiy et al. 2020), which consists of multiple stacked transformer encoder layers. Each encoder layer processes the patch embeddings hierarchically, refining representations at different abstraction levels. After passing through the transformer layers, we consider only the output embedding h_{cls} corresponding to the <*cls*> token as the representation of the entire image. This embedding aggregates information from all the patches and summarizes it. In summary: (a) We split the input image into patches, flatten them into 1D vectors, augment them with positional embeddings to provide spatial information, and include a classification token <*cls*>. (b) We process the resulting sequence of patch embeddings using a transformer-based architecture to capture long-range dependencies and relationships between different regions of the electron micrograph.

Zero-Shot CoT LLMs Prompting: We access LLMs through the LMaaS(Sun et al. 2022) hosted by the cloud-based service provider, which provides a text-based black-box API interaction to send text inputs and receive generated text without access to the underlying model parameters or gradients or the model's internal mechanisms. We utilize open-ended natural language prompts, designed to be flexible and non-restrictive, to instruct LLMs to generate the detailed technical descriptions related to various aspects of nanomaterials, including their structure, properties, and applications. We query LMMs to generate detailed descriptions of nanomaterials by employing a customized prompt template for zero-shot generation tasks, steering them through CoT prompts in a zero-shot context. This process involves extracting pre-existing domain-specific knowledge embedded within the language model parameters acquired during training to generate in-depth, technical descriptions of nanomaterials that are both detailed and comprehensive. The customized CoT prompt format is as follows:

Prompt 1: Introduction: Provide an overview of the nanomaterial category and its significance across various fields. **Prompt 2:** Definition and Structure: Define the nanomaterial category and describe its typical structure at the nanoscale. **Prompt 3:** Synthesis Methods: Examine different methods employed for synthesizing or fabricating nanomaterials within this category. Discuss both their advantages and limitations. **Prompt 4:** Properties: Highlight the unique physical, chemical, and electronic properties exhibited by nanomaterials in this category. Explain how these properties differ from those of bulk materials. **Prompt 5:** Surface Modification: Describe strategies used to modify the surface properties of nanomaterials in this category, including techniques like functionalization, coating, or doping. Explain how these modifications enhance their performance or enable specific applications. **Prompt 6:** Applications: Explore the extensive range of applications wherein nanomaterials from this category find use. Discuss their potential impact on fields such as electronics, energy, medicine, and more.

The structured prompts are designed to facilitate a comprehensive, in-depth exploration of various facets, ranging from fundamental properties to practical applications and potential risks associated with these nanomaterials. Prompting LLMs generates text that responds to and elaborates on the specific aspects mentioned in each prompt.

(**Chatbot's Response**) [Generated Text]

Table 10 shows the sample responses generated by GPT-4 to the natural language prompts on MEMS devices. In the subsequent section, we will describe our approach to integrate technical descriptions into a small-scale LM for fine-tuning on the supervised nanomaterial identification task.

Fine-Tuning Smaller LMs: Our approach employs a smaller pretrained language model (LM) to summarize the technical descriptions generated by a large language model (LLM) on nanomaterials. In our study, we utilize a pre-trained small-scale LM, DeBERTa[1] (He et al. 2020b), an improved version of the BERT architecture. We fine-tune this small-scale LM on the generated technical descriptions for domain-specific customization on the downstream task. This helps the language model learn the statistical relationships between words and phrases in the large corpus of LLM textual outputs, thereby facilitating the generation of context-aware token embeddings. We input the text sequences generated by LLMs (denoted as $\mathcal{S}_{\text{expl}}$) into the LM_{expl} model, which then generates expressive token embeddings as follows:

$$h_{\text{expl}} = \text{LM}_{\text{expl}}(\mathcal{S}_{\text{expl}}) \tag{2}$$

where the deep contextualized embeddings are denoted as $h_{\text{expl}} \in \mathbb{R}^{m \times d}$, where m represents the number of tokens in $\mathcal{S}_{\text{expl}}$ and d is token embedding dimension. We perform weighted average of all token embeddings for a comprehensive representation of the entire text. We use a softmax attention mechanism to compute interpretable attention weights, denoted as α, and subsequently use these weights to sum-pool the token embeddings, encoding the textual descriptions into a fixed-size, high-level text representation as follows:

$$\alpha = \text{softmax}(q); \quad q = \mathbf{u}^T h_{\text{expl}} \tag{3}$$

$$h_{\text{text}} = \sum_{j=0}^{m} \alpha_i h_{\text{expl}}^{(j)} \tag{4}$$

where $\mathbf{u}$ is a learnable vector and the subscript j refers to token. The text-level embedding $h_{\text{text}} \in \mathbb{R}^{(d)}$ encapsulates the relevant and concise information at the core of the domain knowledge in technical descriptions, which is extracted from the general-purpose LLMs for each nanomaterial.

Cross-Modal Alignment Using Multi-Head Self-Attention: We employ the multi-head self-attention mechanism to align image embeddings with their corresponding text embeddings for image-text matching purposes. This approach emphasizes specific aspects or features of the image that is semantically relevant to the textual descriptions, taking into consideration the different facets of the relationship of the cross-domain modalities (both text-level and image embeddings). The mechanism calculates similarity scores between the provided image embedding and all text embeddings. The text embedding with the highest similarity score is considered the best match for the given image embedding. We initially concatenate the text-level embeddings for the different nanomaterial categories to obtain a unified text-level embedding $\mathbf{h}_{\text{text}} = [\mathbf{h}_{\text{text}}^{(1)}, \cdots, \mathbf{h}_{\text{text}}^{(c)}]$, where c refers to the total number of nanomaterials. We compute the value and key projections for the unified text-level embedding, which represents the combined semantic information of all nanomaterials, for each head h as follows:

$$K_{\text{text}}^{h} = \mathbf{h}_{\text{text}} \text{W}_{K_{\text{text}}}^{h}; V_{\text{text}}^{h} = \mathbf{h}_{\text{text}} \text{W}_{V_{\text{text}}}^{h}$$

Similarly, the query projection for image embedding $\mathbf{h}_{cls}$ for each head h is as follows:

$$Q_{cls}^{h} = \mathbf{h}_{cls} \text{W}_{Q_{cls}}^{h}$$

where $\text{W}_{K_{\text{text}}}^{h}$, $\text{W}_{V_{\text{text}}}^{h}$, and $\text{W}_{Q_{cls}}^{h}$ are trainable weight matrices. We now utilize the scaled-dot product attention mechanism(Vaswani et al. 2017) to compute the normalized attention score, which measures the semantic similarity between the image embedding and each text embedding for a specific attention head h as follows:

---

[1] For more information, refer to the DeBERTa model documentation available at https://huggingface.co/docs/transformers/index.

$$\mathrm{A}_w^h = \mathrm{softmax}\left(\frac{Q_{cls}^h (K_{\text{text}}^h)^\top}{\sqrt{d_k}}\right)$$

where d_k denotes the dimensionality of the keys. The attention weights align complementary information from the cross-domain embeddings, focusing on relevant information for cross-modal alignment and capturing richer semantics by allowing the framework to dynamically weigh different parts of the input based on their relevance to the context. We then compute the weighted sum of the value projection as follows:

$$\mathrm{O}_{\text{text}}^h = \mathrm{A}_w^h V_{\text{text}}^h$$

We concatenate the outputs across different heads because it encapsulates perspectives from multiple heads which focus on different aspects or features, thus producing a more comprehensive and robust alignment between the two modalities. We project the outputs to obtain the final representation as follows:

$$\mathrm{O}_{\text{text}} = \left[\mathrm{O}_{\text{text}}^1, \ldots, \mathrm{O}_{\text{text}}^H\right] \mathrm{W}_O$$

where H represents the total number of heads and W_O deontes the trainable weight matrix. We now compute the cosine similarity between the two cross-domain embeddings as follows:

$$\mathrm{Sim} = \frac{\mathrm{O}_{\text{text}} \cdot \mathbf{h}_{cls}}{||\mathrm{O}_{\text{text}}||_2 \times ||\mathbf{h}_{cls}||_2}$$

where sim $\in \mathbb{R}^c$, and we select the best match based on the highest similarity value. The index is determined for the text-level embedding with the highest similarity score as follows:

$$i^* = \mathrm{argmax}_i\,(\mathrm{Sim})$$

Here, i^* is the index of the best-matching text-level embedding. So, the best-matching text-level embedding would be:

$$\mathbf{h}_{\text{text}}^* = \mathbf{h}_{\text{text}}^{(i^*)}$$

This is essentially a matching mechanism that seeks to find the best pairwise alignment among the various text-level embeddings and the image embedding. We utilize backpropagation error in the downstream supervised multi-classification task to fine-tune the ViT and smaller LMs to maximize the pairwise alignment between the complementary image embedding ($\mathbf{h}_{cls}$) and its corresponding text-level embedding ($\mathbf{h}_{\text{text}}^*$). To summarize, $\mathbf{h}_{\text{text}}^*$ incorporates the prior knowledge obtained from LLMs for the appropriate nanomaterial underlying the elecron micrographs as auxiliary information to support multi-modal learning.

Few-Shot LMM Prompting: Few-shot prompting enables LMMs such as GPT-4V to adapt to new tasks without the need for explicit, gradient-based fine-tuning(Brown et al. 2020) using the labeled data. This approach allows LMMs to learn by analogy, utilizing only a few input-output pairs specific to the downstream task. Few-shot prompting leverages the implicit knowledge embedded in pretrained LMM parameters to adapt to new tasks through task-specific demonstrations, thereby avoiding the need to repurpose LMMs with parameter updates. The context-augmented prompt provides task-specific instructions and demonstrations(input-output pairs), enabling LMMs to generate outputs conditioned on the prompt for improved generalization performance on the new, unfamiliar tasks. In the case of nanomaterial identification tasks, few-shot prompting involves creating a context-augmented prompt using a few input-output mappings $(\mathcal{I}_i, \mathcal{Y}_i)$, which are a small number of image-label pairs as demonstrations sampled from the training data relevant to the query(new/unseen) image. These mappings provide relevant context to aid in understanding and classifying unseen images. The task-specific instruction is related to the query image, instructing the multimodal model to predict its associated label. At inference time, given test images $\mathcal{I}_{\text{test}}$, few-shot prompting predicts an output label based on the conditional probability distribution, $\mathbf{P}(\mathcal{Y}_{\text{test}} \mid ((\mathcal{I}_{\text{train}}, \mathcal{Y}_{\text{train}}), \mathcal{I}_{\text{test}}))$. To explore how the quality and quantity of few-shot demonstrations affect the performance in nanomaterial identification tasks, we consider two distinct sampling strategies: “Random” and “Similarity-driven Sampling”. The random approach involves selecting demonstrations (image-label pairs) arbitrarily from the training data without any specific criteria, serving as a baseline for evaluation. In contrast, the similarity-driven sampling strategy employs cosine similarity to identify the top-K images that most closely resemble a given query image within the training data. This strategy operates under the hypothesis that demonstrations which are more representative of the query image’s data distribution can potentially enhance model adaptability and accuracy. By utilizing both the diverse strategies to construct augmented prompts, we aim to provide a comprehensive analysis of how different demonstration sampling methods impact the effectiveness of few-shot learning in nanomaterial identification tasks. Furthermore, the efficacy of the demonstrations is inherently related to the sampling methods used to retrieve the top-K images that align most closely with the query image. To further explore the impact of the volume of demonstrations on performance, we adjust the number of demonstrations K for each query image. In summary, our objective is to provide LMMs with a context-augmented prompt, comprising image-label pairs selected from the training data, along with task-specific instructions that guide the LMMs in predicting the nanomaterial category of the query image. This task evaluates the LMMs’ ability to predict nanomaterial categories based on the prompt without any parameter updates, distinguishing it from traditional supervised learning, where models are fine-tuned on labeled data. For each query image, the LMMs generate a c-dimensional one-hot vector $h_{\text{pred}} \in \mathbb{R}^c$, where c denotes the predefined number of nanomaterial categories. This vector is linearly encoded into a high-dimensional space to produce a prediction embedding $h_{\text{ICL}} \in \mathbb{R}^d$, which encapsulates the LMMs predictions. Here, d represents the embedding dimension and $c \ll d$. An example of an ICL prompt is as follows,

> Below are the input-output pairs (image-label pairs) for the nanomaterial identification task. Predict the nanomaterial category for the query image.

The prediction embedding likely contains valuable information about potential outcomes, allowing the framework to refine its cross-modal representation for better alignment with desired results. Given the complexity of nanomaterials structures and properties, this prediction embedding has the potential to capture some of that complexity, guiding the framework toward correct identification through the integration of prior knowledge and auxiliary information. A general purpose GPT-4V is accessible to ChatGPT Plus subscribers with a usage cap at chat.openai.com. However, it's not currently available for public use through Multimodal Modeling as a Service (MMaaS) — a cloud-based service that accepts both image and textual inputs as prompts. By utilizing black-box GPT-4V through MMaaS as an on-demand service, typically hosted on cloud servers and accessed via an API, users can design task-specific prompts to query pre-trained LMMs for solving multimodal tasks of interest. This is analogous to how users might access LLMs via Language Modeling as a Service (LMaaS(Sun et al. 2022)) to tackle language-specific tasks. GPT-4V is anticipated to become publicly accessible by mid-November 2023. APIs are designed for large-scale and concurrent requests and are ideal for integration into automated systems. Conversely, websites might not efficiently handle numerous interactions in rapid succession, and automating tasks on them could be prohibited. Manually sending inputs for GPT-4V for the entire training dataset would be a daunting and tedious task. Instead, we select a subset of images from the whole training dataset, termed 'Confounding or Ambiguous Micrographs', for few-shot prompting of GPT-4V. The selection process for these images is as follows: The SEM electron micrographs, initially sized at $1024 \times 768 \times 3$ pixels, were downscaled to $224 \times 224 \times 3$ pixels. They were then normalized using z-score normalization and flattened into one-dimensional vectors. Their dimensionality was further reduced using Principal Component Analysis (PCA) before employing the K-Means clustering algorithm. We chose $K = 10$ clusters based on a predefined number of nanomaterial categories based on benchmark dataset. This method identifies images that are challenging to classify by measuring distances from centroids, assessing cluster variance, and calculating the Silhouette Score. Ground truth labels aid in the evaluation of the clustering's effectiveness. We sampled a fixed 10% of the ambiguous images from the entire dataset. For these images, we apply few-shot prompting of GPT-4V to predict labels. The goal is to learn the optimal projection layer weight matrices, which transform the GPT-4V predictions (one-hot vectors) into a high-dimensional space, producing a prediction embedding $h_{\text{ICL}} \in \mathbb{R}^d$ that encapsulates the GPT-4V predictions. The projection layer is subsequently trained through the supervised learning task. This training aims to minimize the cross-entropy loss and obtain optimal weights.

Unified Attention Layer: We compute the cross-modal embedding, denoted as h_{cross}, using a hierarchical multi-head attention mechanism that integrates the original image embedding $\mathbf{h}_{\text{cls}}$, the text-level embedding $\mathbf{h}^*_{\text{text}}$, and the prediction embedding $\mathbf{h}_{\text{ICL}}$. This mechanism offers a robust framework for integrating diverse information from different modalities. As a result, it can produce a more holistic representation that encompasses a wide range of information, potentially improving the performance of nanomaterial identification tasks. In general, the multi-head attention mechanism uses multiple heads to capture different attention patterns, allowing the model to recognize a variety of relationships in the data from multiple subspace representations. Given queries Q, keys K, and values V, the scaled dot-product attention is defined as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

where d_k is the dimensionality of the keys. The *multi-head* attention mechanism employs multiple heads to integrate various attention patterns into a unified representation. Each of these heads utilizes the scaled dot-product attention on distinct linear transformations of the input queries, keys, and values. These transformations are parameterized by matrices $\mathbf{W}_{Q_i}$, $\mathbf{W}_{K_i}$, and $\mathbf{W}_{V_i}$. The final output is derived by concatenating the results from these heads and subjecting it to a subsequent matrix transformation $\mathbf{W}_O$.

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}_O$$
$$\text{where} \quad \text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_{Q_i}, \mathbf{K}\mathbf{W}_{K_i}, \mathbf{V}\mathbf{W}_{V_i})$$

Given the context described earlier, the unified attention layer employs the multi-head attention mechanism in a hierarchical fashion to derive the cross-modal embedding. The procedure consists of two main stages: (1) **Image-Text Attention:** Here, the unified attention layer focuses on the image embedding $\mathbf{h}_{\text{cls}}$ in relation to the text-level embedding $\mathbf{h}^*_{\text{text}}$. The result is an intermediate embedding, $\mathbf{h}_{\text{img-text}}$, which amalgamates details from both image and text modalities through the multi-head attention mechanism. The primary intent of this step is to incorporate relevant textual information guided by the image's context. This can be mathematically described as follows:

$$\mathbf{h}_{\text{img-text}} = \text{MultiHead}(\mathbf{h}_{\text{cls}}, \mathbf{h}^*_{\text{text}}, \mathbf{h}^*_{\text{text}})$$

(2) **Image-Text-Prediction Attention:** During this stage, the previously derived intermediate embedding $\mathbf{h}_{\text{img-text}}$ undergoes further refinement. The unified attention layer aligns this embedding with the prediction embedding $\mathbf{h}_{\text{ICL}}$, computing the final cross-modal representation h_{cross}. This stage aims to combine insights from the intermediate representation with the prediction embedding, creating a comprehensive representation that seamlessly integrates various modalities. This can be represented mathematically as:

$$h_{\text{cross}} = \text{MultiHead}(\mathbf{h}_{\text{img-text}}, \mathbf{h}_{\text{ICL}}, \mathbf{h}_{\text{ICL}})$$

In summary, the unified attention layer uses multi-head attention mechanisms to capture and integrate information from multiple different modalities (image, text, prediction) in a hierarchical manner, resulting in the comprehensive cross-modal embedding that can be used for nanomaterial identification tasks. This mechanism employs multiple sets of learned weight matrices to emphasize various aspects or relationships within the data. Consequently, this approach has the potential to foster robust and enriched embeddings capable of capturing complex patterns. Additionally, it aids in focusing on contextually relevant information and in achieving semantic alignment across different embeddings, thereby enhancing the capacity to identify and utilize important features in the input data. Finally, we linearly transform the final unified cross-modal embedding to obtain a probability distribution p_i over the possible outcomes, as follows:

$$\mathrm{p}_i = \text{softmax}\big(\mathrm{W}h_{\text{cross}}\big)$$

p_i represents the probability distribution across nanomaterial categories. We apply the argmax operation to p_i to determine the most likely nanomaterial category predicted by the framework. In our approach, we concurrently carry out three tasks: (i) we compute image embeddings using an image encoder, (ii) we utilize zero-shot CoT prompting with LLMs to generate technical descriptions of nanomaterials and fine-tune smaller pre-trained LMs on the generated descriptions—subsequently, a softmax attention pooling mechanism is employed to produce text-level embeddings, and (iii) we employ few-shot prompting of LMMs to derive prediction embeddings. Following this, we jointly optimize these different embeddings using a hierarchical multi-head attention mechanism for supervised learning tasks. The overarching goal is to reduce the cross-entropy loss and improve multi-class classification accuracy. Furthermore, the MHA adeptly captures and aligns diverse data sources, making it indispensable for multi-modal integration and analysis. This capability is especially crucial in fields like nanomaterial analysis, where multiple modalities offer complementary insights.

Experiments And Results

Datasets: Our study utilized the SEM dataset(Aversa et al. 2018) to automate nanomaterial identification. The expert-annotated dataset spans 10 distinct categories, representing a range of nanomaterials, such as particles, nanowires, and patterned surfaces, among others. It contains approximately 21,283 electron micrographs. Figure 4 provides a representation of the different nanomaterial categories in the SEM dataset. Despite initial findings(Modarres et al. 2017) on a subset, our research was based on the complete dataset. The original dataset curators(Aversa et al. 2018), did not provide predefined splits for training, validation, and testing, so we employed the k-fold cross-validation method. This strategy facilitated a fair and rigorous comparison with popular baseline models. To further validate our proposed framework, we evaluated it on several open-source material benchmark datasets relevant to our study, encompassing diverse applications. This allowed us to demonstrate the efficacy of our framework and its applicability to a broader range than just the SEM dataset.

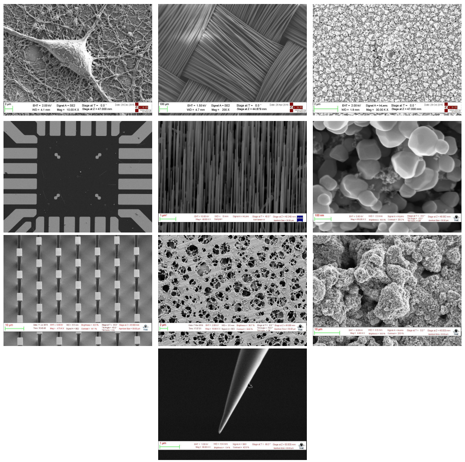


Figure 4: The figure showcases nanomaterials from the SEM dataset(Aversa et al. 2018). In the first, second and third rows (from left to right), we have: *biological, fibers, films*, *MEMS, nanowires, particles*, and *patterned surface, porous sponges, powder*, respectively. Meanwhile, the last row displays: *tips*.

Results: To measure the effectiveness of our proposed framework, we conducted an in-depth analysis contrasting it with popular computer vision baseline models. We compared our framework to supervised learning models, notably Convolutional Neural Networks (ConvNets) and Vision Transformers (al. 2022b,a), and self-supervised approaches like Vision Contrastive Learning (et al. 2020). The results of this analysis are shown in Table 1. To ensure an fair and rigorous comparison, all tests were conducted under uniform settings across different algorithms. We assessed performance using the Top-N accuracy metric, specifically for N values of 1, 2, 3, and 5. Notably, our framework outperformed the best-performing baseline model, T2TViT ((Yuan et al. 2021)), demonstrating a significant 30.50% improvement in Top-1 accuracy and a modest 3.61% gain in Top-5 accuracy. Table 2 presents a comparison between our framework and a selection of supervised learning-based baseline models. This includes several GNN architectures (Rozemberczki et al. 2021; Fey and Lenssen 2019) as well as Graph Contrastive Learning (GCL) algorithms (Zhu et al. 2021). Impressively, our framework establishes a new state-of-the-art benchmark, outperforming all other baselines on the benchmark dataset(Aversa et al. 2018). Figures 5 and 6 shows the radar charts corresponding to the results shown in Tables 1 and 2. The underlying hypothesis of our framework is that ViTs can be employed for initial explorations and the generation of baseline results in this context. Zero-shot CoT prompting of LLMs can be leveraged to enhance the initial outcomes of ViTs by utilizing the implicit domain-specific knowledge embedded within the language model's trainable parameters to obtain expressive cross-modal embeddings. On the other hand, few-shot (in-context) learning of LMMs can be utilized to further refine the framework's predictions by providing demonstrations from the training data, potentially leading to a more robust and accurate predictive framework for nanomaterial category prediction. The experimental findings validate this hypothesis and further advancements in the semiconductor industry—a domain where traditional deep learning techniques often underperform due to their lack of a holistic and nuanced approach. Such shortcomings could hinder breakthroughs in the semiconductor industry.

Related Work: The landscape of computer vision has been profoundly influenced by convolutional networks (ConvNets or CNNs). The pioneering LeNet(LeCun et al. 1998) set the stage for ConvNets, which were subsequently employed in a variety of vision tasks ranging from image classification(Krizhevsky, Sutskever, and Hinton 2017) to semantic segmentation(Long, Shelhamer, and Darrell 2015). Over recent years, groundbreaking architectures like ResNet(He et al. 2016), MobileNet(Howard et al. 2017), and NAS(Zoph and Le 2016; Yang et al. 2020) have further refined the capabilities of CNNs. However, the introduction of vision transformers (ViTs)(Dosovitskiy et al. 2020; Han et al. 2022; Carion et al. 2020; Chen et al. 2021a) marked a paradigm shift, leading to the development of numerous enhanced ViT variants. These advances encompass pyramid architectures(Liu et al. 2021; Wang et al. 2021), local attention mechanisms(Han et al. 2021; Liu et al. 2021), and innovative position encoding methods(Wu et al. 2021b). Drawing inspiration from ViTs, the computer vision community has also delved into the potential of Multilayer Perceptrons (MLP) for vision tasks(Touvron et al. 2022; Tolstikhin et al. 2021). Current vision-based frameworks in the semiconductor manufacturing sector fall short in various aspects, especially when com-

Table 1: The table compares our proposed method to baseline algorithms, such as vision-based supervised convolutional neural networks (ConvNets), vision transformers (ViTs), and self-supervised learning (SSL) algorithms.

	Algorithms	**Top-1**	**Top-2**	**Top-3**	**Top-5**
ConvNets	AlexNet(Krizhevsky, Sutskever, and Hinton 2017)	0.493	0.582	0.673	0.793
	DenseNet(Huang et al. 2017)	0.539	0.750	0.875	0.906
	ResNet(He et al. 2016)	0.512	0.766	0.891	0.906
	VGG(Simonyan and Zisserman 2014)	0.517	0.644	0.717	0.779
	GoogleNet(Szegedy et al. 2015)	0.560	0.844	0.906	0.938
	SqueezeNet(Iandola et al. 2016)	0.436	0.469	0.609	0.656
VSL	Barlowtwins(Zbontar et al. 2021)	0.138	0.250	0.328	0.453
	SimCLR(Chen et al. 2020b)	0.157	0.234	0.359	0.469
	byol(Grill et al. 2020)	0.130	0.234	0.281	0.422
	moco(He et al. 2020a)	0.158	0.188	0.250	0.438
	nnclr(Dwibedi et al. 2021)	0.144	0.266	0.313	0.531
	simsiam(Chen and He 2021)	0.170	0.266	0.391	0.500
ViTs	CCT(Hassani et al. 2021)	0.600	0.781	0.875	0.969
	CVT(Wu et al. 2021a)	0.537	0.750	0.828	0.953
	ConViT(d'Ascoli et al. 2021)	0.582	0.734	0.828	0.938
	ConvVT(Wu et al. 2021a)	0.291	0.563	0.734	0.875
	CrossViT(Chen, Fan, and Panda 2021)	0.466	0.719	0.828	0.938
	PVTC(Wang et al. 2022)	0.567	0.766	0.813	0.922
	SwinT(Liu et al. 2021)	0.675	0.766	0.891	0.938
	VanillaViT(Dosovitskiy et al. 2020)	0.623	0.828	0.859	0.938
	Visformer(Chen et al. 2021b)	0.371	0.578	0.641	0.797
	ATS(Fayyaz et al. 2021)	0.511	0.703	0.828	0.938
	CaiT(Touvron et al. 2021b)	0.616	0.750	0.906	0.938
	DeepViT(Zhou et al. 2021)	0.512	0.734	0.875	0.938
	Dino(Caron et al. 2021)	0.047	0.219	0.391	0.432
	Distallation(Touvron et al. 2021a)	0.516	0.719	0.844	0.938
	LeViT(Graham et al. 2021)	0.597	0.813	0.875	0.953
	MA(He et al. 2021)	0.192	0.288	0.350	0.459
	NesT(Zhang et al. 2022)	0.636	0.828	0.891	0.953
	PatchMerger(Renggli et al. 2022)	0.549	0.719	0.859	0.922
	PiT(Heo et al. 2021)	0.520	0.703	0.828	0.953
	RegionViT(Chen, Panda, and Fan 2021)	0.575	0.797	0.859	0.922
	SMIM(Xie et al. 2021)	0.163	0.297	0.453	0.609
	T2TViT(Yuan et al. 2021)	0.702	0.859	0.906	0.938
	ViT-SD(Lee, Lee, and Song 2021)	0.613	0.766	0.906	0.953
	`CM-EMRL`	**0.9161**	**0.9339**	**0.9691**	**0.9719**

pared to the recently proposed advancements in generative deep learning and multimodal learning. Many existing solutions fail to capitalize on the detailed analysis achievable through the synergy of LLMs, LMMs, and small-scale LMs with electron micrographs. Moreover, the existing frameworks typically analyze electron micrographs (nano images) at a singular modality, through the use of architectures such as ConvNets, ViTs, or MLP-Mixer, missing the opportunities that a multi-modality fusion approach could offer in enhancing classification accuracy. Furthermore, the industry has not fully embraced the utilization of zero-shot CoT LLMs prompting for generating technical descriptions of nanomaterials, which can significantly enhance domain-specific insights essential for nanomaterial identification tasks. Furthermore, the semiconductor manufacturing sector has not fully tapped into the emerging in-context learning capabilities of LMMs with few-shot prompting for predictive nanomaterial analysis, even though these capabilities could significantly enhance the accuracy of nanomaterial predictions. This glaring gap in the integration of image-based and linguistic insights renders current architectures less comprehensive and nuanced, potentially impeding breakthroughs in the semiconductor industry. Instead of relying solely on conventional classification methods, the new framework incorporates both image-based and linguistic insights by leveraging the capabilities of ViTs and LLMs, respectively, as well as the predictive abilities of LMMs. This framework aims to facilitate a more comprehensive and nuanced analysis of electron micrographs, holding significant promise for advancements in the semiconductor industry through automated nanomaterial identification. These advancements highlight the ongoing push for innovation in semiconductor manufacturing, driven by the escalating demand for more powerful and efficient electronic devices.

Conclusion

To conclude, we conducted the first in-depth study aimed at achieving state-of-the-art performance in nanomaterial characterization. We have introduced an innovative framework that employs ViTs as the foundational layer, further enriched through the multi-modal fusion approach of zero-shot CoT prompting of LLMs and refined with few-shot (in-context) learning of LMMs. Our experiments confirm the superiority of this framework, indicating its transformative potential for semiconductor manufacturing in the age of advanced electronic devices.

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.

Bianchi, F. M.; Grattarola, D.; Livi, L.; and Alippi, C. 2021. Graph neural networks with convolutional arma filters. *IEEE transactions on pattern analysis and machine intelligence*.

Bielak, P.; Kajdanowicz, T.; and Chawla, N. V. 2021. Graph Barlow Twins: A self-supervised representation learning framework for graphs. *arXiv preprint arXiv:2106.02466*.

Bresson, X.; and Laurent, T. 2017. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, C.-F.; Panda, R.; and Fan, Q. 2021. Regionvit: Regional-to-local attention for vision transformers. *arXiv preprint arXiv:2106.02689*.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 357–366.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12299–12310.

Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; and Li, Y. 2020a. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, 1725–1735. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; and Tian, Q. 2021b. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 589–598.

d'Ascoli, S.; Touvron, H.; Leavitt, M.; Morcos, A.; Biroli, G.; and Sagun, L. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *arXiv preprint arXiv:2103.10697*.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Deshpande, A. M.; Minai, A. A.; and Kumar, M. 2020. One-shot recognition of manufacturing defects in steel surfaces. *Procedia Manufacturing*, 48: 1064–1071.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, J.; Zhang, S.; Wu, G.; Moura, J. M.; and Kar, S. 2017. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9588–9597.

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

Fayyaz, M.; Kouhpayegani, S. A.; Jafari, F. R.; Sommerlade, E.; Joze, H. R. V.; Pirsiavash, H.; and Gall, J. 2021. Ats: Adaptive token sampling for efficient vision transformers. *arXiv preprint arXiv:2111.15667*.

Fey, M. 2019. Just jump: Dynamic neighborhood aggregation in graph neural networks. *arXiv preprint arXiv:1904.04849*.

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

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12259–12269.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.;

Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; and Wang, Y. 2021. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34: 15908–15919.

Hassani, A.; Walton, S.; Shah, N.; Abuduweili, A.; Li, J.; and Shi, H. 2021. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020a. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020b. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11936–11945.

Holt, D. B.; and Joy, D. C. 2013. *SEM microcharacterization of semiconductors*. Academic Press.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

Kim, D.; and Oh, A. 2022. How to find your friendly neighborhood: Graph attention design with self-supervision. *arXiv preprint arXiv:2204.04879*.

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

Klicpera, J.; Bojchevski, A.; and Günnemann, S. 2018. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, S. H.; Lee, S.; and Song, B. C. 2021. Vision Transformer for Small-Size Datasets. *arXiv preprint arXiv:2112.13492*.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

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.

Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; and Grohe, M. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 4602–4609.

OpenAI. 2023a. GPT-4 Technical Report. arXiv:2303.08774.

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

Renggli, C.; Pinto, A. S.; Houlsby, N.; Mustafa, B.; Puigcerver, J.; and Riquelme, C. 2022. Learning to Merge Tokens in Vision Transformers. *arXiv preprint arXiv:2202.12015*.

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.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*.

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.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Thakoor, S.; Tallec, C.; Azar, M. G.; Munos, R.; Veličković, P.; and Valko, M. 2021. Bootstrapped representation learning on graphs. In *ICLR 2021 Workshop on Geometrical and Topological Representation Learning*.

Thekumparampil, K. K.; Wang, C.; Oh, S.; and Li, L.-J. 2018. Attention-based graph neural network for semi-supervised learning. *arXiv preprint arXiv:1803.03735*.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34: 24261–24272.

Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5314–5321.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media*, 1–10.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021a. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22–31.

Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021b. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10033–10041.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2021. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision). *arXiv preprint arXiv:2309.17421*.

Yang, Z.; Wang, Y.; Chen, X.; Shi, B.; Xu, C.; Xu, C.; Tian, Q.; and Xu, C. 2020. Cars: Continuous evolution for efficient neural architecture search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1829–1838.

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.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhang, Z.; Zhang, H.; Zhao, L.; Chen, T.; Arik, S.; and Pfister, T. 2022. Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding.

Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; and Feng, J. 2021. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.

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

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*.

Zoph, B.; and Le, Q. V. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Technical Appendix

Ablation Study

Figure 3 illustrates an overview of the framework. The proposed framework involves four components: (a) The first component, the electron micrograph encoder, takes an input image and divides it into smaller patches. These patches are transformed into tokens, enriched with positional embeddings for spatial information, and a classification token is added as a separate token to represent the overall image content. The resulting augmented token sequence is fed into a ViT model to generate an embedding that represents the entire image. (b) Next, the zero-shot CoT LLMs prompting technique uses cloud services to access LLMs for generating detailed descriptions about various aspects of nanomaterials. Structured prompts guide the LLMs in generating in-depth descriptions on topics ranging from the fundamental properties of nanomaterials to their practical applications. Subsequently, we fine-tune smaller LMs on the descriptions generated by the LLMs to obtain context-aware token embeddings. We then perform sum-pooling attention mechanism to obtain contextualized text-level embeddings, which capture the core knowledge in the generated texts. (c) Finally, the cross-modal alignment employs the scaled dot-product attention mechanism to match the image embeddings with the corresponding nanomaterial-specific text-level embeddings. This process

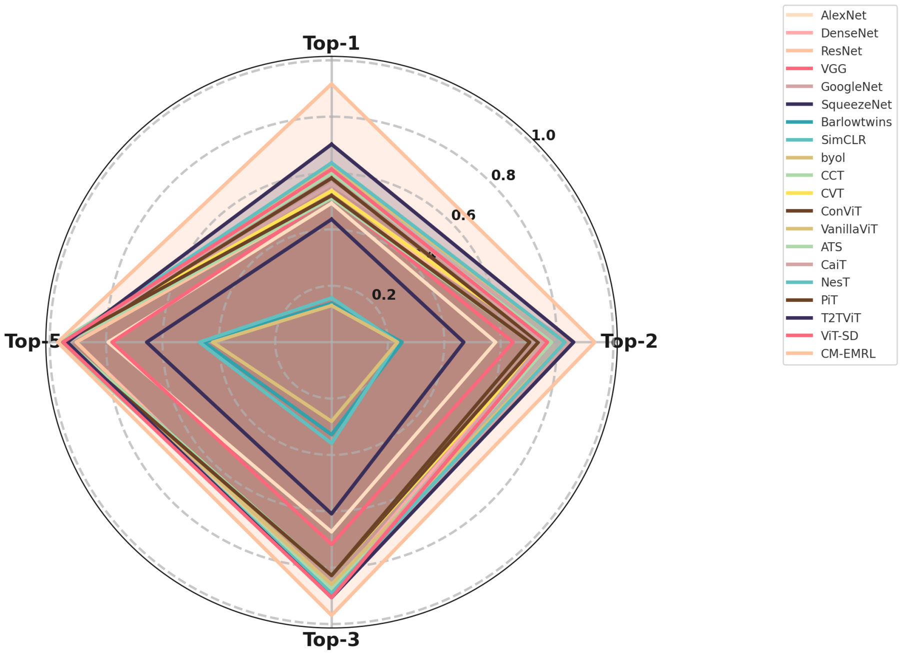


Figure 5: The figure compares our proposed framework to vision-based supervised convolutional neural networks (ConvNets), vision transformers (ViTs), and self-supervised learning (VSL) algorithms on the SEM dataset (Aversa et al. 2018).

Table 2: The table shows the performance comparison of the proposed method with supervised GNNs and self-supervised GCL algorithms on the SEM dataset(Aversa et al. 2018).

	Algorithms	Top-1	Top-2	Top-3	Top-5
GSL	GBT(Bielak, Kajdanowicz, and Chawla 2021)	0.513	0.595	0.686	0.778
	GRACE(Zhu et al. 2020)	0.581	0.646	0.711	0.773
	BGRL(Thakoor et al. 2021)	0.573	0.629	0.671	0.728
	InfoGraph(Sun et al. 2019)	0.560	0.631	0.694	0.756
GNN	APPNP(Klicpera, Bojchevski, and Günnemann 2018)	0.604	0.713	0.792	0.823
	AGNN(Thekumparampil et al. 2018)	0.517	0.733	0.841	0.943
	ARMA(Bianchi et al. 2021)	0.553	0.747	0.848	0.925
	DNA(Fey 2019)	0.593	0.677	0.786	0.891
	GAT(Veličković et al. 2017)	0.507	0.724	0.807	0.914
	GGConv(Li et al. 2015)	0.583	0.778	0.841	0.944
	GraphConv(Morris et al. 2019)	0.533	0.727	0.847	0.961
	GCN2Conv(Chen et al. 2020a)	0.697	0.813	0.867	0.945
	ChebConv(Defferrard, Bresson, and Vandergheynst 2016)	0.547	0.762	0.834	0.896
	GraphConv(Morris et al. 2019)	0.533	0.727	0.847	0.961
	GraphUNet(Gao and Ji 2019)	0.622	0.738	0.866	0.912
	MPNN(Gilmer et al. 2017)	0.643	0.792	0.873	0.959
	RGGConv(Bresson and Laurent 2017)	0.633	0.727	0.886	0.928
	SuperGAT(Kim and Oh 2022)	0.561	0.676	0.863	0.935
	TAGConv(Du et al. 2017)	0.614	0.739	0.803	0.946
	`CM-EMRL`	**0.9161**	**0.9339**	**0.9691**	**0.9719**

highlights image features relevant to the textual descriptions, aiding in the identification of text-level embeddings that correspond to the image. (c) Few-shot prompting enables LMMs to quickly adapt to new tasks without traditional fine-tuning on labeled data. Using a small set of input-output pairs, these models learn tasks by drawing from their vast pre-existing knowledge acquired during training on vast, diverse text corpora. In the context of nanomaterial identification, LMMs utilize a handful of image-label pairs from the training data to classify new, unseen images and obtain prediction embeddings. Demonstrations can be selected either randomly or based on their similarity to the (c) Few-shot prompting enables LMMs to quickly adapt to new tasks without traditional fine-tuning on labeled data. Using a small set of input-output pairs, these models learn tasks by drawing from their vast pre-existing knowledge acquired during training on vast, diverse text corpora. In the context of nanomaterial identification, LMMs utilize a handful of image-label pairs from the train-

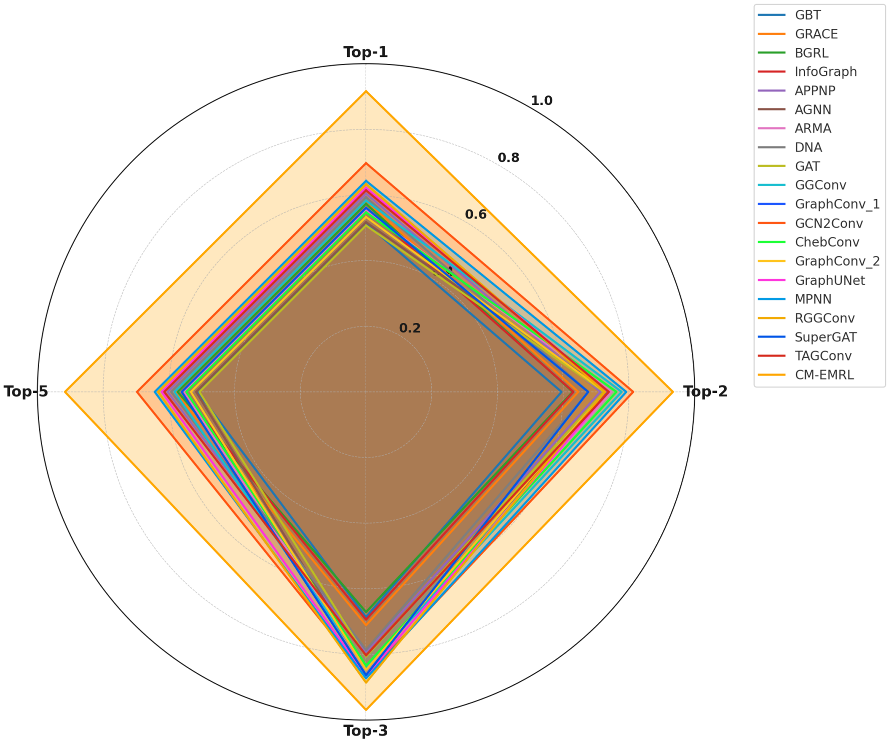


Figure 6: The figure shows the performance comparison of the proposed framework to supervised GNNs and self-supervised GCL algorithms on the SEM dataset (Aversa et al. 2018).

ing data to classify unseen query images and obtain prediction embeddings. Demonstrations can be selected either randomly or based on their similarity to the query image. (d) Finally, the unified attention layer, through the hierarchical multi-head attention mechanism, combines information from the original image embedding, a text-level embedding, and the prediction embedding, optimizing for accuracy in nanomaterial categorization. To evaluate the efficacy of the individual components and validate the design choices for their inclusion in the framework, we conducted an ablation study. In this study, we selectively disabled specific components to create various ablated variants, which were then evaluated using the SEM dataset(Aversa et al. 2018) on nanomaterial identification. Compared to our proposed original framework, which serves as the baseline, the ablated variants exhibited a notable decline in performance, underscoring the importance of the components that were disabled. The ablation study results support the hypothesis that each component is crucial for the framework's peak performance in nanomaterial identification. Ablated variants excluding the zero-shot CoT LLMs prompting with cross-modal alignment, few-shot prompting with LMMs, and unified attention methods are labeled as proposed framework "w/o LLMs", "w/o LMMs", and "w/o MHA"; "w/o" is shorthand for "without". In the case of "w/o MHA", we concatenate the cross-domain embeddings and then transform them through a linear layer to predict the label. The findings from the ablation study are presented in Table 3. The 'w/o LLMs' variant exhibits a significant decrease in performance compared to the baseline, with a noteworthy drop of 19.18% in **Avg-Precision**, underscoring the crucial role of zero-shot CoT LLMs in prompting. This technique extracts detailed nanomaterial technical descriptions from closed-source large language models (LLMs), then fine-tunes smaller language models on these descriptions to produce context-aware token embeddings. It subsequently employs weighted sum-pooling mechanism to derive text-level embeddings. Cross-modal alignment correlates these text embeddings with image embeddings via scaled dot-product attention, thereby synchronizing and matching information across different types of data modalities in a unified representation space.

Algorithms	**Avg-Precision**	**Avg-Recall**	**Avg-F1 Score**
`CM-EMRL`	**0.9157**	**0.9125**	**0.9120**
w/o LLMs	0.7401	0.7936	0.8141
w/o LMMs	0.8160	0.7804	0.8053
w/o MHA	0.6937	0.7180	0.7159

Table 3: In our ablation study, we methodically deactivate individual components to evaluate their unique contributions. The aim is to gauge how specific components influence the framework's overall performance. The experiments underscore the importance of each disabled component, as evidenced by the marked drop in performance metrics when compared to the baseline. These results validate our hypothesis that the joint optimization of the zero-shot CoT LLMs prompting with cross-modal alignment, few-shot prompting with LMMs, and unified attention methods enhances the overall framework's efficacy.

Similarly, the ‘w/o LMMs’ variant performs notably worse than the baseline, exhibiting a drop of 10.88% in **Avg-Precision**. This underscores the significance of few-shot prompting to predict the label of query image, which leverages LMMs to rapidly adapt to new tasks with limited demonstrations (image-label pairs) and generate prediction embeddings. In addition, the ‘w/o MHA’ variant showed a significant performance deterioration compared to the baseline, marked by a 24.24% drop in **Avg-Precision**. This deterioration can largely be attributed to the overly simplified linear operator in the output layer. This further emphasizes the importance of the unified attention layer, which amalgamates insights from image, text, and prediction embeddings using a hierarchical multi-head attention mechanism to optimize nanomaterial categorization accuracy. Similarly, in comparison to the baseline, the ‘w/o LLMs’ variant exhibits a 13.03% decrease in **Avg-Recall**, the ‘w/o LMMs’ variant shows a 14.48% reduction, and the ‘w/o MHA’ variant experiences a significant drop of 21.31% in the same metric. Our ablation study highlighted the significant contributions of each component within our framework. When individual components were omitted, we observed a consistent drop in performance. These findings validate the integral role that each component plays in achieving optimal performance of the holistic framework. In summary, each component serves a specific purpose, and their integration ensures a comprehensive approach to nanomaterial categorization. Their inclusion is justified by the need for detailed analysis, adaptability, and holistic consideration of both visual and textual data in categorizing nanomaterials.

Empirical Insights into Nanomaterial Classification

We conducted comprehensive experiments to assess our proposed framework’s effectiveness in classifying electron micrographs of various nanomaterials, from simple to intricate patterns. Our primary goal is to highlight the framework’s classification capabilities. Nanomaterials exhibit a diverse range of patterns due to differences in composition, morphology, surface properties, crystallinity, and synthesis methods. These patterns, as captured in electron micrographs, reflect the materials’ unique properties, structures, and potential applications. They can depict anything from isolated nanoparticles to complex aggregations, and from crystalline to amorphous structures. Accurate interpretation of these patterns is vital for understanding and leveraging each nanomaterial category’s distinctive qualities. Electron micrographs provide detailed nanoscale insights, revealing structural and morphological features essential for various applications in materials science and related fields. As demonstrated by the experimental results displayed in Figure 7, our framework can effectively generalize across diverse nanomaterials, even those with complex patterns. The figure presents bar plots of the framework’s performance on multiple metrics, color-coded by category. We evaluated its performance on the SEM dataset(Aversa et al. 2018) using standard metrics such as precision (P in $\%$), recall (R in $\%$), and F1-score (F1 in $\%$). These findings underscore the framework’s effectiveness and emphasize its relevance in materials science and nanotechnology. We employ a comprehensive multi-metric evaluation for robust comparison with baseline models, anchored by a confusion matrix that details our framework’s performance in classifying electron micrographs across nanomaterial categories. The matrix includes key metrics for multi-class classification: True Positives (TP), which are micrographs correctly identified as belonging to a category; False Negatives (FN), which are micrographs that belong to a category but are incorrectly classified as not belonging; True Negatives (TN), which are micrographs correctly identified as not belonging to a category; and False Positives (FP), which are micrographs incorrectly identified as belonging to a category. From these metrics, we calculate precision (the ratio of correctly classified micrographs to the total classified as belonging to a category, TP / (TP + FP)), recall (the proportion of actual micrographs from a category that were correctly classified, TP / (TP + FN)), and the F1-score, which is the harmonic mean of precision and recall, providing an overall measure of effectiveness in micrograph categorization. Thus, our multi-metric approach, anchored by the confusion matrix and its derived metrics, ensures a rigorous evaluation of our framework. This approach facilitates a detailed understanding of our model’s effectiveness in categorizing electron micrographs across various nanomaterial categories. It is crucial to highlight that the SEM dataset exhibits significant class imbalance. Notably, our framework achieves higher classification scores for nanomaterial categories with a substantial number of labeled instances, outperforming its performance on categories with fewer instances. The notable success with fewer labeled instances can be attributed to our proposed framework’s reduced reliance on nanomaterial-specific relational inductive biases, which distinguishes it from conventional methods. In conclusion, our extended experiments have bolstered our confidence in the framework’s ability to generalize and accurately categorize various nanomaterials using electron micrographs. We believe that these advancements will greatly benefit the wider community by accelerating materials characterization and related research.

Experimental Setup

The SEM dataset (Aversa et al. 2018) comprises electron micrographs with a resolution of $1024 \times 768 \times 3$ pixels. For our analysis, we resized these to $224 \times 224 \times 3$ pixels. During preprocessing, we normalized the images, adjusting their mean and covariance to a consistent value of 0.5 across all channels, ensuring values between [-1, 1]. These downscaled and normalized images were segmented into distinct, non-overlapping patches, treated as a sequence of tokens with a 32-pixel resolution. Both the patch dimension (d_{pos}) and the position embedding dimension (d) were set to 64. We employed a 10-fold cross-validation for evaluation, training for 50 epochs with an initial learning rate of 10^{-3} and a batch size of 48. The unifying (cross-modal) attention layer had the number of attention heads (H) set to 4 and the key/query/value dimensionality (d_h) to 16. To optimize the `CM-EMRL` framework’s performance, we used early stopping on the validation set and a learning rate scheduler, decreasing the rate by 50% if the validation loss did not decrease for five epochs. The Adam optimization algorithm (Kingma and Ba 2014) was used to fine-tune the framework’s parame-

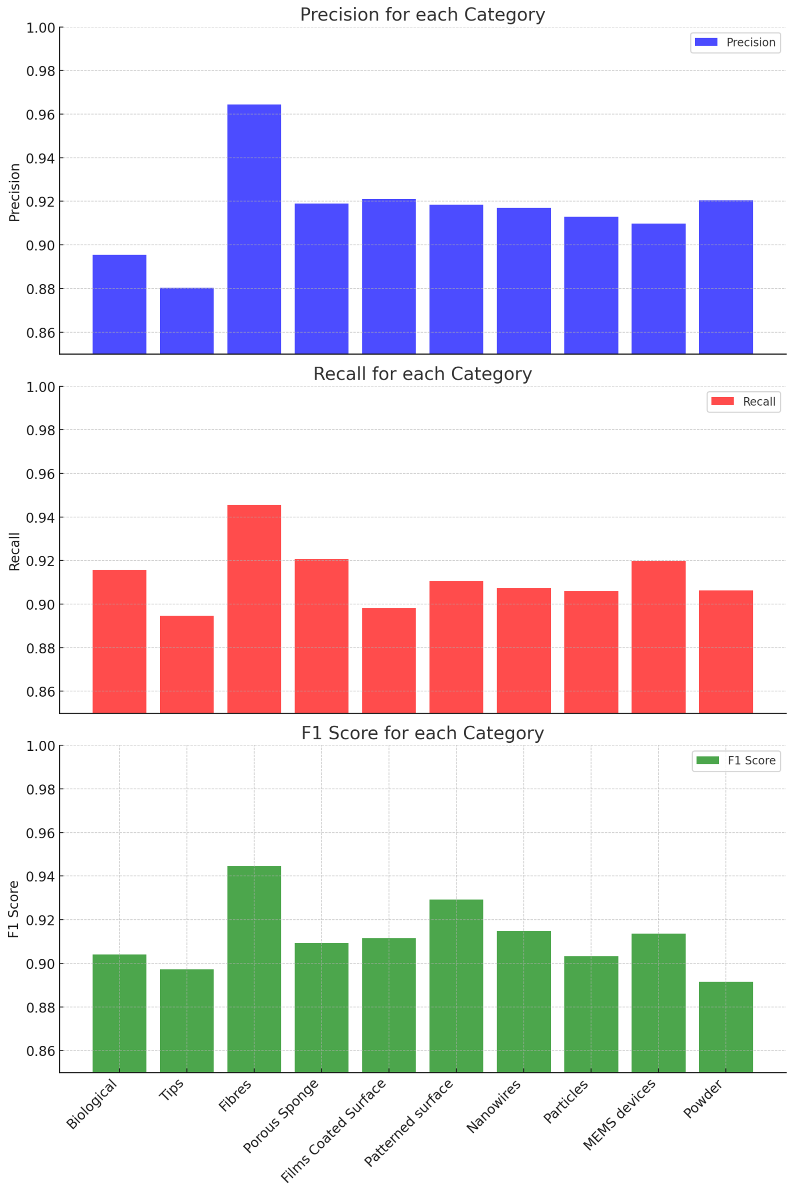


Figure 7: The bar plots illustrate the precision, recall, and F1-score metrics for the framework's performance across various nanomaterial categories, as depicted by electron micrographs within the SEM dataset. Each category is represented with a distinct light shade for clarity.

ters. Our proposed framework seeks to enhance multi-class classification precision by combining the strengths of large multimodal models (LMMs) such as GPT-4V, large language models (LLMs) like GPT-4, and smaller language models (LMs). The methodology utilizes LLMs to generate detailed technical descriptions for nanomaterials, capturing crucial linguistic insights critical for nanomaterial identification. We access LLMs, such as GPT-4, through a Language Model as a Service (LaMaaS) platform using text-based APIs, with GPT-4's maximum token sequence output being 4096. GPT-4 with Vision (GPT-4V) extends GPT-4 by adding visual processing capabilities, allowing it to analyze image inputs alongside text. While GPT-4 is limited to text processing, GPT-4V can handle both text and image inputs, enabling applications like visual question answering. We employ few-shot prompting with GPT-4V, using a small number of demonstrations (image-label pairs included in the prompts) to guide its understanding and to predict the label of a query image. To optimize resource utilization, we trained our deep learning models, which are built upon the PyTorch framework, on two V100 GPUs, each with 8 GB of GPU memory. Given the significant computational demands of prompting Large Language Models (LLMs) and Large Multimodal Models (LMMs), we repeated each experiment twice and reported the average results.

Baseline Algorithms

Our baseline methods are organized into four main categories. First, we employ Graph Neural Networks (GNNs) for the supervised multi-class classification of vision graphs (Rozemberczki et al. 2021; Fey and Lenssen 2019). Second, we utilize Graph Contrastive Learning (GCL) methods (Zhu et al. 2021), which involve creating multiple correlated graph views through stochastic augmentations of the input graph data to maximize mutual information between these views. GCL methods aim to enhance the similarities between positive graph views and reduce dissimilarities between negative graph views sampled from different images. Typically, these methods use the Graph Attention Network (GAT) (Veličković et al. 2017) to learn node-level embeddings, with graph-level embeddings obtained via sum-pooling of node embeddings. For classification, we apply supervised learning using the Random Forest (RF) algorithm, which utilizes these embeddings to predict nanomaterial categories. We then evaluate the efficiency of these unsupervised embeddings by assessing the RF model's accuracy with holdout data. Third, we employ supervised learning with Convolutional Neural Networks (ConvNets) for the classification of electron micrographs (al. 2022b,a). Lastly, Vision Transformers (ViTs) are used for supervised classification, utilizing sequences of image patches from electron micrographs as input (al. 2022b,a), while Vision Contrastive Learning (VCL) techniques (et al. 2020) are self-supervised algorithms designed for contrastive learning in computer vision, utilizing the ResNet architecture for feature extraction. As for our data representation, we construct vision graphs from electron micrographs using the Top-K nearest neighbor search, where patches are treated as nodes, and edges represent pairwise associations between semantically similar neighboring nodes. We employ a 32-pixel patch size and choose K=5 for simplicity to avoid multi-scale vision graphs with varying patch resolutions.

Hyperparameter Studies

We carried out an extensive hyperparameter tuning for our framework, specifically examining the embedding dimension (d) and batch size (b). We evaluated d values from the set $\{32, 64, 128, 256\}$ and b values from the set $\{32, 48, 64, 96\}$. Using the random-search approach, we measured performance based on Top-1 classification accuracy on the validation set. This thorough analysis determined the optimal settings for our framework to be $d = 64$ and $b = 48$, with a corresponding Top-1 accuracy of 0.9161.

(d, b)	(32, 48)	(64, 48)	(128, 48)	(256, 48)
Accuracy	0.8997	0.9161	0.8996	0.9061

Table 4: Experimental findings of the hyperparameter study: Set 1.

(d, b)	(64,32)	(64, 48)	(64, 64)	(64,96)
Accuracy	0.9017	0.9161	0.9036	0.9023

Table 5: Experimental findings of the hyperparameter study: Set 2.

Benchmarking with open-source material datasets

- The NEU-SDD dataset[2] ((Deshpande, Minai, and Kumar 2020)) comprises 1,800 grayscale images captured through electron microscopy, showcasing various surface imperfections in hot-rolled steel strips. This comprehensive dataset is categorized into six types of surface defects—*pitted surfaces, scratches, rolled-in scale, crazing, patches, and inclusions*, with each category represented by 300 images standardized to a resolution of 200×200 pixels. Illustrative examples from these defect categories are presented in Figure 8. Furthermore, we conducted a detailed comparative study using a range of established algorithms to evaluate the proposed method's performance, with a particular focus on multi-class classification tasks for identifying these surface defects.
- The CMI dataset[3] encompasses 600 high-definition electron microscope images that display varying stages of metal panel deterioration. Corrosion experts have meticulously labeled these images following the ASTM-D1654 standard criteria, which employ a scale ranging from 5 to 9 to denote the severity or extent of corrosion. The dataset provides 120 unique images for each level of the corrosion grade defined by this scale, with each image offering a detailed resolution of 512×512 pixels. Figure 9 showcases a set of example images for each corrosion category. We evaluate the effectiveness of our proposed multiclass classification method by benchmarking its ability to accurately assign corrosion grades against a range of established algorithms.
- The KTH-TIPS dataset[4] is a detailed collection of 810 electron micrograph images, each measuring 200×200 pixels, and featuring ten different types of materials. The dataset is notably diverse, including textures such as *sponge, orange peel, styrofoam, cotton, cracker, linen, brown bread, sandpaper, crumpled aluminum foil, and corduroy*, captured under various conditions of lighting, orientation, and scale. Figure 10 displays a representative set of these images, providing a visual overview of the dataset's range. To assess the effectiveness of our proposed technique, we conducted a thorough comparative study, benchmarking it against established algorithms for classifying the various textures or materials represented in the KTH-TIPS dataset.

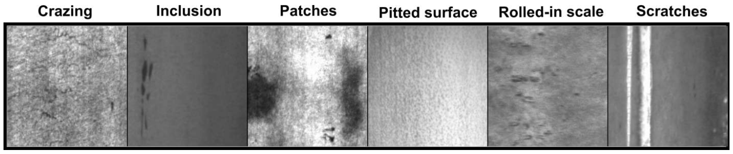


Figure 8: The NEU-SDD dataset comprises six categories of defects in hot-rolled steel strips, as detailed in reference 2((Deshpande, Minai, and Kumar 2020)).

Table 6 displays CoT prompts that focus on various types of surface defects. Each prompt is tailored to initiate a comprehensive exploration of the defects characteristics, causes, and impacts. The structured prompts ensure a detailed and systematic exploration, providing clarity and comprehensive coverage. Consequently, this approach yields precise information and a nuanced understanding of the defects implications. Similarly, Table 7 presents a series of CoT prompts designed

---

[2] Datasource: http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html

[3] https://arl.wpi.edu/corrosion_dataset

[4] https://www.csc.kth.se/cvap/databases/kth-tips/index.html

for an in-depth examination of metal panel corrosion. These prompts guide the research towards a nuanced exploration of different corrosion grades and their effects on metal panels. This structured approach facilitates the study, analysis, and practical application of findings, such as the automatic rating of corrosion in metal panels.

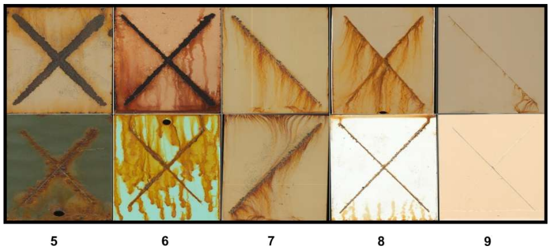


Figure 9: The CMI dataset comprises electron micrographs categorized into five distinct corrosion ratings, as detailed in 3. A rating of 9 indicates the least corrosion, while a rating of 5 indicates the most.

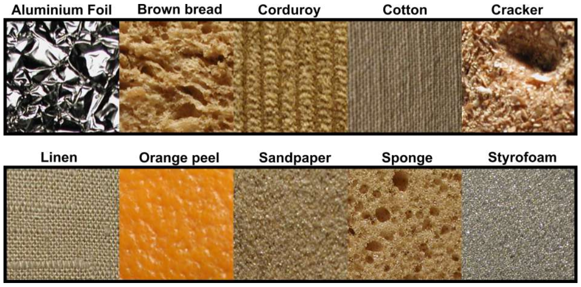


Figure 10: The KTH-TIPS dataset consists of electron micrographs of ten unique materials, which are elaborated upon in reference 4.

The notion that a single, universal CoT prompt can address all tasks with GPT-4 is inaccurate when it comes to generating technical descriptions of nanomaterials as metadata. Specialized prompts tailored to specific needs are crucial for effectively utilizing AI in the analysis of nanomaterials across different diverse datasets. Customized prompting is essential in AI, as demonstrated by the use of GPT-4 to generate prompts(questions) specific to nanomaterial categories in various datasets, as opposed to relying on a single universal CoT prompt for all across various datasets.

Table 6: Customized CoT prompts for detailed analysis of material surface defects.

Prompt 1: Overview: Briefly describe the specific material surface defect and its impact on material performance. **Prompt 2:** Characteristics: Define the defect and its identifying features. **Prompt 3:** Formation: Discuss the formation mechanisms of the defect. **Prompt 4:** Detection: List the primary techniques for defect detection. **Prompt 5:** Effects: Explain the defect's effects on material properties. **Prompt 6:** Mitigation: Outline strategies to mitigate the defect. **Prompt 7:** Engineering: Describe surface engineering techniques to control the defect. **Prompt 8:** Case Studies: Provide examples where defect management improved material use.

Table 8 presents prompts designed to thoroughly cover a wide range of information exploring the characteristics and applications of various materials and providing detailed technical descriptions of these materials. Table 9 offers a detailed juxtaposition of the efficacy of our method against several standard approaches, with evaluations conducted on multiple datasets. The outcomes of these experiments reveal that our approach consistently attains unparalleled results, highlighting its effectiveness and reliability.

Table 7: Customized CoT prompts for detailed analysis of material panel corrosion.

Prompt 1: *Corrosion Grading Overview:* Summarize the numerical corrosion grade system. **Prompt 2:** *Grade Characteristics:* Detail key features of each corrosion grade. **Prompt 3:** *Influencing Factors:* Discuss factors that influence corrosion grading. **Prompt 4:** *Deterioration Mechanisms:* Describe deterioration mechanisms by grade. **Prompt 5:** *Property Impact:* Examine corrosion's impact on metal properties. **Prompt 6:** *Mitigation Strategies:* Outline preventive measures for each grade. **Prompt 7:** *Grade Progression Analysis:* Analyze corrosion grade progression over time. **Prompt 8:** *Rehabilitation Approaches:* Guide on repair or replace decisions by grade. **Prompt 9:** *Corrosion Management Case Studies:* Present case studies on corrosion management. **Prompt 10:** *Economic and Safety Considerations:* Discuss economic and safety implications.

Table 8: Customized prompts for detailed material analysis.

Prompt 1: Contextual Overview: Introduce the material's origin, common use, and relevance. **Prompt 2:** Properties: Discuss the material's physical and chemical characteristics. **Prompt 3:** Production: Outline the processes of material preparation or manufacturing. **Prompt 4:** Structure: Examine the material's structural features and their implications. **Prompt 5:** Modification: Describe possible modifications to enhance the material's properties. **Prompt 6:** Longevity: Analyze the material's durability, degradation, and environmental impact. **Prompt 7:** Applications: Explore diverse applications and uses of the material. **Prompt 8:** Economic Impact: Reflect on the material's economic significance and societal influence. **Prompt 9:** Safety: Address health and safety considerations related to the material. **Prompt 10:** Future Outlook: Speculate on potential future developments and research directions.

	Algorithms	**NEU-SDD**	**CMI**	**KTH-TIPS**
Baselines	ResNet	0.906	0.928	0.941
	GoogleNet	0.936	0.928	0.929
	SqueezeNet	0.955	0.943	0.963
	VanillaViT	0.962	0.968	0.972
	MultiFusion-LLM	**0.988**	**0.979**	**0.983**

Table 9: The table displays a comparative analysis of the performance of our framework in relation to a number of benchmark algorithms across various datasets.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	**LLM-retrieved text**
Introduction: Provide an overview of the MEMS and its significance in various fields?	Introduction: Micro-Electro-Mechanical Systems (MEMS) combine miniaturized mechanical and electromechanical elements that are made using the techniques of microfabrication. When these structures are taken down to the nanoscale, they enter a realm often referred to as Nano-Electro-Mechanical Systems (NEMS). These systems, constructed of nanomaterials, display properties and capabilities that are vastly different from bulk materials. Nanomaterials in MEMS: Nanomaterials are materials with individual units or structures sized between 1 and 100 nanometers in at least one dimension. This size gives them unique mechanical, electrical, and optical properties. Examples include carbon nanotubes, graphene, quantum dots, and metallic nanoparticles. In MEMS applications, nanomaterials offer: (a) Enhanced Mechanical Properties: High strength-to-weight ratio and improved flexibility. (b) Advanced Electrical Conductivity: Enhanced charge transfer and quantum tunneling properties, (c) Sensitivity: Due to their high surface area-to-volume ratio, they are incredibly sensitive to environmental changes, making them ideal for sensing applications, and (d) Thermal Conductivity: Some nanomaterials exhibit exceptional heat conduction properties. Significance across Various Fields: MEMS nanomaterials have made inroads in various fields due to their unique attributes: (a) Medical and Healthcare: Used in implantable devices, drug delivery systems, and advanced diagnostic tools. For example, MEMS sensors can detect minute changes in physiological parameters, (b) Telecommunications: Nanomaterials enhance the performance of micro-sized antennas, resonators, and optical switches, (c) Consumer Electronics: Accelerometers in smartphones, tiny microphones, and miniature cameras all leverage MEMS technology, (d) Aerospace and Defense: Advanced sensors for navigation, pressure sensing in harsh environments, and micro propulsion systems, (e) Energy: Harvesting ambient energy (e.g., vibrations) to power devices. Nanomaterials also find applications in efficient solar cells and batteries, (f) Environmental Monitoring: Nanoscale sensors can detect minute concentrations of pollutants or pathogens in air and water, and (g) Automotive: MEMS sensors help in real-time tire pressure monitoring, optimizing air/fuel mixtures, and safety systems like airbags. Conclusion: The category of MEMS nanomaterials has paved the way for groundbreaking advancements in various fields. These tiny yet robust systems are redefining the borders of what's feasible, leading to the miniaturization of devices, enhancement of performance, and new applications previously deemed unattainable. Their influence spans from the phones in our pockets to the aerospace technologies that traverse our skies, making MEMS nanomaterials a central player in modern science and technology.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	**LLM-retrieved text**
Definition and Structure: Define the nanomaterial category and describe its typical structure at the nanoscale?	Definition of MEMS: MEMS, or Micro-Electro-Mechanical Systems, are miniaturized devices that combine both mechanical and electronic components. They are made using microfabrication techniques, originally derived from the semiconductor industry. MEMS devices integrate functionalities ranging from sensing to actuating in systems at the micrometer scale. When these devices or structures are reduced to the nanoscale, they often fall under the category of Nano-Electro-Mechanical Systems (NEMS). Structure of MEMS at the Nanoscale: MEMS structures, when reduced to the nanoscale, can have distinct characteristics due to quantum effects and increased surface-to-volume ratios. A typical MEMS structure at the nanoscale may include: (a) Substrate Layer: Often made of silicon or another semiconductor material. This forms the base of the MEMS device. (b) Thin Film Layers: These can be conductive, insulating, or piezoresistive depending on the MEMS device's function. Materials often include metals, oxides, or polymers. (c) Mechanical Structures: These can be beams, membranes, cantilevers, or other movable elements that interact with the environment, either by moving in response to external stimuli or by causing movement to achieve a specific function. (d) Electrodes: Used to apply electrical signals to the mechanical structures or to read electrical signals resulting from the motion of these structures. (e) Sensing Elements: At the nanoscale, these elements can be extremely sensitive due to their high surface area-to-volume ratio. They can be designed to detect changes in pressure, temperature, chemical concentration, and other parameters. (f) Actuation Elements: These convert energy (often electrical) into mechanical action. Examples at the nanoscale include nano-tweezers or nano-switches, and (g) Interconnects: Tiny pathways that provide electrical connectivity among different components of the MEMS device. In conclusion, MEMS devices, when miniaturized to the nanoscale, retain their integrated electronic and mechanical functionalities but can exhibit unique properties and enhanced performance due to the nanoscale effects. The structure involves a combination of mechanical elements, sensing and actuation capabilities, and electronic interconnects, all intricately designed and layered on a substrate.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	**LLM-retrieved text**
Synthesis Methods: Examine different methods employed for synthesizing or fabricating nanomaterials within this category. Discuss both their advantages and limitations :	Synthesis Methods for Nanomaterials in MEMS: Nanomaterials used in MEMS devices can be fabricated or synthesized using various techniques. These techniques differ in their mechanisms, complexity, and outcomes. Here's an examination of some common methods, along with their advantages and limitations: (1) Top-down Fabrication: (a) Description: This involves scaling down bulk materials to the desired nano-dimensions using external forces or tools. Techniques like photolithography, electron beam lithography, and nano-imprint lithography fall under this category. (b) Advantages: Direct integration with existing semiconductor processes. Precise control over dimensions and patterns. (c) Limitations: Physical constraints limit the achievable dimensions. Potential damage or defects in the structure during the process. (2) Bottom-up Fabrication: (a) Description: This method assembles nanomaterials atom-by-atom or molecule-by-molecule. Techniques include chemical vapor deposition (CVD), molecular beam epitaxy (MBE), and atomic layer deposition (ALD). (b) Advantages: Can achieve true nanoscale dimensions. Potential for defect-free structures. Greater control at the atomic or molecular level. (c) Limitations: May require high temperatures or ultra-high vacuum conditions. Some processes can be slower and more resource-intensive. (3) Sol-Gel Synthesis: (a) Description: A wet chemical technique where metal alkoxides undergo hydrolysis and polymerization reactions to form a colloidal solution (sol) that acts as a precursor for an inorganic network (gel). (b) Advantages: Can be performed at relatively low temperatures. Versatility in synthesizing various metal oxides. Allows doping with other elements for added functionalities. (c) Limitations: Controlled evaporation is crucial; otherwise, cracks can form. Purity of the end-product may be a concern. (4). Chemical Reduction: (a) Description: This involves reducing metal ions in solution to form metal nanoparticles. Common reducing agents include sodium borohydride or hydrazine. (b) Advantages: Simple and cost-effective. Can produce a wide range of metal nanoparticles. (c) Limitations: Size and shape control can be challenging. Some reducing agents can be hazardous. (5) Hydrothermal/Solvothermal Synthesis: (a) Description: This involves the crystallization of materials at high temperatures and pressures in water or organic solvents, respectively. (b) Advantages: Can synthesize a wide range of nanostructures. Offers good control over particle size and morphology. (c) Limitations: Requires specialized equipment for high pressures and temperatures. Scale-up can be challenging. In conclusion, the choice of synthesis or fabrication method for nanomaterials in MEMS largely depends on the desired properties, dimensions, and applications of the end product. While some methods offer precise control over dimensions, others might be more suited for bulk synthesis. Advancements in these techniques continue, aiming to address their limitations and make the production of nanomaterials more efficient and scalable.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	**LLM-retrieved text**
Properties: Highlight the unique physical, chemical, and electronic properties exhibited by nanomaterials in this category. Explain how these properties differ from those of bulk materials.	Nanomaterials, due to their reduced dimensions, exhibit distinct physical, chemical, and electronic properties that often differ remarkably from their bulk counterparts. Here's a highlight of these unique attributes and a brief comparison with bulk materials: (1). Physical Properties: (a) Surface Area-to-Volume Ratio: Nanomaterials have a much larger surface area relative to their volume. This increased ratio can enhance processes that occur at the surface, such as catalysis or adsorption. Bulk vs Nano: In bulk materials, the majority of atoms are inside the material and not on the surface. In nanomaterials, a significant fraction of atoms or molecules reside on the surface, leading to pronounced surface effects. (b) Mechanical Strength: Some nanomaterials, like carbon nanotubes, exhibit significantly higher strength compared to their bulk forms. Bulk vs Nano: The enhanced strength in nanoscale materials can be attributed to fewer defects as compared to bulk materials, resulting in stronger atomic or molecular interactions. 2. Chemical Properties: (a) Reactivity: Nanomaterials tend to be more reactive due to their increased surface area and the higher number of surface atoms. Bulk vs Nano: Bulk materials, with fewer surface atoms relative to their volume, generally exhibit lower reactivity compared to their nanoscale counterparts. (b) Catalytic Activity: Many nanomaterials serve as effective catalysts because of their enhanced surface reactivity. Bulk vs Nano: In bulk catalysts, only the surface atoms or molecules can interact with reactants, but in nanomaterials, a greater fraction of the material can participate in the catalytic process. 3. Electronic Properties: (a) Quantum Confinement: When the dimensions of materials are reduced to the nanoscale, electrons are confined in a smaller space, leading to discrete energy levels. This can affect electronic and optical properties. Bulk vs Nano: In bulk semiconductors, electrons have a continuous range of energy states, while in quantum dots (a type of nanomaterial), the energy states are quantized. (b) Electrical Conductivity: The conductivity of nanomaterials can be higher due to fewer grain boundaries and defects, or it can be altered by quantum effects. Bulk vs Nano: In bulk materials, grain boundaries and defects often impede electron flow, which may not be the case in some nanomaterials. (c) Band Gap Modulation: The band gap of nanomaterials can be tuned based on their size, especially evident in semiconductor quantum dots. Bulk vs Nano: Bulk semiconductors have a fixed band gap, while in nanomaterials, especially quantum dots, the band gap can vary based on the size of the particle. In conclusion, the properties of nanomaterials are inherently influenced by their reduced dimensions, leading to a set of characteristics that are distinct from their bulk counterparts. These unique attributes are central to the expanding applications of nanomaterials in MEMS and other technological fields, unlocking functionalities and performances that were previously unattainable with bulk materials.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	LLM-retrieved text
Surface Modification: Describe strategies used to modify the surface properties of nanomaterials in this category, including techniques like functionalization, coating, or doping. Explain how these modifications enhance their performance or enable specific applications.	Surface modification plays a pivotal role in tuning the properties of nanomaterials, given the significant fraction of atoms or molecules that reside on their surfaces. By tweaking the surface properties, the performance, compatibility, or functionality of nanomaterials can be optimized. Here are some common strategies: (1) Functionalization: (a) Description: Functionalization involves the attachment of specific functional groups, molecules, or polymers to the surface of nanomaterials. This can be achieved through covalent bonding, electrostatic interactions, or other chemical or physical methods, (b) Applications and Performance Enhancement: Biocompatibility: Functionalization with specific organic molecules can make nanomaterials more compatible with biological systems, facilitating their use in medical MEMS devices. (c) Improved Solubility: Functionalizing nanoparticles with hydrophilic groups can enhance their dispersibility in aqueous solutions. (d) Sensing Applications: By attaching specific molecules to the surface, nanoparticles can be tailored to detect certain chemicals or biological molecules selectively. (2) Coating: (a) Description: This involves enclosing or covering the nanomaterial with another material. Common coatings include polymers, metals, oxides, or other nanoparticles. (b) Applications and Performance Enhancement: (i) Protection: Coatings can protect sensitive nanomaterials from environmental factors like oxidation or moisture. (ii) Enhanced Stability: Encapsulation can prevent aggregation of nanoparticles, ensuring consistent behavior and performance. (iii) Tailored Optical Properties: Coating can modify the optical properties, like fluorescence, of nanoparticles. This is beneficial for imaging or display applications. (3) Doping: (a) Description: Doping introduces impurity atoms into the nanomaterial, deliberately modifying its electronic properties. For example, introducing phosphorus or boron atoms into silicon can change its conductivity type. (b) Applications and Performance Enhancement: (i) Modulation of Electronic Properties: Doping can convert semiconductors from being p-type (positive) to n-type (negative) or vice versa, essential for creating electronic components like transistors. (ii) Enhanced Sensing Capabilities: Doping can make nanomaterials more responsive to certain chemicals, enhancing their sensitivity in sensing applications. (iii) Improved Catalytic Activity: Doped nanomaterials can exhibit enhanced or selective catalytic behaviors, useful in chemical synthesis or environmental remediation. In conclusion, the surface modification strategies like functionalization, coating, and doping offer a versatile toolkit to fine-tune the properties of nanomaterials used in MEMS. By strategically altering their surface, these nanomaterials can be optimized for specific applications or to exhibit desired behaviors, expanding their utility and effectiveness in various technological domains.

Table 10: The table presents the question-answer pairs retrieved from the LLM(GPT-4) for natural language queries about MEMS devices.

Query prompt	LLM-retrieved text
Applications: Explore the extensive range of applications wherein nanomaterials from this category find use. Discuss their potential impact on fields such as electronics, energy, medicine, and more.	Nanomaterials used in MEMS have spurred a revolution across a multitude of fields due to their unique properties. Let's explore some of these applications and their impact on various sectors: (1) Electronics: (a) Nano-transistors: As the demand for smaller and more powerful electronic devices grows, there's a shift towards using nanomaterials like graphene and carbon nanotubes to create transistors at the nanoscale, allowing for more densely packed circuits and improved device performance. (b) Memory Storage: Nanomaterials enable the creation of high-density memory storage devices, leading to compact storage solutions with greater capacities. (c) Sensors: Nanomaterial-based sensors can detect minute quantities of chemicals or physical changes, making them suitable for a range of applications, from environmental monitoring to electronic noses. (2) Energy: (a) Solar Cells: Nanomaterials like quantum dots and nanostructured semiconductors can be used to improve the efficiency of solar cells by enhancing light absorption and charge transport. (b) Batteries and Supercapacitors: Nanomaterials increase the surface area for chemical reactions in batteries, leading to devices with higher energy storage capacities and faster charge/discharge rates. (c) Fuel Cells: Nanocatalysts improve the efficiency of fuel cell reactions, paving the way for cleaner energy sources. (3) Medicine: (a) Drug Delivery: Nanoparticles can be engineered to target specific cells or tissues in the body, allowing for targeted drug delivery, which can reduce side effects and improve therapeutic outcomes. (b) Imaging: Nanoparticles with unique optical properties, such as gold nanoparticles or quantum dots, can be used for enhanced imaging techniques, aiding in the diagnosis and monitoring of diseases. (c) Biosensors: Nanoscale sensors can detect minute concentrations of biomolecules, enabling early disease detection and monitoring. (4) Environmental Applications: (a) Water Purification: Nanomaterials like carbon nanotubes and nanostructured membranes can effectively filter out contaminants from water, addressing global clean water challenges. (b) Air Purification: Nanocatalysts can break down harmful pollutants in the air, contributing to cleaner urban environments. (c) Environmental Monitoring: Nanosensors can detect trace amounts of environmental pollutants, aiding in monitoring and mitigation efforts. (5) Photonics and Optoelectronics: (a) Displays: Quantum dots and other nanoscale phosphors can be used in display technologies, offering brighter and more energy-efficient screens. (b) Lasers: Nanomaterials can be used to produce lasers with specific wavelengths, useful in communications and medical procedures. (c) Photodetectors: Nanoscale detectors can respond to a broader range of light wavelengths with higher sensitivities, improving optical communication systems. In conclusion, the nanomaterials in MEMS have an extensive range of applications that touch almost every aspect of modern life. From enhancing the performance of electronic devices to pioneering medical breakthroughs and addressing environmental challenges, these materials hold the promise of reshaping the future of technology and improving the quality of life on a global scale.