MICROWORLD: Empowering Multimodal Large Language Models to Bridge the Microscopic Domain Gap with Multimodal Attributed Property Graph

Manyu Li **Ruian He** **Chenxi Ma*** **Weimin Tan*** **Bo Yan***
College of Computer Science and Artificial Intelligence
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, China
manyuli24@m.fudan.edu.cn
{rahe16,cxma,wmtan,byan}@fudan.edu.cn

Abstract

Multimodal large language models (MLLMs) show remarkable potential for scientific reasoning, yet their performance in specialized domains such as microscopy remains limited by the scarcity of domain-specific training data and the difficulty of encoding fine-grained expert knowledge into model parameters. To bridge the gap, we introduce MICROWORLD, a framework for structured-knowledge augmentation with a complementary visual retrieval pathway, using a *multimodal attributed property graph* (MAPG) without domain-specific fine-tuning. MICROWORLD extracts biomedical entities and relations via scispaCy or LLM-based triplet mining, aligns images and entities in a shared embedding space using Qwen3-VL-Embedding, and assembles a knowledge graph comprising approximately 106K nodes and 2.62M typed edges spanning eight relation categories. At inference time, a graph-augmented retrieval pipeline matches query entities to the MAPG and injects structured knowledge context into the MLLM prompt. On MicroVQA, MICROWORLD raises Qwen3-VL-8B-Instruct accuracy from 48.8% to **67.1%** (+18.3 percentage points; 37.5% relative), exceeding *zero-shot* GPT-5 (59.4%). The identical retrieved context raises GPT-5 to **71.5%**, and GPT-5's own entity extraction reaches 73.3%. On MicroBench, 8B Single-Pass improves overall accuracy by **4.0 percentage points**. The benefits are greatest for knowledge-intensive reasoning; fine-grained visual discrimination and the additional visual-pathway contribution remain limited. Code and accompanying resources: https://github.com/ieellee/MicroWorld.

1 Introduction

The rapid advancement of artificial intelligence is fundamentally reshaping how we understand the natural world and conduct biomedical research [1–3]. Foundation models have driven breakthroughs across a wide spectrum of scientific domains, from predicting protein structures with atomic accuracy [4] to accelerating drug discovery [5] and enabling automated disease diagnosis [6]. In microscopy, modern imaging modalities such as cryo-electron microscopy, super-resolution fluorescence, and high-throughput confocal imaging are generating data at an unprecedented scale and resolution [7, 8], far outpacing the capacity of human experts to analyze, interpret, and derive actionable insights. This widening gap between data acquisition and scientific understanding calls for intelligent systems that can reason about microscopy images at the level of domain experts, assisting in experimental result interpretation, hypothesis generation, and experimental design, and

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

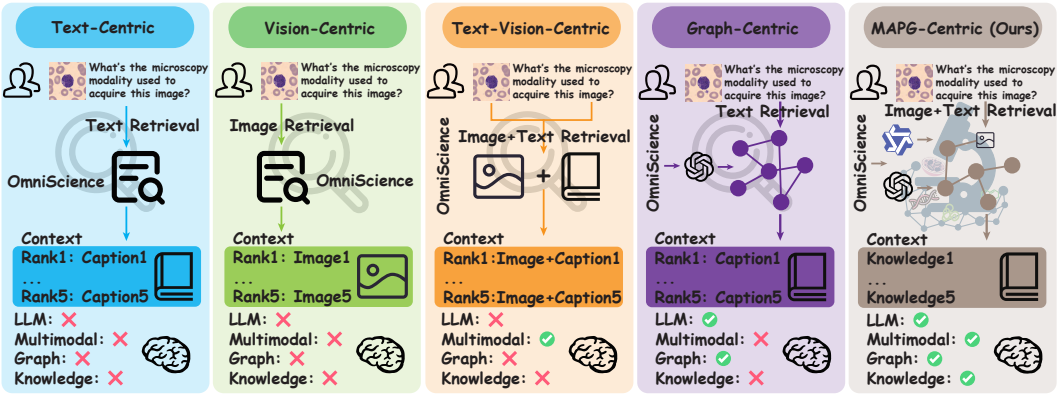


Figure 1: Comparison with prior RAG paradigms. While **Text-Centric** , **Vision-Centric** , and **Text-Vision-Centric** methods rely respectively on captions, images, or a multimodal combination of both, **Graph-Centric** RAG builds a caption-based structured graph but lacks domain-specific modeling. Conversely, our **MAPG-Centric** approach is tailored for biological microscopy and naturally integrates visual modalities. **LLM**: Utilization of a Large Language Model. **Multimodal**: Use of multimodal features during retrieval. **Graph**: Construction of a structured retrieval space. **Knowledge**: Provision of specialized biological microscopy knowledge.

ultimately participate in a *lab-in-the-loop* paradigm that tightly integrates AI reasoning into the scientific discovery cycle [9, 10].

However, building domain-specialized MLLMs for microscopy faces a fundamental tension. Fine-tuning a foundation model on a narrow scientific sub-field demands massive volumes of high-quality, expert-annotated instruction data, which is prohibitively expensive and labor-intensive to curate in microscopy, where each annotation requires deep domain expertise [11, 12]. Moreover, models fine-tuned on limited domain data often suffer from poor generalization and catastrophic forgetting of broader capabilities [13, 14]. An alternative paradigm is retrieval-augmented generation (RAG), which enriches the input context with relevant external knowledge at inference time, circumventing the need for extensive domain-specific training [15–17]. Recent studies have reinforced this insight: for specialized scientific domains with scarce instruction-tuning data, augmenting the reasoning context of general-purpose MLLMs can be more effective and robust than fine-tuning [18, 19], as it preserves the model's broad reasoning capabilities while injecting domain-specific knowledge precisely where it is needed. This raises a natural question: ***can we leverage a scientific literature database to construct a multimodal knowledge base that enhances the reasoning capabilities of MLLMs in the microscopy domain without supervised fine-tuning?***

A key enabler for this vision is the emergence of large-scale scientific corpora. OmniScience [20] has assembled over 1.5 million image–caption pairs sourced from top-tier journals (with an average impact factor exceeding 12) and high-citation preprint servers (arXiv, bioRxiv, medRxiv), and further employed MLLMs to reconstruct high-quality, self-contained captions. Among this vast collection, approximately 692K pairs are related to microscopy, encompassing rich interactions among domain-specific entities: biological structures, imaging techniques, staining protocols, morphological properties, and organism types. These densely interconnected scientific descriptions present a unique opportunity to construct a comprehensive, structured knowledge base for microscopy that goes beyond flat image-caption pair retrieval to capture the relational fabric of the domain.

In this paper, we introduce MICROWORLD, a framework that bridges the microscopic domain gap for MLLMs by constructing and leveraging a multimodal attributed property graph (MAPG) tailored for microscopy to introduce domain knowledge. Our main contributions are as follows:

- **A large-scale multimodal knowledge graph for microscopy.** We construct MICROWORLD, a multimodal attributed property graph comprising approximately 106K entity and image nodes interconnected by 2.62M typed edges spanning eight relation categories. By integrating biomedical entity extraction, LLM-based relation mining, and vision–language alignment via Qwen3-VL-Embedding, MICROWORLD captures both the semantic structure and visual characteristics of the microscopy domain within a unified graph representation.

- **Knowledge-augmented microscopy reasoning without supervised fine-tuning.** MICROWORLD raises Qwen3-VL-8B-Instruct from 48.8% to 67.1% on MicroVQA, above zero-shot GPT-5 (59.4%). Giving GPT-5 the identical context reaches 71.5%, demonstrating that the knowledge also benefits a stronger backbone.
- **A dynamic, extensible knowledge infrastructure.** MICROWORLD is designed as a living knowledge system that supports incremental graph augmentation from heterogeneous scientific corpora. Its modular architecture allows seamless integration of new data sources, enabling the knowledge graph to evolve alongside the rapidly advancing microscopy literature and to be adapted to diverse experimental contexts without retraining.

2 Related work

2.1 Multimodal large language models in microscopy

Recent advances in multimodal large language models (MLLMs) have demonstrated remarkable emergent capabilities in understanding and reasoning over complex data, showing immense potential as intelligent assistants for accelerating biological discoveries [12]. To facilitate this progress, OmniScience [20] curated a large-scale dataset of images and captions from premier journals and conferences in biological sciences, and reconstructed self-contained captions to improve data quality. However, in the microscopy sub-domain of biological sciences, the scarcity of large-scale, high-quality annotated data makes it extremely challenging to fine-tune a dedicated microscopy foundation model [12]. To evaluate MLLM capabilities in this domain, several benchmarks have been proposed: MicroBench [21] focuses on fundamental perception-level tasks, while MicroVQA [22] targets more advanced scientific reasoning abilities. In this work, we construct MICROWORLD, a comprehensive framework that substantially enhances MLLM reasoning performance on microscopy tasks.

2.2 Retrieval-augmented generation in the biomedical field

Retrieval-augmented generation (RAG) has been extensively explored for natural images, with representative approaches spanning text-centric methods such as FLMR [23], vision-centric methods including VisRAG [24] and ColPali [25], as well as text-vision-centric frameworks like REVEAL [26]. In the biomedical domain, MasonNLP [27] leverages a generalist, instruction-tuned large language model coupled with a RAG framework to integrate in-domain textual and visual exemplars, demonstrating the benefits of retrieval-based knowledge grounding. MMed-RAG [28] further introduces a versatile multimodal RAG system specifically designed to enhance the factual accuracy of MLLMs in medical contexts. Following the development of MMed-RAG, RAG methodologies for medical VQA have continuously evolved and improved [29, 30]. Despite these advances, the application of RAG in microscopy remains largely unexplored. Our proposed MICROWORLD bridges this gap by introducing a retrieval-augmented pipeline tailored for microscopy, offering new insights for future research in this specialized domain.

2.3 Graph-structured knowledge in multimodal large language models

Graph-structured representations offer a natural and expressive means of encoding relational knowledge, capturing latent associations among entities that flat image-caption pair stores or unstructured corpora cannot easily reveal [31–33]. GraphRAG [34] pioneers a graph-based approach to question answering over private text corpora, constructing a community-level summarization of an LLM-generated knowledge graph to handle queries that require holistic reasoning across the entire dataset, scaling effectively with both the generality of user questions and the volume of source text. LightRAG [35] further advances this direction by incorporating graph structures into text indexing and retrieval, employing a dual-level retrieval paradigm that enables comprehensive information discovery at both fine-grained entity-level and coarse-grained topic-level granularities. Inspired by these developments, we propose MICROWORLD, which carefully models the rich relational structure inherent in microscopy data, including associations among specimens, imaging modalities, staining techniques, and morphological features to construct a multimodal attributed property graph, thereby uncovering latent connections that can guide more informed and accurate scientific reasoning. Figure 1 shows the comparison between our method and prior RAG paradigms.

3 Method

We present MICROWORLD, a framework that constructs a *multimodal attributed property graph* (MAPG) from scientific image–caption corpora and leverages it to augment MLLM reasoning on microscopy VQA. As shown in Figure 2, the pipeline consists of three stages: (i) *triplet extraction and multimodal embedding* (§3.1), (ii) *graph construction* (§3.2), and (iii) *knowledge-grounded inference* (§3.3).

Notation. Let $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$ be N image–caption pairs. We define the MAPG as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \boldsymbol{\phi}, \boldsymbol{\psi})$, where $\mathcal{V} = \mathcal{V}_e \cup \mathcal{V}_x$ contains entity nodes $\mathcal{V}_e$ and image nodes $\mathcal{V}_x$; $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ are typed edges over relation vocabulary $\mathcal{R}$; $\boldsymbol{\phi} : \mathcal{V} \to \mathbb{R}^d$ maps nodes to embeddings; and $\boldsymbol{\psi} : \mathcal{V}_e \to \mathcal{A}$ assigns structured attributes. Each entity node has one of five semantic types: STRUCTURE, METHOD, CONDITION, PROPERTY, or ORGANISM; image nodes belong to the separate set $\mathcal{V}_x$. Given a VQA instance $(x_q, q, \{a_k\}_{k=1}^{K})$, the goal is to predict k^* by conditioning on subgraph knowledge from $\mathcal{G}$.

3.1 Stage 1: Triplet extraction and multimodal embedding

This stage transforms the corpus $\mathcal{D}$ (filtered from OmniScience via the method in Appendix G) into structured triplets and dense embeddings.

Entity and relation extraction. For each caption c_i, we apply `scispaCy` [36] NER with noun-chunk augmentation to extract biomedical entities, each linked to MeSH identifiers and classified into one of the five semantic types via heuristic rules (details in Appendix D). We then prompt GPT-5 [37] to extract relation triplets $\{(h, r, t)\}$ constrained to five biologically meaningful types: `observed_by`, `has_property`, `located_in`, `interacts_with`, and `part_of` (see Figure 2). Additionally, for each caption with m entities, we add $\binom{m}{2}$ undirected `co_occurs_with` edges:

$$\mathcal{E}_{\text{co}} = \big\{(e_j,\ \texttt{co_occurs_with},\ e_k) \mid 1 \le j < k \le m,\ e_j, e_k \in \text{Entities}(c_i)\big\}. \tag{1}$$

Vision–language embedding. We employ Qwen3-VL-Embedding [38] to compute ℓ_2-normalized d-dimensional embeddings for images ($\mathbf{h}_x$), image–caption pairs ($\mathbf{h}_m$), and entity names ($\mathbf{h}_e$), using task-specific instructions. These embeddings are used in Stage 2 for visual similarity edges and in Stage 3 for retrieval.

3.2 Stage 2: Graph construction

This stage assembles the extracted triplets and embeddings into the MAPG $\mathcal{G}$.

Entity deduplication and graph assembly. An entity registry merges surface forms with identical canonical names (after lowercasing, punctuation removal, and whitespace collapsing), unifying aliases, MeSH IDs, and provenance. For each image x_i, we create an image node $v_{x_i} \in \mathcal{V}_x$ connected to its caption entities via `shown_in` edges: $\mathcal{E}_{\text{vis}} = \big\{(v_{x_i},\ \texttt{shown_in},\ v_e) \mid v_e \in \text{Entities}(c_i)\big\}$. The full edge set $\mathcal{E} = \mathcal{E}_{\text{bio}} \cup \mathcal{E}_{\text{co}} \cup \mathcal{E}_{\text{vis}} \cup \mathcal{E}_{\text{vsim}}$ uses bidirectional adjacency for non-`shown_in` relations.

Entity description enrichment. We query NCBI Entrez Utilities [39] for authoritative definitions with type-dependent lookup strategies, falling back to GPT-5-generated definitions when NCBI results are absent or too short (<40 chars).

Visual similarity edges. We augment $\mathcal{G}$ with `visually_similar` edges between image pairs whose embedding cosine similarity exceeds $\sigma = 0.85$:

$$\mathcal{E}_{\text{vsim}} = \big\{(v_{x_i},\ \texttt{visually_similar},\ v_{x_j}) \mid \boldsymbol{\phi}(v_{x_i})^\top \boldsymbol{\phi}(v_{x_j}) \ge \sigma,\ i < j\big\}. \tag{2}$$

Structural precomputation. To support efficient online retrieval, we precompute (i) adaptive k-hop neighborhoods (expanding to 2-hop for low-degree nodes, capping at 50 neighbors for hubs) and (ii) a hybrid similarity score $S(v_i, v_j) = \alpha \cdot J(v_i, v_j) + (1 - \alpha) \cdot C(v_i, v_j)$ that fuses Jaccard structural similarity with cosine embedding similarity (default $\alpha = 0.5$). Full formulations are in Appendix D. The graph also supports incremental augmentation from external corpora (Appendix D.5). Table 6 in the Appendix summarizes all eight relation types.

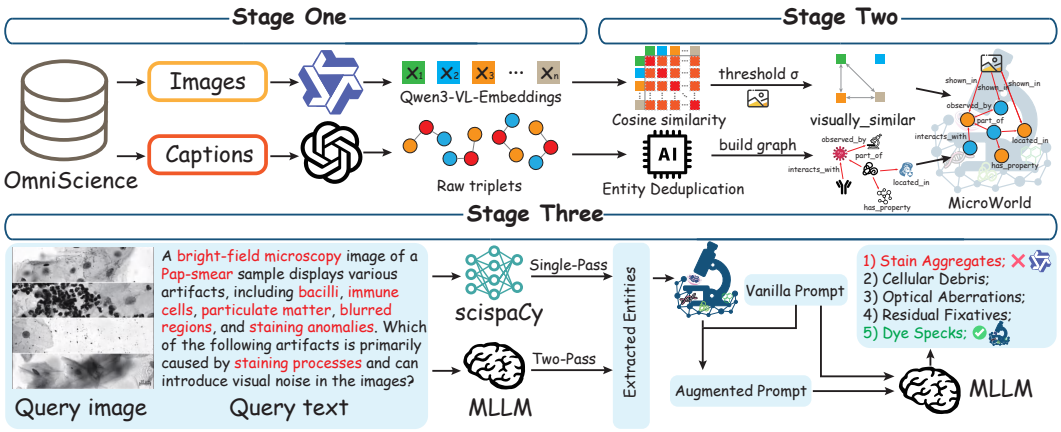


Figure 2: Overview of the MICROWORLD framework. **Stage 1**: Images and captions from OmniScience are processed via scispaCy NER and LLM-based relation extraction to produce raw triplets, while Qwen3-VL-Embedding computes dense vision–language representations. **Stage 2**: Extracted entities undergo deduplication and are assembled into a multimodal attributed property graph (MAPG) with eight relation types; image embeddings are compared via cosine similarity to add `visually_similar` edges above threshold σ. **Stage 3**: At inference time, entities are extracted from the query via scispaCy (Single-Pass) or the MLLM itself (Two-Pass), matched to the MAPG, and the retrieved knowledge context is injected into an augmented prompt for the final answer.

3.3 Stage 3: Knowledge-grounded inference

At inference time, given a query (x_q, q), we retrieve relevant subgraph context and inject it into the MLLM prompt (Algorithm 1 in the Appendix).

Single-pass inference. ScispaCy NER extracts entity mentions from q, which are matched to the KG via cascaded exact-alias and fuzzy matching ($\rho_{\text{fuzzy}} = 0.72$); overly generic entities (frequency $> \rho_{\text{skip}}$) are filtered. We encode x_q with the same Qwen3-VL-Embedding encoder and retrieve nearest image nodes among the 20K stored vectors, then expand over `visually_similar` and `shown_in` edges; up to three visual entities supplement the text matches. The query need not be a graph node. For each matched entity, we assemble a knowledge block with its definition, semantic type, and k-hop neighbors, then prepend the assembled context $\mathcal{C}$ to the question.

Two-pass inference. The MLLM first extracts up to 8 domain-specific entities from (x_q, q) (Round 1), which are then matched against the KG. A fresh single-turn prompt with the retrieved knowledge is used for the final answer (Round 2), avoiding multi-turn context degradation.

4 Experiments

4.1 Experimental setup

Dataset. We evaluate on MicroVQA and MicroBench. MicroVQA is a microscopy visual question answering benchmark containing 1,042 multiple-choice questions spanning three scientific reasoning tasks: *Hypothesis Generation* (420 samples), *Perception* (392 samples), and *Experiment Proposal* (230 samples). MicroBench is a microscopy visual question answering benchmark that contains 17,315 images of multiple-choice questions that focus on *Perception*.

Backbone models. We evaluate MICROWORLD with two families of multimodal large language models at three scales: **Qwen3-VL** [40] (2B, 4B, 8B) and **InternVL3.5** [41] (2B, 4B, 8B).

Inference modes. We compare two MICROWORLD inference paradigms: *Single-Pass*, where scispaCy extracts entities from the question before the MLLM forward pass (§3.3), and *Two-Pass*,

Table 1: MicroVQA accuracy (%). "Baseline" is zero-shot without KG context. GPT-5 "shared" uses identical Qwen3-VL-8B Two-Pass context; "own" uses GPT-5 entity extraction. GPT-5 overall scores are independently reported from task scores. InternVL3.5 uses frozen Qwen3-VL prompts. Best commercial and per-size open-source results are **bolded**. *Results from MicroVQA leaderboard.

Model	Mode	Hyp. Gen.	Perception	Exp. Prop.	Overall
Commercial Models (zero-shot / MAPG-augmented)					
o1* [42]	Baseline	50.2	53.0	55.4	52.8
Claude Sonnet 4.5* [43]	Baseline	56.4	49.6	55.1	54.4
o4-mini* [44]	Baseline	56.1	50.4	57.9	55.6
o3* [44]	Baseline	60.5	53.5	61.5	59.3
GPT-5* [37]	Baseline	58.9	53.9	63.3	59.4
GPT-5	MAPG shared	72.5	71.0	70.4	71.5
GPT-5	MAPG own	**74.3**	**73.2**	**71.7**	**73.3**
2B scale					
InternVL3.5-Instruct	Baseline	42.9	43.4	49.6	44.5
InternVL3.5-Instruct	Single-Pass	49.5	49.2	58.3	51.3
InternVL3.5-Instruct	Two-Pass	53.3	55.4	**63.5**	56.3
Qwen3-VL-Instruct	Baseline	42.4	44.9	53.9	45.9
Qwen3-VL-Instruct	Single-Pass	51.9	51.8	56.1	52.8
Qwen3-VL-Instruct	Two-Pass	**60.0**	**57.4**	61.7	**59.4**
4B scale					
InternVL3.5-Instruct	Baseline	43.1	45.4	50.4	45.6
InternVL3.5-Instruct	Single-Pass	43.8	45.9	43.0	44.4
InternVL3.5-Instruct	Two-Pass	48.1	51.0	50.9	49.8
Qwen3-VL-Instruct	Baseline	46.4	47.4	44.8	46.4
Qwen3-VL-Instruct	Single-Pass	55.2	55.1	53.0	54.7
Qwen3-VL-Instruct	Two-Pass	**62.9**	**66.6**	**63.0**	**64.3**
8B scale					
InternVL3.5-Instruct	Baseline	38.1	42.6	47.4	41.8
InternVL3.5-Instruct	Single-Pass	55.0	55.6	53.0	54.8
InternVL3.5-Instruct	Two-Pass	64.0	63.5	62.2	63.4
Qwen3-VL-Instruct	Baseline	48.6	49.2	48.3	48.8
Qwen3-VL-Instruct	Single-Pass	58.3	60.5	57.4	58.9
Qwen3-VL-Instruct	Two-Pass	**67.9**	**68.4**	**63.5**	**67.1**

where the MLLM first extracts entities, followed by KG retrieval and a single-turn knowledge-augmented answer (§3.3).

RAG baselines. We compare against six representative RAG methods spanning different modality focuses: *text-centric*: FLMR [23]; *vision-centric*: VisRAG [24] and ColPali [25]; *vision-text*: REVEAL [26]; *graph-centric*: GraphRAG [34] and LightRAG [35]. All baselines use the same filtered 20K OmniScience corpus, backbone, prompt template, and decoding configuration. We sweep top-$k \in \{1, 3, 5\}$ and report each method's best setting. Indexing and retrieved content differ by method: MAPG additionally supplies entity definitions and typed relations, so shared source data do not imply identical knowledge content.

Metrics. We report per-task accuracy and overall accuracy of MicroVQA and MicroBench samples. In Appendix F, we also report 95th-percentile (P95) latency and throughput (queries per second, QPS) for both single-pass and two-pass inference.

Implementation details. Local-model experiments use 4×NVIDIA RTX 3090 GPUs with multi-GPU parallel inference. Default hyperparameters are: $M_{\text{text}} = 6$, $M_{\text{vis}} = 3$, $\rho_{\text{skip}} = 0.08$, $\rho_{\text{compact}} = 0.04$, maximum context length 6,000 characters, $M_{\text{nbr}} = 4$ neighbors, and a `max_2hop` budget of 3.

Table 2: RAG comparison at 8B on MicroVQA: same filtered-20K corpus, backbone, prompt template, and decoding; best top-$k \in \{1, 3, 5\}$ per baseline. Type: T=text, V=vision, G=graph. MICROWORLD uses Two-Pass and additional entity definitions and relations. 2B/4B results: Appendix E. Best results are **bolded**.

Method	Type	k	Hyp. Gen.	Perception	Exp. Prop.	Overall
MICROWORLD (Ours)	G+V+T	–	**67.9**	**68.4**	**63.5**	**67.1**
FLMR	T	1	48.6	50.0	46.5	48.7
VisRAG	V	1	47.1	50.3	47.8	48.5
ColPali	V	1	47.1	49.2	48.3	48.2
GraphRAG	G	1	46.9	51.8	46.1	48.6
LightRAG	G	3	45.5	48.7	47.8	47.2
REVEAL	T+V	3	46.2	45.9	50.0	46.9

4.2 Main results

Table 1 reports results across backbone models and inference modes. **(1) MICROWORLD boosts zero-shot performance.** Compared to the zero-shot baseline (no KG context), MICROWORLD Two-Pass achieves large improvements across all models: Qwen3-VL-2B-Instruct improves from 45.9% to 59.4% (+13.5 pp), 4B from 46.4% to 64.3% (+17.9 pp), and 8B from 48.8% to 67.1% (+18.3 pp). InternVL3.5 shows similar gains, with 8B jumping from 41.8% to 63.4% (+21.6 pp). **(2) Two-Pass consistently outperforms Single-Pass.** Two-Pass yields 5.0–9.6 pp higher accuracy than Single-Pass across all six backbones; Single-Pass also gains +6.9 to +10.1 pp over baseline on Qwen3-VL. **(3) Backbone capability affects knowledge use.** Qwen3-VL Two-Pass improves from 59.4% (2B) through 64.3% (4B) to 67.1% (8B). InternVL3.5 is non-monotonic (56.3%, 49.8%, 63.4%), so size alone does not determine performance. **(4) Cross-model transfer is effective.** InternVL3.5-8B achieves 63.4% using frozen prompts originally generated for Qwen3-VL, only 3.7 pp below the native Qwen3-VL-8B result (67.1%). This supports the transferability of the structured knowledge context across these two model families. **(5) GPT-5 also benefits from the same knowledge.** The identical 8B Two-Pass context raises GPT-5 from 59.4% to 71.5% (+12.1 pp); its own Round-1 extraction reaches 73.3%. Thus, the 67.1% open-source result exceeds zero-shot GPT-5, while MAPG-augmented GPT-5 remains stronger.

4.3 Comparison with RAG baselines

Tables 2, 7 show comparison results with RAG baselines (based on the same raw data as MICROWORLD). MICROWORLD outperforms all RAG baselines by a substantial margin across all three model scales. On the 2B backbone, MICROWORLD achieves 59.4% overall, surpassing the best baseline (ColPali at 46.9%) by **12.5 pp**. On the 4B backbone, the gap widens to **16.6 pp** (64.3% vs. GraphRAG at 47.7%). On the 8B backbone, MICROWORLD reaches 67.1%, exceeding the best baseline (FLMR at 48.7%) by **18.4 pp**. The advantage holds across all three task types, consistent with the value of entity-centric definitions and relation-linked knowledge. This comparison does not isolate graph structure from differences in knowledge content. Among baselines, graph-centric methods (GraphRAG, LightRAG) generally outperform text-only (FLMR) and vision-only (VisRAG) at the 4B scale, while at 8B text-centric FLMR emerges as the strongest baseline, suggesting larger models extract more value from raw textual context.

4.4 Ablation study

Table 3 reports one-at-a-time (OAT) ablations on eight retrieval settings (Qwen3-VL, Two-Pass), with Δ relative to default. The joint 8B text-only control is analyzed separately below. **(1) Context length matters for larger models.** Increasing `max_context_chars` from 6K to 8K improves 8B accuracy by 1.6 pp (to 68.7%), while reducing it to 4K consistently hurts performance across all sizes (-1.5 to -3.0 pp). Larger models can effectively utilize longer context windows. **(2) Visual entities are valuable at scale.** Removing visual entities (`max_visual_entities`$= 0$) reduces 4B and 8B accuracy by 1.5 and 1.3 pp, but improves 2B by 0.8 pp. Increasing visual entities to 5 benefits the 8B model (+1.8 pp) but hurts the 2B model (-1.0 pp), suggesting that smaller models

Table 3: MicroVQA Two-Pass ablations. Δ is the difference from default (pp). The joint 8B text-only control is reported separately in Appendix I.4.

		2B		4B		8B	
Parameter	**Value**	**Acc.**	Δ	**Acc.**	Δ	**Acc.**	Δ
Default	–	59.4	–	64.3	–	67.1	–
`max_text_entities`	3	59.9	+0.5	62.5	−1.8	64.6	−2.5
	10	59.7	+0.3	64.6	+0.3	66.1	−1.0
`max_context_chars`	4000	57.9	−1.5	61.3	−3.0	64.5	−2.6
	8000	59.6	+0.2	66.1	+1.8	**68.7**	**+1.6**
`max_neighbors`	3	59.3	−0.1	64.3	+0.0	67.7	+0.6
	6	59.0	−0.4	64.3	+0.0	67.6	+0.5
`freq_skip_ratio`	0.06	58.3	−1.1	64.7	+0.4	66.4	−0.7
	0.12	59.9	+0.5	64.2	−0.1	67.9	+0.8
`freq_compact_ratio`	0.02	59.4	+0.0	65.5	+1.2	67.9	+0.8
	0.08	60.0	+0.6	64.5	+0.2	66.5	−0.6
`max_visual_entities`	0	60.2	+0.8	62.8	−1.5	65.8	−1.3
	5	58.4	−1.0	64.6	+0.3	**68.9**	**+1.8**
`max_2hop`	0	**62.0**	**+2.6**	63.7	−0.6	67.9	+0.8
	2	60.4	+1.0	63.8	−0.5	67.9	+0.8
`visually_similar`	-	60.7	+1.3	63.4	−0.9	68.6	+1.5

may struggle with additional visual context. **(3) 2-hop expansion shows diminishing returns.** Disabling 2-hop neighbors surprisingly improves the 2B model by 2.6 pp, while the effect on 4B and 8B is marginal ($<$1 pp). This indicates that deeper graph traversal introduces noise for smaller models. **(4) Neighbor count is robust.** Varying `max_neighbors` between 3 and 6 has minimal impact ($\leq$0.6 pp), suggesting the default value of 4 is near-optimal. **(5) Visual similarity edges show mixed effects.** Disabling `visually_similar` edges ($\sigma = 0.85$, §3.2) yields +1.3 pp on 2B and +1.5 pp on 8B, but -0.9 pp on 4B. The model-dependent effects suggest a trade-off between relevance, context length, and retrieval noise.

Joint and format controls. The 8B text-only KG reaches 66.0% versus full MAPG's 67.1% on MicroVQA, and 58.15% versus 58.54% on MicroBench (Two-Pass). These modest visual contributions are non-additive relative to the individual ablations. A separate 2-hop $\times$ visual-edge control yields a difference-in-differences of -0.5 pp (Appendix I.2). Rendering identical entities, definitions, and relations as a flat paragraph yields 66.9% versus 67.1% with structured blocks, suggesting limited sensitivity to presentation format (Appendix I.4).

4.5 Knowledge graph scale study

We subsample the MAPG at six scales (10%–90%) and compare against the full graph; Figure 3 shows results across all three model sizes. **(1) Accuracy generally improves with scale.** For all three model sizes, Two-Pass accuracy shows a clear upward trend from 10% to 90%, with Qwen3-VL-8B improving from 52.4% (10%) to 68.1% (90%) before settling at 67.1% (100%). The largest gains occur between 10% and 50% (+9.3 pp for 8B), after which improvement flattens within this selected corpus. **(2) The 90% scale sometimes surpasses 100%.** For 2B (62.8% vs. 59.4%) and 8B (68.1% vs. 67.1%), the 90% graph slightly outperforms the full graph. This may be attributed to the reduced noise from infrequent or low-quality entities in the remaining 10%. **(3) Experiment Proposal benefits most from scale.** This task shows the sharpest per-scale gains, while Perception grows more gradually, consistent with Experiment Proposal requiring broader contextual knowledge that denser graphs supply.

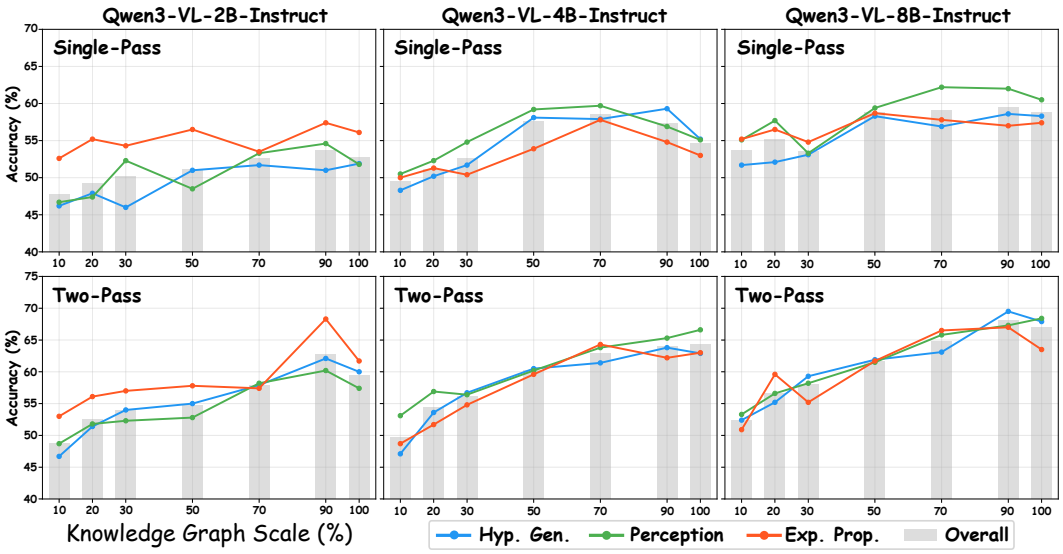


Figure 3: Effect of MAPG scale on Qwen3-VL-Instruct (2B/4B/8B), Single-Pass and Two-Pass, on MicroVQA. Overall and task accuracy generally increase as the graph is randomly subsampled from 10% to 100%, with the largest gains between 10% and 50%. Some 90% subgraphs outperform the full graph, potentially due to reduced retrieval noise.

Table 4: Per-task accuracy (%) on MicroBench. *Coarse-grained* tasks: Modality, Sub-modality, Domain, Sub-domain, Stain. *Fine-grained* tasks: Classification. Best result per model in column is **bolded**.

Model	Method	*Coarse-Grained Perception*					*Fine-Grained*	*Aggregate*	
		Modality	Sub-mod.	Domain	Sub-dom.	Stain	Classif.	Coarse	Overall
Qwen3-VL-2B	Baseline	**79.5**	47.2	51.4	**57.7**	**49.1**	**34.3**	57.3	53.8
	Single-Pass	63.4	46.2	53.9	53.2	46.4	32.5	52.8	49.7
	Two-Pass	77.5	**59.2**	**56.3**	54.1	45.8	33.8	**59.0**	**55.2**
Qwen3-VL-4B	Baseline	83.0	62.0	44.0	**57.9**	52.8	36.2	60.2	54.1
	Single-Pass	**85.6**	**68.0**	**55.8**	55.6	**54.2**	37.3	**64.1**	**60.0**
	Two-Pass	82.5	62.4	44.3	**57.9**	52.2	**38.8**	60.1	56.9
Qwen3-VL-8B	Baseline	80.7	67.1	53.7	52.6	56.9	34.3	62.4	58.1
	Single-Pass	**83.8**	**69.8**	**67.1**	52.6	**61.3**	33.8	**67.2**	**62.1**
	Two-Pass	80.8	66.6	54.4	**53.4**	56.8	**36.2**	62.6	58.5

4.6 Cross-benchmark evaluation: MicroBench

To assess transfer beyond MicroVQA, we evaluate on MicroBench [21], which was not used in corpus filtering (Table 4). MICROWORLD's gains concentrate in coarse-grained perception, where Single-Pass yields the strongest improvements at 4B (coarse: +3.9 pp, overall: 60.0%, +5.9 pp) and 8B (+4.8 pp, 62.1%, +4.0 pp), while Two-Pass leads at 2B (+1.7 pp, 55.2%, +1.4 pp overall). Fine-grained classification benefits are comparatively modest ($\leq$+2.6 pp at 4B), as this sub-task is more vision-centric and less amenable to external knowledge augmentation. Each scale has a configuration that improves on baseline, but the gains depend on task and mode.

Table 5: Corpus-selection controls on MicroVQA.

Graph	Def. coverage (%)	Accuracy (%)	Gain (pp)
Random	18.4	55.7	+6.9
Random (+ top-6K definitions)	53.6	56.0	+7.2
Filtered (seed 42)	82.8	59.3	+10.5

4.7 Corpus-selection controls

Table 5 compares 5K-pair graphs (Qwen3-VL-8B, Two-Pass), with gains relative to the 48.8% zero-shot baseline. Random sampling without MicroVQA contact retains +6.9–7.2 pp gains. Filtered (seed 42) reaches 59.3%, but unequal definition coverage prevents attributing its advantage solely to corpus alignment (Appendix I.1).

4.8 Qualitative case study

We analyze Qwen3-VL-8B-Instruct on MicroVQA, examining how knowledge context (KC) flips incorrect answers to correct ones (*upgrades*) and vice versa (*downgrades*). Full case illustrations are provided in Appendix H.

Upgrade mechanisms. We identify five recurring mechanisms through which KC improves reasoning (Figures 5–9): **(1) Terminology disambiguation**: KC provides precise definitions that help the model distinguish semantically similar options (e.g., "dye specks" vs. "stain aggregates" in Pap smear artifact identification); **(2) Domain-specific knowledge injection**: KC supplies entity-specific facts unavailable in the model's parametric memory (e.g., the role of sample purification in interpreting cryo-EM asymmetry of CHIKV); **(3) Reasoning level elevation**: KC elevates the model from local mechanistic explanations to broader biological insights (e.g., recognizing structural plasticity of alphavirus geometry rather than local protein conformational changes); **(4) Misconception correction**: KC corrects oversimplified textbook-level reasoning with domain-precise knowledge (e.g., lysosome ring arrangements reflecting multi-organellar defense assemblies rather than mere phagosome fusion); **(5) Logical chain correction**: KC rules out an incorrect causal chain (e.g., interpreting energy-intensive actin polymerization as energy conservation).

Downgrade mechanisms. We also identify five failure modes (Figures 10–14): **(1) Knowledge overriding visual evidence**: generic KC introduces textual noise that overrides the model's originally correct visual judgment; **(2) Context mismatch**: KC retrieved from a different experimental context (e.g., in vitro quality control advice) is misapplied to a distinct biological scenario (e.g., in vivo wound healing); **(3) Ambiguous KC signals**: conflicting information within the retrieved context leads the model to favor the more "novel" but incorrect explanation; **(4) Knowledge-induced over-reasoning**: detailed procedural knowledge causes the model to prefer a superficially direct but scientifically flawed strategy over a methodologically sound approach; **(5) Visual–textual confidence imbalance**: detailed textual knowledge leads the model to discount discriminative image features. These failures motivate context relevance filtering, confidence-aware knowledge integration, and visual–textual evidence arbitration.

5 Conclusion

We presented MICROWORLD, a framework that constructs a multimodal attributed property graph (MAPG) from scientific image–caption corpora and uses it to augment MLLM reasoning for microscopy VQA. Graph-augmented prompting raises Qwen3-VL-8B to 67.1% on MicroVQA, exceeding zero-shot GPT-5 (59.4%); the identical context raises GPT-5 to 71.5%. The current gains are predominantly text-driven and strongest for knowledge-intensive reasoning. Cross-model transfer and smaller MicroBench and random-corpus gains support applicability beyond the main setting, with limits from corpus alignment, extraction noise, and fine-grained visual discrimination. Qualitative analysis further reveals both the promise of knowledge-grounded reasoning and its failure modes (detailed in Appendix H), offering actionable insights for future systems. We also discuss ethics considerations in Appendix A, discussions and more information in Appendix B, C.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Program of China (Grant No. 2025YFF0514500) and Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G11). The computations in this research were performed using the CFFF platform of Fudan University.

The authors declare no competing interests.

References

[1] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[2] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[3] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

[4] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[5] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[6] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

[7] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018.

[8] Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer. Deep learning massively accelerates super-resolution localization microscopy. *Nature biotechnology*, 36(5):460–468, 2018.

[9] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[10] Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86, 2023.

[11] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[12] Shanghang Zhang, Gaole Dai, Tiejun Huang, and Jianxu Chen. Multimodal large language models for bioimage analysis. *nature methods*, 21(8):1390–1393, 2024.

[13] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*, 2023.

[14] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[16] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

[17] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068, 2024.

[18] Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*, 2025.

[19] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

[20] Haoyi Tao, Chaozheng Huang, Nan Wang, Han Lyu, Linfeng Zhang, Guolin Ke, and Xi Fang. Omniscience: A large-scale multi-modal dataset for scientific image understanding. *arXiv preprint arXiv:2602.13758*, 2026.

[21] Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket R Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. Micro-bench: A microscopy benchmark for vision-language understanding. *Advances in Neural Information Processing Systems*, 37:30670–30685, 2024.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19552–19564, 2025.

[23] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023.

[24] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

[25] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

[26] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023.

[27] AHM Rezaul Karim and Ozlem Uzuner. Masonnlp at mediqa-wv 2025: Multimodal retrieval-augmented generation with large language models for medical vqa. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 84–94, 2025.

[28] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.

[29] Yinan Wu, Yuming Lu, Yan Zhou, Yifan Ding, Jingping Liu, and Tong Ruan. Mkgf: A multi-modal knowledge graph based rag framework to enhance lvlms for medical visual question answering. *Neurocomputing*, 635:129999, 2025.

[30] Ankan Deria, Komal Kumar, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, and Imran Razzak. Medmo: Grounding and understanding multimodal large language model for medical images. *arXiv preprint arXiv:2602.06965*, 2026.

[31] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *Advances in neural information processing systems*, 31, 2018.

[32] Michelle M Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine and healthcare. *Nature biomedical engineering*, 6(12):1353–1369, 2022.

[33] Belinda Mo, Kyssen Yu, Joshua Kazdan, Joan Cabezas, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. Kggen: Extracting knowledge graphs from plain text with language models. *arXiv preprint arXiv:2502.09956*, 2025.

[34] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[35] Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3), 2024.

[36] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pages 319–327, 2019.

[37] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[38] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

[39] Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10–D17, 2021.

[40] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[41] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[42] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[43] Anthropic. Claude sonnet 4.5 system card, 2025.

[44] OpenAI. Openai o3 and o4-mini system card, 2025.

[45] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. spacy: Industrial-strength natural language processing in python. 2020.

[46] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384, 2013.

[47] John W. Ratcliff and David Metzener. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, page 46, 1988.

[48] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[49] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

[50] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

A Ethics

Ethical use of biomedical data. Our work uses image–caption pairs sampled from OmniScience, which aggregates scientific figures and captions from biomedical literature. We do not introduce patient-level data or collect personally identifiable information. Access to a publication does not by itself establish permission to redistribute its figures.

Data licensing and usage. We intend to license eligible MICROWORLD-derived assets under Creative Commons Attribution–NonCommercial–ShareAlike 4.0 (CC BY-NC-SA 4.0), subject to upstream permissions. This does not relicense source images or texts: their original attribution, copyright, and redistribution conditions remain applicable. Redistribution must preserve provenance and source-specific terms; where permission does not allow image redistribution, source identifiers and reconstruction instructions provide an alternative. Code licensing is separate from the data license. Users must verify the terms applicable to their intended use.

B Limitations, discussions and future work

Text-centric knowledge representation. At the current 20K graph scale and corpus composition, MICROWORLD is predominantly text-driven: the 8B joint text-only control loses only 1.1 pp on MicroVQA and 0.39 pp on MicroBench relative to full MAPG. Fine-grained morphological classification benefits less than terminology- and context-intensive reasoning (Section 4.6). Nearest-neighbor noise and the composite-figure nature of OmniScience limit the visual pathway; splitting figures into panels may improve specificity.

Benchmark-aligned corpus selection. MeSH-derived terms are weighted using MicroVQA question frequencies, so corpus selection is aligned with this benchmark even without accessing answers or options. The random-5K controls retain +6.9–7.2 pp gains, while filtered-5K reaches 59.3% versus random-5K's 56.0% and filtered-20K's 67.1% (Appendix I.1). The approximately 3.4 pp fixed-size gap also reflects unequal definition coverage (82.8% versus 53.6%). The approximately 7.8 pp filtered-5K-to-20K gap is likewise a descriptive scale comparison. Random-20K remains unevaluated. Transfer to distributions unlike MicroVQA is supported at smaller observed gains, including +4.0 pp on MicroBench at 8B Single-Pass.

Scale of MICROWORLD. The graph uses 20K image–caption pairs from OmniScience because of computation and API-budget constraints. The scale study shows flattening only within this selected corpus. Expansion requires further extraction, indexing, storage, and embedding resources.

Limited Evaluation of Larger Open-Source MLLMs. Our open-source evaluation is restricted to Qwen3-VL and InternVL3.5 up to 8B parameters. GPT-5+MAPG improves further (71.5–73.3%), but does not characterize larger open-source backbones or establish a universal interaction between model scale and external knowledge.

Retrieval quality bottleneck. Automatic relation extraction and definition generation can introduce errors, while entity disambiguation and image retrieval can add mismatched context. Our downgrade cases show that even factually valid knowledge may override correct visual judgments; current retrieval does not explicitly arbitrate conflicting evidence.

Broader impact. MICROWORLD may help laboratories use external scientific knowledge without domain-specific fine-tuning. Its 8B result exceeds zero-shot GPT-5 on MicroVQA, while GPT-5 given the same context is stronger. With a prebuilt graph, open-source inference can run locally without sending query images to a commercial API; graph construction still uses GPT-5 for extraction and some definitions. Applying this pipeline to other scientific domains remains a direction for validation. The graph is automatically extracted from literature, so its outputs should support expert judgment rather than serve as authoritative scientific or clinical conclusions. Source-publication biases, extraction errors, and licensing restrictions also constrain responsible use.

Future work. Several promising directions emerge from the above limitations. **First**, splitting composite figures into panels and indexing representative crops alongside textual descriptions may

reduce visual retrieval noise. **Second**, *retrieval quality refinement* via reranking, retrieval-augmented filtering, or uncertainty-aware context selection would help mitigate the downgrade failure modes. **Third**, scaling MICROWORLD to the full OmniScience corpus with efficient approximate nearest-neighbor indexing is a natural next step to improve coverage. Finally, evaluating larger open-source backbones, random-20K graphs, and controls with matched definition coverage would better characterize model-scale effects and corpus-alignment dependence.

C Concerns on data leakage

We distinguish benchmark distribution alignment from answer leakage, and clarify the provenance and nearest-neighbor semantics of image retrieval.

Corpus selection uses MicroVQA term frequencies. The weighting in Eq. 8 uses term frequencies from MicroVQA questions. This makes the selected corpus benchmark-aligned and limits claims about general microscopy transfer; it is separate from using ground-truth answers or answer options.

(a) The vocabulary $\mathcal{T}$ originates from MeSH ontology, not from MicroVQA labels. The domain vocabulary $\mathcal{T}$ is derived entirely from biomedical entity mentions extracted via MeSH identifiers (a standardized biomedical ontology) and subsequently filtered by GPT-5 to retain microscopy-relevant terms (Appendix G). MicroVQA ground-truth answers and answer options are *never* accessed during vocabulary construction.

(b) Question frequencies create distribution alignment. $\mathrm{tf}_{\mathcal{Q}}(t)$ weights the domain vocabulary toward terms frequent in MicroVQA. MicroBench is not used in this filtering. The random-corpus controls characterize the dependence on corpus selection without eliminating the alignment confound (Table 5; Appendix I.1).

Nearest-neighbor image retrieval. The RETRIEVESIMILARIMAGES$(x_q, \mathcal{G})$ step in Algorithm 1 encodes an unseen query and retrieves existing OmniScience image nodes by embedding similarity. It does not insert benchmark images, questions, answers, or options into the graph, nor require exact image identity.

(a) MicroVQA images are in-house and independent of OmniScience. The MicroVQA benchmark [22] contains in-house laboratory images rather than images drawn from the OmniScience literature corpus. OmniScience, by contrast, aggregates compositional figure panels extracted from open-access biomedical publications. The MAPG image nodes are constructed from the selected OmniScience records, not from benchmark images. Their captions and linked entities supply external context for nearest-neighbor retrieval.

Summary. The graph is built from external literature without benchmark answers or options. Corpus selection nevertheless uses MicroVQA question frequencies and therefore aligns with its distribution.

D Method details

D.1 Entity type classification

A heuristic classifier assigns each entity to one of the five semantic types based on keyword matching and spaCy [45] entity labels:

$$\tau(e) = \begin{cases} \textsc{Method} & \text{if } e \cap \mathcal{K}_{\text{method}} \neq \emptyset, \\ \textsc{Property} & \text{if } e \cap \mathcal{K}_{\text{property}} \neq \emptyset, \\ \textsc{Organism} & \text{if label}(e) \in \{\textsc{Taxon}, \textsc{Species}\}, \\ \textsc{Condition} & \text{if label}(e) \in \{\textsc{Disease}, \textsc{Phenotype}\}, \\ \textsc{Structure} & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{K}_{\text{method}}$ and $\mathcal{K}_{\text{property}}$ are curated keyword sets for experimental methods and morphological properties, respectively. To ensure reproducibility, we provide part of the keyword sets below:

Algorithm 1 Visual–Semantic Subgraph Retrieval (Single-Pass / Two-Pass)

Require: Query image x_q, question text q, graph $\mathcal{G}$, MLLM $\mathcal{F}$, parameters $M_{\text{text}}, M_{\text{vis}}, \rho_{\text{skip}}$, mode $\in \{\textsc{Single}, \textsc{Two}\}$
Ensure: Knowledge context string $\mathcal{C}$, answer a

// — Entity extraction (mode-dependent) —
if mode $=$ TWO **then**
  $\mathcal{E}_q \leftarrow \mathcal{F}.\textsc{ExtractEntities}(x_q, q)$ ▷ MLLM extracts up to 8 domain entities (Round 1)
else
  $\mathcal{E}_q \leftarrow \textsc{SciSpaCy-NER}(q) \cup \textsc{NounChunks}(q)$ ▷ Offline NER for single-pass
end if
// — Textual pathway: match entities to KG —
$\mathcal{M}_{\text{text}} \leftarrow \emptyset$
for $e \in \mathcal{E}_q$ **do**
  $v \leftarrow \textsc{AliasMatch}(e, \mathcal{G})$ ▷ Exact alias lookup
  if $v = \text{NULL}$ **then**
    $v \leftarrow \textsc{FuzzyMatch}(e, \mathcal{G}, \rho_{\text{fuzzy}}{=}0.72)$ ▷ Fall back to Ratcliff/Obershelp fuzzy match
  end if
  if $v \neq \text{NULL} \wedge \text{freq}(v) \leq \rho_{\text{skip}}$ **then**
    $\mathcal{M}_{\text{text}} \leftarrow \mathcal{M}_{\text{text}} \cup \{v\}$
  end if
  if $|\mathcal{M}_{\text{text}}| \geq M_{\text{text}}$ **then break**
  end if
end for
// — Visual pathway: nearest image nodes + graph expansion —
$\mathcal{I}_{\text{nn}} \leftarrow \textsc{RetrieveSimilarImages}(x_q, \mathcal{G})$ ▷ Same encoder; nearest stored image vectors
$\mathcal{I}_{\text{sim}} \leftarrow \mathcal{I}_{\text{nn}} \cup \{v' \mid v \in \mathcal{I}_{\text{nn}}, (v, \texttt{visually_similar}, v') \in \mathcal{E}_{\text{vsim}}\}$
$\mathcal{V}_{\text{vis}} \leftarrow \emptyset$ ▷ Candidate entity set from visual pathway
for $v_{\text{img}} \in \mathcal{I}_{\text{sim}}$ **do**
  $\mathcal{V}_{\text{vis}} \leftarrow \mathcal{V}_{\text{vis}} \cup \{u \mid (v_{\text{img}}, \texttt{shown_in}, u) \in \mathcal{E}_{\text{vis}}\}$ ▷ Collect caption entities via `shown_in` edges
end for
$\mathcal{V}_{\text{vis}} \leftarrow \textsc{RankByDescAvail}(\mathcal{V}_{\text{vis}})[:M_{\text{vis}}]$ ▷ M_{vis} is the scalar cap; $\mathcal{V}_{\text{vis}}$ is the entity set
// — Merge, expand, and format —
$\mathcal{M} \leftarrow \mathcal{M}_{\text{text}} \cup \mathcal{V}_{\text{vis}}$
for $v \in \mathcal{M}$ **do**
  Retrieve precomputed $\mathcal{N}_k(v)$ (Eq. 4) and descriptions $\boldsymbol{\psi}(\cdot)$
end for
$\mathcal{C} \leftarrow \textsc{FormatKnowledgeBlocks}(\mathcal{M}, \mathcal{N}_k, \boldsymbol{\psi})$
// — Answer generation —
if mode $=$ TWO **then**
  $a \leftarrow \mathcal{F}.\textsc{Answer}(x_q, \mathcal{C} \oplus q)$ ▷ Fresh single-turn prompt with retrieved context (Round 2)
else
  $a \leftarrow \mathcal{F}.\textsc{Answer}(x_q, \mathcal{C} \oplus q)$
end if
return $\mathcal{C}$, a

- $\mathcal{K}_{\text{method}}$: *microscopy, imaging, staining, fluorescence, immunofluorescence, confocal, electron, cryo-EM, SEM, TEM, FISH, FRET, super-resolution, STORM, PALM, TIRF, phase-contrast, brightfield, widefield, two-photon, CRISPR, sequencing, PCR, Western blot, immunohistochemistry, flow cytometry, mass spectrometry, spectroscopy, tomography, diffraction, crystallography, NMR, patch-clamp, electrophysiology, transfection, transduction, knockout, knockdown, overexpression, fixation, permeabilization, embedding, sectioning*.
- $\mathcal{K}_{\text{property}}$: *morphology, shape, size, length, width, diameter, thickness, density, intensity, brightness, contrast, texture, granularity, elongation, roundness, circularity, polarity, symmetry, distribution, localization, co-localization, overlap, pattern, structure, boundary, membrane, lumen, condensation, aggregation, fragmentation, branching, tubular, punctate, filamentous, ring-shaped, rod-shaped, spherical, irregular*.

Matching is performed case-insensitively on tokenized entity surface forms. Entities matching neither $\mathcal{K}_{\text{method}}$ nor $\mathcal{K}_{\text{property}}$ keywords, and without a recognized Taxon/Species or Disease/Phenotype label, default to STRUCTURE.

D.2 Vision–language embedding details

We employ Qwen3-VL-Embedding [38] to compute d-dimensional embeddings for three modalities:

- **Image embeddings** $\mathbf{h}_x \in \mathbb{R}^d$: Each image x_i is encoded with the instruction *"Represent this microscopy image for knowledge graph node retrieval"*.
- **Mixed embeddings** $\mathbf{h}_m \in \mathbb{R}^d$: Each image–caption pair (x_i, c_i) is jointly encoded with instruction *"Represent this image-text pair for biomedical knowledge retrieval"*.
- **Entity embeddings** $\mathbf{h}_e \in \mathbb{R}^d$: Each entity canonical name is encoded with instruction *"Represent this biomedical entity name for visual alignment"*.

All embeddings are ℓ_2-normalized: $\boldsymbol{\phi}(v) = f_\theta(v)/\|f_\theta(v)\|_2$.

D.3 K-hop neighbor precomputation

We precompute neighborhood structures using an adaptive k-hop expansion strategy based on node degree:

$$\mathcal{N}_k(v) = \begin{cases} \mathcal{N}_1(v) \cup \mathcal{N}_2(v) & \text{if } \deg(v) \leq 3, \\ \mathcal{N}_1(v) & \text{if } 3 < \deg(v) \leq 20, \\ \mathcal{N}_1(v)\big|_{\leq 50} & \text{if } \deg(v) > 20, \end{cases} \tag{4}$$

where $\mathcal{N}_1(v)$ and $\mathcal{N}_2(v)$ denote the 1-hop and 2-hop neighbor sets of v, and $|_{\leq n}$ indicates truncation to at most n neighbors. The three cases correspond to: *(i)* sparse nodes that receive 2-hop expansion to ensure sufficient context; *(ii)* medium-degree nodes ($4 \leq \deg(v) \leq 20$) whose full 1-hop neighborhood is used as-is; and *(iii)* hub nodes ($\deg(v) > 20$) that are truncated to the 50 highest-scoring neighbors (ranked by the hybrid similarity score S, Eq. 6) to prevent context overflow.

D.4 Hybrid similarity ranking

For each pair of entity nodes (v_i, v_j), we compute the Jaccard [46] structural similarity over non-image 1-hop neighbors:

$$J(v_i, v_j) = \frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}, \tag{5}$$

and the cosine embedding similarity $C(v_i, v_j) = \boldsymbol{\phi}(v_i)^\top \boldsymbol{\phi}(v_j)$. The fused ranking score is:

$$S(v_i, v_j) = \alpha \cdot J(v_i, v_j) + (1 - \alpha) \cdot C(v_i, v_j), \tag{6}$$

where $\alpha = 0.5$ by default. For each entity, we retain the top-K most similar nodes.

D.5 Graph augmentation with external corpora

To scale knowledge coverage, we support merging the primary MAPG (MICROWORLD) with a supplementary graph from the external corpus. The merging follows three principles: **(1) primary-first**: nodes and descriptions from the primary graph are never overwritten; **(2) entity-only augmentation**: supplementary image nodes and `shown_in` edges are excluded to preserve frequency statistics; **(3) embedding alignment**: embeddings from both graphs are concatenated with shared index files. This yields $\mathcal{G}^+$ with broader entity coverage while maintaining visual retrieval integrity.

D.6 Inference details

Entity frequency definition. The entity frequency $\text{freq}(v)$ is the fraction of graph-construction image–caption pairs in which entity v appears:

$$\text{freq}(v) = \frac{|\{d \in \mathcal{D} : v \in \text{Entities}(d)\}|}{|\mathcal{D}|}, \tag{7}$$

Table 6: Complete relation schema of the multimodal attributed property graph.

Relation Type	Source	Direction	Semantics
`observed_by`	LLM extraction	Directed	Structure $\rightarrow$ imaging method
`has_property`	LLM extraction	Directed	Entity $\rightarrow$ morphological property
`located_in`	LLM extraction	Directed	Entity $\rightarrow$ spatial location
`interacts_with`	LLM extraction	Directed	Functional/physical interaction
`part_of`	LLM extraction	Directed	Sub-component $\rightarrow$ whole
`co_occurs_with`	Co-occurrence	Undirected	Same-caption co-occurrence
`shown_in`	Visual link	Directed	Image $\rightarrow$ depicted entity
`visually_similar`	Embedding sim.	Undirected	Image visual similarity ($\geq \sigma$)

where $\mathcal{D}$ is the set of all image-caption pairs used to construct the MAPG and Entities(d) denotes the entity set extracted from image-caption pair d. Values are pre-computed over the 20k image-caption pairs corpus and stored in the entity registry. The skip threshold $\rho_{\text{skip}} = 0.08$ discards entities appearing in more than 8% of corpus pairs (e.g. *cells*, *protein*); the compact threshold $\rho_{\text{compact}} = 0.04$ triggers abbreviated representation for moderately common entities.

Textual pathway. We apply scispaCy NER to the question q, yielding candidate entity mentions. Each mention is matched to the KG through: **(1)** exact alias lookup against the normalized alias index; **(2)** fuzzy matching via the Ratcliff/Obershelp algorithm [47] with cutoff $\rho_{\text{fuzzy}} = 0.72$. Entities with corpus frequency exceeding ρ_{skip} are discarded as overly generic.

Visual pathway. We encode x_q with the same image encoder and instruction used for the stored 20K image nodes, then perform nearest-neighbor retrieval by cosine similarity. The returned node set is expanded over `visually_similar` edges, and `shown_in` links yield caption entities (Algorithm 1). Entities are ranked by description availability and frequency, and at most $M_{\text{vis}} = 3$ are added. An empty candidate set contributes no visual entities; the query itself need not be a stored node.

Context assembly. For each matched entity, we construct a knowledge block containing: **(1)** canonical name and semantic type; **(2)** NCBI/LLM-derived definition; **(3)** up to M_{nbr} 1-hop neighbors with relation types and descriptions; **(4)** 2-hop context for low-degree nodes ($|\mathcal{N}_1(v)| \leq 3$), capped by `max_2hop`.

Frequency-aware formatting. Entities with moderate frequency ($\rho_{\text{compact}} < \text{freq}(v) \leq \rho_{\text{skip}}$) receive compact representations (truncated to 120 characters).

E RAG baseline results at 2B and 4B scales

Table 7 reports the full RAG baseline comparisons at 2B and 4B model scales, complementing the 8B results in Table 2 of the main text.

F Inference latency analysis

MICROWORLD's Two-Pass inference pipeline introduces an entity extraction step before the main knowledge-augmented reasoning call. We benchmark its end-to-end per-query latency against a Single-Pass baseline to characterize the practical overhead and the effect of entity extraction batch size on latency and throughput.

Setup. All experiments are conducted on a single NVIDIA RTX 3090 24 GB GPU. *Single-Pass* runs spaCy before standard MLLM inference. *Two-Pass* first performs a batched entity extraction call (on average 6–8 biomedical entities per query), retrieves the relevant KG subgraph, and then performs the knowledge-augmented reasoning call. We sweep the entity extraction batch size $B \in \{1, 2, 4, 8, 16\}$ and report mean latency, 95th-percentile (P95) latency, and throughput in queries per second (QPS).

Table 7: Comparison with RAG baselines at 2B and 4B scales (best top-k per method). Type: T=text, V=vision, T+V=text-vision, G=graph. MICROWORLD uses Two-Pass mode. Best results per column within each scale are **bolded**.

Method	Type	k	Hyp. Gen.	Perception	Exp. Prop.	Overall
2B scale						
MICROWORLD (Ours)	G+V+T	–	**60.0**	**57.4**	**61.7**	**59.4**
FLMR	T	3	41.9	43.6	50.0	44.3
VisRAG	V	3	45.6	42.2	50.2	45.3
ColPali	V	5	41.0	45.3	60.4	46.9
REVEAL	T+V	1	45.5	44.6	50.0	46.2
GraphRAG	G	1	44.0	46.7	47.8	45.9
LightRAG	G	3	40.0	42.6	54.3	44.1
4B scale						
MICROWORLD (Ours)	G+V+T	–	**62.9**	**66.6**	**63.0**	**64.3**
FLMR	T	5	44.2	47.7	46.5	46.0
VisRAG	V	5	41.7	46.2	49.5	45.1
ColPali	V	3	44.8	48.2	48.3	46.8
REVEAL	T+V	3	43.8	48.0	47.8	46.3
GraphRAG	G	3	46.7	48.5	48.3	47.7
LightRAG	G	5	45.0	51.5	45.7	47.6

Table 8: **Inference latency comparison between Single-Pass and Two-Pass on MicroVQA (B = entity extraction batch size).** Mean and P95 latency (ms) and throughput (QPS) are reported over 64 test samples on a single NVIDIA RTX 3090 24 GB GPU. "—" indicates GPU out of memory (OOM).

Method	B	Qwen3-VL-2B			Qwen3-VL-4B			Qwen3-VL-8B		
		Mean (ms)	P95 (ms)	QPS	Mean (ms)	P95 (ms)	QPS	Mean (ms)	P95 (ms)	QPS
Single Pass	—	649.6	1339.3	1.54	659.6	1338.4	1.52	651.3	1338.4	1.54
Two-Pass	1	2503.9	2916.8	0.40	1930.2	2568.6	0.52	2227.7	2818.8	0.45
Two-Pass	2	1280.2	1652.2	0.78	1332.3	1664.0	0.75	1526.8	1828.0	0.65
Two-Pass	4	789.1	1618.0	1.27	790.4	1475.7	1.27	902.6	1274.2	1.11
Two-Pass	8	473.3	834.8	**2.11**	489.9	793.8	**2.04**	—	—	—
Two-Pass	16	278.3	340.8	**3.59**	—	—	—	—	—	—

Results. Table 8 reports the full latency sweep. Three findings stand out. **(1) Batching drastically reduces Two-Pass overhead.** At B=1, Two-Pass processes entities sequentially and incurs a 2.9–3.9× mean-latency penalty over Single Pass. Increasing B to 4 brings mean latency within $\sim$25% of Single Pass for the 2B and 4B models; the 8B overhead is 38.6%. **(2) Two-Pass surpasses Single-Pass throughput at large batch sizes.** At B=8, Two-Pass reaches 2.11 QPS (2B) and 2.04 QPS (4B), outperforming the corresponding Single-Pass throughput (1.54 and 1.52 QPS) by $\sim$34%. This counterintuitive result arises because a single batched entity-extraction call amortizes prompt-processing overhead over multiple entities, making the two-LLM-call pipeline cheaper per query than a single un-batched call when the batch is large enough. **(3) Model size has limited impact on Single-Pass latency.** Single-Pass mean latency is nearly identical across the 2B, 4B, and 8B variants (649–660 ms), suggesting that, under this hardware configuration, the bottleneck may lie in memory bandwidth rather than arithmetic throughput for these model sizes.

G Domain-relevant corpus selection from OmniScience

The OmniScience corpus contains a broad collection of biological science image–caption pairs spanning diverse sub-fields. To construct a supplementary knowledge graph that is maximally relevant to microscopy reasoning, we design a weighted TF-IDF [48] scoring algorithm to select a compact, domain-focused subset from OmniScience. The core idea is to prioritize image-caption pairs that densely cover the specialized vocabulary of the microscopy domain (as reflected in MeSH

identifiers aided with GPT-5) while penalizing image-caption pairs dominated by generic biomedical terms.

Terminology extraction. We first extract a domain vocabulary $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$ by collecting all biomedical entity mentions from the MeSH identifiers filtered by GPT-5. Representative high-frequency terms include *cells*, *microscopy*, *protein*, *fluorescence*, *staining*, *mitochondrial*, and *confocal*, which constitute the core vocabulary of microscopic sciences.

Query-side term weighting. For each term $t \in \mathcal{T}$, we compute a query-side weight that balances its importance in the target domain against its generality in the source corpus:

$$w_t = \mathrm{tf}_{\mathcal{Q}}(t) \times \mathrm{IDF}_{\mathcal{S}}(t), \quad \text{where} \quad \mathrm{IDF}_{\mathcal{S}}(t) = \log \frac{|\mathcal{S}|}{1 + |\{d \in \mathcal{S} : t \in d\}|}, \tag{8}$$

where $\mathrm{tf}_{\mathcal{Q}}(t)$ is the term frequency of t in the MicroVQA question set $\mathcal{Q}$, $\mathcal{S}$ denotes the full OmniScience corpus, and $|\mathcal{S}|$ is the total number of image-caption pairs. This formulation assigns higher weights to terms that are both frequent in MicroVQA and rare in OmniScience, thereby emphasizing domain-specific terminology over generic biomedical vocabulary.

Image-caption Pairs scoring. For each candidate image-caption pair $d \in \mathcal{S}$, we compute a relevance score by aggregating the weighted contributions of all matching domain terms, modulated by a diversity bonus:

$$\mathrm{score}(d) = \underbrace{\sum_{t \in d \cap \mathcal{T}} w_t \cdot \log\big(1 + \mathrm{tf}_d(t)\big)}_{\text{weighted term relevance}} \times \underbrace{\log\big(|\{t \in \mathcal{T} : t \in d\}| + 1\big)}_{\text{diversity bonus}}, \tag{9}$$

where $\mathrm{tf}_d(t)$ is the local term frequency of t in image-caption pair d. The logarithmic dampening of local term frequency prevents individual high-frequency terms from dominating the score. The diversity bonus $\log(|\{t \in \mathcal{T} : t \in d\}| + 1)$ encourages the selection of image-caption pairs that cover a broader range of MicroVQA concepts, favoring comprehensiveness over narrow specialization.

Subset selection. We rank all OmniScience image-caption pairs by their relevance score in descending order and select the top-K image-caption pairs to form the filtered subset $\mathcal{S}^* \subset \mathcal{S}$:

$$\mathcal{S}^* = \underset{d \in \mathcal{S}}{\text{top-K}}\ \mathrm{score}(d). \tag{10}$$

In practice, we set $K = 20{,}000$. Figure 4 provides a detailed analysis of the filtering outcome across four dimensions: **(1)** the top-50 MicroVQA term frequency comparison across the three corpora (MicroVQA, full OmniScience, and the selected 20k subset), demonstrating that the filtered subset exhibits a term distribution substantially closer to MicroVQA than the unfiltered corpus; **(2)** the Zipf's law distribution [49, 50], confirming that the filtered corpus preserves the rank–frequency profile of MicroVQA; **(3)** the relevance score distribution with the selection threshold; and **(4)** the vocabulary coverage–relevance trade-off curve, showing that $K = 20{,}000$ achieves 89.2% coverage of the MicroVQA domain vocabulary, representing a favorable balance between knowledge breadth and corpus compactness.

H Case study

To gain deeper insight into how MICROWORLD's knowledge context (KC) influences model reasoning beyond aggregate accuracy numbers, we conduct a systematic qualitative analysis of Qwen3-VL-8B-Instruct on MicroVQA under the Two-Pass inference paradigm. Specifically, we compare the model's predictions with and without KC augmentation, and categorize every answer flip into one of two groups: *upgrades* (baseline incorrect $\rightarrow$ KC-augmented correct) and *downgrades* (baseline correct $\rightarrow$ KC-augmented incorrect). From the full set of flipped answers, we select five representative cases from each group that illustrate distinct and recurring mechanisms.

For each case, we present the original question, the microscopy image, the baseline and KC-augmented predictions, the relevant knowledge retrieved from the MAPG, and a mechanistic analysis

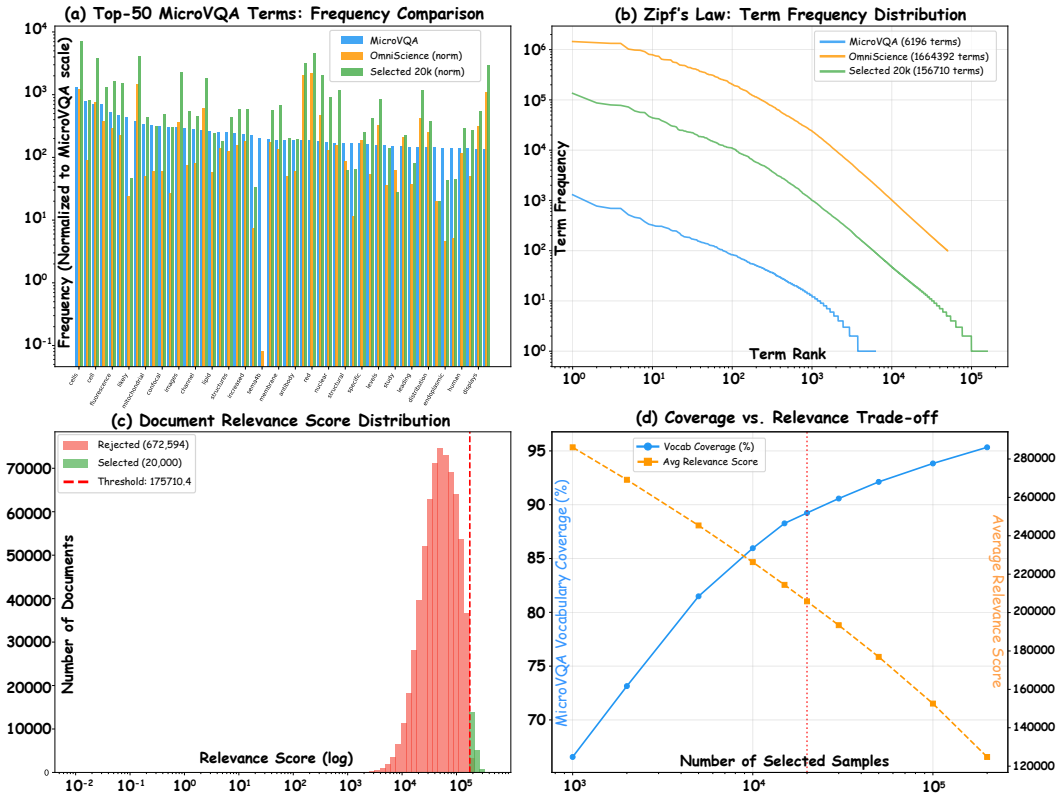
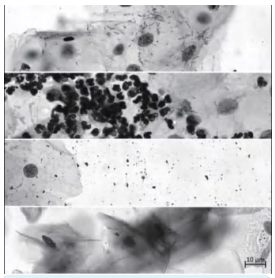


Figure 4: **Corpus filtering analysis for MICROWORLD subset selection. (a) Top-50 MicroVQA Terms: Frequency Comparison.** Term frequency distributions of the top-50 MicroVQA terms across three corpora, MicroVQA questions, the full OmniScience corpus, and the filtered 20k subset, showing that the selected subset closely mirrors the MicroVQA term distribution whereas the unfiltered corpus diverges substantially. **(b) Zipf's Law: Term Frequency Distribution.** Rank–frequency plots for the three corpora, confirming that the filtered subset preserves the power-law rank–frequency profile of MicroVQA and does not distort its linguistic statistics. **(c) Image-caption pairs Relevance Score Distribution.** Histogram of TF-IDF-based relevance scores over all OmniScience image-caption pairs, with the selection threshold (top-$K = 20{,}000$) indicated; the heavy-tailed distribution motivates threshold-based selection to retain only domain-aligned image-caption pairs. **(d) Coverage vs. Relevance Trade-off.** MicroVQA vocabulary coverage as a function of subset size K; $K = 20{,}000$ achieves 89.2% coverage of MicroVQA domain vocabulary while maintaining corpus compactness, representing a favorable balance between knowledge breadth and indexing cost.

explaining *why* the knowledge helped or hurt. The upgrade cases illustrate five mechanisms: *terminology disambiguation*, *domain-specific knowledge injection*, *reasoning level elevation*, *misconception correction*, and *logical chain correction*. The downgrade cases illustrate five failure modes: *knowledge overriding visual evidence*, *context mismatch*, *ambiguous KC signals*, *knowledge-induced over-reasoning*, and *visual–textual confidence imbalance*. These patterns provide actionable insights for improving knowledge-augmented reasoning systems in future work.

H.1 Upgrade cases: Knowledge enabling correct reasoning

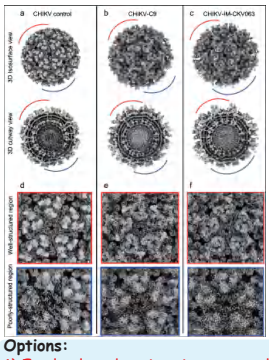


Options:
1) Stain Aggregates; ✗
2) Cellular Debris;
3) Optical Aberrations;
4) Residual Fixatives;
5) Dye Specks; ✓

Questions:
A bright-field microscopy image of a Pap-smear sample displays various artifacts, including bacilli, immune cells, particulate matter, blurred regions, and staining anomalies. Which of the following artifacts is primarily caused by staining processes and can introduce visual noise in the images?

Relevant Biological Knowledge
1. **Pap smear** [structure]
(matched from: \"Pap-smear\")_\n Definition: A Pap smear is a cytology sample collected from the cervix, typically containing exfoliated squamous and glandular epithelial cells suspended on a glass slide or in liquid-based cytology medium. ... Dye specks are non-biological particles generated during staining and are a common source of visual noise. ...
2. **rod-shaped bacteria** [organism]
(matched from: \"bacilli\")_\n Definition: Rod-shaped bacteria are small prokaryotic cells with an elongated morphology commonly seen as tiny short rods in brightfield, phase contrast, or fixed-sample microscopy. ...
3. **immune cells** [organism]\n Definition: Immune cells are host-derived cells of the innate and adaptive immune systems that detect, engulf, signal against, or kill pathogens and infected cells. In microscopy contexts, the term may refer broadly to leukocytes or more specific populations such ...[truncated]

Figure 5: **Upgrade Case 1** (Perception): *Terminology disambiguation.* **Question**: In a Pap-smear bright-field image, which artifact is primarily caused by staining and introduces visual noise? **Baseline** selects Option 1 (Stain Aggregates) ✗; **KC-augmented** selects Option 5 (Dye Specks) ✓. The KG provides: *"Dye specks are non-biological particles generated during the staining process and are a common source of visual noise."* While the baseline hesitates between two semantically similar options and defaults to the more visually prominent one, KC supplies a precise terminological definition that resolves the ambiguity.

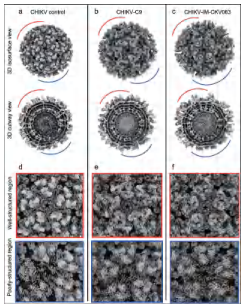


Options:
1) Employ low-dose imaging protocols to reduce sample alterations; ✗
2) Refine contrast enhancement methods to improve visibility of details;
3) Optimize sample growth conditions prior to cryoEM analysis;
4) Adjust image processing algorithms to standardize outputs
5) Implement enhanced purification steps to maintain sample integrity; ✓

Questions:
During a cryoEM evaluation of chikungunya virus (CHIKV) particles, scientists have noted unexpected asymmetry in the structural output. Consider which of the following strategies would likely help determine if this asymmetry results from an artifact involved in the imaging process.

Relevant Biological Knowledge
1. **cryogenic electron microscopy** [method]\n _(matched from: \"cryoEM\")_\n Definition: Cryogenic electron microscopy is an electron imaging method in which biological or other beam-sensitive specimens are rapidly vitrified and imaged at cryogenic temperatures, preserving near-native structure without chemical fixation or staining. It is commonly used for single-particle analysis, cryo-electron tomography, and microcrystal electron diffraction, and is often abbreviated cryoEM. ...
2. **Chikungunya virus** [organism]\n Definition: Chikungunya virus (CHIKV) is an enveloped, positive-sense single-stranded RNA virus in the genus Alphavirus, family Togaviridae. It is transmitted primarily by Aedes mosquitoes and causes chikungunya fever, typically characterized by fever, rash, and severe joint pain. In structural biology and microscopy, CHIKV is studied for its icosahedral virion architecture and envelope glycoproteins involved in host-cell entry. Chikungunya virus is an alphavirus whose mature virion is classically described as having icosahedral organization. In structural studies, deviation from this expected icosahedral framework can indicate virion plasticity or noncanonical particle architecture rather than only local epitope-level conformational change. As an enveloped alphavirus studied by cryoEM, CHIKV particles can show reconstruction variability when preparations contain damaged, incomplete, or heterogeneous virions. In structural virology, chikungunya virus is an alphavirus typically described as having an ordered, roughly icosahedral virion organization with surface glycoprotein spikes arranged over a symmetric particle. CryoEM observations that do not show the expected icosahedral/20-faced framework can be interpreted as evidence of virion structural plasticity or adaptable geometry under specific conditions, including antibody binding. CHIKV is an enveloped alphavirus whose virion is typically treated as having icosahedral organization in structural studies. \n...[truncated]

Figure 6: **Upgrade Case 2** (Experiment Proposal): *Domain-specific knowledge injection.* **Question**: Unexpected asymmetry observed in cryo-EM of Chikungunya virus (CHIKV), which strategy best determines whether this is an artifact? **Baseline** selects Option 1 (low-dose imaging to reduce radiation damage) ✗; **KC-augmented** selects Option 5 (enhanced purification to maintain sample integrity) ✓. The KG provides CHIKV-specific knowledge: *"CHIKV particles can show reconstruction variability when preparations contain damaged, incomplete, or heterogeneous virions. Sample quality and purification are therefore important controls."* The baseline applies generic cryo-EM reasoning (radiation damage as the default artifact source), whereas KC injects entity-specific facts indicating that sample purification quality is the key control variable for this enveloped virus.

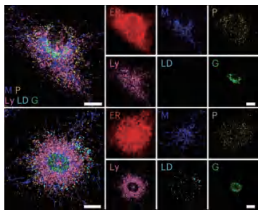


Options:
1) Virus surface proteins undergo structural changes upon antibody exposure, potentially affecting symmetries.; ✗
2) Traditional detection methods may inaccurately portray viruses due to inherent technical constraints.;
3) The geometric integrity of alphaviruses such as chikungunya could be more adaptable than initially thought.; ✓
4) Data interpretation techniques might compromise reliability when assessing certain antibody responses.;
5) The standard cryoEM technique might occasionally overestimate viral structural variations.;

Questions:
The investigative application of cryo-electron microscopy (cryoEM) to chikungunya virus structures, exposed to neutralizing antibodies, led to an unexpected conclusion: absence of the usual 20-sided framework. What might this suggest about the observed viral configuration under these specific conditions?

Relevant Biological Knowledge
1. **cryogenic electron microscopy** [method]\n _(matched from: \"cryo-electron microscopy\")_\n Definition: Cryogenic electron microscopy is an electron imaging method in which biological or other beam-sensitive specimens are rapidly vitrified and imaged at cryogenic temperatures, preserving near-native structure without chemical fixation or staining. It is commonly used for single-particle analysis, cryo-electron tomography, and microcrystal electron diffraction, and is often abbreviated cryoEM.
2. **Chikungunya virus** [organism]\n Definition: Chikungunya virus (CHIKV) is an enveloped, positive-sense single-stranded RNA virus in the genus Alphavirus, family Togaviridae. It is transmitted primarily by Aedes mosquitoes and causes chikungunya fever, typically characterized by fever, rash, and severe joint pain. In structural biology and microscopy, CHIKV is studied for its icosahedral virion architecture and envelope glycoproteins involved in host-cell entry. Chikungunya virus is an alphavirus whose mature virion is classically described as having icosahedral organization. As an enveloped alphavirus studied by cryoEM, CHIKV particles can show reconstruction variability when preparations contain damaged, incomplete, or heterogeneous virions. Sample quality and purification are therefore important controls when interpreting unexpected asymmetry. In structural virology, chikungunya virus is an alphavirus typically described as having an ordered, roughly icosahedral virion organization with surface glycoprotein spikes arranged over a symmetric particle. CryoEM observations that do not show the expected icosahedral/20-faced framework can be interpreted as evidence of virion structural plasticity or adaptable geometry under specific conditions. CHIKV is an enveloped alphavirus whose virion is typically treated as having icosahedral organization in structural studies. Therefore, reports of asymmetric features in cryo-EM reconstructions often require contr\n...[truncated]",

Figure 7: **Upgrade Case 3** (Perception): *Reasoning level elevation.* **Question**: After neutralizing antibody treatment, CHIKV loses icosahedral symmetry. What does this imply? **Baseline** selects Option 1 (surface protein conformational changes affect symmetry) ✗; **KC-augmented** selects Option 3 (alphavirus geometric integrity is more adaptive than previously thought) ✓. The baseline produces a locally correct but suboptimal "mechanistic" explanation (protein conformational change), while KC's alphavirus structural knowledge elevates reasoning to a higher-level biological insight, that viral geometric integrity itself exhibits structural plasticity, rather than being merely a consequence of local protein changes.

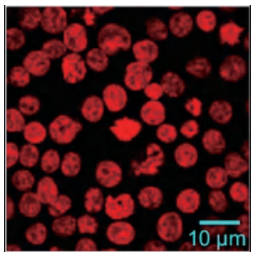


Questions:
A fluorescence microscopy image depicts bone marrow-derived macrophages stained for six organelles before and after exposure to heat-killed S. aureus. After exposure, lysosomes are observed to position themselves in a circular pattern. What is the most likely reason for this reorganization of lysosomes?

Options:
1) To enhance the secretion of lysosomal enzymes;
2) To assemble multi-organellar units that aid in adapting to foreign pathogens; ✅
3) To increase the fusion rate with phagosomes; ❌
4) To initiate apoptotic signaling pathways

Relevant Biological Knowledge
1. **BMDM** [cell_type]\n _(matched from: \"bone marrow-derived macrophages\")_\n Definition: Bone marrow-derived macrophages (BMDMs) are primary immune cells isolated from murine bone marrow and differentiated in vitro using colony-stimulating factor 1 (CSF-1). These cells serve as a physiologically relevant model for studying innate immunity, phagocytosis, and inflammatory signaling pathways due to their ability to polarize into distinct functional states. In microscopy research, BMDMs are frequently utilized to visualize dynamic cellular processes such as lysosomal trafficking, pathogen engulfment, and cytoskeletal reorganization through high-resolution live-cell imaging and immunofluorescence techniques.
2. **lysosome** [structure]\n _(matched from: \"lysosomes\")_\n Definition: A lysosome is a membrane-bound acidic organelle found in animal cells that contains hydrolytic enzymes for degrading macromolecules, damaged organelles, and engulfed material. It functions in intracellular digestion, recycling, autophagy, and cellular homeostasis. In microscopy, lysosomes are typically seen as small punctate vesicular structures in the cytoplasm and can be labeled with markers such as LAMP1, LAMP2, or LysoTracker. In infection or microbial-challenge imaging, lysosomes may undergo conspicuous spatial reorganization, including clustering or ring-like arrangements, as part of broader organelle remodeling. Such patterns can reflect participation in multi-organellar defensive assemblies. In infection or stress-imaging experiments, altered lysosome number, size, or distribution can reflect a broad adaptive response and is not by itself specific for autophagy. \n... [truncated]",

Figure 8: **Upgrade Case 4** (Hypothesis Generation): *Misconception correction.* **Question**: After macrophage exposure to heat-killed *S. aureus*, lysosomes exhibit ring-like arrangements. What is the underlying cause? **Baseline** selects Option 3 (increased phagosome fusion rate) ✗; **KC-augmented** selects Option 2 (assembly of multi-organellar cooperative defense units) ✓. The KG definition of lysosomes explicitly states: *"ring-like arrangements reflect participation in multi-organellar defensive assemblies and should not automatically be interpreted only as increased lysosome–phagosome fusion."* The baseline relies on textbook-level reasoning (lysosome reorganization ≈ increased fusion), while KC provides a domain-precise correction that the ring pattern reflects broader organelle remodeling for cooperative defense.



Questions:
In a confocal microscopy image, cells deficient in mtDNA display mitochondria-associated membranes (MAM) and show actin predominantly at the cell periphery. What is the most probable cause for this distinct actin distribution?

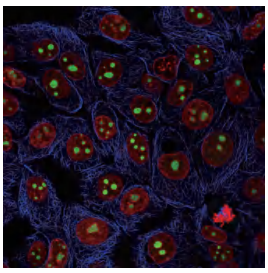
Options:
1) Increased cortical actin due to external stimuli promoting cell surface reorganization; ✅
2) Localized actin adaptations for efficient internal nutrient transport;
3) Redistribution of actin to conserve energy resources under stress; ❌
4) Enhanced actin dynamics to maintain structural equilibrium on the cell perimeter;
5) Reinforced actin positioning to support intracellular signaling pathways targeting organelles",

Relevant Biological Knowledge
1. **mitochondria-ER contact site** [structure]\n _(matched from: \"mitochondria-associated membranes\")_\n Definition: A mitochondria-ER contact site is a region where the endoplasmic reticulum membrane lies in close apposition to the mitochondrial membrane without fusion. These sites mediate inter-organelle communication, including calcium transfer, lipid exchange, and organelle dynamics, and are often quantified in microscopy studies.
2. **actin** [structure]\n Definition: Actin is a highly conserved cytoskeletal protein that polymerizes into microfilaments and plays central roles in cell shape, motility, adhesion, cytokinesis, and intracellular organization. In microscopy and cell biology, actin usually refers to filamentous actin (F-actin) or total cellular actin stained to visualize the cytoskeleton, often enriched at the cell cortex and stress fibers. Multiple actin isoforms exist in eukaryotic cells, including cytoplasmic and muscle-specific forms. In fluorescence microscopy, actin concentrated along the cell edge usually corresponds to cortical actin beneath the plasma membrane. Such peripheral enrichment is commonly associated with cell spreading, adhesion, and surface reorganization. Actin abundance or fluorescence intensity is commonly used as a downstream cytoskeletal readout in confocal microscopy. Changes in actin signal after receptor-ligand stimulation are often evaluated after perturbing the receptor to test whether the receptor is required for the response. In spheroid dispersion assays, actin staining is commonly used to visualize changes in cell morphology and migration, including elongation, protrusions, and outward spreading from a compact cell cluster.\n Direct relations:\n • SIRT1 [gene] (relation: co-occurs with)\n SIRT1 (Sirtuin 1) is a NAD+-dependent class III histone deacetylase\n...[truncated]",

Figure 9: **Upgrade Case 5** (Hypothesis Generation): *Logical chain correction.* **Question**: In mtDNA-deficient cells, actin concentrates at the cell periphery. What is the most likely cause? **Baseline** selects Option 3 (actin redistribution to conserve energy) ✗; **KC-augmented** selects Option 1 (external stimuli promote cell surface remodeling, increasing cortical actin) ✓. The baseline follows the plausible but incorrect causal chain: mitochondrial dysfunction → energy deficit → energy conservation. KC about actin and cytoskeleton reveals that actin polymerization is itself energy-intensive, invalidating the "conservation" hypothesis. The correct explanation is that cortical actin enrichment reflects normal cell surface remodeling, which KC helps the model recognize.

H.2 Downgrade cases: Knowledge inducing incorrect reasoning

While MICROWORLD improves overall accuracy substantially, a non-trivial fraction of answer flips go in the wrong direction. Understanding these failure modes is crucial for guiding future improvements. The five cases below illustrate how retrieved knowledge can mislead reasoning through different mechanisms, from overriding correct visual judgments with irrelevant textual noise to misapplying knowledge from a different experimental context. Notably, several of these failures stem not from incorrect knowledge in the graph, but from the model's inability to properly weigh, contextualize, or arbitrate between visual evidence and textual knowledge.

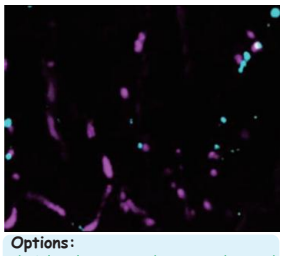


Options:
1) Nucleoli; ✓
2) Nuclear speckles;
3) Nuclear pores;
4) Chromatin aggregates ✗

Questions:
A confocal fluorescence microscopy image of cultured human MCF-7 cells is shown. The cells were fixed with 4% PFA and stained for nuclei (DAPI, blue), microtubules (tubulin antibody, red), and a protein of interest (green) using an indirect ICC/IF staining protocol. Large green patches are observed in most nuclei. To which subcellular structure does the protein of interest most likely localize?

Relevant Biological Knowledge
1. **MCF-7** [cell_line]\n Definition: MCF-7 is a human breast adenocarcinoma cell line derived from the pleural effusion of a 69-year-old female, characterized by its estrogen receptor-positive (ER+) status and luminal A molecular subtype.
2. **DAPI stain** [method]\n _(matched from: \"DAPI\")_\n Definition: DAPI stain is a fluorescent nuclear counterstain used to label DNA, especially cell nuclei, in fixed-cell and tissue fluorescence microscopy.
3. **tubulin** [structure]\n Definition: Tubulin is a family of globular proteins, primarily alpha- and beta-tubulin, that polymerize to form microtubules in eukaryotic cells. Microtubules are major components of the cytoskeleton and are essential for cell shape, intracellular transport, mitosis, and cilia/flagella function. In fluorescence microscopy, 'tubulin' staining usually labels the microtubule network or mitotic spindle. In immunofluorescence microscopy, tubulin staining visualizes the microtubule cytoskeleton and can show organization, spindle formation, or cell shape context. \n...[truncated]",

Figure 10: **Downgrade Case 1** (Perception): *Knowledge overriding visual evidence.* **Question**: Identify the subcellular localization of green patches within MCF-7 cell nuclei. **Baseline** correctly selects Option 1 (Nucleoli) ✓, based on direct visual assessment of bright, discrete intranuclear foci consistent with nucleolar morphology. **KC-augmented** incorrectly selects Option 4 (Chromatin aggregates) ✗. KC contains no information about the specific protein's localization, but provides generic knowledge about MCF-7 cells, DAPI staining, and confocal microscopy. This textual noise causes the model to second-guess its visual judgment, reinterpreting the signal pattern as "not confined to discrete structures" and switching to chromatin aggregates. **Lesson**: when KC lacks query-relevant knowledge, generic context can erode the model's visual confidence.



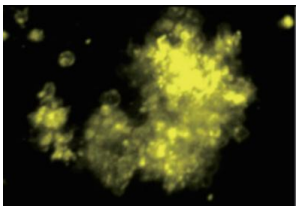
Options:
1) Colocalization indicates enhanced wound closure through potentially novel biochemical interactions.; ✅
2) The overlapping signals are likely due to misalignment of imaging channels, without biological value.;
3) The presence of these biomarkers could represent cross-reactivity, complicating precise labeling.; ❌
4) This pattern suggests a misinterpretation, potentially leading to incorrect assumptions in data analysis.;
5) There is a heightened immune response involving CD41 elements, triggering migratory behavior.

Questions:
In images of hepatic tissue obtained via fluorescence microscopy, we observe colocalization of magenta-marked CD41 elements and yellow-labeled Ly6G structures at the boundary of tissue repair. Without the specific visual information, what might this colocalization imply about interactions at the injury site?

Relevant Biological Knowledge
1. **CD41** [structure]\n Definition: This gene encodes a member of the integrin alpha chain family of proteins. The encoded preproprotein is proteolytically processed to generate light and heavy chains that associate through disulfide linkages to form a subunit of the alpha-IIb/beta-3 integrin cell adhesion receptor. This receptor plays a crucial role in the blood coagulation system, by mediating platelet aggregation. Mutations in this gene are associated with platelet-type bleeding disorders, which are characterized by a failure of platelet aggregation, including Glanzmann thrombasthenia.

2. **Ly6G** [structure]\n Definition: Ly6G is a glycosylphosphatidylinositol (GPI)-anchored cell- surface protein widely used as a canonical marker of murine neutrophils and neutrophil-derived material. In microscopy, Ly6G labeling most often identifies neutrophil plasma membranes, cell bodies, protrusions, extracellular vesicles, or migrasome-like structures. It is primarily a marker reagent in imaging and flow cytometry rather than a universal structural organelle marker. Ly6G is used primarily as a neutrophil marker in mouse immunostaining. Detection of Ly6G signal in platelets or other non-neutrophil cells is atypical and, unless independently validated, should raise concern for antibody cross- reactivity or nonspecific staining before concluding true expression or uptake. When Ly6G is imaged in unwounded liver, signal is usually low or sparse relative to injured tissue. Minor spatial overlap with vascular or platelet-associated markers may occur from proximity or projection in tissue sections and is not sufficient evidence of a functional neutrophil-platelet interaction. This interpretation is often more plausible than de novo Ly6G expression by platelets. Ly6G is widely used as a neutrophil marker in mouse immunofluorescence microscopy. \n \n...[truncated]",

Figure 11: **Downgrade Case 2** (Perception): *Context mismatch.* **Question**: CD41/Ly6G co-localization at a hepatic injury border, what does it indicate? **Baseline** correctly selects Option 1 (novel biochemical interaction enhancing wound healing) ✓. **KC-augmented** incorrectly selects Option 3 (antibody cross-reactivity artifacts) ✗. KC states: *"Detection of Ly6G signal in platelets should raise concern for antibody cross-reactivity."* This advice originates from in vitro platelet preparation quality control, but the question concerns in vivo tissue repair imaging, where CD41–Ly6G co-localization reflects genuine platelet–neutrophil cooperative repair. The model misapplies quality-control guidance from a different experimental context.

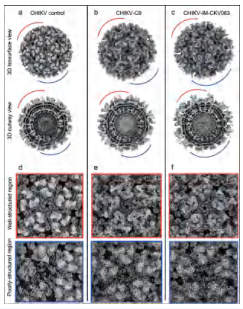


Options:
1) Antibody cross-reactivity causing Ly6G binding to unwanted targets; ✅
2) Internal synthesis of Ly6G in platelets under certain stimuli;
3) Platelets engulfing neutrophil byproducts causing Ly6G presence; ❌
4) Technical imaging errors leading to misinterpretation;
5) Experimental artifacts from chemical reagents used in analysis;

Questions:
A confocal microscopy image displays eukaryotic cells with CD41 tagged in turquoise and Ly6G in lavender. In the cells, the CD41 label identifies platelets clearly visible in image segment 0. Surprisingly, segment 1 with Ly6G label displays localized signal within these platelets. What hypothesis could explain this unexpected Ly6G representation in platelets?

Relevant Biological Knowledge
1. **Ly6G** [structure]\n Definition: Ly6G is a glycosylphosphatidylinositol (GPI)-anchored cell-surface protein widely used as a canonical marker of murine neutrophils and neutrophil-derived material. In microscopy, Ly6G labeling most often identifies neutrophil plasma membranes, cell bodies, protrusions, extracellular vesicles, or migrasome-like structures. It is primarily a marker reagent in imaging and flow cytometry rather than a universal structural organelle marker. Ly6G is used primarily as a neutrophil marker in mouse immunostaining. Detection of Ly6G signal in platelets or other non-neutrophil cells is atypical and, unless independently validated, should raise concern for antibody cross-reactivity or nonspecific staining before concluding true expression or uptake. When Ly6G is imaged in unwounded liver, signal is usually low or sparse relative to injured tissue. Minor spatial overlap with vascular or platelet-associated markers may occur from proximity or projection in tissue sections and is not sufficient evidence of a functional neutrophil-platelet interaction. This interpretation is often more plausible than de novo Ly6G expression by platelets. Ly6G is widely used as a neutrophil marker in mouse immunofluorescence microscopy. Apparent overlap of Ly6G with platelet markers such as CD41 can reflect true platelet-neutrophil aggregates or platelet-associated neutrophil-derived vesicles. \n
\n...[truncated]",

Figure 12: **Downgrade Case 3** (Hypothesis Generation): *Ambiguous KC signals.* **Question**: Unexpected Ly6G signal detected in isolated platelets, what is the most likely explanation? **Baseline** correctly selects Option 1 (antibody cross-reactivity) ✓. **KC-augmented** incorrectly selects Option 3 (platelet phagocytosis of neutrophil-derived material) ✗. KC simultaneously contains: (i) *"should raise concern for antibody cross-reactivity"* (supporting Option 1) and (ii) *"can reflect true platelet–neutrophil aggregates or platelet-associated neutrophil-derived vesicles"* (supporting Option 3). Faced with conflicting signals, the model favors the more "novel" biological explanation from the KG over the more fundamental but correct answer.

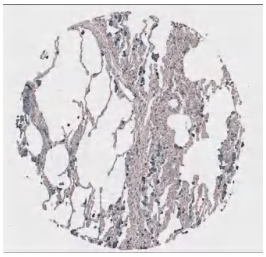


Questions:
In a cryo-electron microscopy (cryoEM) investigation of chikungunya virus (CHIKV) particles, asymmetrical structures were observed both in the absence and presence of neutralizing antibodies. To assess whether this asymmetry is a result of artifacts introduced during symmetry-free data processing, which experimental method would be most suitable?

Options:
1) Implement a different image reconstruction algorithm that enforces symmetrical constraints during data processing.; ✗
2) Conduct site-directed mutagenesis to modify the virus particles and observe resulting structural changes.;
3) Perform simulations with icosahedrally symmetric particles overlaid with noise and assess the effects of symmetry relaxation.; ✓
4) Utilize focused ion beam (FIB) milling to prepare samples and reduce potential artifacts in imaging.;
5) Integrate single-particle analysis techniques combined with cross-validation against known symmetric structures.

Relevant Biological Knowledge
1. **cryogenic electron microscopy** [method]\n _(matched from: \"cryo-electron microscopy\")_\n Definition: Cryogenic electron microscopy is an electron imaging method in which biological or other beam-sensitive specimens are rapidly vitrified and imaged at cryogenic temperatures, preserving near-native structure without chemical fixation or staining. It is commonly used for single-particle analysis, cryo-electron tomography, and microcrystal electron diffraction, and is often abbreviated cryoEM. In structural virology, cryoEM is commonly used to evaluate whether virions conform to expected symmetries such as icosahedral organization. Failure to observe the expected symmetry can support conclusions about particle heterogeneity or structural flexibility. In TEM, cryoEM images of unstained proteins embedded in vitreous ice often appear relatively grainy and low in contrast, with diminished definition compared with negative staining. Background may include diffuse amorphous density from the ice rather than sharply outlined stain-filled silhouettes.
2. **Chikungunya virus** [organism]\n Definition: Chikungunya virus (CHIKV) is an enveloped, positive-sense single-stranded RNA virus in the genus Alphavirus, family Togaviridae. It is transmitted primarily by Aedes mosquitoes and causes chikungunya fever, typically characterized by fever, rash, and severe joint pain. In structural biology and microscopy, CHIKV is studied for its icosahedral virion architecture and envelope glycoproteins involved in host-cell entry. Chikungunya virus is an alphavirus whose mature virion is classically described as having icosahedral organization. In structural studies, deviation from this expected icosahedral framework can indicate virion plasticity or noncanonical particle architecture rather than only local epitope-level conformational change. As an enveloped alphavirus studied by cryoEM, CHIKV particles can show reconstruction variability when preparations contain damaged, incomplete, or heterogeneous virions. Sample quality and purification are therefore important controls when interpreting unexpected asymmetry. In structural virology, chikungunya virus is an alphavirus typically described as having an ordered, roughly icosahedral virion organization with surface glycoprotein spikes arranged over a symmetric particle. CryoEM observations that do not show the expected icosahedral/20-faced framework can be interpreted as evidence of virion structural plasticity or adaptable geometry under specific conditions, including antibody binding. CHIKV is an enveloped alphavirus whose virion is typically treated as having icosahedral organization in structural studies. Therefore, reports of asymmetric features in cryo-EM reconstructions often require contr\n...[truncated]",

Figure 13: **Downgrade Case 4** (Experiment Proposal): *Knowledge-induced over-reasoning.* **Question**: How to verify whether the observed cryo-EM asymmetry in CHIKV is an artifact? **Baseline** correctly selects Option 3 (simulate icosahedral particles with added noise to evaluate symmetry relaxation effects) ✓. **KC-augmented** incorrectly selects Option 1 (use reconstruction algorithms with enforced symmetry constraints) ✗. The baseline correctly reasons that forward simulation is the methodologically sound approach to test the artifact hypothesis. KC provides detailed cryo-EM processing knowledge, leading the model to prefer a superficially more direct strategy (enforced symmetry). However, enforcing symmetry only eliminates asymmetry without determining its origin, a critical distinction between "testing a hypothesis" and "avoiding the problem."



Questions:
A light microscopy image of liver tissue is stained with hematoxylin (blue) and diaminobenzidine (DAB, brown) to visualize specific proteins. Several areas of the tissue display dense clusters of small cells with spherical, dense nuclei and minimal cytoplasm. What is the most likely cell type forming these clusters, and what is the biological significance of their accumulation?

Options:
1) Hepatocytes; indicating enhanced metabolic activity;
2) Kupffer cells; suggesting increased phagocytic activity; ✗
3) Lymphocytes; indicating an inflammatory reaction; ✓
4) Stellate cells; implying fibrosis development

Relevant Biological Knowledge
1. **mouse liver** [structure]\n _(matched from: \"liver tissue\")_\n Definition: Mouse liver is the liver organ of the mouse, a major abdominal organ composed mainly of hepatocytes along with biliary cells, endothelial cells, Kupffer cells, stellate cells, and immune cells. It performs key functions including metabolism, detoxification, bile production, glycogen storage, and regulation of lipid and protein homeostasis. In biological and microscopy research, mouse liver is commonly studied to investigate hepatic architecture, vascular organization, inflammation, infection, regeneration, and metabolic disease. In fluorescence microscopy, mouse liver images often label multiple compartments simultaneously, including immune cells such as neutrophils or Kupffer cells and vascular structures such as blood vessels or sinusoids.
2. **minimal cytoplasm** [property]\n Definition: Minimal cytoplasm refers to cells in which the cytoplasmic compartment is scant and occupies very little space relative to the nucleus, producing a high nucleus-to-cytoplasm ratio. Morphologically, these cells appear small, crowded, and basophilic, with nuclei that dominate the cell profile and only a thin rim of surrounding cytoplasm visible on light micr\n...[truncated]

Figure 14: **Downgrade Case 5** (Perception): *Visual–textual confidence imbalance.* **Question**: Identify small, densely clustered cells in liver tissue. **Baseline** correctly selects Option 3 (lymphocytes indicating inflammatory response) ✓, accurately recognizing morphological features: small, round, high nuclear-to-cytoplasmic ratio, and dense clustering. **KC-augmented** incorrectly selects Option 2 (Kupffer cells indicating increased phagocytic activity) ✗. KC provides detailed knowledge about Kupffer cells (liver-resident macrophages), which induces the model to reinterpret the observed cells as Kupffer cells despite their morphology being inconsistent (Kupffer cells are larger and irregularly shaped). The presence of domain knowledge reduces the model's trust in its own visual evidence.

I Supplementary experimental controls

This appendix details the corpus-selection protocol and completed controls for graph retrieval, backbone transfer, and knowledge presentation.

I.1 Corpus-selection protocol

Table 5 compares 5K-pair graphs on MicroVQA using Qwen3-VL-8B, Two-Pass. We uniformly sampled 5K image–caption pairs from OmniScience without MicroVQA contact and reran triplet extraction, graph construction, definition generation, and retrieval-augmented inference. Random uses the initial definition set; Random (+ top-6K definitions) adds definitions for the top-6K corpus-frequent entities in the same graph. Filtered (seed 42) uniformly subsamples the vocabulary-filtered 20K corpus, rebuilding the graph from the sampled pairs' original GPT-5 triplets and reusing their definitions. Both random conditions use no MicroVQA information; the filtered condition inherits benchmark-aligned corpus selection.

All gains use the 48.8% zero-shot reference. The unrounded filtered and definition-enriched random accuracies are 59.31% and 55.95%, respectively, a 3.36-point difference. Their unequal definition coverage (82.8% versus 53.6%) prevents attributing this difference solely to vocabulary alignment. The approximately 7.8-point gap between filtered-5K and the full filtered-20K result (67.1%) is a descriptive scale comparison. These controls support benefits beyond benchmark-aligned selection but do not isolate its causal contribution.

I.2 2-hop and visual-edge control

We vary the 2-hop neighbor budget (`max_2hop`) and `visually_similar` edges in a complete 2×2 control on MicroVQA (1,042 questions, Qwen3-VL-8B, Two-Pass). A budget of 3 caps the number of 2-hop neighbors; it does not denote three-hop traversal.

Table 9: 2-hop $\times$ visual-edge control.

2-hop budget	Visual edges	Accuracy (%)
3 (default)	On	67.1
0	On	67.9
3	Off	68.6
0	Off	68.9

Removing 2-hop expansion changes accuracy by +0.8 pp, removing visual edges by +1.5 pp, and removing both by +1.8 pp. The difference-in-differences is $68.9 - 68.6 - 67.9 + 67.1 = -0.5$ pp. These observed differences describe non-additivity.

I.3 GPT-5 context protocol

The two MAPG-augmented GPT-5 conditions in Table 1 cover all 1,042 MicroVQA questions using the same prompt format as the open-source runs. Shared context directly reuses the retrieved context supplied to Qwen3-VL-8B Two-Pass, yielding 71.5%. Own extraction uses GPT-5 for Round-1 entity extraction before retrieval from the same MAPG, yielding 73.3%. The 1.8-point difference compares retrieval conditioned on different query extractors, with the answer model held fixed.

I.4 Text-only and knowledge-format controls

Table 10 reports the completed Qwen3-VL-8B Two-Pass controls on MicroVQA. Text-only KG removes visual entities and `visually_similar` edges while retaining the text pathway. The flat-paragraph condition retains the same retrieved entities, definitions, and relations as full MAPG and changes only their presentation.

On MicroBench, full MAPG and text-only KG reach 58.54% and 58.15%, respectively, a 0.39-point difference under the same 8B Two-Pass configuration. These results indicate modest visual-pathway contributions in the evaluated settings. The joint text-only change on MicroVQA (-1.1 pp) differs

Table 10: MicroVQA knowledge controls (8B, Two-Pass).

Condition	Accuracy (%)	Δ (pp)
Full MAPG, structured blocks	67.1	0.0
Text-only KG	66.0	-1.1
Full MAPG, flat paragraph	66.9	-0.2

from the sum of the individual visual-entity and visual-edge changes ($-1.3+1.5 = +0.2$ pp; Table 3). This comparison varies different factors from Table 9.

The 0.2-point flat-format difference holds knowledge content fixed and supports a limited sensitivity to presentation format in this setting.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All of the datasets involved in this work are open-source and accessible. All the experimental results are reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All of the codes and datasets involved in this work are open-source and accessible.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Our method is training-free, we have specified all the test details in Appendix D.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not report error bars or confidence intervals in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed broader impacts in Appendix A, B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No model in this paper has a high risk of misuse. The collected datasets are all open-source and accessible.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets utilized in this paper have been properly attributed to their creators or original owners. We have explicitly stated all licenses and terms of use, ensuring full compliance with their requirements.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and data are available at https://github.com/ieellee/MicroWorld.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our research method utilizes LLM as a crucial component to implement the innovation we proposed. We have described the usage of LLMs in Section 3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.