ATOMIC: A Domain-Specific Vision-Language Model for Transmission Electron Microscopy

Chong-ren Tu[1], Hung-wei Hsueh[2], and Shu-han Hsu[2]*

[1] Graduate Program of Artificial Intelligence, National Cheng Kung University, Tainan, Taiwan
[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
shhsu@gs.ncku.edu.tw

Abstract. Vision–language models (VLMs) achieve strong performance on general-domain tasks but can degrade when visual distributions, terminology, and reasoning requirements diverge from general-domain pretraining assumptions, including in scientific imaging. We study this challenge in materials-science transmission electron microscopy (TEM), a data-scarce setting that requires nanoscale visual grounding, modality-specific interpretation, and specialized materials-science knowledge. We introduce ATOMIC (Assistant for TEM Oriented Multimodal Instruction and Conversation), a TEM-oriented vision–language model adapted using peer-reviewed, literature-derived data without manual annotation. Our pipeline extracts figure–caption pairs, isolates TEM subfigures, and generates multimodal instruction data via GPT-based instruction generation. The training strategy separates two complementary supervision signals: VisionGround, which enforces image-only grounding, and DomainContext, which injects caption-conditioned scientific context. Their combination, Blend, yields consistent improvements across perception- and knowledge-intensive queries. ATOMIC-7B-Blend achieves 75.2% accuracy on TEM-MCQ and 37.5% recall-oriented Answer Word Coverage (AWC) on TEM-VQA, narrowing the gap to GPT-4o on AWC (40.4%) while enabling local deployment. On the TEM subset of MatCha, an independent external benchmark, ATOMIC-7B-Blend+FT outperforms LLaVA-v1.5-7B+FT by 18.1 percentage points under identical fine-tuning conditions. We release TEM-MCQ/TEM-VQA annotations, article URLs, figure IDs, normalized crop coordinates for sub-figure reconstruction, the data-curation pipeline, and model weights to support reproducibility within copyright constraints. These results show that structured literature-derived supervision offers an effective data-centric strategy for adapting VLMs to materials-science TEM and may provide a reproducible foundation for future studies of VLM adaptation in related data-scarce scientific-imaging domains.

Keywords: Vision-Language Model · Scientific Image Understanding · Instruction Tuning · Transmission Electron Microscopy

* Corresponding author.

1 Introduction

Scientific domain shift in vision–language models. Large-scale vision–language models (VLMs) achieve strong performance on general-domain tasks [1, 24, 25, 39, 42], yet their behavior under domain shift—especially in scientific imaging—remains insufficiently understood. Scientific domains differ from natural-image corpora along multiple axes: (i) *distribution shift* in visual statistics (e.g., modality-dependent contrast mechanisms, grayscale textures, atomic-scale lattices), (ii) *vocabulary shift* toward specialized terminology, and (iii) *reasoning shift* requiring experimental interpretation rather than object-centric recognition. Together, these shifts make scientific imaging a challenging setting for evaluating and adapting vision–language models.

Materials-science TEM as a domain-shift setting. Transmission electron microscopy (TEM) provides a representative testbed for studying scientific domain shift in VLMs [3, 26, 37]. Widely used in materials science and semiconductor research, TEM enables atomic-scale characterization of lattice structures, interfaces, and nanoscale defects. While deep learning methods have been applied to TEM for specific tasks such as denoising [17], defect detection [7, 11, 33], and segmentation [12, 15, 32, 38], these approaches typically optimize single-purpose outputs and do not support interactive, language-conditioned interpretation of scientific images. Furthermore, acquiring large-scale annotated datasets for TEM remains difficult, creating a fundamental bottleneck for adapting and evaluating vision–language models in this domain.

Research question. This setting raises the following question, which we study in the context of materials-science TEM:

> *How can vision–language models be adapted to a scientific imaging domain where both visual statistics and domain vocabulary lie outside the distributions seen during multimodal pretraining, particularly when labeled data are scarce?*

Our approach. ATOMIC addresses this challenge from a *data-centric* perspective, demonstrating that structured literature mining combined with a GPT-based self-instruct data generation strategy [28, 36] can induce domain-aligned multimodal representations without architectural modification. A key design decision is the separation of two complementary training signal sources: *VisionGround*, which enforces pure visual grounding, and *DomainContext*, which injects domain context from parent figure captions. ATOMIC is designed for open, locally deployable inference, offering an API-free alternative that avoids sending proprietary microscopy data to cloud-based services.

Findings. To evaluate the effectiveness of this structured pipeline, we construct TEM-MCQ and TEM-VQA as reproducible evaluation datasets spanning visual

perception, scientific reasoning, and experimental methodology. ATOMIC-7B-Blend achieves 75.2% accuracy on TEM-MCQ and outperforms all locally deployable baselines [10, 20, 22] on TEM-VQA across all evaluation metrics. We further evaluate on the TEM subset of MatCha [18], an independent external benchmark, where ATOMIC-7B-Blend+FT improves over LLaVA-v1.5-7B+FT by 18.1 percentage points. While frontier models (GPT-4o, Gemini-2.5-Flash-Lite) serve as upper-bound references, ATOMIC narrows the open-ended TEM-VQA AWC gap while enabling local deployment without API dependency.

Contributions.

- **Data-centric adaptation framework for TEM VLMs.** We demonstrate a practical route for reducing annotation barriers in data-scarce TEM VLM adaptation: peer-reviewed materials-science literature provides a scalable source of multimodal supervision, and systematic figure–caption mining can construct TEM-specific training data without manual annotation.
- **Complementary training signals from scientific literature.** We show that *VisionGround*, which enforces image-only grounding, and *DomainContext*, which injects caption-derived scientific context, address distinct capability dimensions under TEM domain shift. Their combination, *Blend*, yields more consistent performance across perception- and knowledge-intensive queries.
- **Evaluation protocol for TEM VLM adaptation.** We establish a replicable TEM evaluation dataset construction pipeline spanning visual perception, scientific reasoning, and experimental methodology across multiple-choice and open-ended formats, enabling controlled assessment of both closed-ended decision making and free-form scientific explanation.

2 Related Work

General-purpose and domain-specific vision–language models. Large-scale VLMs have advanced through cross-modal alignment and instruction tuning, including BLIP-2 [20], InstructBLIP [10], LLaVA [23], and LLaVA-v1.5 [22]. However, because these general-purpose models are trained largely on natural-image corpora, their behavior under materials-science microscopy domain shift—where visual distributions, specialized terminology, and scientific reasoning requirements depart from general-domain pretraining—remains insufficiently characterized [5, 34]. Domain-specific VLMs and biomedical image–text foundation models, such as LLaVA-Med [19] and BiomedCLIP [41], illustrate a parallel line of domain-adaptive multimodal training for specialized image distributions. These systems target biomedical images and clinical or biomedical language rather than TEM-specific materials characterization. ATOMIC instead focuses on data-centric adaptation for materials-science TEM, holding the LLaVA-v1.5 architecture fixed to isolate the effect of supervision design rather than architectural changes.

Instruction tuning with GPT. LLM-generated instruction data has become widely used for instruction tuning [4], beginning with Self-Instruct [36] and Alpaca [35] and further strengthened by GPT-4-generated data [28]. This recipe was adopted in multimodal settings by LLaVA [23] and LLaVA-Med [19]. For TEM, however, generic visual-instruction templates are insufficient: interpretation depends on modality-specific contrast, lattice and defect morphology, crystallographic cues, and experimental context from expert captions. GPT-based generation in such specialized settings can also introduce hallucination risk [40]. ATOMIC addresses this by conditioning generation on expert-written captions and separating image-only *VisionGround* supervision from caption-conditioned *DomainContext* supervision, enabling analysis of how visual evidence and domain context support TEM VLM adaptation.

Scientific imaging and multimodal reasoning. Machine learning methods in microscopy have largely focused on task-specific objectives. Although effective, these approaches optimize single-purpose outputs and do not support interactive, language-conditioned interpretation of microscopy evidence. More recently, MicroscopyGPT [8] explored VLM-based atomic structure prediction from simulated STEM images of 2D materials, representing an important step toward language-conditioned microscopy understanding. Its task formulation differs from ours: MicroscopyGPT requires chemical-formula input and produces POSCAR-format atomic structures, whereas TEM-MCQ and TEM-VQA evaluate natural-language responses to TEM images. Direct quantitative comparison would therefore require substantial task reformulation. Our work complements this direction by studying open-ended VLM adaptation across multiple TEM modalities, with evaluation spanning visual perception, scientific reasoning, and experimental methodology in both multiple-choice and free-form VQA formats.

3 Method

3.1 ATOMIC Training Framework

We build ATOMIC on LLaVA-v1.5 without modifying the underlying architecture, focusing on data-centric adaptation for materials-science TEM. Our training design decomposes supervision into two complementary signals: *VisionGround*, derived from TEM images for visual grounding, and *DomainContext*, derived from figure captions for domain-specific scientific context. This decomposition enables controlled analysis of how structured supervision affects vision–language alignment under TEM domain shift.

The core objective is to bridge the mismatch between general-domain VLM pretraining and TEM interpretation. General vision encoders have limited exposure to TEM-specific visual primitives, such as lattice fringes, diffraction patterns, and STEM contrast, while generic instruction data rarely captures microscopy terminology or reasoning patterns. We therefore use structured literature-derived supervision to evaluate whether data-centric supervision design can improve TEM VLM adaptation while keeping the base architecture fixed.

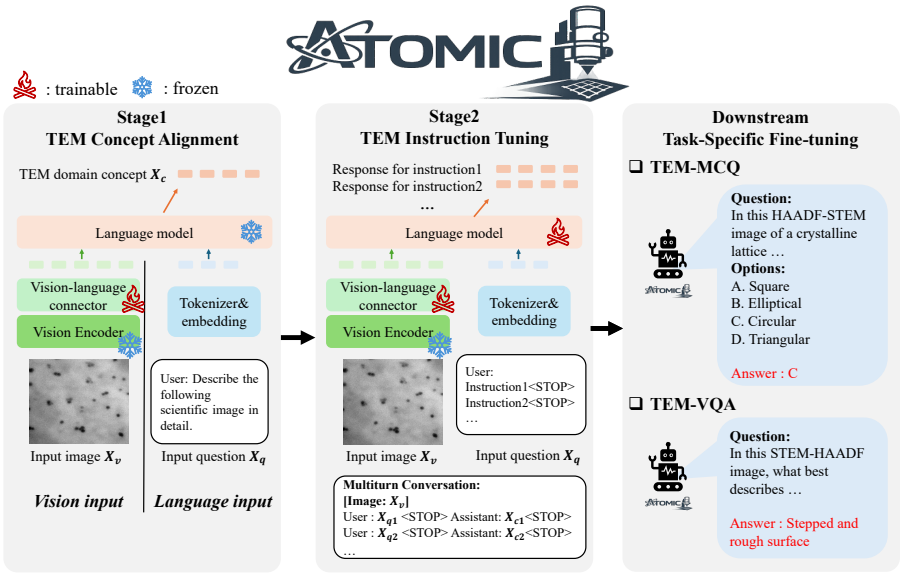

Fig. 1: Overview of ATOMIC model training pipeline. Stage 1 (TEM Concept Alignment) trains only the Vision-language connector while keeping the Vision Encoder and Language model frozen. Stage 2 (TEM Instruction Tuning) unfreezes the Language model for joint training using multi-turn conversational data. The resulting ATOMIC model can be adapted to downstream tasks such as TEM-MCQ and TEM-VQA.

We follow LLaVA-v1.5, using Vicuna-7B [6] as the language backbone and CLIP-ViT-L-336px [29] as the vision tower. The resulting model, ATOMIC, is adapted for materials-science TEM, as summarized in Fig. 1.

Stage 1: Image–Text Alignment. Stage 1 freezes the Vision Tower and language model, training only the MLP Connector to project visual features into the language space and establish image–text correspondence.

Stage 2: Visual Instruction Tuning. Stage 2 keeps the Vision Tower frozen while jointly training the language model and MLP Connector on multi-turn visual conversations, enabling image-conditioned instruction following.

Downstream: Task-Specific Fine-Tuning. After Stage 2, we fine-tune ATOMIC on downstream tasks including visual question answering (VQA) and multiple-choice questions (MCQ), adapting the model to open-ended and multiple-choice response formats.

3.2 Data Source: Peer-Reviewed Figure–Caption Pairs

We crawl TEM-related articles from four Nature-series journals (*Nature*, *Nature Communications*, *Nature Materials*, and *Nature Nanotechnology*) and extract

figure elements along with their corresponding captions using Selenium-based browser automation and HTML parsing, yielding 34,073 figure–caption pairs. Captions provide expert-written context (materials, processing conditions, imaging mode) that is difficult to annotate manually and is essential for generating domain-aligned training data. For each figure, we retain article URL, figure ID, caption association, and sub-figure crop metadata to support reconstruction without redistributing publisher-owned images.

3.3 TEM Sub-Figure Extraction and Modality Verification

Raw paper figures are often compound, mixing TEM panels with plots, schematics, and other non-TEM characterization panels. We isolate TEM regions with a two-stage pipeline: YOLOv11 [16] localizes candidate subfigures, and ResNet-50 [13] retains four commonly used TEM modalities (CTEM, HRTEM, STEM, SAED). This filters 34,073 figure–caption pairs to 9,513 pairs and 32,564 modality-verified TEM subimages. Before instruction generation, we split these into 29,950 images for Stages 1–2 and 2,614 held-out images for evaluation dataset construction; no held-out evaluation image is used to generate Stage 1 or Stage 2 training data. The complete data construction pipeline is shown in Fig. 2.

3.4 Stage 1: TEM-Oriented Image–Text Alignment

Stage 1 trains only the vision–language connector to map visual tokens into the language space while keeping the vision encoder and language model frozen. A central challenge in TEM is the semantic gap between pretrained visual features and domain-specific terminology.

To address this, we generate single-turn alignment targets with GPT using a two-dimensional target-generation design: response granularity (Brief vs. Detail) and conditioning regime (VisionGround vs. DomainContext):

- **Brief vs. Detail.** Brief responses provide concise visual alignment signals, forcing the MLP Connector to map visual tokens to core semantics, while Detail responses provide fine-grained feature representations, ensuring alignment richness.
- **VisionGround vs. DomainContext.** VisionGround conditions solely on the TEM image to enforce image-grounded alignment and avoid textual shortcuts. DomainContext additionally incorporates the parent figure caption when generating the target response, injecting domain-consistent terminology and expert-level contextual information that pure visual observation cannot provide. The parent caption is used as supervision-generation context, not as an auxiliary input supplied to the evaluated model.

For each training sub-image, we generate four alignment targets: Brief and Detail responses under VisionGround and DomainContext conditioning. These targets form the VisionGround-60k, DomainContext-60k, and Blend-120k Stage 1 datasets in Tab. 1.

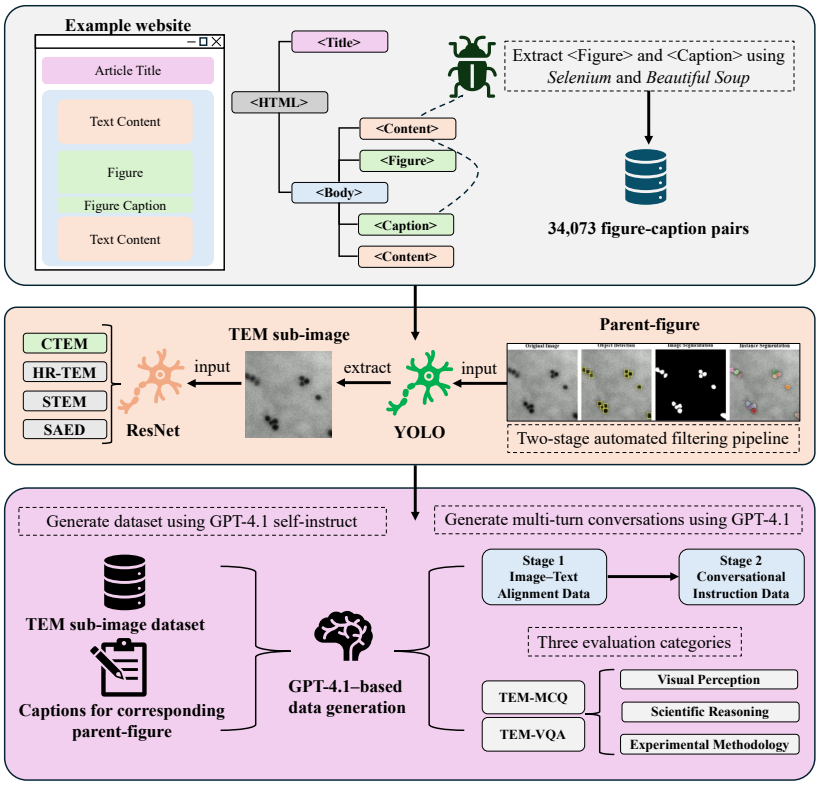

Fig. 2: Overview of data construction pipeline. Scientific articles are automatically converted into multimodal training data through figure–caption mining, modality-aware filtering, and GPT-based instruction generation. The resulting dataset comprises visual grounding signals and domain-context signals, enabling controlled analysis of training signal composition for vision–language domain adaptation.

3.5 Stage 2: Multi-Turn TEM Instruction Tuning

Stage 2 unfreezes the language model and jointly trains it with the MLP Connector on GPT-generated multi-turn conversations derived from Stage 1 Detail responses. We retain the VisionGround/DomainContext split when forming the Stage 2 datasets summarized in Tab. 1. Detail responses are selected over Brief responses because they provide richer visual descriptions and, for DomainContext, caption-derived scientific context for generating meaningful multi-turn interactions. This stage adapts the model's response style to microscopy-specific discourse, enabling interactive question answering and scientifically grounded explanations. Captions influence Stage 2 only through DomainContext-derived generated responses; at inference time, the model receives the image and prompt, not the parent caption.

Table 1: Dataset versions and their composition.

	Version	Composition	Size
Stage 1	VisionGround-60k	Brief + Detail	60k
	DomainContext-60k	Brief + Detail	60k
	Blend-120k	VisionGround-60k + DomainContext-60k	120k
Stage 2	VisionGround-30k	Converted from VisionGround Detail	30k
	DomainContext-30k	Converted from DomainContext Detail	30k
	Blend-60k	VisionGround-30k + DomainContext-30k	60k

3.6 Evaluation Dataset Construction

We construct TEM-MCQ and TEM-VQA using 2,614 TEM subimages held out from all training stages. TEM-MCQ enables objective accuracy measurement through closed-ended multiple-choice questions, while TEM-VQA assesses open-ended scientific dialogue capability through free-form responses. For each image, an LLM generates multiple questions spanning three capability dimensions: *visual perception* (observable morphology, contrast, and texture), *scientific reasoning* (structure, defects, mechanisms, and property inference), and *experimental methodology* (imaging mode identification, artifacts, calibration, and acquisition considerations). When domain context is required to make a question well-posed, we apply *context injection* by embedding the minimal necessary non-answer-revealing information directly into the question stem.

TEM-VQA is formed by converting a randomly selected half of TEM-MCQ questions into open-ended VQA by removing answer choices and requiring free-form responses. This paired design preserves the visual distribution and capability taxonomy while changing the response format. TEM-MCQ and TEM-VQA comprise 23,526 QA pairs; dataset statistics are reported in Sec. 4.

Leakage Control. During evaluation-dataset generation, the GPT-based generator has access to the image and parent caption to support factually grounded question generation. At inference time, however, all evaluated models receive only the TEM image and task prompt. Parent captions, article metadata, source indices, and reconstruction metadata are never supplied to the evaluated model. This design preserves realistic scientific question content while preventing the evaluated model from solving the task through textual source cues alone.

4 Experiments

4.1 Experimental Setup

Dataset Split. We divide the 32,564 TEM subimages into two subsets: 29,950 images are used for Stage 1 and Stage 2 training, and 2,614 images are reserved for TEM-MCQ/TEM-VQA construction as described in Sec. 3.6. No image from

Table 2: Hyperparameter configurations for each training stage.

Hyperparameter	**Stage 1**	**Stage 2**	**Task FT**
Learning Rate	1e-3	2e-5	2e-5
Batch Size (per device)	32	16	16
Gradient Accumulation	1	1	1
Epochs	1	3	1
LR Scheduler	Cosine	Cosine	Cosine
Warmup Ratio	0.03	0.03	0.03
Max Sequence Length	2,048	2,048	2,048
DeepSpeed Stage	ZeRO-2	ZeRO-3	ZeRO-3

the 2,614-image pool is used during Stage 1 or Stage 2 training. The training data is further organized into multiple versions according to different data generation strategies, as detailed in Sec. 3.4 and Sec. 3.5.

The evaluation-construction pool is further divided into a downstream fine-tuning split and a final test split. The test split (TEM-VQA: 2,349 pairs; TEM-MCQ: 2,358 pairs) is used for the final evaluation of all models, and all experimental results are reported on this final test split exclusively. The task-training split (TEM-VQA: 9,405 pairs; TEM-MCQ: 9,414 pairs; 18,819 pairs in total) is reserved exclusively for the downstream task fine-tuning (+FT) stage to train ATOMIC-7B-Blend+FT, and is kept fully independent from the Stage 1 and Stage 2 adaptation data.

Training Setup. All models are trained on 4 × NVIDIA RTX PRO 6000 Blackwell GPUs with BF16 mixed precision, using the DeepSpeed [30, 31] framework for distributed training. The hyperparameter configurations for the three training stages are summarized in Tab. 2.

Using the Blend dataset as a reference, Stage 1 training (120k samples) takes approximately 2 hours, Stage 2 training (60k samples) approximately 4.5 hours, and downstream fine-tuning approximately 45 minutes.

In this work, we train two model scales: the ATOMIC-7B series uses Vicuna-7B-v1.5 as the base language model, and the ATOMIC-13B series uses Vicuna-13B-v1.5 as the base language model. Both series adopt CLIP-ViT-L-336px as the Vision Tower and undergo the complete Stage 1 and Stage 2 training pipeline. For downstream task fine-tuning (+FT), we initialize from the Stage 2 checkpoint of ATOMIC-7B-Blend and further fine-tune on the TEM-MCQ and TEM-VQA training data, yielding ATOMIC-7B-Blend+FT. Frontier VLMs are evaluated zero-shot as upper-bound API references and are not task fine-tuned.

Evaluation Metrics. For the open-ended question answering task (TEM-VQA), we adopt the following four evaluation metrics.

Table 3: Overall performance on TEM-VQA and TEM-MCQ.

Model	TEM-VQA				TEM-MCQ
	BLEU-1 [27]	**ROUGE-L** [21]	**METEOR** [2]	**AWC**	**Accuracy**
Frontier VLMs					
GPT-4o [14]	**21.9%**	**26.1%**	**29.3%**	**40.4%**	**92.3%**
Gemini-2.5-Flash-Lite [9]	17.8%	22.2%	26.0%	37.7%	90.3%
ATOMIC (Ours)					
ATOMIC-13B-Blend	11.7%	15.5%	21.7%	36.6%	**76.5%**
ATOMIC-7B-Blend	**12.2%**	**16.2%**	**22.8%**	**37.5%**	75.2%
Baselines					
LLaVA-v1.5-13B [22]	10.5%	15.0%	17.8%	31.0%	74.0%
LLaVA-v1.5-7B [22]	9.6%	13.6%	18.2%	33.3%	70.9%
BLIP-2 [20]	9.7%	14.3%	9.1%	12.3%	73.5%
InstructBLIP [10]	9.0%	14.4%	9.1%	12.6%	61.6%

BLEU-1 [27] measures the unigram precision between the model's response and the reference answer, reflecting the degree of lexical overlap.

ROUGE-L [21] computes recall based on the Longest Common Subsequence (LCS), reflecting the degree of structural similarity between the model's response and the reference answer at the sequence level.

METEOR [2] extends BLEU with stemming and synonym matching, capturing overlap beyond exact n-gram matches.

Answer Word Coverage (AWC) is a reference-token coverage metric that measures the proportion of reference-answer tokens covered by the model response. It indicates whether the response includes expected technical content, rather than serving as a full semantic-quality metric. AWC is defined as:

$$\mathrm{AWC} = \frac{\sum_{w \in V_{\mathrm{ref}}} \min(\mathrm{count}_{\mathrm{ref}}(w),\, \mathrm{count}_{\mathrm{hyp}}(w))}{\sum_{w \in V_{\mathrm{ref}}} \mathrm{count}_{\mathrm{ref}}(w)}, \tag{1}$$

where V_{ref} is the set of unique reference tokens, and $\mathrm{count}_{\mathrm{ref}}(w)$ and $\mathrm{count}_{\mathrm{hyp}}(w)$ denote token counts in the reference answer and model response. Because TEM-VQA reference answers are often short and terminology-dense, while model responses can vary substantially in length, AWC complements BLEU, ROUGE-L, and METEOR by emphasizing coverage of expected technical content rather than exact phrasing or fluency. We further examine its correlation with TEM-specialist scores in Sec. 4.4.

For the closed-ended multiple-choice task (TEM-MCQ), we report the accuracy of each of the three capability dimensions — Visual Perception, Reasoning, and Experimental Methodology — as well as the overall average accuracy.

4.2 Main Results

Table 3 presents results across all models. GPT-4o and Gemini-2.5-Flash-Lite are included as zero-shot frontier references rather than locally deployable baselines.

Table 4: Performance on the TEM subset of MatCha [18], an independent materials-characterization benchmark not generated by our pipeline. All scores are accuracy (%).

Model	Processing Correlation (PC) (n=77)	Morphology Analysis (MA) (n=90)	Structure Analysis (SA) (n=165)	Property Analysis (PA) (n=6)	Defect Type Classification (DTC) (n=33)	ALL (n=371)
Frontier VLMs						
GPT-4o [14]	**83.1**	**73.3**	**72.1**	66.7	**57.6**	**73.3**
Gemini-2.5-Flash-Lite [9]	83.1	60.0	67.9	**100.0**	54.5	68.5
ATOMIC (Ours)						
ATOMIC-7B-Blend+FT	**72.7**	**63.3**	**52.7**	**50.0**	**21.2**	**56.6**
ATOMIC-7B-Blend	59.7	47.8	47.9	33.3	21.2	47.7
Baselines						
LLaVA-v1.5-7B+FT	35.1	46.7	39.4	50.0	18.2	38.5
LLaVA-v1.5-7B [22]	22.1	44.4	39.4	33.3	18.2	35.0
BLIP-2 [20]	16.9	25.6	33.9	33.3	21.2	27.2
InstructBLIP [10]	59.7	20.0	32.7	0.0	24.2	34.0

They achieve TEM-MCQ accuracy of 92.3%/90.3% and TEM-VQA AWC of 40.4%/37.7%. Among locally deployable models, ATOMIC-7B-Blend achieves the best TEM-VQA performance and ATOMIC-13B-Blend the highest MCQ accuracy (76.5%), both outperforming all locally deployable baselines. ATOMIC-7B-Blend reaches 37.5% AWC, narrowing the gap to GPT-4o (40.4%) while enabling local, API-free deployment for privacy-sensitive proprietary microscopy workflows.

To assess generalization beyond our internal evaluation set, we evaluate on the TEM subset of MatCha [18], an independent benchmark not generated by our pipeline. ATOMIC-7B-Blend+FT achieves 56.6% overall accuracy, outperforming LLaVA-v1.5-7B+FT by 18.1 percentage points under identical fine-tuning conditions, suggesting that staged TEM adaptation provides a more effective task-fine-tuning initialization than the corresponding general-domain LLaVA-v1.5-7B checkpoint. Without task fine-tuning, ATOMIC-7B-Blend reaches 47.7%, outperforming all non-fine-tuned local baselines. Given the small PA category ($n = 6$), its per-category accuracy is reported for completeness and should be treated as descriptive rather than conclusive.

4.3 Ablation Study

To evaluate the impact of different supervision strategies, we compare VisionGround, DomainContext, and Blend, as well as the effect of downstream task fine-tuning (+FT). The detailed results are presented in Tab. 5 and Tab. 6.

Effect of Data Strategy (VisionGround vs. DomainContext vs. Blend). Combining the results of Tab. 5 and Tab. 6, VisionGround and DomainContext exhibit complementary strengths across tasks and capability dimensions.

Blend consistently improves robustness by combining image-only grounding with caption-conditioned domain context.

On TEM-VQA, DomainContext outperforms VisionGround on both Reasoning (25.0 vs. 23.7) and Experimental Methodology (18.3 vs. 17.7), while VisionGround performs slightly better on Visual Perception (22.2 vs. 20.7). This pattern is intuitive: open-ended question answering requires the model to generate free-form responses containing accurate terminology and experimental context, and the caption-derived materials-science context in DomainContext data helps the model produce more precise descriptions for reasoning and experimental methodology questions. Visual Perception questions, by contrast, primarily test directly observable image features, where the image-only training signal of VisionGround holds a natural advantage. Blend improves consistency across all three dimensions by integrating both supervision sources.

On TEM-MCQ, the patterns for Visual Perception and Experimental Methodology are consistent with TEM-VQA — DomainContext outperforms VisionGround on Experimental Methodology (73.3% vs. 71.4%), while VisionGround performs better on Visual Perception (64.8% vs. 62.9%). However, a noteworthy reversal is observed in the Reasoning dimension: VisionGround (86.1%) outperforms DomainContext (83.0%), which is opposite to the pattern observed in TEM-VQA. We attribute this discrepancy to the fundamentally different demands the two tasks place on reasoning capability: MCQ Reasoning questions require the model to directly discriminate between answer options based on visual evidence, relying more heavily on precise visual feature judgment; whereas VQA Reasoning questions require the generation of free-form text incorporating reasoning processes, relying more on domain knowledge expression at the language level. This pattern highlights the distinct strengths of VisionGround and DomainContext, and further explains why the Blend strategy achieves more consistent performance across both tasks.

Effect of Model Scale (7B vs. 13B). Scaling ATOMIC-Blend from 7B to 13B yields only a small TEM-MCQ gain (75.2% to 76.5%) and does not improve TEM-VQA, where ATOMIC-7B-Blend remains stronger. These results suggest that, in this data-scarce TEM setting, performance is not governed by param-

Table 5: Ablation results on TEM-MCQ.

	Model	Perception	Reasoning	Methodology
ATOMIC Variants	ATOMIC-7B-Blend+FT	**84.6%**	**94.3%**	**94.5%**
	ATOMIC-7B-Blend	66.5%	85.6%	73.3%
	ATOMIC-7B-DomainContext	62.9%	83.0%	73.3%
	ATOMIC-7B-VisionGround	64.8%	86.1%	71.4%
	ATOMIC-13B-Blend	66.5%	87.7%	75.3%
Baselines	LLaVA-v1.5-13B [22]	65.5%	83.8%	72.5%
	LLaVA-v1.5-7B [22]	60.9%	84.2%	67.6%

Table 6: Ablation results on TEM-VQA. Scores averaged over BLEU-1, ROUGE-L, METEOR, and AWC.

	Model	Perception	Reasoning	Methodology
ATOMIC Variants	ATOMIC-7B-Blend+FT	**41.9%**	**36.9%**	**47.9%**
	ATOMIC-7B-Blend	22.5%	25.1%	19.0%
	ATOMIC-7B-DomainContext	20.7%	25.0%	18.3%
	ATOMIC-7B-VisionGround	22.2%	23.7%	17.7%
	ATOMIC-13B-Blend	21.0%	24.6%	18.5%
Baselines	LLaVA-v1.5-13B [22]	17.0%	21.6%	17.2%
	LLaVA-v1.5-7B [22]	16.9%	21.9%	17.2%

eter scaling alone; carefully structured domain-specific supervision is a major contributor to TEM VLM adaptation.

Effect of Downstream Task Fine-tuning (Blend vs. Blend+FT). Downstream task fine-tuning yields the largest gains across both closed-ended and open-ended tasks. These gains reflect both task-format adaptation and the benefit of ATOMIC's Stage 1/2 TEM-adapted initialization. On the independent TEM subset of MatCha (Tab. 4), ATOMIC-7B-Blend+FT outperforms LLaVA-v1.5-7B+FT by 18.1 percentage points under identical fine-tuning conditions, suggesting that staged TEM adaptation provides a more effective initialization for task-specific fine-tuning than the corresponding general-domain LLaVA-v1.5-7B checkpoint.

4.4 Human Evaluation and Data Quality

Human evaluation of open-ended responses. A TEM domain specialist scored 810 responses to 270 TEM-VQA questions (0–10) across GPT-4o, ATOMIC-7B-Blend, and LLaVA-v1.5-7B. As shown in Tab. 7, AWC achieves the highest Pearson correlation with specialist scores in every model subset, and all AWC correlations are statistically significant ($p < 0.05$). This supports AWC as a useful reference-token coverage metric for TEM-VQA, while TEM-specialist scores provide an independent assessment of overall response quality. Human average scores (GPT-4o: 6.27, ATOMIC: 5.75, LLaVA: 4.57) are consistent with the ranking established by automatic metrics.

GPT training data quality. A TEM domain specialist independently scored 962 Stage 2 QA pairs (479 VisionGround / 483 DomainContext) on a factual-accuracy rubric (0–10), where $\geq$7 indicates high-quality responses and $\leq$3 indicates clear factual errors. Of the 962 pairs, 961 (99.9%) were rated $\geq$7 and none were rated $\leq$3. This sampled audit indicates a low observed severe factual-error rate in the generated TEM instruction data. It should not be interpreted as an error-free guarantee for the entire generated corpus.

Table 7: Pearson correlations between automatic metrics and TEM-specialist scores across three models. * denotes $p < 0.05$.

Model	Human Avg.	AWC	BLEU-1	ROUGE-L	METEOR
GPT-4o	6.27	**+0.516***	+0.398*	+0.422*	+0.478*
ATOMIC-7B-Blend	5.75	**+0.394***	+0.259*	+0.286*	+0.328*
LLaVA-v1.5-7B	4.57	**+0.424***	+0.214*	+0.141	+0.325*

Failure mode analysis. Drawing on the TEM specialist scores, we observe three systematic failure categories in ATOMIC: inaccurate quantitative scale estimation, difficulty in periodic structure interpretation, and overgeneralization of local contrast differences. As these failure patterns are also observed in LLaVA-v1.5-7B, they may reflect limitations not fully addressed by data-level adaptation alone, suggesting that further improvements may require architectural advances in addition to data-level adaptation.

4.5 ATOMIC Response Examples

Figure 3 presents a qualitative comparison of model outputs on a representative TEM image. Given the prompt *"Describe the following scientific image in detail,"* general-purpose models fail to recognize the domain-specific nature of the input: LLaVA-v1.5-7B describes the image as *"possibly a microscope slide"* with *"visually intriguing patterns,"* while InstructBLIP likens the content to *"a snowy landscape"* caused by *"dirt, dust, or other substances."* BLIP-2 produces a degenerate response with no meaningful content. In contrast, ATOMIC correctly identifies the image as a transmission electron microscopy image and accurately describes the key visual features — nanoparticle distribution, size uniformity, and dispersion — using domain-appropriate terminology. This qualitative comparison illustrates the benefit of TEM-specific supervision for interpreting TEM images that general-purpose VLMs otherwise describe using natural-image priors.

5 Conclusion

This work introduces ATOMIC, a TEM-oriented vision–language model built via a data-centric framework that converts peer-reviewed materials-science literature into structured multimodal training data without manual annotation. A key finding is that VisionGround and DomainContext signals address distinct capability dimensions under TEM domain shift, and that their combination yields consistent improvements across task types and formats. ATOMIC outperforms all locally deployable baselines on TEM-MCQ and TEM-VQA, with further gains from downstream fine-tuning. Evaluation on the TEM subset of MatCha provides additional evidence beyond our internal evaluation set, with ATOMIC-7B-Blend+FT outperforming LLaVA-v1.5-7B+FT by 18.1 percentage points under

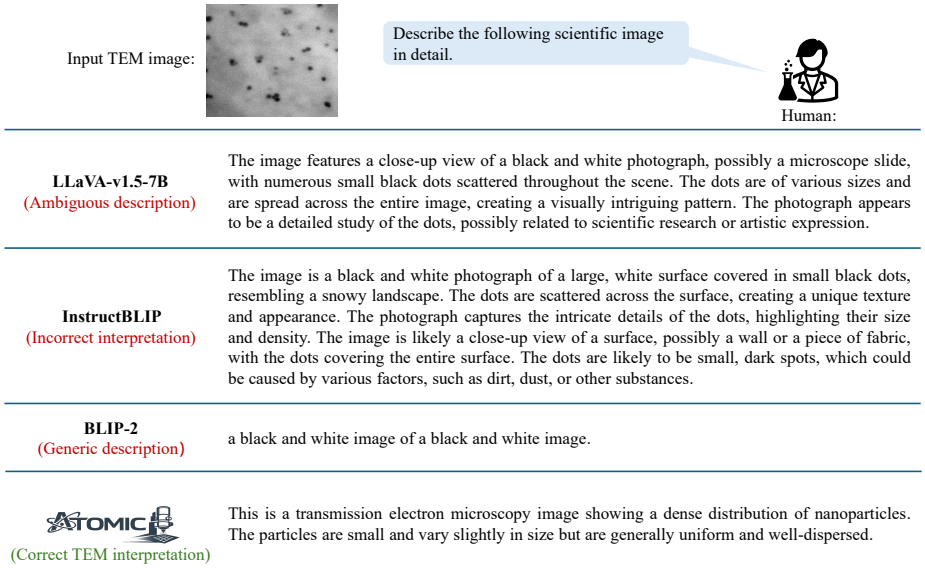

Fig. 3: Qualitative comparison of model responses to the prompt *"Describe the following scientific image in detail"* on a representative TEM image. General-purpose models fail to recognize the TEM domain, producing generic or erroneous descriptions. ATOMIC correctly identifies the imaging modality and accurately characterizes the nanoparticle distribution using domain-appropriate terminology.

identical conditions. Overall, our results suggest that, for materials-science TEM, structured literature-derived supervision is an effective data-centric strategy for VLM adaptation without architectural modification. More broadly, ATOMIC offers a reproducible, TEM-grounded template that may inform data-centric VLM adaptation in related scientific-imaging settings.

Limitations and Dataset Availability. Our literature-derived training data and internal evaluation datasets are derived from published figures, which introduces distribution bias toward well-curated, publication-quality images and may underrepresent challenging acquisition conditions in real-world laboratory settings. To support reproducibility within copyright constraints, we release TEM-MCQ/TEM-VQA annotations, article URLs, figure IDs, normalized crop coordinates for sub-figure reconstruction, the complete data-curation pipeline, and fine-tuned model weights at `https://github.com/SemiMIRTLab/ATOMIC.git`.

Acknowledgements

We are grateful for the support from the National Science and Technology Council (Taiwan, ROC) through Project NSTC 112-2221-E-006-151-MY3, NVIDIA Academic Grant, and Google Research Scholar Award.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (NeurIPS) **35**, 23716–23736 (2022)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
3. Chen, K., Barnard, A.: Advancing electron microscopy using deep learning. Journal of Physics: Materials **7**(2), 022001 (2024)
4. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision (ECCV). pp. 370–387. Springer (2024)
5. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: Wsi-vqa: Interpreting whole slide images by generative visual question answering. In: European Conference on Computer Vision (ECCV). pp. 401–417. Springer (2024)
6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`, accessed: 2026-06-29
7. Cho, P., Wood, A., Mahalingam, K., Eyink, K.: Defect detection in atomic resolution transmission electron microscopy images using machine learning. Mathematics **9**(11), 1209 (2021)
8. Choudhary, K.: Microscopygpt: generating atomic-structure captions from microscopy images of 2d materials with vision-language transformers. The Journal of Physical Chemistry Letters **16**(27), 7028–7035 (2025)
9. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
10. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems (NeurIPS) **36**, 49250–49267 (2023)
11. Gui, C., Zhang, Z., Li, Z., Luo, C., Xia, J., Wu, X., Chu, J.: Deep learning analysis on transmission electron microscope imaging of atomic defects in two-dimensional materials. IScience **26**(10) (2023)
12. Gumbiowski, N., Loza, K., Heggen, M., Epple, M.: Automated analysis of transmission electron micrographs of metallic nanoparticles by machine learning. Nanoscale advances **5**(8), 2318–2326 (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
15. Jacobs, R., Shen, M., Liu, Y., Hao, W., Li, X., He, R., Greaves, J.R., Wang, D., Xie, Z., Huang, Z., et al.: Performance and limitations of deep learning semantic

segmentation of multiple defects in transmission electron micrographs. Cell Reports Physical Science **3**(5) (2022)
16. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), `https://github.com/ultralytics/ultralytics`, accessed: 2026-06-29
17. Kim, J., Rhee, J., Kang, S., Jung, M., Kim, J., Jeon, M., Park, J., Ham, J., Kim, B.H., Lee, W.C., et al.: Self-supervised machine learning framework for high-throughput electron microscopy. Science Advances **11**(14), eads5552 (2025)
18. Lai, Z., Zheng, Y., Cai, Z., Lyu, H., Yang, J., Liang, H.Q., Hu, Y., Wang, B.: Can multimodal LLMs see materials clearly? a multimodal benchmark on materials characterization. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 4384–4404. Association for Computational Linguistics, Suzhou, China (Nov 2025)
19. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems (NeurIPS) **36**, 28541–28564 (2023)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML). pp. 19730–19742. PMLR (2023)
21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the ACL Workshop on Text Summarization Branches Out. pp. 74–81 (2004)
22. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems (NeurIPS) **36**, 34892–34916 (2023)
24. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision (ECCV). pp. 216–233. Springer (2024)
25. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems (NeurIPS) **35**, 2507–2521 (2022)
26. Pai, C.W., Hsueh, H.W., Hsu, S.H.: Reliability of deep learning models for scanning electron microscopy analysis. Machine Learning: Science and Technology **6**(4), 045008 (2025)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
28. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
30. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC20). pp. 1–16. IEEE (2020)

31. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD). pp. 3505–3506 (2020)
32. Shen, M.H., Chang, W.C., Chu, W.H., Cheng, Y.H., Tseng, S.W., Hsu, S.H.: A lightweight data-augmented deep learning framework for real-time instance segmentation in liquid-phase in situ transmission electron microscopy. ACS Measurement Science Au **6**(2), 476–487 (2026)
33. Shen, M., Li, G., Wu, D., Yaguchi, Y., Haley, J.C., Field, K.G., Morgan, D.: A deep learning based automatic defect analysis framework for in-situ tem ion irradiations. Computational Materials Science **197**, 110560 (2021)
34. Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., Zhang, Y., Wan, D., Lan, X., Zheng, M., et al.: Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In: European Conference on Computer Vision (ECCV). pp. 56–73. Springer (2024)
35. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023)
36. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (volume 1: long papers). pp. 13484–13508 (2023)
37. Xing, F., Xie, Y., Su, H., Liu, F., Yang, L.: Deep learning in microscopy image analysis: A survey. IEEE Transactions on Neural Networks and Learning Systems **29**(10), 4550–4568 (2017)
38. Ye, Z.W., Hsueh, H.W., Hsu, S.H.: Diversity-based two-phase pruning strategy for maximizing image segmentation generalization with applications in transmission electron microscopy. Machine Learning: Science and Technology **6**(2), 025021 (2025)
39. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9556–9567 (2024)
40. Zhang, J., Wang, T., Zhang, H., Lu, P., Zheng, F.: Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In: European Conference on Computer Vision (ECCV). pp. 196–213. Springer (2024)
41. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
42. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)