SEM-VQA: STRUCTURED VISUAL QUESTION ANSWERING FOR SCIENTIFIC UNDERSTANDING OF SCANNING ELECTRON MICROSCOPE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding scanning electron microscope (SEM) images is central to materials science, yet progress on applying multimodal large language models (MLLMs) to this domain is limited by the scarcity of large-scale, open-ended visual question answering (VQA) datasets grounded in real scientific imagery. We introduce SEM-VQA, a pipeline for constructing and evaluating SEM-focused VQA data that enforces image-grounded reasoning. Using LLM-assisted annotation and prompting constraints requiring answers to be derivable from the image alone, we build a corpus of 224,026 question-answer-visual-evidence triplets over 47,981 SEM micrographs, organised into four levels of visual understanding: observation, detection, identification, and interpretation. We construct a held-out benchmark of 300 images via embedding-based clustering and use it to evaluate open-source and proprietary MLLMs under an LLM-as-judge protocol. Fine-tuning a compact 2B-parameter Qwen3.5 model on our corpus raises its judged score from 3.33 to 3.83 (5-point scale), a 14.9% relative gain that closes roughly 76% of the gap to GPT-5.4-mini (3.98). The fine-tuned model significantly outperforms Claude-Haiku-4.5 and, by a narrower margin, Gemini-3.1-Flash, though a modest but significant deficit remains relative to GPT-5.4-mini; it also leads all evaluated systems across every automatic metric (token-F1, BERTScore-F1, STS, ROUGE-L, METEOR, BLEU). These results show that targeted, structured dataset construction, rather than model scale alone, can substantially close the gap between small open models and large proprietary systems on specialised scientific visual reasoning. We release our dataset, benchmark, and evaluation protocol.

1 INTRODUCTION

Scanning electron microscopy (SEM) is a primary tool for examining the microstructure of materials. Features such as grain morphology, porosity, crack, phase-related structures, and fracture morphology are interpreted directly from micrographs and are central to materials characterization and interpreting them requires domain knowledge. For a multimodal model, however, understanding an SEM image is not simply an object-recognition problem: useful interpretation requires recognizing visible features, localizing them, assigning domain-specific terminology when justified, and reasoning about their microstructural significance. This raises a practical question for scientific multimodal learning: can current vision-language models reliably answer open-ended (free-form) questions about real SEM micrographs, and what data and evaluation methodology are needed to measure and improve such capability? A major obstacle is the lack of large-scale, materials-specific VQA data grounded in real micrographs. Existing scientific multimodal datasets provide substantial scale, but are not specialized to materials characterization or SEM imagery, while materials-focused resources are considerably smaller and are predominantly designed around constrained evaluation formats. This limits both supervised adaptation and systematic evaluation of models on the fine-grained visual concepts that arise in microstructural analysis. In particular, a useful SEM-VQA resource must capture more than whether a feature is present: it should support questions ranging from broad visual description and feature recognition to technical identification and microstructural interpretation, while restricting unsupported scientific claims.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce **SEM-VQA**, a large-scale corpus for open-ended question answering over SEM micrographs. The corpus contains 224,026 question–answer–visual-evidence triplets spanning 47,981 SEM micrographs drawn from the peer-reviewed materials literature. Question–answer pairs are generated using a single prompt specification appendix A, with `gemini-2.5-flash` as the generation model. The prompt organizes questions into a four-level sec 3.2.2, prompt-defined progression of visual understanding—*Observation*, *Detection*, *Identification*, and *Interpretation*—and requires question sets to be supported by the image, vary across features or aspects of the microstructure, and avoid unsupported or redundant content. Importantly, the generator receives the micrograph together with its source caption and summary. We therefore do not treat the resulting questions as inherently image-only; instead, the prompt explicitly requires image-grounded answerability while using the accompanying text as contextual information during generation.

We further construct **rep300** sec 4.1, a held-out benchmark containing 300 representative SEM images and 1,497 question–answer pairs. The images are selected using embedding-based clustering to cover the corpus's visual range and are held out from the training, validation, and test partitions. We evaluate eleven multimodal large language models, spanning open-weight and proprietary systems, under a common free-form answering and scoring protocol. Because reference-based metrics can be sensitive to lexical overlap and output length, we report six automatic metrics alongside a blind LLM-as-judge evaluation. The evaluation is therefore interpreted using both reference-based similarity measures and a judge designed to assess substantive agreement with the reference answer.

Finally, we investigate whether domain-specific supervision can improve the performance of a compact multimodal model. Fine-tuning Qwen3.5-2B on SEM-VQA increases its judged score from 3.33 to 3.83 on the five-point scale, corresponding to a 14.9% relative improvement. The result demonstrates that targeted SEM-specific supervision can substantially improve scientific visual question answering in a compact open-weight model, motivating the use of structured domain data alongside model-scale comparisons.

In summary, this paper makes the following contributions:

- We introduce **SEM-VQA**, a large-scale free-form VQA corpus containing 224,026 question–answer–visual-evidence triplets over 47,981 SEM micrographs.
- We define a **four-level, prompt-defined taxonomy** of visual understanding (Observation, Detection, Identification, and Interpretation), together with grounding and anti-hallucination constraints that regulate technical terminology and unsupported inference.
- We construct **rep300**, a held-out benchmark of 300 representative SEM images and 1,497 question–answer pairs, and evaluate eleven open-weight and proprietary multimodal models using a common free-form evaluation protocol.
- We develop an evaluation framework combining reference-based automatic metrics with blind LLM-based judging, and show that fine-tuning a compact Qwen3.5-2B model on SEM-VQA substantially improves its judged performance.

2 Related Work

Existing VQA benchmarks for scientific imagery fall into three groups. Biological-microscopy benchmarks such as MicroVQA (Burgess et al., 2025) (1,042 questions), its scaled successor MicroVQA++ (Li et al., 2025) (up to 26,000 questions with chain-of-thought rationales), and SPUR (Ding et al., 2026) (4,264 questions on immunofluorescence and blot imagery) are all multiple-choice and outside the materials domain.

Materials-science benchmarks are closer in domain but remain small and almost universally multiple-choice: MatCha (Lai et al., 2025) (1,500 questions across SEM/TEM/XRD/AFM), MatVQA (Wu et al.) (1,325 questions), MatQnA (Weng et al., 2025) (4,968 questions across ten characterization techniques, only 811 SEM-derived, and only its multiple-choice half is scored), and MaCBench (Alampara et al., 2025) (1,155 multiple-choice/numeric questions across chemistry and materials). CSMBench (Zheng et al., 2026) pairs 1,041 figures with both an open-ended description task and multiple-choice, but reports no combined item count. The closest prior work is MATRIX (McGrath et al., 2026), the only other free-form, rubric-judged benchmark on

SEM/XRD/EDS/TGA data; its paper reports two irreconcilable evaluation-set sizes (470 in Figure 1 vs. 300 in Appendix F.2), so we cite both rather than choosing one.

Large-scale general-science VQA resources — MMSci (Li et al., 2024) (1.08M questions, 72 disciplines), SciGraphQA (Li & Tajbakhsh, 2023) and SPIQA (Pramanick et al., 2024) (both CS-paper charts/plots), and ViTAR (Chen et al., 2025) (medical-imaging-derived) — are far larger but not aligned with our domain.

SEM-VQA is, to our knowledge, the largest materials-science VQA benchmark to date and the only one offering free-form question–answer pairs at scale for SEM micrographs: 47,981 images and 224,026 QA pairs spanning metals, ceramics, composites, and polymers.

Materials- and microscopy-specific VQA benchmarks. A cluster of recent benchmarks target visual question answering over materials-characterization imagery, and are uniformly small, expert-curated, and evaluation-only. MicroVQA (Burgess et al., 2025) is 1,042 multiple-choice questions over 255 biological-microscopy image sets, built from $\sim$600 expert-hours of raw QA writing followed by an agentic shortcut-removal pass (RefineBot) that drops model accuracy by 35–42 points relative once naive language shortcuts are closed. MatCha (Lai et al., 2025) is the closest domain match to our corpus — 1,500 MCQs over 1,260 SEM/TEM/XRD/AFM figures organized along a four-stage processing$\rightarrow$morphology$\rightarrow$structure$\rightarrow$ property chain, with model accuracy degrading monotonically along that chain.MatVQA (Wu et al.) (1,325 MCQs over 378 images) and MaCBench (Alampara et al., 2025) (1,155 questions across nine characterization tasks, per its own Appendix Table A.1 total) report similarly small, heavily shortcut-audited benchmarks; MaCBench's ablations show models score higher when the same information is given as text rather than as an image, implicating visual encoding, not domain knowledge, as the bottleneck. MatQnA (Weng et al., 2025) is the largest of this group (4,968 QA across ten characterization techniques, SEM: 811); its raw data includes 2,749 free-form (subjective) items alongside 2,219 MCQs, but the paper reports results only on the MCQ split, stating that objective questions "provide more standardized evaluation criteria and reduce potential bias from subjective scoring." The free-form items were never scored or evaluated, so in practice, and for the purposes of this comparison, MatQnA functions as a closed-ended (MCQ) benchmark, not a free-form one. CSMBench (Zheng et al., 2026) and MATRIX (McGrath et al., 2026) are the two benchmarks in this group that use long-form, judge-scored answers rather than MCQ, and are accordingly the closest methodological precedents for our evaluation protocol MATRIX is additionally the only one of these benchmarks that closes the loop from benchmark to post-training, fine-tuning on roughly 3,056 aligned SEM image–text pairs and reporting the largest gains on SEM specifically. Across this group, scale tops out at 4,968 QA pairs and 1,260 images; ours is 224,026 QA pairs over 47,981 SEM micrographs, and is a training corpus rather than an evaluation-only benchmark.

Large auto-generated scientific VQA corpora. Outside the materials-characterization literature, several corpora demonstrate that a single strong model can generate hundreds of thousands of scientific QA pairs without per-item human authorship. SciGraphQA (Li & Tajbakhsh, 2023) generates 657,000 multi-turn QA pairs over 295,000 scientific-graph figures by prompting a text-only model with the figure's caption and surrounding paper text (the figure image itself is not shown to the generator), then validates a 3,000-item sample with an LLM judge on a 0–10 scale (86% $\geq$ 8.5). MMSci (Li et al., 2024) is the largest corpus in this space (742,273 figures, over one million training instances across captioning and figure–caption matching tasks) and demonstrates the clearest downstream transfer result available: a 7B model fine-tuned on its auto-generated data outperforms GPT-4o and PhD experts on its own benchmark, and separately improves a non-VQA materials task (crystal-structure generation) after visual pretraining. SPIQA (Pramanick et al., 2024) is the closest precedent for our exact format — 270,194 free-form (non-MCQ) QA pairs over scientific-paper figures, generated by a single model from the figure, caption, and surrounding text, validated with a pilot study rather than exhaustive human review, and evaluated with a purpose-built continuous judge metric (L3Score). SPUR (Ding et al., 2026) is methodologically the strictest of this group: its generator is shown *only* the figure panel, with captions and body text explicitly withheld, and every candidate question is additionally required to survive a blind text-only filter (10 attempts, discard if ≥ 5 are answered correctly) before expert review. Our corpus sits inside this cluster by construction method and scale, and differs from SPIQA and MMSci mainly in domain (SEM micrographs rather

than paper figures generally) and from SPUR in one specific respect: our generator is given the image together with its source caption and a summary (Section 3.2.3), not the image alone.

Why a large corpus, and why free-form. No existing resource combines the three properties this corpus has together. The large auto-generated corpora (SciGraphQA, MMSci, SPIQA, SPUR) are not materials-specific — they cover CS-paper graphs, cross-disciplinary figures, and biological micrographs, respectively, not SEM or materials characterization. The materials- and microscopy-specific benchmarks (MicroVQA, MatCha, MatVQA, MaCBench, MatQnA, CSMBench), conversely, are uniformly small (at most a few thousand items) *and* scored, where scored at all, as multiple-choice: MatQnA is the only one that even constructs free-form items, and it neither scores nor reports on them, so in practice every materials-specific benchmark in this literature is closed-ended. There is accordingly no existing large, free-form question-answering corpus for a materials-characterization modality; this is the specific gap our corpus fills. MicroVQA++ (Li et al., 2025) motivates both choices directly. Its authors identify "the lack of high-quality multimodal microscopy datasets" as the key bottleneck limiting scientific reasoning in MLLMs, and show that careful large-scale data construction lets a 4B-parameter open model match GPT-5 on their benchmark after fine-tuning the argument for scale. For format, their own training ablation (Table 6) finds that SFT on free-form QA scores markedly lower than SFT on the same content converted to MCQ when evaluated on an MCQ benchmark (42.4 vs. 59.4 average accuracy); they attribute this gap in their own words to what each format teaches, not to one being invalid: "QA-only supervision improved factual grounding and descriptive precision but did not fully train option discrimination." Read against CSMBench's Observation 4 that MCQ accuracy and open-ended description quality dissociate within the same model, this is the case for free-form supervision as a complementary target, not a strictly inferior one: MCQ training teaches option discrimination, which our format does not test or train, while free-form supervision and evaluation teach and test factual grounding and articulation, which an MCQ arm cannot measure at all.

Positioning. Two properties separate our corpus from every benchmark above. First, scale: at 224,026 QA pairs it is roughly 45$\times$ the largest materials-specific benchmark (MatQnA) and within the same order of magnitude as the largest general-science corpora (SciGraphQA, MMSci, SPIQA), while being, to our knowledge, the largest QA corpus built on a single microscopy modality. Second, purpose: MicroVQA, MatCha, MatVQA, MaCBench, MatQnA, and CSMBench are evaluation-only — none but MATRIX reports fine-tuning on its own data, and MATRIX's post-training set is roughly 70$\times$ smaller than ours. The gap this corpus targets is the one MaCBench's authors state directly: reliable multimodal scientific assistants require advances in *training-data curation*, not evaluation alone.

3 Dataset Construction

3.1 Source micrographs

The source micrographs used in this work are obtained from MatSciFig (Ghosh et al., 2026), a panel-level materials image–text dataset curated from 14,810 open-access (CC-BY/CC-BY-NC) peer-reviewed literature of the materials domain.The dataset contains 391,606 single image-caption-summary triplets along with broad category and sub-category labels.47,981 distinct images labelled "SEM" were retained for VQA construction. Each selected panel is associated with its corresponding figure sub-caption and contextual information generated from the source article as a summary. These image captions and summary provide the scientific context required for constructing question–answer pairs. The resulting collection covers a diverse range of materials systems and microstructural morphologies, providing the basis for developing a domain-specific VQA corpus for understanding SEM images.

3.2 Generation protocol

The question–answer pairs were generated by `gemini-2.5 flash` under a single prompt specification, `v7 _four _level _grounded`, recorded in every record. Each request supplies the micrograph, its caption, and its summary, and requests strict JSON containing up to five pairs, with

no minimum requirement; the number of question images genuinely supported is entirely determined by the generator. Every pair carries a `question`, an `answer`,`level` label (Section 3.2.2), a free-text `feature` label, and a `visual_evidence` string stating what in the image supports the answer. The generator was chosen over five alternatives on a human-rated pilot, and decoding used provider defaults throughout.

The prompt requests *up to* five questions and instructs the model to produce only as many as the image genuinely supports, with an explicit instruction not to pad. The constraint binds only weakly: 68.5% of images received five questions, 30.0% received four, 1.5% received three or fewer, and none received zero. The corpus contains 224,026 pairs over 47,981 images, a mean of 4.67 per image. Schema adherence is high but imperfect — 99.975% of records carry a canonical `level` label.

3.2.1 Question Diversity and Content Relevance

Beyond the question-count constraint, the generation protocol requires diversity and scientific relevance within each image-level question set. Questions must cover different microstructural features or different aspects of the same microstructure rather than repeat the same visual fact in different wording. The prompt illustrates this requirement through complementary questions targeting, for example, grain morphology, visible defects, and phase contrast. It further defines an explicit redundancy criterion: "Two questions that teach the same visual fact are one question; drop the weaker."

The protocol also prioritises scientifically meaningful questions about microstructural features over superficial image-level or administrative questions. Questions concerning image metadata or presentation elements, such as "is a scale bar present?" or "is there text?", are discouraged to keep the generated question set focused on domain-relevant visual content.

3.2.2 Four-Level Visual Understanding Taxonomy

The question–answer pairs were generated under a four-level taxonomy of progressive visual understanding Observation $\rightarrow$ Detection $\rightarrow$ Identification $\rightarrow$ Interpretation designed to reflect the progression from broad visual description to feature recognition, technical identification, and microstructural interpretation.

Observation. Provides a single broad summary of the most salient visible features, with at most one observation question per image. **Detection.** Identifies which features are present, their appearance, and location. Open, discovery-style questions are preferred over closed confirmation questions. **Identification.** Assigns technical identities to visible features. Technical terminology is permitted only when corroborated by the caption or summary (Section 3.2.3). **Interpretation.** Explains the significance, morphology, distribution, or relationships of features, while avoiding unsupported claims about processing, composition, manufacturing route, or properties. Furthermore, the generation process does not require every image to support all four levels or impose a fixed number of questions at each level. Instead, the generator selects the levels supported by the visible features of the input micrograph and the available contextual information. The 224,026 pairs are distributed across levels as follows: detection 97,409 (43.5%), interpretation 49,922 (22.3%), observation 47,924 (21.4%) and identification 28,744 (12.8%), with 27 pairs (0.01%) carrying non-canonical labels (capitalisation variants are folded into their level). Identification is the smallest because it is the only level gated on external corroboration, and observation's share matches its one-per-image cap given a mean of 4.67 questions per image.

Overall, the four-level taxonomy provides a structured framework for organising questions according to the type of visual understanding they require. It distinguishes direct visual description and feature recognition from technical identification and higher-level interpretation, while incorporating evidence constraints for claims that extend beyond directly observable characteristics. The taxonomy is therefore used as a prompt-defined organisational framework for question generation and analysis, rather than as a validated measure of cognitive difficulty or a requirement that every micrograph exhibit all four levels.

3.2.3 GROUNDING AND ANTI-HALLUCINATION CONSTRAINTS

Each question–answer pair must be grounded in the input micrograph and answerable from the image alone. The caption and summary serve only as contextual references and cannot be quoted or rephrased as substitutes for visual inspection. **Terminology and inference.** Directly visible features such as cracks, grains, pores, particles, and fracture features may be named from visual evidence. Technical terms such as precipitates, twins, dendrites, carbide networks, and specific phases require corroboration from the caption or summary; otherwise, features are described by their observable morphology. Unsupported features, measurements, compositions, counts, or interpretations must not be introduced, and uncertainty must be acknowledged. **Measurement and relational constraints.** Quantitative descriptions avoid false precision and use only coarse ranges supported by a visible scale bar. Paper-added annotations such as arrows, labels, and boxes cannot be assigned technical meaning. Comparison and counting questions are generated only when multiple distinct or countable features are visibly present **Abstention under insufficient evidence.** Finally, the prompt provides an explicit abstention mechanism: when an image is too ambiguous, low-quality, blurry, or out of focus to support any real question, the generator is instructed to return an empty question list rather than fabricate detail to make the image appear clearer than it is.

3.3 BENCHMARKING AND SELECTION OF THE VQA GENERATION MODEL

We selected the VQA generation LLM model through a human-evaluated pilot study involving six candidate LLMs. The generated pairs were evaluated using five dimensions: image grounding, answer validity, answerability from the image, hallucination, and overall QA quality. Grounding assesses whether the answer is supported by the micrograph's visual content. Answer validity concerns the scientific correctness of the response, while answerability assesses whether the answer can be determined from the image. Hallucination captures unsupported or fabricated visual or scientific claims. Overall QA quality provides a holistic assessment of relevance, completeness, and usefulness of the question–answer pair for training and benchmarking of MllMs for the materials domain SEM-VQA

Among the evaluated models, `gemini-2.5-flash` achieved the highest favourable rating across evaluation dimensions. It was also the only model with no hallucinated pairs among the rated outputs and generated the second-most question-answer pairs per image. We therefore selected `gemini-2.5-flash` for large-scale VQA annotation generation. The complete pilot design, comparative results, and schema-compliance analysis are provided in Appendix C.

4 EXPERIMENTS

4.1 REPRESENTATIVE BENCHMARK CONSTRUCTION

From the corpus described in Section 3 224,026 question-answer-visual evidence triplets over 47,981 distinct SEM micrographs — we draw a held-out benchmark, *rep300*, of 300 images and 1,497 question–answer pairs, used for all evaluation reported below.

Rather than sampling uniformly at random, the 300 images were selected to cover the corpus's visual range. DINOv3 CLS-token embeddings were extracted over the panel pool, ℓ_2-normalised and projected to 50 dimensions by PCA, and partitioned by k-means; within each cluster, images were further stratified by distance to the cluster centroid, and an equal quota was drawn from every cluster–band cell. A uniform random sample would reproduce the corpus's own frequency profile, so image types that are rare in the source would appear in the benchmark only in proportion to their rarity. Drawing a fixed quota from each cell instead represents sparsely populated regions of the embedding space on the same terms as dense ones, which is the property we want of an evaluation set. Visual diversity mirrors that of the corpus it is drawn from.

The 300 rep300 images are held out from both the training and test partitions: membership is verified by content hash, and no rep300 image occurs in either split. This is a deliberate design choice made to prevent evaluation leakage, rather than an incidental property of the sampling procedure.

4.2 EVALUATION-METRICS

Evaluation of MCQ-based VQA is straightforward because accuracy is well defined, whereas free-form scoring remains less established. We therefore use two complementary approaches: reference-based automatic metrics and an LLM-as-judge protocol.

Automatic metrics. We compute six reference-based metrics for each prediction: token-F1, BERTScore-F1 (Zhang et al., 2019), ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), BLEU (Papineni et al., 2002), and semantic textual similarity (STS). BERTScore-F1 uses `RoBERTa-large`, while STS is computed as cosine similarity between 384-dimensional `all-MiniLM-L6-v2` embeddings (Reimers & Gurevych, 2019).

LLM-as-judge. We use Gemini-3.5-Flash as a blind judge under the five-point rubric in Section B. System identity, item ordering, and question-level labels are withheld during judging. The judge scores substantive agreement with the reference answer while ignoring differences in length and phrasing. It additionally flags `unsupported_specifics`, defined as specific details not supported by the reference answer.

4.3 BENCHMARK EVALUATION RESULTS

We benchmark eleven MLLMs, three proprietary systems (GPT-5-mini, Gemini-3.1-Flash, Claude-Haiku-4.5), and eight open-weight models spanning 2B–8B parameters across four architectural families — Qwen3VL-2B/4B/8B, InternVL3.5-2B/4B/8B, InternVL3-2B, and Qwen3.5-2B. The families differ in language-model generation as well as scale: InternVL3 is built on Qwen2, InternVL3.5 on Qwen3, and Qwen3.5-2B on the current generation.

4.3.1 PROMPT FORMAT AS AN EVALUATION CONTROL

Every system in Section 4.3 is evaluated with the same structured prompt, which specifies the imaging modality and requests `Evidence: ...` followed by `Answer: ....` Compared with the bare-question format, this prompt improves token-F1, BERTScore-F1, ROUGE-L, and BLEU for all seven open-weight models, with mean gains of $+0.089$, $+0.021$, $+0.082$, and $+0.016$, respectively. It also reduces the prediction-to-reference length ratio from 2.3–7.9 to 0.55–1.92; METEOR is the only metric that decreases consistently (-0.046 on average). We therefore use the structured prompt for all zero-shot evaluations; the three proprietary systems were evaluated only under this format.

4.3.2 ZERO-SHOT COMPARISON ON AUTOMATIC NLP METRICS

Table 1 reports the eleven systems under the structured prompt.Increasing scale from 2B to 8B produces only small changes in the main reference-based metrics (at most 0.013), with no consistent trend across families. Among the ten format-comparable systems, open-weight models have the highest values on four of six metrics, while STS and METEOR are highest for Claude-Haiku-4.5. Given the limited sensitivity of reference-based metrics to scientific answer quality, these results are interpreted descriptively and motivate the complementary blind judge-based evaluation.

4.4 FINE-TUNING FOR MATERIALS SEM IMAGES

The competitiveness of small open-weight models on automatic metrics (Section 4.3.2) motivates fine-tuning one ofthem. We select Qwen3.5-2B as the most recent language backbone in that size class: the InternVL family evaluated here is built on earlier Qwen generations — InternVL3 on Qwen2 and InternVL3.5 on Qwen3 — so training Qwen3.5-2B ensures any gain is attributable to adaptation rather than to a newer backbone. Training proceeds in two stages, both using LoRA adapters with the base weights frozen. In **stage 1** the model is trained to produce the panel summary from the image alone, a captioning objective that adapts it to SEM imagery before any question answering. Two stage-1 variants are trained, differing only in whether the vision encoder is adapted alongside the language model or left frozen. In **stage 2** the model is trained on the VQA corpus, with the vision encoder frozen in every run. To isolate what stage 1 contributes, a third condition skips it entirely and trains a fresh stage-2 adapter directly on the base model. These three initialisations —

Table 1: Zero-shot comparison under the evidence prompt format, grouped by parameter scale. All metrics are computed on the `Answer:` span over the 1,492 (image, question) pairs answered by every system; no fine-tuned system appears here. *Len.* is the ratio of mean prediction length to mean reference length.

System	Params	Token-F1	BERTSc.	STS	ROUGE-L	METEOR	BLEU	Len.
Small open-source								
`Qwen3VL-2B`	2B	0.285	0.875	0.640	0.238	0.204	0.045	1.92
`InternVL3.5-2B`	2B	0.293	0.878	0.616	0.250	0.179	0.043	0.63
`InternVL3-2B`	2B	**0.320**	**0.888**	0.648	**0.270**	0.204	**0.051**	0.71
`Qwen3.5-2B`	2B	0.254	0.865	0.659	0.194	**0.267**	0.029	3.65
Medium open-source								
`Qwen3VL-4B`	4B	0.295	0.878	0.614	0.251	0.172	0.042	0.62
`InternVL3.5-4B`	4B	0.289	0.877	0.606	0.246	0.172	0.042	0.58
Large open-source								
`Qwen3VL-8B`	8B	0.286	0.874	0.601	0.240	0.176	0.040	0.74
`InternVL3.5-8B`	8B	0.301	0.881	0.621	0.257	0.178	0.043	0.55
Proprietary MLLMs								
`GPT-5-mini`	—	0.279	0.869	0.611	0.226	0.144	0.024	0.83
`Gemini-3.1-Flash`	—	0.295	0.877	0.598	0.253	0.172	0.042	0.60
`Claude-Haiku-4.5`	—	0.302	0.876	**0.665**	0.239	0.250	0.042	2.06

fresh adapter (B1), continued stage-1 adapter with vision frozen (B2), and continued stage-1 adapter with vision adapted (B3) — form one axis of the design.

The second axis is the stage-2 supervision target. The image and question presented to the model are identical in both; only the system prompt and the target differ. Under **Format A** the target is the answer alone. Under **Format B** the target is the visual evidence and the answer together, as `Evidence: ...\nAnswer: ...`, so the evidence string recorded during corpus construction (Section 3.2.2) is produced at inference rather than discarded after training. Crossing three initialisations with two targets gives the six cells reported below. Full training configurations are given in Appendix E. Systems are evaluated with the automatic metrics of Section 4.2 and with the LLM judge. A detailed analysis of the 2×3 fine-tuning grid, including the effects of initialization and supervision format and the comparison between automatic metrics and judge scores, is provided in Appendix E.8.

4.5 Automatic metrics and judged agreement

Twelve systems were scored on the held-out benchmark by six automatic metrics and by an LLM judge applying the answer-agreement rubric of Appendix B blind to system identity. The judge returned a usable score for 99.88% of submitted predictions; per-system judged counts range from 1,484 to 1,492. Restricting to the 1,472 items judged for every system changes no system's mean by more than 0.003 and leaves the ranking unchanged.

No automatic metric recovers the judged ranking. Across the twelve systems, correlations with judged agreement are moderate at best: token-F1 $r = 0.54$ ($\rho = 0.49$), ROUGE-L 0.53 (0.46), BLEU 0.53 (0.37), STS 0.49 (0.47), METEOR 0.47 (0.37), BERTScore 0.46 (0.48). These are upper bounds: six systems are Qwen3.5 fine-tunes differing only in stage-2 initialisation and target format (spanning 0.046 judged, 0.0103 token-F1), and these near-replicates inflate the fitted trend. On the nine more diverse systems available before they were added, correlations drop to token-F1 $r = 0.41$ ($\rho = 0.23$) and BERTScore $r = 0.31$ ($\rho = 0.17$). The disagreements are large, not noise: GPT-5-mini ranks **11th of 12 on token-F1 but 1st on the judge**, while zero-shot InternVL3-2B ranks 7th and last. Semantic metrics (BERTScore, STS) do no better than token overlap.

The metrics measure stylistic conformity. All six fine-tuned systems produce answers at 0.93–0.94 of reference length, versus 0.60–3.65 for all others. Fine-tuning installs the references' surface form, which lexical overlap rewards directly. The effect is strongest on STS (0.792–0.797 fine-tuned vs. 0.598–0.665 otherwise), a much wider gap than BERTScore shows (0.919–0.920 vs. 0.865–

Table 2: All metrics on the held-out benchmark, answer-only, ordered by judged agreement. Suffixes *-A* and *-B* denote Format A and Format B; *B1–B3* denote the stage-2 initialisation. *len* is the mean prediction length as a fraction of the reference. Judge score is 1–5.

System	Group	Judge	Flag%	BERTSc.-F1	STS	tok-F1	ROUGE-L	BLEU	len
GPT-5-mini	prop. 0-shot	3.983	45.8	0.8685	0.6112	0.2785	0.2255	0.024	0.83
Qwen3.5-2B SFT B2-B	fine-tuned	3.823	26.4	0.9191	0.7919	0.4979	0.4353	0.157	0.93
Qwen3.5-2B SFT B1-A	fine-tuned	3.816	27.3	0.9200	0.7956	0.5056	0.4426	0.164	0.94
Qwen3.5-2B SFT B2-A	fine-tuned	3.815	26.5	0.9196	0.7954	0.5042	0.4386	0.162	0.94
Qwen3.5-2B SFT B3-A	fine-tuned	3.800	25.2	0.9199	0.7968	0.5076	0.4435	0.166	0.94
Qwen3.5-2B SFT B1-B	fine-tuned	3.789	25.7	0.9189	0.7916	0.4973	0.4343	0.156	0.93
Qwen3.5-2B SFT B3-B	fine-tuned	3.777	26.3	0.9189	0.7918	0.4976	0.4356	0.159	0.93
Gemini-3.1-Flash	prop. 0-shot	3.749	30.6	0.8770	0.5984	0.2948	0.2532	0.042	0.60
Claude-Haiku-4.5	prop. 0-shot	3.473	67.0	0.8764	0.6646	0.3018	0.2390	0.042	2.06
Qwen3.5-2B	open 0-shot	3.330	86.3	0.8646	0.6589	0.2541	0.1944	0.029	3.65
InternVL3.5-2B	open 0-shot	3.055	35.9	0.8782	0.6157	0.2931	0.2498	0.043	0.63
InternVL3-2B	open 0-shot	2.868	31.9	0.8877	0.6485	0.3196	0.2702	0.051	0.71

0.888). Length alone does not explain the inversion: GPT-5-mini (23.8 words) and InternVL3-2B (19.9 words) are similarly short yet sit at opposite ends of both rankings. We therefore make no claims from automatic metrics alone, since on this corpus they largely reward resemblance to the reference distribution, which fine-tuning confers by construction. **Judge family bias.** The judge and one system are both Gemini models. Gemini-3.1-Flash moves only $+2$ places from token-F1 to judged ranking, whereas GPT-5-mini moves $+11$ to first, so we find no detectable self-preference. **Limitations.** The 1,492 items come from 300 images, so the effective sample is closer to 300. We used a single judge; rubric variants were validated (Appendix B), but not across judge families at this scale. The judge sees no images and scores agreement with the reference, so a correct answer that contradicts a flawed reference is counted as disagreement.

5 Conclusion

We introduced a corpus of 224,026 question–answer–visual-evidence triplets over 47,981 SEM micrographs, spanning four levels of visual understanding (observation, detection, identification, interpretation) and generated under constraints requiring answers to be derivable from the image alone. From it we held out a 300-image, 1,497-pair benchmark, stratified over a DINOv3 embedding to span the corpus's visual range, and evaluated twelve systems with six automatic metrics and a blind LLM judge (`gemini-3.5-flash`).

Structured data, rather than scale, closes most of the gap between a small open model and proprietary systems. Fine-tuning a 2B Qwen3.5 raises its judged score from 3.33 to 3.82 ($+14.9\%$), closing $\sim$76% of the gap to GPT-5-mini (3.98). It surpasses Claude-Haiku-4.5 ($t = +8.39$) and Gemini-3.1-Flash ($t = +2.03$) but remains below GPT-5-mini ($t = -4.02$). Supervising visual evidence alongside answers costs nothing measurable ($|t| \leq 1.1$) while producing grounded evidence, and continuing a stage-1 adapter gives no benefit over training from the backbone.

Methodologically, reference-overlap metrics do not recover the judged ranking. Token-F1 correlates with the judge at only $r = 0.54$ (0.41 excluding near-replicate fine-tunes), and GPT-5-mini ranks 11th on token-F1 but 1st on the judge. The cause is stylistic conformity: fine-tuned systems match reference length (0.93–0.94) and phrasing, which overlap metrics reward directly. Even within the fine-tuning grid, the highest token-F1 cell is not the judge's top choice. We caution against ranking scientific VQA systems on such metrics alone.

Limitations and release. Results rest on a single 300-image benchmark, so narrow margins are provisional. We used one judge, from the same vendor family as one evaluated system, and it scores agreement with the reference without seeing the image. Because questions are generated from captions, image-only answerability is a design goal rather than a guarantee. Multi-judge protocols and expert human evaluation are natural next steps. We release the corpus, benchmark, evaluation protocol, per-system predictions, and all eight LoRA adapters (two stage-1, six stage-2), allowing every fine-tuned system to be reproduced without retraining.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript for language editing, including grammar correction, phrasing, and improving readability. AI tools were also used for code debugging and troubleshooting during the implementation of the computational experiments. All scientific content, methodological decisions, experimental design, results, interpretations, and final conclusions were reviewed and verified by the authors.

ETHICS STATEMENT

This work involves no human or animal subjects and no collection of personal or sensitive information. The SEM micrographs are derived from open-access peer-reviewed materials literature and are used subject to their source licensing and attribution requirements. Automatically generated annotations may contain errors or unsupported interpretations; we therefore apply explicit grounding and anti-hallucination constraints and recommend expert review for scientific use. The dataset may also reflect biases present in the underlying published literature and imaging practices. .

REPRODUCIBILITY STATEMENT

The SEM-VQA corpus (224,026 QA pairs over 16,192 SEM images) and its held-out SEM-VQA-Bench evaluation subset are available at `https://huggingface.co/datasets/uhbuvbuvu/sem-vqa` and `https://huggingface.co/datasets/uhbuvbuvu/sem-vqa-bench` (anonymized for review). Code, training configurations, and the fine-tuned adapters described in Section **??** are available at `https://anonymous.4open.science/r/SEMVQA-286D` (anonymized for review). Both datasets are released under CC BY 4.0. All links will be de-anonymized upon acceptance.

ACKNOWLEDGMENTS

We Acknowledge Lamda AI for compute resources. Department of Material and metallurgy IIT ROORKEE Department o .

REFERENCES

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, 5(10):952–961, 2025.

Satanjeev Banerjee and Alon Lavie. 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.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19553–19564. IEEE, 2025.

Kaitao Chen, Shaohao Rui, Yankai Jiang, Jiamin Wu, Qihao Zheng, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Think twice to see more: Iterative visual reasoning in medical vlms. *arXiv preprint arXiv:2510.10052*, 2025.

Junpeng Ding, Zichen Tang, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, et al. Decoding scientific experimental images: The spur benchmark for perception, understanding, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12859–12882, 2026.

Subham Ghosh, Shubham Tiwari, Mohammad Ibrahim, and Abhishek Tewari. Unlocking the visual record of materials science: A large-scale multimodal dataset from scientific literature. *arXiv preprint arXiv:2606.29667*, 2026.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jingpu Yang, Hong-Qing Liang, Yan Hu, and Benyou Wang. Can multimodal LLMs see materials clearly? a multimodal benchmark on materials characterization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4384–4404, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.235. URL https://aclanthology.org/2025.findings-emnlp.235/.

Manyu Li, Ruian He, Chenxi Ma, Weimin Tan, and Bo Yan. Microvqa++: High-quality microscopy reasoning dataset with weakly supervised graphs for multimodal large language model. *arXiv preprint arXiv:2511.11407*, 2025.

Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Delia McGrath, Curtis Chong, Rohil Kulkarni, Gerbrand Ceder, and Adeesh Kolluru. Matrix: A multimodal benchmark and post-training framework for materials science. *arXiv preprint arXiv:2602.00376*, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Yonghao Weng, Liqiang Gao, Linwu Zhu, and Jian Huang. Matqna: A benchmark dataset for multi-modal large language models in materials characterization and analysis. *arXiv preprint arXiv:2509.11335*, 2025.

Sifan Wu, Huan Zhang, Yizhan Li, Farshid Effaty, Amirreza Ataei, and Bang Liu. Seeing beyond words: Matvqa for challenging visual-scientific reasoning in materials science, 2025. *URL https://arxiv. org/abs/2505.18319*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuting Zheng, Zijian Chen, and Qi Jia. Benchmarking cross-scale perception ability of large multimodal models in material science. *arXiv preprint arXiv:2603.19327*, 2026.

A SEM-VQA GENERATION PROMPT

```
PROMPT_TEMPLATE = """You are an expert Metallurgist and Materials
Scientist annotating VQA training data that will teach a
vision-language model to read SEM microstructure the way you do.
The model will only ever see the image at inference -- never this
caption or summary -- so every question must be answerable from
the image alone. It's fine if your answer agrees with the
caption/summary (they describe what's visible, so agreement is a
grounding signal) -- what's NOT allowed is phrasing a question so
it can be answered by quoting or rephrasing the caption/summary
text without looking at the image.

## GOAL
Teach the model to recognize and reason about the microstructural
features actually present in this image. Prefer scientifically
meaningful questions (about real microstructural features --
cracks, pores, grains, boundaries, particles, precipitates,
fracture morphology, phases, surface texture) over superficial
administrative questions (is a scale bar present? is there text?).
Generate only questions that are strongly supported by the image
-- if an image genuinely supports only one simple question,
generate only that one.

## HOW MANY QUESTIONS
Decide how many this image genuinely supports, up to 5. There is
no minimum. Do NOT pad. The questions on one image should VARY by
covering DIFFERENT features or DIFFERENT aspects of the
microstructure (e.g. one on grain morphology, one on a visible
defect, one on phase contrast) -- not the same feature re-asked in
different words. Two questions that teach the same visual fact are
one question; drop the weaker.

## THE FOUR LEVELS (use whichever the image supports; they build
on each other)
1. OBSERVATION -- a broad "what do you see / describe the
microstructure" summary. At most one. A good observation answer
summarizes the most salient VISIBLE microstructural features in
decreasing order of importance. Mention only features that are
actually visible. Do NOT pad the answer with generic statements
like "this is an SEM image" or "the image shows a microstructure."
2. DETECTION -- identify what features are present, what they look
like, and where they are. PREFER open, discovery-style questions
that make the answerer RECOGNIZE the feature rather than confirm a
named one:
    - good: "What defects or discontinuities are visible in the
    microstructure?"
    - good: "What features are present along the grain
    boundaries?"
    - good (detail of a visibly present feature): "What is the
    morphology of the particles?", "Which region contains the most
    porosity?", "Which area shows rougher surface texture?"
```

```
    - avoid: "Are there pores?", "Is a precipitate visible?" --
    these name the answer inside the question, so the model only
    has to confirm it, not recognize it. A closed "Is X present?"
    question is allowed ONLY when the answer is genuinely in doubt
    (the feature could plausibly be absent) -- never ask it about
    a feature that obviously dominates the image, since that makes
    the question rhetorical and every such answer is
    "yes."Localization (where / which region) is an important
    microscopy skill -- use it where the image supports it.
3. IDENTIFICATION -- naming what a feature IS (e.g. "the dark
network is a carbide network", "this is the alpha phase"). Only
name a technical constituent if the caption/summary corroborates
it (see grounding rule). If the caption/summary does NOT confirm
the technical term, describe the feature by its appearance only
(shape, contrast, arrangement) and do NOT state a technical name
-- naming an unconfirmed constituent as fact is hallucination.
4. INTERPRETATION -- reasoning about a feature's significance or
behavior. This is the most error-prone level, so it is
constrained:
    - Interpretation MAY cover: fracture mode (brittle vs
    ductile), feature morphology, phase continuity/connectivity,
    particle or defect distribution (uniform vs clustered), defect
    severity, and spatial relationships between features.
    - Interpretation MUST NOT infer processing history (heat
    treatment, cooling rate), manufacturing / deposition method
    (e.g. "typical of thermal-sprayed coatings"), composition, or
    mechanical properties (strength, hardness, toughness) -- these
    are NOT visible in an SEM image, and stating them (including
    via "typical of..." phrasing) is hallucination even if it
    sounds plausible.
    - If the image alone cannot fully confirm an interpretation,
    say so explicitly in the answer.

## GROUNDING RULE (direct vs technical)
    - Directly visible features -- crack, grain, grain boundary,
    pore, void, particle, agglomerate, surface roughness, fracture
    surface, dimple, cleavage facet, corrosion pit -- need only
    visual evidence. Name them freely when you see them.
    - Technical identifications -- carbide network, twin,
    dendrite, precipitate, secondary phase, phase boundary,
    lamellae, inclusion, specific named phases -- require the
    caption/summary to confirm the term before you state it as
    fact. If unconfirmed, describe the feature by its appearance
    only and do NOT state the technical name -- do not launder a
    caption term into a confident visual claim.

## STRICT ANTI-HALLUCINATION RULES
    - Base every answer ONLY on what is actually visible. Never
    invent features, phases, grains, compositions, measurements,
    counts, or defects that are not clearly visible.
    - Never state something as fact if you are inferring or
    guessing it. Prefer qualitative, directly-observable claims
    ("the surface shows a porous, granular texture") over
    precise-sounding but unverifiable ones ("47% porosity").
```

```
    - Prefer qualitative size language ("a few micrometers", "tens
    of micrometers", "sub-micron"). Give a numeric size ONLY as a
    coarse range you can justify by directly laying the scale bar
    against the feature, and say you used the scale bar -- never
    state a narrow or single-value measurement (e.g. "2.7 um"),
    since that implies a precision the image cannot support.
    - Do NOT assign technical meaning to paper-added annotations
    (arrows, dashed/colored lines, labels, boxes). You may note
    that a mark is present, but do not state what it "represents"
    or "indicates" (e.g. "the dashed line marks a prior grain
    boundary") -- that meaning comes from the caption, not the
    image, and the trained model will never see the caption.
    - Do NOT generate a comparison question unless the image
    actually contains multiple distinct objects or regions to
    compare. Do NOT generate a counting question unless multiple
    countable objects are actually visible.
    - Agreement between your answer and the caption/summary is a
    GOOD grounding signal, not a shortcut -- but never phrase a
    question so it is answerable from the caption/summary text
    alone.
    - If the image is too ambiguous, low-quality, blurry, or out
    of focus to support any real question, return an empty list.
    Never fabricate detail to make the image seem clearer than it
    is.

## CONTEXT (reference only -- never quote it back as the answer)
Caption: {subcaption}
Summary: {summary}

## OUTPUT (strict JSON, no markdown fences)
{{"questions": [
  {{"question": "...", "answer": "...",
    "level": "observation | detection | identification |
    interpretation",
    "feature": "the specific feature this question is about, or
    \\"none\\" for broad observation",
    "visual_evidence": "one sentence naming the visible evidence
    that supports the answer"}}
]}}
If nothing in the image supports a real question, return
{{"questions": []}}.
"""
```

B JUDGE PROMPT FOR COMPARISON OF ANSWER

```
You are an expert materials scientist grading an AI's answer to a
microscopy image question against a reference answer. You do not
have the image -- judge only whether the predicted answer is
factually/semantically consistent with the reference answer and
conveys the same scientific information.
Judge meaning, not presentation:
    - Ignore differences in phrasing, terminology and sentence
    structure.
    - Ignore LENGTH. A one-sentence answer and a paragraph
    conveying the same scientific findings receive the same score.
    Don't reward elaboration.

Question: {question}
Reference answer: {answer}
```

```
Predicted answer: {pred}

Score the prediction 1-5:
5 = Equivalent -- conveys all the substantive content of the
reference, no contradictions, no unsupported specifics.
4 = Substantially equivalent -- captures the main finding; omits a
secondary detail or is slightly less specific. No contradictions.
3 = Partially equivalent -- captures some of the reference's
content but misses a substantive part, or hedges where the
reference is definite.
2 = Largely disagrees -- identifies a different primary feature,
phase or mechanism, or the central claim conflicts with the
reference, though some minor content overlaps.
1 = Contradicts or unrelated -- the central claim is incompatible
with the reference, or the answer does not address the question.

SEPARATELY, check for unsupported specifics: does the prediction
state any specific detail (a number, angle, element name,
direction of change, named phase, defect type, location) that is
NOT supported by the reference answer AND is not a reasonable,
verifiable elaboration of it? Flag this independent of the score
above.

unsupported_detail must contain ONLY specifics the prediction ADDS
that the reference neither states nor contradicts. Do NOT record
omissions or missing details, and do NOT record details that
CONTRADICT the reference -- both are already reflected in the 1-5
score. If the prediction's specifics are all either present in the
reference or in conflict with it, set
unsupported_specifics to false.

Output strict JSON only:
{"score": <1-5>, "verdict": "Equivalent|Substantially
equivalent|Partially equivalent|Largely disagrees|Contradicts or
unrelated",
 "unsupported_specifics": true|false, "unsupported_detail":
 "<specific
 unsupported claim the prediction ADDS, or empty string>",
 "justification": "<1-2 sentences>"}
```

C Benchmarking and Selection of the VQA Generation Model

The VQA generation model was selected via a human-evaluated pilot rather than on capability or throughput alone, aiming to identify a model producing QA pairs that are visually grounded, scientifically valid, answerable from the image, and minimally prone to unsupported claims. Six candidate models generated QA pairs for a shared pool of 35 SEM images using A.Five models produced questions for all 35 images; `deepseek-v4-flash` returned an empty list for one (`img6644_A.jpg`), which the prompt permits when an image cannot support grounded questions, so this was treated as a valid response rather than a failure.

Table 3 gives the rubric exactly as it was put to the annotator: the question asked of each question–answer pair, the scale offered, and the response we count as favourable in Table 4. Three of the five dimensions use a three-point scale, and we count only the top point — `average` and `partially` are counted as *not* favourable. This is a deliberately strict reading, and it is what gives the comparison its resolution; we return to the consequences below.

For a controlled comparison, the analysis was restricted to the 34 images common to all six models. This ensured that each candidate was evaluated on identical image content. The resulting compari-

Table 3: The rating rubric as presented to the annotator. Each question–answer pair was scored on all five dimensions. The final column gives the response counted as favourable in Table 4; the abbreviated column headings there correspond to the dimensions in the same order.

Dimension	Question asked of each pair	Scale	Favourable
Image grounding	Is the answer supported by visual evidence in the SEM image?	Poor / Average / Good	Good
Scientific validity	Is the answer scientifically correct and appropriate?	Poor / Average / Good	Good
Answerability	Can the question be answered from the SEM image?	Yes / Partially / No	Yes
Hallucination	Does the answer contain unsupported or fabricated claims?	Yes / No	No
QA quality	Overall, how relevant, clear, complete and useful is the answer?	Poor / Average / Good	Good

Table 4: Six-model VQA generation pilot evaluated on the 34 images common to all six models. The table reports the number of generated question–answer pairs (n), the percentage of pairs receiving a favourable rating on each evaluation dimension, and the mean number of questions generated per image. All pairs in the comparison were rated on all five dimensions.

Model	n	Grounded	Valid	Answerable	Halluc.-free	QA quality	Q/image
`gemini-2.5-flash`	151	98.0%	97.4%	99.3%	100.0%	95.4%	4.44
`gemini-3.5-flash-lite`	96	95.8%	94.8%	94.8%	99.0%	86.5%	2.82
`gemini-3.1-flash-lite`	114	95.6%	95.6%	99.1%	99.1%	86.0%	3.35
`gpt-5.4-mini`	125	92.8%	88.0%	96.8%	95.2%	67.2%	3.68
`deepseek-v4-flash`	137	91.2%	92.7%	94.9%	96.4%	75.9%	4.03
`mistral-large-3`	168	82.1%	79.2%	85.7%	93.5%	67.3%	4.94

son contained 791 rated question–answer pairs. All models were evaluated on the same 34 images, but each generated its own questions, differing in type, number, and difficulty. So this isn't a strictly paired comparison — reported percentages reflect each model's output quality under pilot conditions, not performance on identical items. Results should be treated as descriptive evidence for model selection, not definitive relative rankings.

`gemini-2.5-flash` achieved the highest favourable-rating percentage across all five evaluation dimensions within its model-specific generated outputs: 98.0% for grounding, 97.4% for validity, 99.3% for answerability, 100.0% for hallucination-free responses, and 95.4% for overall QA quality. The largest margin over the second-highest model was observed for overall QA quality (8.9 percentage points), followed by grounding (2.2), validity (1.8), and answerability (0.2).

These results motivated the selection of `gemini-2.5-flash` for large-scale annotation generation under the evaluated prompt and sampling conditions. Although `mistral-large-3` produced the highest mean number of questions per image (4.94), it obtained lower favourable-rating percentages for grounding, validity, answerability, and hallucination-free responses. Thus, within this pilot, generation volume alone was not a reliable proxy for annotation quality.

Because the models generated different question–answer pairs and different numbers of questions per image, the reported percentages represent model-conditional quality estimates rather than strictly paired comparisons. The results are descriptive rather than statistically conclusive, given the limited image pool, model-specific output distributions, and single-annotator evaluation.

The selected model generated 4.44 questions per image during the pilot and 4.67 during full-scale generation, an absolute difference of 0.23 questions per image. This suggests broadly similar generation volume, but does not establish comparable annotation quality, question distributions, or feature coverage across the two settings.

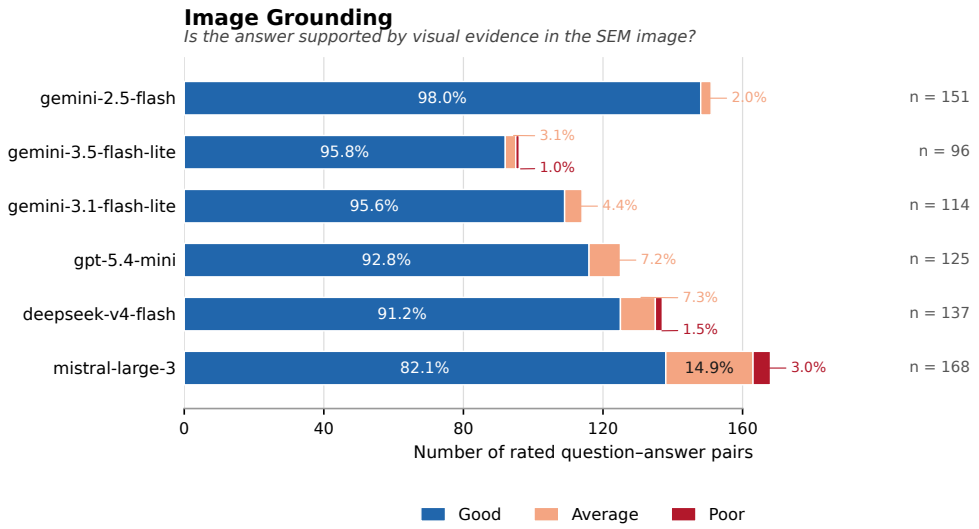

Figure 1: Generator pilot, image grounding: is the answer supported by visual evidence in the SEM image? Bars show the distribution of the annotator's Poor/Average/Good ratings for each candidate model; bar length is the number of rated pairs, which differs by model.

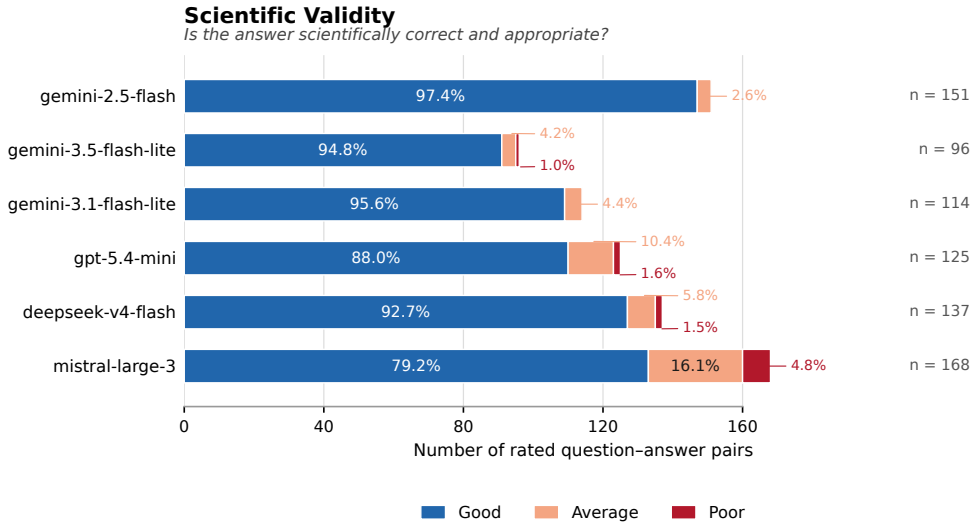

Figure 2: Generator pilot, scientific validity: is the answer scientifically correct and appropriate? Scale and layout as in Figure 1.

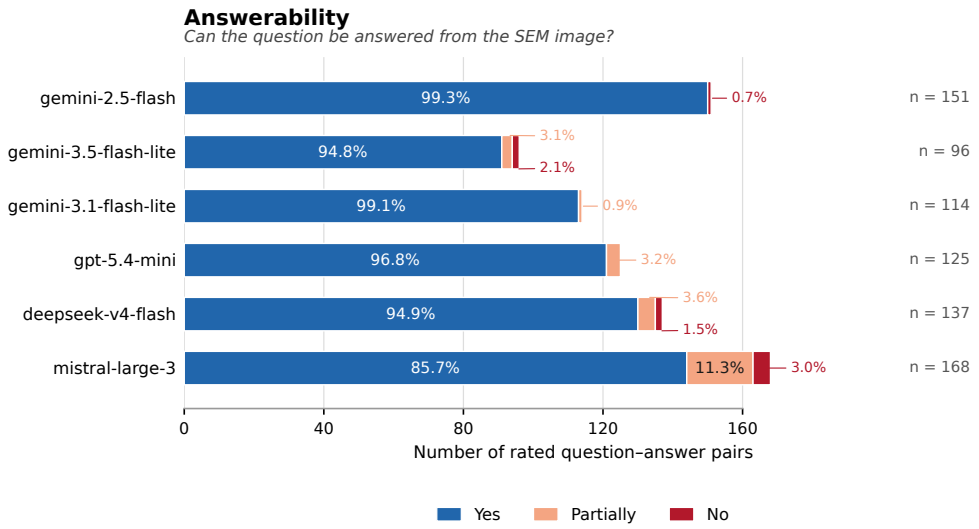

Figure 3: Generator pilot, answerability: can the question be answered from the SEM image? Rated Yes/Partially/No.

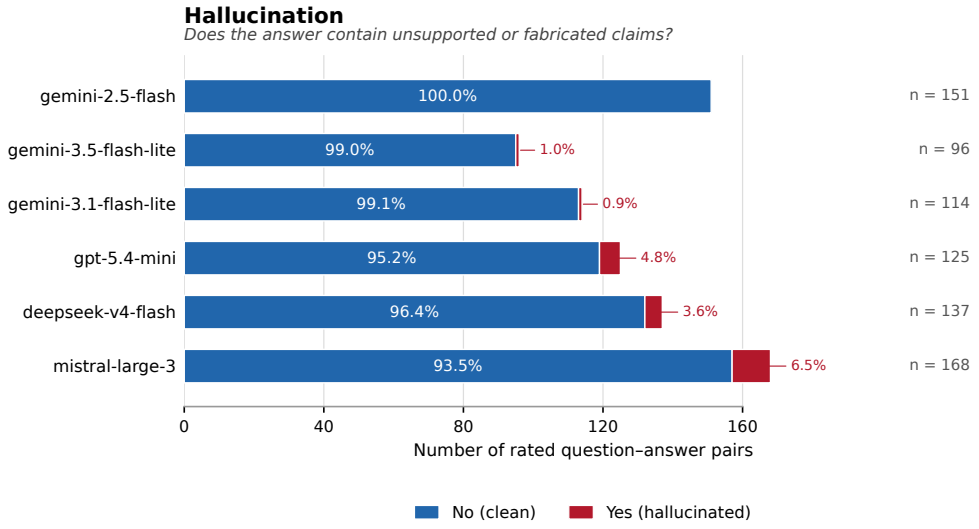

Figure 4: Generator pilot, hallucination: does the answer contain unsupported or fabricated claims? Rated Yes/No, so this dimension has two response categories rather than three.

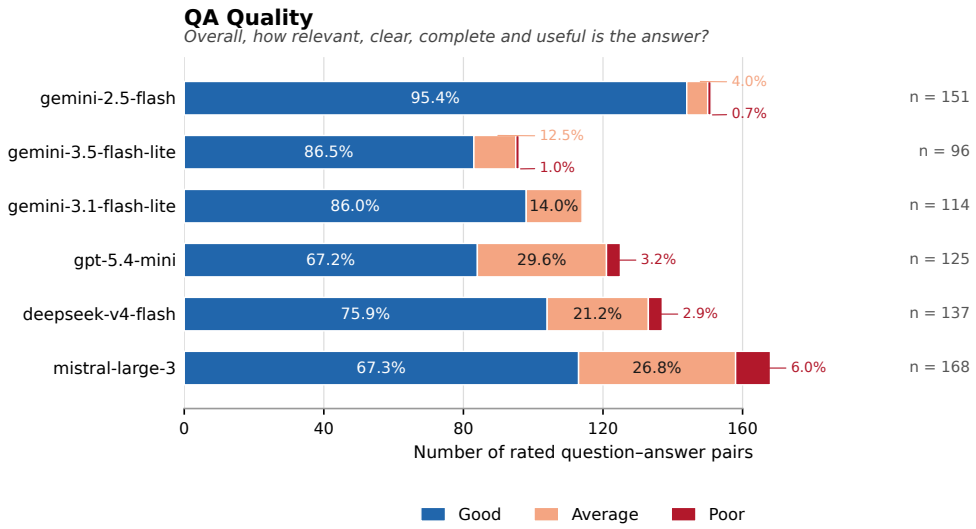

Figure 5: Generator pilot, overall QA quality: how relevant, clear, complete and useful is the answer? This is the composite judgement on which the candidates separate most widely.

D REPRESENTATIVE BENCHMARK CONSTRUCTION

Embedding. Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ image set of released corpus. Each panel is encoded by the DINOv3 CLS token (`facebook/dinov3-vitb16-pretrain-lvd1689m`, ViT-B/16), $\varphi : \mathcal{X} \to \mathbb{R}^{768}$, and the result is ℓ_2-normalised and linearly projected to 50 dimensions by PCA with projection W fitted on the normalised embeddings:

$$z_i = \frac{\varphi(x_i)}{\|\varphi(x_i)\|_2}, \qquad u_i = W z_i \in \mathbb{R}^{50}.$$

Normalising before projection makes the representation depend on embedding direction rather than magnitude, so that subsequent Euclidean distances approximate cosine dissimilarity.

Clustering. k-means is fitted on $\{u_i\}_{i=1}^{N}$, minimising the within-cluster sum of squares

$$\min_{\mathcal{C}_1,\ldots,\mathcal{C}_k} \sum_{c=1}^{k} \sum_{i \in \mathcal{C}_c} \|u_i - \mu_c\|_2^2, \qquad \mu_c = \frac{1}{|\mathcal{C}_c|} \sum_{i \in \mathcal{C}_c} u_i,$$

over a range of k, each fit scored by mean silhouette, Davies–Bouldin and Calinski–Harabasz indices; $k = 5$ is retained.

The sweep is informative less for the k it selects than for what it reveals about the geometry. The mean silhouette never rises above 0.08 at any k: a value near zero means the average point lies about as close to a neighbouring cluster as to its own, so no choice of k recovers well-separated groups. This is the expected outcome for SEM appearance, which varies continuously with magnification, detector mode and surface topography rather than falling into discrete visual classes. We therefore treat the partition strictly as a *stratification device* — a tessellation of a continuous space into regions of comparable occupancy — and not as a taxonomy. No semantic reading is attached to any cluster, centroids are not treated as prototypes, and no result depends on the clusters being meaningful, only on their jointly covering the space.

Radial banding. Stratifying on cluster identity alone would be insufficient for exactly this reason: within a region of a continuum, a sample may still concentrate near the centre of mass and miss the transitional material at the edges, which is where the modes of variation overlap and where the hardest cases sit. The construction therefore adds a second, orthogonal stratification on radial

position. For image i with assignment $c(i)$, define the centroid distance and its within-cluster rank percentile

$$d_i \;=\; \|u_i - \mu_{c(i)}\|_2, \qquad \rho_i \;=\; \frac{\big|\{\, j \in \mathcal{C}_{c(i)} : d_j \le d_i \,\}\big|}{|\mathcal{C}_{c(i)}|} \;\in\; (0, 1],$$

and assign a band $\beta_i \in \{\text{CORE}, \text{MID}, \text{PERIPHERAL}\}$ by

$$\beta_i \;=\; \begin{cases} \text{CORE} & \rho_i \le 0.10, \\ \text{MID} & 0.10 < \rho_i \le 0.70, \\ \text{PERIPHERAL} & \rho_i > 0.70. \end{cases}$$

Defining the bands on the percentile ρ_i rather than on d_i itself is what makes this work uniformly: the split is exact in every cluster regardless of that cluster's size, radius or density, whereas a fixed distance threshold would place almost all of a tight cluster in one band and almost all of a diffuse one in another. The asymmetric $10/60/30$ allocation deliberately over-weights the periphery relative to a median split, reflecting that the outer region of each cluster is both larger in volume and more heterogeneous than its core.

Cell design. The pool is partitioned into cells $\mathcal{P}_{c,\beta} = \{\, i : c(i) = c,\ \beta_i = \beta \,\}$, one per cluster–band combination, and a fixed quota of images is drawn from each; with five clusters, three bands and a quota of twenty, this yields the 300. Filling every cell to the same quota, rather than allocating proportionally to cell population, is the step that makes the benchmark representative in the intended sense: proportional allocation would simply reproduce the corpus's frequency profile and defeat the purpose, whereas a fixed quota guarantees that sparsely populated regions are represented on equal terms with dense ones. Balance over cluster $\times$ band is therefore exact by construction rather than achieved in expectation, and remains so under any subsequent reweighting of the corpus.

Selection within a cell. Each image i carries the attribute sets induced by its own questions: the cognitive levels L_i and the free-text feature labels F_i (Section 3.2). Given the set S already selected in a cell, with covered attributes $\Lambda(S) = \bigcup_{j \in S} L_j$ and $\Phi(S) = \bigcup_{j \in S} F_j$, define the marginal coverage gain and the radial separation

$$\nu(i \mid S) \;=\; \big|L_i \setminus \Lambda(S)\big| \;+\; \big|F_i \setminus \Phi(S)\big|, \qquad g(i \mid S) \;=\; \min_{j \in S} \big|d_i - d_j\big|,$$

with $g(i \mid \emptyset) = \infty$. Selection is greedy under the lexicographic objective

$$i^\star \;=\; \operatorname{argmax}_{i \in \mathcal{P}_{c,\beta} \setminus S} \big(\, \nu(i \mid S),\ g(i \mid S) \,\big),$$

ties resolved by a seeded uniform draw. The two components serve different ends: ν pushes the cell's contents toward semantic variety, while g is one-dimensional farthest-point sampling on the centroid-distance scale, which keeps the picks spread across the band instead of collecting at whichever end the candidate ordering happens to favour. Ordering them lexicographically rather than combining them in a weighted sum avoids having to choose a scale factor between a set cardinality and a Euclidean distance. A previously constructed 100-image subset is supplied as an exclusion set, so the two benchmarks are disjoint by construction.

One property of this objective is worth stating, since it bears on how much the greedy stage should be credited. Because F_i is drawn from an uncontrolled free-text vocabulary 3.2 whose cardinality is large relative to any cell, almost every candidate contributes some previously unseen label, so ν is positive at nearly every step and discriminates weakly; the tiebreak g is correspondingly seldom reached. The radial spread of the selected set is therefore attributable principally to the cell design rather than to the greedy rule layered on it — the stratification is doing the work, and it is sufficient for the purpose. Per-step selection statistics and the realised coverage over levels, features and the centroid-distance range are reported in the released construction record.

E FINE-TUNING CONFIGURATION

All six fine-tuned cells share a single training pipeline; they differ only in the two axes of Section **??**. This appendix records the configuration in full so the runs can be reproduced or audited.

Table 5: Effect of the prompt format on the seven open-weight models evaluated in both conditions. *plain* supplies the bare question; *evidence* additionally states the imaging modality and requests an `Evidence:`/`Answer:` structure. Scoring is identical in both cases; the `Answer:` span covers the same 1,492 pairs, so the differences lie in what the models produce, not in how they are measured. Bold marks the better of the two formats for each model and metric. *Len.* is the ratio of mean prediction length to mean reference length.

System	Prompt	Token-F1	BERTSc.-F1	ROUGE-L	METEOR	BLEU	Len.
`Qwen3VL-2B`	plain	0.160	0.844	0.133	**0.231**	0.019	7.43
	evidence	**0.285**	**0.875**	**0.238**	0.204	**0.045**	1.92
`Qwen3VL-4B`	plain	0.166	0.843	0.134	**0.221**	0.016	6.88
	evidence	**0.295**	**0.878**	**0.251**	0.172	**0.042**	0.62
`Qwen3VL-8B`	plain	0.156	0.840	0.125	**0.221**	0.017	7.92
	evidence	**0.286**	**0.874**	**0.240**	0.176	**0.040**	0.74
`InternVL3.5-2B`	plain	0.239	0.868	0.198	**0.223**	0.037	2.65
	evidence	**0.293**	**0.878**	**0.250**	0.179	**0.043**	0.63
`InternVL3.5-4B`	plain	0.225	0.867	0.186	**0.226**	0.032	3.02
	evidence	**0.289**	**0.877**	**0.246**	0.172	**0.042**	0.58
`InternVL3.5-8B`	plain	0.229	0.865	0.190	**0.216**	0.035	2.33
	evidence	**0.301**	**0.881**	**0.257**	0.178	**0.043**	0.55
`InternVL3-2B`	plain	0.269	0.874	0.216	**0.272**	0.041	3.20
	evidence	**0.320**	**0.888**	**0.270**	0.204	**0.051**	0.71
Mean change		+0.089	+0.021	+0.082	−0.046	+0.016	−3.95
Models improved		7/7	7/7	7/7	0/7	7/7	—

E.1 Adaptation

The base model is Qwen3.5-2B. All training is parameter-efficient: the base weights stay frozen and LoRA adapters are inserted with rank $r = 16$, $\alpha = 32$ and dropout 0.05, no bias terms, no RSLoRA, and standard initialisation. Adapters are applied to twelve module types in the language model, covering both attention paths of this architecture and the feed-forward block:

- grouped-query attention: `q_proj`, `k_proj`, `v_proj`, `o_proj`
- linear attention: `in_proj_a`, `in_proj_b`, `in_proj_qkv`, `in_proj_z`, `out_proj`
- feed-forward: `gate_proj`, `up_proj`, `down_proj`

Expanded over the network this resolves to 186 adapted modules. The vision-adapted stage-1 variant additionally targets four module types inside the vision tower — `qkv`, `attn.proj`, `linear_fc1`, `linear_fc2` — raising the count to 284.

E.2 Two-stage schedule

Stage 1: image to summary. The model is trained to generate the panel summary from the image alone. Each record supplies an image and its summary; no question is involved. Two variants are trained, identical except for whether the vision tower receives adapters (`vision_mode`: *frozen* against *LoRA*).

Stage 2: visual question answering. The model is trained on the QA corpus under one of the two supervision targets. **The vision tower is frozen in every stage-2 run**, so it varies only within stage 1. Stage 2 is initialised in one of three ways: from a fresh adapter on the base model (B1), or by continuing the stage-1 adapter from either the vision-frozen (B2) or vision-adapted (B3) variant.

Table 6: Training configuration. Values not listed separately are shared between stages.

	Stage 1 (summary)	Stage 2 (VQA)
Epochs	2	1
Learning rate	1×10^{-4}	7×10^{-5}
Warmup steps	80	188
Training examples	42,989	200,675
Validation examples	512	1,024
Steps per epoch	1,344	6,272
Total optimiser steps	2,688	6,272
Max sequence length	1,536	1,024
Max pixels	1,048,576	262,144
Vision tower	frozen / LoRA	frozen
Shared		
Per-device batch size	16	
Gradient accumulation	2	
Effective batch size	32	
LoRA r / α / dropout	16 / 32 / 0.05	
Optimiser	AdamW (fused)	
β_1, β_2, ϵ	0.9, 0.999, 1×10^{-8}	
Weight decay	0.0	
LR schedule	cosine	
Gradient clipping	1.0	
Precision	bfloat16	
Attention kernel	SDPA	
Gradient checkpointing	enabled	
Min pixels	200,704 (448^2)	
Seed	42	

Table 7: Training partition, in images. Both stages draw on the same images, so the partition is stated once: stage 1 uses one summary per image and stage 2 the question–answer pairs attached to that image.

Split	Images
Train	42,989
Validation	2,345
Test	2,347
Total	47,681

E.3 Training splits

Fine-tuning uses the partition prepared by the collaborating group over the same corpus. Images are split once and both stages reuse that partition, so a panel seen in stage 1 is never evaluated in stage 2:

Two properties of the partition were checked rather than assumed. Every pair of splits is disjoint in images, and the split is **grouped by source figure**: all panels of one figure fall on the same side, so near-duplicate panels of a single figure cannot straddle a boundary. Both were verified directly rather than inferred from the construction procedure.

The 300 rep300 benchmark images are held out from all three splits. The partition covers 47,681 images and the benchmark a further 300, together accounting for the 47,981 images of the released corpus. We confirmed zero image overlap between the benchmark and each of the training, validation and test splits, so no fine-tuned system in this paper was trained or validated on an image it is evaluated on.

E.4 Objective

Training minimises the standard token-level cross-entropy of causal language modelling, with no label smoothing. The loss is computed on the **target span only**: the collator clones the input ids into

Table 8: The two stage-2 supervision targets. `visual_evidence` and `answer` are fields of the corpus record; the evidence field is present on 200,665 of the 200,675 stage-2 records and is simply unused by Format A, so no data regeneration was required.

	Format A	Format B
System prompt	"... Answer the question about the image accurately and concisely, grounding your answer in the visible microstructural evidence."	"... Answer the question about the image accurately and concisely, and give the visible microstructural evidence that supports your answer."
User turn	image + `question`	image + `question` (identical)
Target	`answer`	`Evidence: <visual_evidence>` \n `Answer: <answer>`

the labels and masks every prompt position with -100, as it does padding. The image tokens, the system prompt and the question therefore contribute no gradient; only the tokens the model is asked to generate do. This is what the two supervision targets change — under Format A the loss covers the answer tokens, under Format B the evidence and answer tokens together.

E.5 Supervision targets

The image and question presented to the model are byte-identical under both formats. Only the system prompt and the target differ.

The prompt does not state the output schema in either case: a model supervised on targets that all carry the markers learns the format from the data. The `Evidence:` and `Answer:` markers are functional rather than cosmetic, since the scoring pipeline locates the answer span by keying on them (Section 4.2). Moving to Format B raises the median target length from 27 to 48 words (307 to 382 tokens under this processor, 428 at maximum), which remains inside the 1,024-token limit, so no target is truncated.

E.6 Convergence and environment

Stage 1 (vision frozen) ends at a training loss of 1.540 and a validation loss of 1.748 after 2,688 optimiser steps; stage 2 from that initialisation ends at 0.995 and 1.005 after 6,272 steps. Validation loss is monitored every 200 steps and is at its minimum at the final checkpoint in both stages, so neither run is reported from an early-stopped point.

Training was performed on a single NVIDIA GH200 (96 GB) under CUDA 13.0, with Python 3.12.14, PyTorch 2.14.0, Transformers 5.17.0 and PEFT 0.20.0. This differs from the environment used for evaluation in this paper; the two were run by different groups on different hardware, and only the released prediction files pass between them.

E.7 Transfer to adjacent materials tasks

All results in the main text are computed on *rep300*, which is drawn from the same generator as the training corpus. To ask whether adaptation to that corpus survives outside it, we evaluate the same systems on the free-form split of four external materials-science benchmarks, none of which contributed to training.

Scoring, and why these numbers stay in the appendix. Each prediction is scored by an LLM judge under the same five-point rubric as the main results, but by a *different* judge: a locally hosted Qwen3.5-4B rather than `gemini-3.5-flash`. Two consequences follow. Absolute values are not on the same scale as Table 2 and must not be compared across the two, and the judge shares a model family with every system it scores, so the self-preference control we were able to apply in the main text is unavailable here. We therefore read this table only as a *relative* comparison within one family, against the zero-shot model as the baseline, and place it in the appendix rather than among the primary results.

Table 9: Transfer to four external materials benchmarks, free-form split (1,041+416+2,749+250 = 4,456 items per system). *Judge* columns are mean scores from a locally hosted Qwen3.5-4B under the rubric of Section B; they are not comparable to the judged scores in the main text. Automatic metrics are means over the four datasets. The zero-shot model is the baseline; **bold** marks cells where a fine-tuned system falls *below* it.

	Judge (Qwen3.5-4B)					Automatic (mean)			
System	csm.	mac.	matqna	matrix	mean	BERTSc.	STS	ROUGE-L	METEOR
Qwen3.5-2B zero-shot	1.757	2.642	2.893	3.048	2.614	0.826	0.523	0.145	0.137
B1-B	1.911	2.889	3.233	**2.480**	2.851	0.836	**0.433**	**0.131**	**0.120**
B2-B	1.880	2.846	3.168	**2.932**	2.824	0.836	**0.439**	**0.139**	**0.133**
B3-B	**1.508**	2.829	3.192	**2.744**	2.740	0.836	**0.431**	**0.131**	**0.126**
B2-A	**1.753**	2.760	3.121	**2.316**	2.722	0.847	**0.447**	0.156	**0.124**
B3-A	**1.707**	2.822	3.087	**2.348**	2.699	0.846	**0.442**	0.155	**0.123**

Transfer is partial, and on two datasets it is negative. On MATRIX the zero-shot model outscores *every* fine-tuned system, by 0.12 to 0.73 judged points; the best adapted cell does not recover the baseline. On CSMBENCH three of the five fine-tuned cells sit at or below the zero-shot score, and B3-B falls to 1.508 against a baseline of 1.757. Gains are confined to MACBENCH and MATQNA, where every fine-tuned cell improves. Averaging over the four datasets hides this: the mean moves from 2.614 to 2.70–2.85 and reads as uniform improvement, which the per-dataset columns show it is not. Adaptation to SEM micrographs therefore transfers to adjacent materials tasks selectively rather than generally, and on some of them it costs accuracy.

The automatic metrics disagree with each other on transfer. BERTScore-F1 rises for every fine-tuned system (0.826 → 0.836–0.847), while STS falls for every one of them (0.523 → 0.431–0.447), and METEOR falls as well. Two embedding-based metrics, computed over identical predictions, order the systems oppositely. This is the dissociation of Section 4.5 reappearing out of domain: fine-tuning moves outputs toward the surface form of the SEM reference answers, which BERTScore rewards and STS — the more style-sensitive of the two on this data — penalises once the reference is drawn from a different corpus. It reinforces the main text's caution rather than qualifying it: on this evidence, no single reference-overlap metric can be relied on to establish whether domain adaptation has helped.

Scope. Only the Qwen3.5 family is scored here, so the table carries no proprietary or cross-architecture baseline and cannot say what a good score on these datasets is. The B1-A cell has no external predictions and is absent. A re-scoring with the main text's judge would remove the family-bias caveat and place these results on the primary scale; we did not run it.

E.8 FINE-TUNING RESULTS

Table 10 reports the six cells. Every cell is judged by `gemini-3.5-flash` under the rubric of Section B and scored by the automatic metrics of Section 4.2 on the same 1,492 pairs as the zero-shot comparison.

Fine-tuning is what moves the judge. Against the zero-shot Qwen3.5-2B baseline (3.330), every cell gains between +0.447 and +0.493 judged points. The best cell, B2-B at 3.823, places the fine-tuned 2B model above Gemini-3.1-Flash (+0.074) and Claude-Haiku-4.5, and 0.160 below GPT-5-mini. The judge's `unsupported_specifics` flag falls from 86.3% on the zero-shot model to 25.2–27.3% across the six cells, so the gain is not obtained by adding unsupported detail.

Neither design axis separates the cells. The six cells span 0.046 judged points in total — under a tenth of the fine-tuning effect itself. Pooling initialisations, Format A averages 3.810 against Format B's 3.796, a difference of 0.014: supervising the visual evidence alongside the answer neither helps nor hurts judged answer quality at the resolution this benchmark affords. Pooling formats, the three initialisation marginals are 3.803 (B1), 3.819 (B2) and 3.789 (B3), a spread of 0.030. Continuing a stage-1 summary adapter does not beat training a fresh adapter on the raw backbone, and

Table 10: The 2×3 fine-tuning grid on the held-out benchmark, ordered by judged agreement. Format A supervises the answer alone, Format B the evidence and answer together; the initialisation axis differs only in the stage-1 adapter each run continues from. Judge score is 1–5; *Flag%* is the rate of the judge's `unsupported_specifics` flag; *len* is mean prediction length as a fraction of the reference.

Initialisation	Fmt	Judge	Flag%	tok-F1	BERTSc.-F1	STS	ROUGE-L	METEOR	len
B2 (stage-1, vision frozen)	B	3.823	26.4	0.4979	0.9191	0.7919	0.4353	0.3899	0.93
B1 (raw backbone)	A	3.816	27.3	0.5056	0.9200	0.7956	0.4426	0.3987	0.94
B2 (stage-1, vision frozen)	A	3.815	26.5	0.5042	0.9196	0.7954	0.4386	0.3975	0.94
B3 (stage-1, vision adapted)	A	3.800	25.2	0.5076	0.9199	0.7968	0.4435	0.4005	0.94
B1 (raw backbone)	B	3.789	25.7	0.4973	0.9189	0.7916	0.4343	0.3907	0.93
B3 (stage-1, vision adapted)	B	3.777	26.3	0.4976	0.9189	0.7918	0.4356	0.3935	0.93

adapting the vision tower during stage 1 does not beat leaving it frozen. The curriculum motivating stage 1 does not, on this benchmark, convert into judged gain.

Metric-best and judge-best are different cells. B3-A attains the highest token-F1 (0.5076), ROUGE-L (0.4435), STS (0.7968) and METEOR (0.4005) of the six, yet ranks fourth on the judge; B2-B is judged best while sitting sixth of six on token-F1. The dissociation of Section 4.5 therefore persists inside a single architecture, initialisation family and training corpus, where backbone or data differences cannot account for it. Because all six cells produce answers at 0.93–0.94 of the reference length, the residual metric differences among them reflect lexical overlap at a resolution the judge does not treat as a difference in answer quality.

E.9 Schema compliance

Adherence to the requested schema is high but not perfect. The `level` field carries one,of the four canonical labels verbatim in 99.975% of records; the remaining 57 use capitalisation variants (30 records), compound labels such as `detection | identification` (10), or categories outside the taxonomy entirely (`comparison`, 12; `localization`, 4), and one is empty. Eleven records have an empty `visual_evidence` field and one has neither `level` nor `feature`. The `feature` field is free text and was not constrained to a vocabulary: it takes 72,308 distinct values, with the sentinel `none` in 17.8% of records.