Decoding Scientific Experimental Images: The SPUR Benchmark for Perception, Understanding, and Reasoning

Junpeng Ding Zichen Tang Haihong E* Mengyuan Ji Yang Liu Haolin Tian
Haiyang Sun Pengqi Sun Yang Xu Yichen Liu Haocheng Gao Zijie Xi
Ruomeng Jiang Peizhi Zhao Rongjin Li Yuanze Li Jiacheng Liu
Zhongjun Yang Jintong Chen Siying Lin
Beijing University of Posts and Telecommunications

bupt-reasoning-lab.github.io/SPUR

BUPT-Reasoning-Lab/SPUR BUPT-Reasoning-Lab/SPUR

Abstract

We introduce **SPUR**, a comprehensive benchmark for scientific experimental image *perception*, *understanding*, and *reasoning*, comprising 4,264 question-answering (QA) pairs derived from 1,084 expert-curated images. SPUR features three key innovations: (1) **Panel-Level Fine-Grained Perception**: evaluating the visual perception of multimodal large language models (MLLMs) across three dimensions (numerical, morphological, and information localization) on six fine-grained panel types; (2) **Cross-Panel Relation Understanding**: utilizing complex images with an average of 14.3 panels per sample to evaluate MLLMs' ability to decipher intricate cross-panel relations; (3) **Expert-Level Reasoning**: assessment of qualitative and quantitative reasoning across five experimental paradigms to determine if models can infer conclusions from evidence as human experts do. Comprehensive evaluation of 20 MLLMs and four multimodal Chain-of-Thought (MCoT) methods reveals that current models fall significantly short of the expert-level requirements for scientific image interpretation, underscoring a critical bottleneck in AI for Science (AI4S) research.

1 Introduction

Recently, multimodal large language models (MLLMs) have demonstrated remarkable capabilities in ***perceiving*** and ***understanding*** scientific images across diverse domains, including statistical charts (Li et al., 2024b; Roberts et al., 2024; Wang et al., 2024c), tables (Zheng et al., 2024b), biomedical images (Lozano et al., 2024), model schematics (Burgess et al., 2025), and chemical diagrams (Li et al., 2025). This establishes their profound potential for AI for Science (AI4S) (Xu and Peng, 2025). The multimodal ***reasoning*** capacities of MLLMs are further enhanced through multimodal Chain-of-Thought (MCoT) methods, which amplify structured reasoning via (1) prompt-based methods (Zheng et al., 2023; Wu et al., 2024), (2) plan-based decomposition (Zheng et al., 2024a; Gao et al., 2024), and (3) training-based frameworks (Wang et al., 2025a; Zhang et al., 2025).

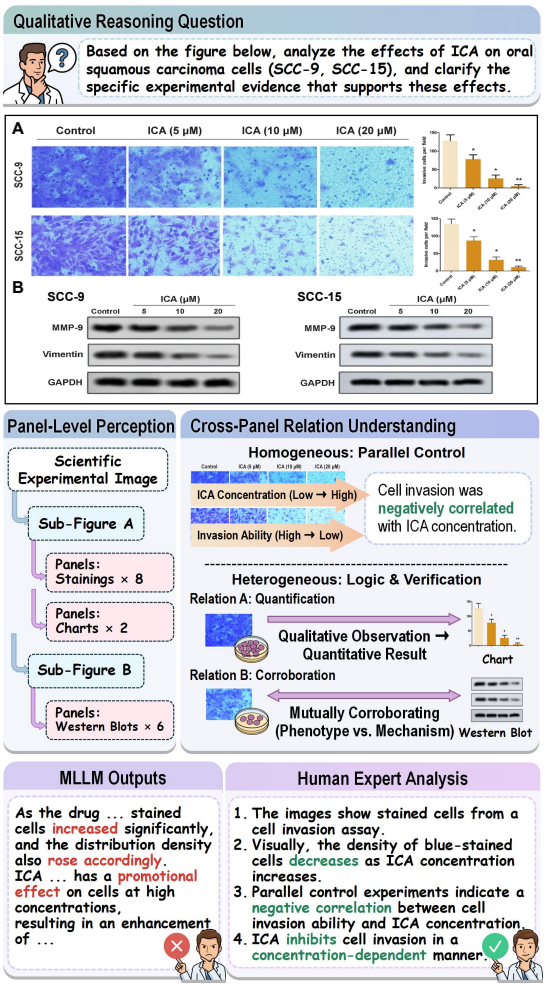


Figure 1: Qualitative Reasoning (Qual.) on a complex multi-panel image. To reach correct conclusions, models must perform (1) panel-level fine-grained perception of morphological features, (2) cross-panel relation understanding, and (3) expert-level reasoning. Current MLLMs struggle with visual trends, leading to hallucinated conclusions compared to human expert analysis.

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12859–12882

Benchmark	Domain	# Samples (k)	Panel-Level Perception		Cross-Panel Understanding			Expert-Level Reasoning		
			Gran.	Dimension	# Panels/Image	Desc.	Rel.	Paradigm	Qual.	Quant.
Non-Academic Images										
ScienceQA	Science	21.2	Figure	N+IL	1.0	✘	✘	VQA	✔	✘
MMMU	Science	11.5	Figure	N+IL	2.5	✘	✘	VQA	✔	✘
M3CoT	Science	11.3	Figure	N+IL	1.0	✘	✘	VQA	✔	✘
EMMA	Science	2.7	Figure	N+IL	1.3	✘	✘	VQA	✔	✔
Academic Images										
MMSci	Science	742.3	Sub-fig.	N+IL	7.4	✔	✘	Caption	✘	✘
SciAssess	Science	6.9	Figure	M+IL	3.4	✘	✘	VQA	✔	✘
SFE	Science	0.8	Sub-fig.	N+M+IL	2.3	✔	✘	VQA	✔	✘
Text2Analysis	Statistics	2.2	Figure	N+IL	1.0	✘	✘	VQA	✔	✔
EvoChart	Statistics	1.2	Figure	N+IL	1.0	✘	✘	VQA	✔	✔
MISS-QA	CS	1.5	Sub-fig.	N+IL	1.6	✘	✘	VQA	✔	✘
SPIQA	CS	270	Sub-fig.	N+IL	2.7	✘	✘	VQA	✔	✘
OmniMedVQA	Medicine	127.9	Sub-fig.	M+IL	1.0	✘	✘	VQA	✔	✘
MicroVQA	Medicine	1.0	Panel	N+M	1.9	✔	✔	Exp. Design	✔	✘
SPUR (ours)	**Experiment**	**4.3**	**Panel**	**N+M+IL**	**14.3**	✔	✔	**Exp. Reason**	**36.1%**	**63.9%**

Table 1: Comparison of SPUR and related benchmarks. For *perception* features, **Gran.**: Granularity; **N**: Numerical; **M**: Morphological; **IL**: Information Localization. For *understanding* capabilities, **Desc.**: Information Description; **Rel.**: Relation Understanding. For *reasoning* tasks, **Qual.**: Qualitative; **Quant.**: Quantitative.

Despite these advancements, complex multi-panel scientific experimental images, which are ubiquitous in scholarly publications, remain underexplored. As shown in Figure 1, such images systematically arrange multiple panels to present experimental processes and results. MLLMs are expected to emulate human experts in achieving: (1) **fine-grained perception** of individual panels, (2) **cross-panel understanding** of intricate relations, and (3) **qualitative and quantitative reasoning** toward scientific conclusions. However, as summarized in Table 1, current scientific reasoning benchmarks exhibit critical limitations in meeting real-world AI4S demands:

- **Absence of Scientific Experimental Images.** Experimental images (e.g., cellular staining images and Western blots) exhibit unique visual characteristics that pose severe fine-grained perception challenges. While benchmarks like ScienceQA (Lu et al., 2022), M3CoT (Chen et al., 2024b), and MMMU (Yue et al., 2024) aggregate diverse scientific imagery for VQA tasks, they severely underrepresent experimental images.

- **Lack of Cross-Panel Relation Understanding.** Interpreting experimental images requires comprehending panel correlations (e.g., control vs. treatment groups), which is a core capability in scientific workflows. However, existing benchmarks like MMSci (Li et al., 2024c), SciAssess (Cai et al., 2025), and SFE (Zhou et al., 2025) predominantly feature isolated images with limited complexity, averaging fewer than eight panels per sample.

- **Neglect of Quantitative Reasoning.** Quantitative reasoning, involving precise analysis of proportions, magnitudes, and metrics, poses substantially greater challenges than qualitative conclusion judgments. Although pioneering works address qualitative aspects (e.g., experimental design in MicroVQA (Burgess et al., 2025)), they largely neglect MLLMs' quantitative reasoning capabilities for deriving scientific conclusions.

To bridge these gaps, we propose **SPUR**, a benchmark for multimodal ***perception***, ***understanding***, and ***reasoning*** on scientific experimental images (Figure 2), comprising 1,084 images and 4,264 QA pairs. SPUR offers three key advantages:

- **Cross-Disciplinary Generalization.** SPUR is the first MLLM benchmark dedicated to scientific experimental image reasoning, filling a critical gap in AI4S evaluation. Distinct from existing benchmarks, it exclusively sources high-quality experimental images from open-access PubMed Central (PMC)[1] papers, covering seven diverse disciplines and five classic experimental paradigms. Notably, these images share consistent visual structures, layout rules, and reasoning logic across fields, ensuring strong cross-disciplinary commonality.

[1] https://pmc.ncbi.nlm.nih.gov

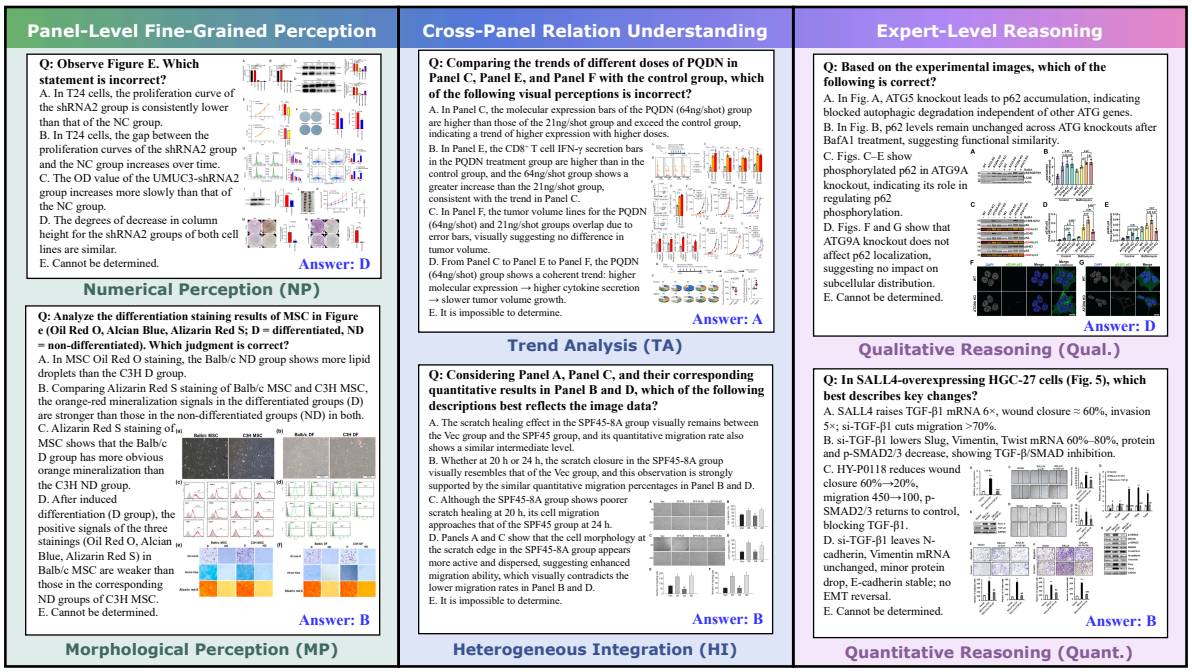


Figure 2: Representative QA pairs across SPUR's three cognitive stages. See Appendix A for more examples.

- **Fine-Grained Panels and Complex Cross-Panel Relations.** SPUR utilizes highly complex images containing an average of 14.3 panels per sample. These are categorized into six fine-grained types (i.e., charts, Western blots, and four subtypes of staining images). Further, we model cross-panel relations, including isomorphic relations (e.g., dose-response trends across staining images) and heterogeneous relations (e.g., Western blot vs. subcellular staining co-validation).

- **Multi-Dimensional Evaluation Framework.** We establish a three-stage evaluation framework (*Perception* → *Understanding* → *Reasoning*) with seven core tasks: (1) The *Perception* stage focuses on **panel-level analysis**, evaluating Numerical Perception (NP, e.g., estimate kinetic curve values), Morphological Perception (MP, e.g., cell invasion extent identification), and Information Localization (IL); (2) The *Understanding* stage deciphers **cross-panel relations**, testing Trend Analysis (TA) for isomorphic panels and Heterogeneous Integration (HI) across diverse panels; (3) The *Reasoning* stage requires **expert-level judgments**, assessing both Qualitative Reasoning (Qual., e.g., directional conclusions) and Quantitative Reasoning (Quant.).

We evaluate eight proprietary and 12 open-source MLLMs, along with four training-free MCoT methods (i.e., DDCoT (Zheng et al., 2023), VoT (Wu et al., 2024), VIC (Zheng et al., 2024a), and Cantor (Gao et al., 2024)), revealing three key findings:

- **MLLMs underperform in the *perception* → *understanding* → *reasoning* pipeline.** Only `Gemini 3 Pro Preview` exceeds 60% accuracy, indicating that current models fall significantly short of rigorous AI4S requirements

- **Critical deficiencies exist across all evaluation stages.** Numerical Perception (NP) yields the poorest results (most models $< 60\%$). Although Morphological Perception (MP) performs better, it exhibits category bias and limited generalization. Cross-Panel Relation Understanding remains a core bottleneck, with accuracy plummeting as relation complexity increases in Trend Analysis (TA). Furthermore, Quantitative Reasoning (Quant.) significantly trails Qualitative Reasoning (Qual.) by 12.76%–31.41%, failing to meet expert-level demands.

- **MCoT methods lack consistent gains in AI4S visual reasoning.** Prompt-based MCoT methods merely augment reasoning steps without mitigating perceptual limitations. Consequently, they often amplify perceptual hallucinations rather than enhance reasoning outcomes. This indicates that robust panel-level perception is an absolute prerequisite for effective AI4S visual reasoning.

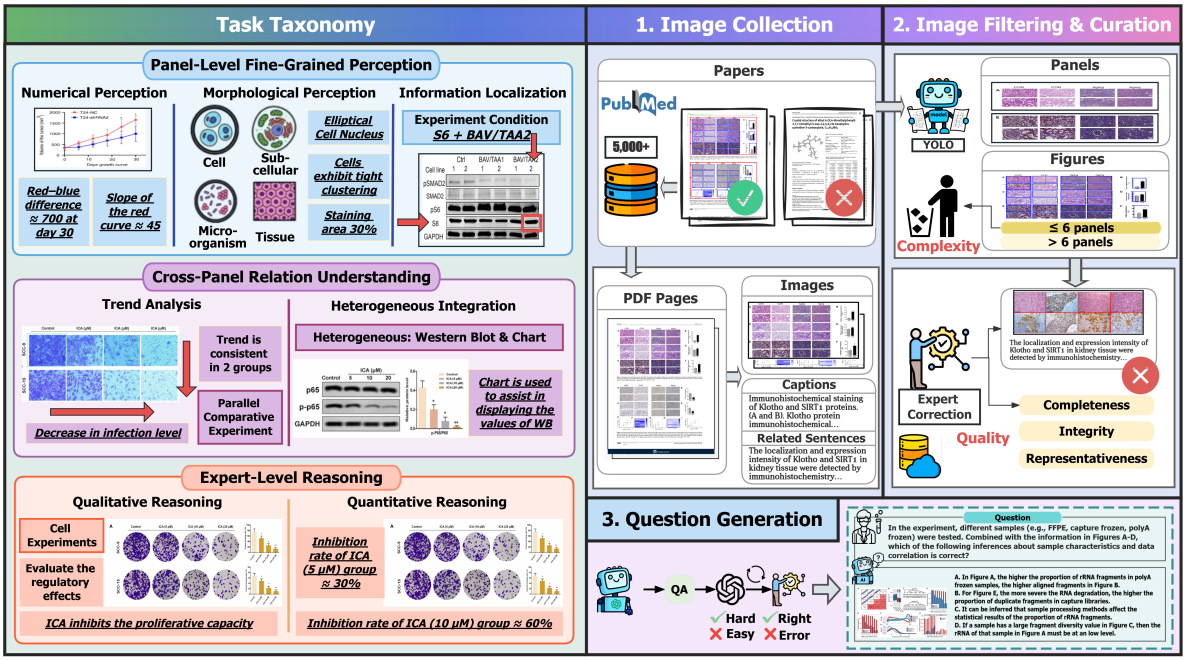


Figure 3: Overview of SPUR. **(Left)** The hierarchical task taxonomy comprising seven core tasks. **(Right)** The three-stage QA pair curation pipeline.

2 SPUR Benchmark

We introduce SPUR to evaluate MLLMs' ***perception***, ***understanding***, and ***reasoning*** capabilities on scientific experimental images through seven core tasks. SPUR comprises 4,264 multiple-choice questions (MCQs) paired with 1,084 scientific images and texts from PMC, spanning seven disciplines, five experimental paradigms, and six fine-grained panel types.

2.1 Task Taxonomy

As shown in Figure 3, SPUR structures its seven tasks across the *Perception* → *Understanding* → *Reasoning* stages:

Panel-Level Fine-Grained Perception. This stage identifies and parses visual features within individual panels (e.g., stained preparations) to establish foundational visual cognition:

- **Numerical Perception (NP).** Quantifies absolute levels and differentiation of visual features.
- **Morphological Perception (MP).** Analyzes biological structure morphology in stained preparations (e.g., cell shape and tissue architecture).
- **Information Localization (IL).** Maps panels to their corresponding experimental conditions.

Cross-Panel Relation Understanding. This stage extracts implicit relations across multi-panel frameworks (e.g., causal, comparative, and argumentative patterns):

- **Trend Analysis (TA).** Interprets directional changes and experimental content across isomorphic panels.
- **Heterogeneous Integration (HI).** Aligns and reasons across disparate panel types through cross-modal synthesis (e.g., information integration and abstract-concrete mapping).

Expert-Level Reasoning. This stage derives scientific inferences from multi-panel visual evidence integrated with experimental context:

- **Qualitative Reasoning (Qual.).** Synthesizes visual cues, domain knowledge, and experimental design logic to interpret biological significance.
- **Quantitative Reasoning (Quant.).** Conducts mathematical verification and evaluation of intergroup differences, effect sizes, and statistical significance via numerical comparison, ratio calculation, and hypothesis testing.

2.2 Dataset Construction and Annotation

As shown in Figure 3, SPUR was built through a three-stage process (see Appendix B for details):

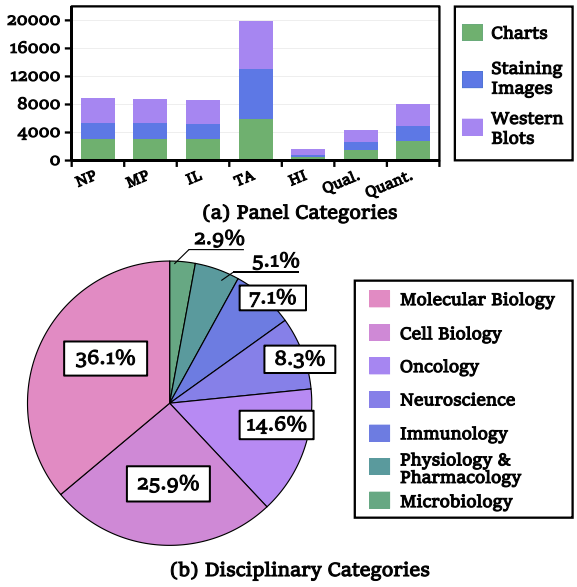


Figure 4: Distribution of **(a)** Panel Categories and **(b)** Disciplinary Categories in SPUR.

Image Collection. We systematically curated over 5,000 open-access papers from PMC, applying two selection criteria: publication within the last decade and a source journal impact factor (IF) > 3.0. Using automated PDF parsing, we extracted 5,632 scientific images. These were manually annotated with related sentences, contextual captions, and disciplinary classifications to form the raw dataset.

Image Filtering and Curation. To ensure question complexity, images underwent a dual-filtering process. First, a YOLO-based panel detector segmented images and excluded those with ≤ 6 panels (removing 77.6% of candidates). Second, experts verified experimental workflow completeness and annotation accuracy, discarding 14.2% of the remaining set for lacking methodological rigor.

Question Generation. Domain experts developed specialized prompt templates aligned with our task hierarchy, enabling `GPT-4o` to generate 7,608 candidate QA pairs.

2.3 Data Quality Assurance

Textual Shortcut Elimination. We implemented a statistical filter to eliminate text-dependent questions. For each QA pair, `GPT-4o` answered text-only queries 10 times under randomized conditions. Questions with ≥ 5 correct responses were discarded (removing 21.2% of candidates), ensuring the retained items require genuine visual reasoning. Subsequently, the remaining `(question, options, answer)` triplets underwent expert review for biomedical validity and task alignment, with 28.9% rejected primarily due to factual inaccuracies or task mismatch.

Expert Review. We implemented a rigorous quality control system through standardized expert review protocols. Four domain specialists (> 40 peer-reviewed publications each) and two senior experts (> 100 publications) conducted multi-tiered validation. Each sample underwent independent dual-expert review to flag inadequate information or ambiguous answers, with senior arbitrators resolving any discrepancies.

Mitigating Data Leakage. During Question Generation, we strictly prohibited deriving questions and options directly from captions or related text. Instead, they were designed exclusively based on panel visuals, demanding fine-grained perception and cross-panel reasoning. Furthermore, filtering out text-only answerable questions prevents models from relying on pre-trained knowledge, common sense, or superficial textual patterns.

2.4 Data Statistics

As shown in Table 2 and Figure 4, SPUR comprises 4,264 expert-curated MCQs derived from 1,084 images, encompassing over 15,500 panels across seven disciplines. Averaging 14.3 panels per image across six fine-grained types, it enables complex cross-panel relations that significantly surpass existing benchmarks in structural complexity. Task distribution includes 636 NP, 634 MP, 621 IL, 1,357 TA, 130 HI, 567 Qual., and 319 Quant. samples. The benchmark's high difficulty is reflected by lengthy options (averaging 54.6 words) and detailed reasoning evidence (58.8 words), demanding robust multimodal capabilities well beyond general benchmarks (See Appendix B.4 for more details).

Property	**Value**
Data Source	
# Experimental Paradigms/Images	5/1,084
Avg. Panels/Image	14.3
Multiple-Choice Question	
Total Samples	**4,264**
# Perception (NP/MP/IL)	636/634/621
# Understanding (TA/HI)	1,357/130
# Reasoning (Qual./Quant.)	567/319
Avg. Words (Question/Options/Evidence)	20.8/54.6/58.8

Table 2: Key statistics of SPUR.

Model	Panel-Level Perception				Cross-Panel Understanding			Expert-Level Reasoning			Overall
	NP	MP	IL	Micro Avg.	TA	HI	Micro Avg.	Qual.	Quant.	Micro Avg.	
Proprietary MLLMs											
Gemini 3 Pro Preview	**61.26**	**67.74**	**59.67**	**62.92**	51.04	59.23	51.77	**90.31**	58.90	**70.29**	**60.57**
Claude 3.7 Sonnet (thinking)	59.67	64.32	57.45	60.50	51.30	60.80	52.12	87.58	**59.96**	69.93	59.52
Gemini 2.5 Pro Preview	56.47	62.97	56.47	58.65	53.30	61.54	54.02	86.54	57.94	68.24	59.00
GPT-5.1	58.73	61.72	54.47	58.33	51.18	50.78	51.15	86.52	56.36	67.23	57.68
OpenAI o4-mini-high	59.24	64.50	55.34	59.72	48.37	59.23	49.32	84.33	56.36	66.44	57.50
Doubao-Seed-1.6	53.30	63.51	56.61	57.81	47.83	56.92	48.62	80.31	53.89	63.43	55.77
GPT-4o	45.91	53.81	52.27	50.64	53.21	63.57	54.12	71.16	54.80	60.73	53.95
Grok 4.1 Fast	47.33	55.99	52.98	52.09	43.70	52.31	44.45	73.44	48.94	57.79	50.61
Open-Source MLLMs											
GLM-4.5V	57.70	61.99	57.65	59.12	55.71	68.46	56.83	80.94	58.48	66.59	59.87
Ministral 3 14B	50.88	61.40	56.56	56.25	**57.79**	**70.00**	**58.86**	72.50	56.81	62.49	58.48
Ministral 3 8B	51.57	57.03	57.03	55.19	57.49	66.15	58.25	70.85	53.53	59.77	57.21
Llama 4 Maverick	51.73	59.78	56.61	56.03	51.88	58.46	52.46	84.64	57.02	67.01	57.06
Qwen3-VL-30B-A3B-Thinking	53.48	58.00	55.27	53.99	54.85	61.24	55.41	75.16	54.17	61.75	56.10
InternVL3-78B	46.30	51.97	49.84	49.36	49.52	61.24	50.54	75.24	51.06	59.80	51.94
Mistral Small 3.1	42.74	51.92	53.48	49.36	51.82	56.92	52.28	72.82	47.56	56.84	51.94
Qwen3-VL-30B-A3B-Instruct	44.18	53.31	51.05	49.50	50.41	63.08	51.51	66.88	51.41	57.00	51.76
Gemma 3 27B	38.36	46.43	48.30	44.33	48.59	57.69	49.39	64.14	51.38	55.95	48.44
Qwen2.5-VL-72B	38.10	45.34	49.11	44.38	51.87	61.90	52.82	73.10	52.51	59.95	48.21
LLaVA-v1.5-13B	33.05	28.11	34.15	31.75	34.52	44.96	35.43	62.19	35.58	45.20	35.97
LLaVA-OneVision-7B	26.22	35.71	40.68	31.49	34.14	40.63	34.79	48.59	35.65	40.34	34.47

Table 3: Performance comparison of proprietary and open-source MLLMs across three cognitive stages and seven core tasks on SPUR. The best and second-best scores are highlighted in **bold** and underlined, respectively.

3 Experiments

3.1 Experimental Setup

Models. We evaluate 20 state-of-the-art MLLMs with visual capabilities (eight proprietary and 12 open-source), spanning diverse architectural paradigms. Model details are provided in Appendix C.1.

Metrics. Model performance is assessed by accuracy on MCQs across all seven tasks.

MCoT Methods. We evaluate two categories of MCoT methods encompassing four specific approaches: (1) *Prompt-based* methods (i.e., DDCoT (Zheng et al., 2023) and VoT (Wu et al., 2024)) using designed prompts for rationale generation; (2) *Plan-based* methods (i.e., VIC (Zheng et al., 2024a) and Cantor (Gao et al., 2024)) enabling dynamic thought exploration during inference.

Research Questions. We investigate five core questions: **RQ1**: Do MLLMs exhibit expert-level perception-to-reasoning capabilities on scientific experimental images? **RQ2**: Can MLLMs achieve precise panel-level perception? **RQ3**: How effectively do MLLMs understand complex cross-panel relations? **RQ4**: Do MLLMs perform expert-level qualitative and quantitative reasoning? **RQ5**: Do MCoT methods enhance scientific reasoning performance?

3.2 Overall Performance (RQ1)

Table 3 presents the comprehensive results. Our main findings are summarized as follows:

Overall model performance remains inadequate. Except for `Gemini 3 Pro Preview`, all models achieve $< 60\%$ accuracy on SPUR. This substantial gap indicates that MLLMs' perception, understanding, and reasoning capabilities fall far below expert-level requirements for AI4S tasks.

Limited panel-level perception constrains reasoning. Open-source models consistently underperform proprietary counterparts in fine-grained perception tasks. All models scoring $< 63\%$ highlights fundamental perceptual limitations across mainstream MLLMs.

Cross-panel understanding presents critical bottlenecks. Models attain significantly lower accuracy in Trend Analysis (TA) than in other tasks, with most showing the poorest performance here. This underscores that comprehending complex cross-panel relations remains a core challenge.

Reasoning tasks reveal substantial disparities. While overall reasoning accuracy exceeds understanding tasks, Quantitative Reasoning (Quant.) lags Qualitative Reasoning (Qual.) by 12.76%–31.41%. This confirms that current MLLMs cannot meet rigorous quantitative analysis demands.

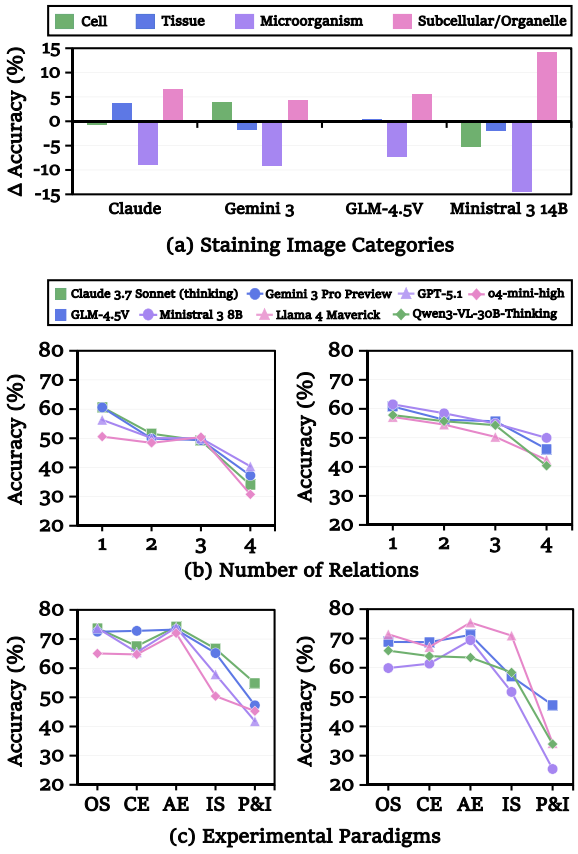


Figure 5: Fine-grained results based on (a) Staining Image Categories, (b) Number of Relations, and (c) Experimental Paradigms. For (a), the Y-axis represents the absolute accuracy deviation from the model's Micro Avg. within the Panel-Level Fine-Grained Perception stage. Further details are provided in Appendix C.2.

3.3 Panel-Level Perception Results (RQ2)

Numerical Perception (NP) shows the weakest performance. Most MLLMs struggle with precise numerical content. Among proprietary models, `Gemini 3 Pro Preview` leads at 61.26%, while for open-source ones, only `GLM-4.5V` (57.70%) approaches mid-tier proprietary performance. This deficiency constrains quantitative reasoning.

Weak Information Localization (IL) impedes understanding. IL yields the lowest scores among perception tasks, where even the top-performing `Gemini 3 Pro Preview` achieves only 59.67%. This prevents accurate panel-to-condition linking, hindering cross-panel relation construction and subsequent reasoning.

Morphological Perception (MP) exhibits high but unstable accuracy. As shown in Figure 5(a), fine-grained analysis across four staining categories reveals a significant imbalance. For instance, `Ministral 3 14B` scores 70.52% on Subcellular/Organelle images but only 42.80% for Microorganisms. Such fluctuations indicate limited generalizability and training data biases.

3.4 Cross-Panel Understanding Results (RQ3)

Models struggle with Trend Analysis (TA) in isomorphic panels. Figure 5(b) confirms that TA accuracy inversely correlates with relation complexity. As the number of relations increase from one to four, `Claude 3.7 Sonnet (thinking)`'s accuracy drops from 60.70% to 34.00%, reflecting an inability to handle multi-relation logic.

Heterogeneous Integration (HI) yields significantly higher performance. This stems from academic contexts: relations across disparate experiments typically involve supportive confirmatory connections, whereas isomorphic panels often generate diverse, contradictory trends, creating inherently more challenging questions.

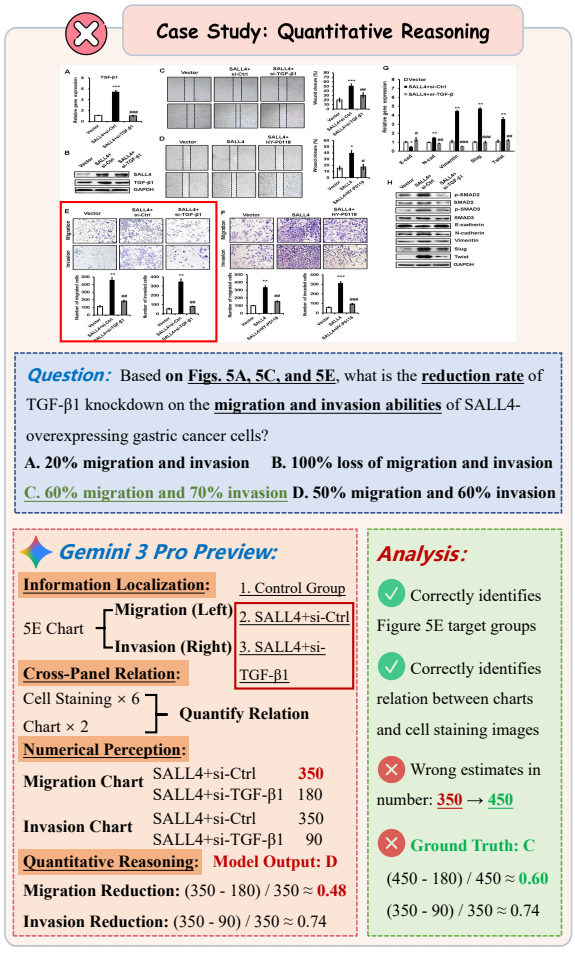


Figure 6: Error analysis of a Quantitative Reasoning (Quant.) task. This case demonstrates how numerical perception errors by `Gemini 3 Pro Preview` propagate into cascaded reasoning failures.

Model	Panel-Level Perception				Cross-Panel Understanding			Expert-Level Reasoning			Overall
	NP	MP	IL	Micro Avg.	TA	HI	Micro Avg.	Qual.	Quant.	Micro Avg.	
GLM-4.5V											
Direct	**57.70**	61.99	**57.65**	**59.12**	**55.71**	**68.46**	**56.83**	**80.94**	**58.48**	**66.59**	**59.87**
DDCoT	47.11	47.67	42.83	45.88	45.24	56.35	46.19	71.52	53.27	59.93	48.90
VoT	55.82	**62.30**	**57.65**	58.59	53.65	60.77	54.27	78.44	57.77	65.23	58.47
VIC	35.50	33.87	32.09	33.83	27.20	29.92	27.43	34.59	36.52	35.83	32.02
Cantor	53.41	58.69	52.96	55.04	51.23	60.00	52.00	77.12	56.61	64.05	55.59
Ministral 3 14B											
Direct	**50.88**	**61.40**	**56.56**	**56.25**	**57.79**	**70.00**	**58.86**	72.50	**56.81**	**62.49**	**58.48**
DDCoT	47.22	52.30	47.97	49.17	51.75	61.24	52.58	**74.61**	48.48	57.97	52.19
VoT	50.00	59.05	55.68	54.89	56.67	67.97	57.66	71.92	55.14	61.24	57.17
VIC	40.10	40.00	37.34	39.16	36.93	43.08	37.47	52.50	42.55	46.15	40.02
Cantor	45.32	49.92	45.94	47.06	37.77	43.41	38.27	59.55	49.36	53.07	45.23
InternVL3-78B											
Direct	46.30	51.97	49.84	49.36	49.52	61.24	50.24	**75.24**	**51.06**	**59.80**	51.94
DDCoT	47.86	54.62	**55.05**	52.84	51.71	**62.20**	52.62	70.74	50.71	57.86	53.64
VoT	47.08	52.93	51.46	50.48	50.15	62.02	51.18	73.67	50.00	58.57	52.40
VIC	40.16	46.60	40.97	42.58	39.51	48.84	40.33	55.62	39.29	45.20	42.35
Cantor	**49.53**	**56.87**	54.27	**53.54**	**53.87**	60.00	**54.40**	71.75	47.32	56.11	**54.37**
Qwen3-VL-30B-A3B-Instruct											
Direct	44.18	53.31	51.05	49.50	50.41	**63.08**	51.51	66.88	51.41	57.00	51.76
DDCoT	41.04	47.00	45.41	44.77	43.58	60.47	45.05	63.95	43.79	51.08	46.04
VoT	46.70	**56.08**	**53.95**	**52.22**	**51.25**	60.77	**52.09**	**68.44**	**51.59**	57.67	**53.31**
VIC	39.08	43.46	40.77	41.12	38.65	48.84	39.56	55.84	41.58	46.81	41.76
Cantor	**49.19**	55.20	51.64	52.02	46.82	52.71	47.34	71.79	50.18	**58.15**	51.63
Gemma 3 27B											
Direct	38.36	46.63	48.30	44.33	**48.59**	**57.69**	**49.39**	64.14	51.38	55.95	48.44
DDCoT	37.38	41.05	42.09	40.16	45.13	55.04	45.98	63.63	49.37	54.56	45.19
VoT	37.08	**48.58**	**48.63**	**44.75**	48.11	56.15	48.82	**66.35**	**55.16**	**59.20**	**49.16**
VIC	31.97	30.17	28.43	30.21	26.01	31.54	26.50	26.33	34.40	31.48	29.17
Cantor	**42.61**	44.08	40.81	42.51	45.79	53.08	46.43	62.50	39.93	48.08	45.03

Table 4: Performance comparison of the direct prompting baseline and four MCoT methods across five base models. The best scores within each model block are highlighted in **bold**.

3.5 Expert-Level Reasoning Results (RQ4)

Quantitative Reasoning (Quant.) presents greater challenges than Qualitative Reasoning (Qual.). Qual. assesses directional conclusions (e.g., "A promotes B"), while Quant. evaluates magnitude (e.g., "A stimulates B by 50%"). As shown in Figure 5(c), models achieve higher accuracy on Observational Studies (OS), Cell Experiments (CE) and Animal Experiments (AE) than on Interventional Studies (IS) and Pathology & Imaging (P&I). This discrepancy arises because IS/P&I paradigms typically require complex quantitative analysis and intricate logic, demanding stronger reasoning.

Insights guide future AI4S research. As illustrated in Figure 6, numerical perception provides foundational support, while cross-panel understanding enables experiment interpretation. Future AI4S development necessitates mature quantitative capabilities. To truly automate scientific research, MLLMs must strengthen precise, multi-step quantitative reasoning over visual evidence.

3.6 MCoT Results (RQ5)

MCoT methods fail to consistently enhance performance. As shown in Table 4, prompt-based approaches merely augment reasoning steps. Without accurately perceiving fine-grained visual content (e.g., numerical values and cross-panel relations), CoT guidance does not guarantee improved outcomes and can sometimes be detrimental.

Plan-based MCoTs partially improves perception. For `Qwen3-VL-30B-A3B-Instruct`, Cantor increased perception tasks accuracy from 49.50% to 52.02% and reasoning tasks accuracy from 57.00% to 58.15%. This demonstrates that predefined observation protocols can mitigate perceptual limitations, thereby enabling reasoning improvements and suggesting optimized strategies for scientific experimental image tasks.

To further decouple the effects of perception on reasoning, we selected `Qwen3-VL-30B-A3B-Instruct` to calculate the reasoning accuracy distribution across different MCoT methods. This analysis encompasses 423 reasoning questions,

Condition	Direct	DDCoT	VoT	VIC	Cantor
Perception Correct	71.66	82.66	98.59	65.66	79.65
Δ vs. Direct	–	(↑ 11.00)	(↑ 26.93)	(↓ 6.00)	(↑ 7.99)
Perception Incorrect	32.40	23.68	9.32	30.30	40.23
Δ vs. Direct	–	(↓ 8.72)	(↓ 23.08)	(↓ 2.10)	(↑ 7.83)

Table 5: Reasoning accuracy (%) of MCoT methods decoupled by perception correctness (evaluated on `Qwen3-VL-30B-A3B-Instruct`). Δ denotes the absolute deviation from the direct prompting baseline.

with quantitative results presented in Table 5.

MCoT significantly improves reasoning when perception is correct. Except for VIC, which ignores visual content during plan design, all MCoT methods achieved accuracy gains of 7.99% to 26.93% over direct prompting for the subset with correct perception. This strongly demonstrates that MCoT methods deliver substantial reasoning gains provided the MLLM initially obtains accurate visual information.

MCoT amplifies errors when perception is incorrect. Conversely, on the subset with incorrect perception, most MCoT methods yielded lower reasoning accuracy than direct prompting (declining by 2.10% to 23.08%). The only exception is Cantor, which intrinsically aids the model's perceptual ability. This confirms our initial observation: if models cannot accurately parse fine-grained content, additional CoT guidance amplifies perceptual hallucinations rather than correcting them.

4 Related Work

Scientific Image Reasoning Existing scientific image benchmarks cover diverse categories including statistical charts (Xu et al., 2025; Huang et al., 2025; Liu et al., 2025b), schematic diagrams (Li and Tajbakhsh, 2023), microscopy images (Burgess et al., 2025), biomedical/chemical images (Laurent et al., 2024; Burgess et al., 2025; Lozano et al., 2024; Hu et al., 2024; Li et al., 2025), and general science images (Wang et al., 2024a; Yue et al., 2025; He et al., 2024b; Jassim et al., 2024; Feng et al., 2025). However, these benchmarks exhibit critical gaps in scientific experimental image understanding. Specifically, experimental images remain severely underrepresented in prominent benchmarks such as ScienceQA (Lu et al., 2022), M3CoT (Chen et al., 2024b), and MMMU (Yue et al., 2024), which predominantly feature non experimental scientific imagery. Furthermore, current benchmarks overlook cross-panel relation understanding, a fundamental aspect of experimental image interpretation. Major benchmarks including MMSci (Li et al., 2024c), SciAssess (Cai et al., 2025), SFE (Zhou et al., 2025), and M3SciQA (Li et al., 2024a) focus on isolated, low complexity images, limiting comprehensive evaluation. In contrast, SPUR establishes a three-stage framework (*Perception* → *Understanding* → *Reasoning*) with seven tasks designed to address these limitations.

Multimodal Chain-of-Thought MCoT methods facilitate step-by-step reasoning for MLLMs in zero-shot and few-shot settings (Zhang et al., 2024; Wang et al., 2024b, 2025b; Chen et al., 2024a,c; Liu et al., 2025a; He et al., 2024a; Qin et al., 2024; Fei et al., 2024a,b). Existing approaches include prompt-based methods (Zheng et al., 2023; Wu et al., 2024), plan-based decomposition (Zheng et al., 2024a; Gao et al., 2024), and training-based frameworks (Wang et al., 2025a; Zhang et al., 2025). However, MCoT remains unverified for scientific image reasoning tasks.

5 Conclusion

We introduce SPUR, a benchmark evaluating MLLMs' scientific experimental image *perception*, *understanding*, and *reasoning* capabilities, alongside an assessment of various MCoT methods. Experiments reveal a significant performance gap between MLLMs and expert-level reasoning: all models except `Gemini 3 Pro Preview` fall short of the 60% accuracy threshold. Bridging this gap, particularly in quantitative reasoning, remains a critical challenge for future research. We expect SPUR to catalyze the development of more robust scientific image reasoning technologies within the AI4S community.

Limitations

SPUR utilizes an MCQ format to evaluate MLLMs, mitigating shortcut biases through rigorous option design and filtering. However, this format provides limited visibility into the internal reasoning processes of the models. Furthermore, while SPUR is curated from open-access PMC articles across seven life science disciplines, the experimental paradigms and core challenges (i.e., panel-level fine-grained perception, cross-panel relation understanding, and expert-level reasoning) are representative of broader scientific contexts. Nevertheless, extending these findings to non-biological fields may require additional adaptation and dataset expansion.

Ethical Considerations

SPUR complies with ACL ethics guidelines. This study involved no human subjects or animal experimentation. All data were collected from open-source repositories in accordance with relevant usage licenses, ensuring that privacy is preserved and no personally identifiable information is included. SPUR is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0), and the associated codebase is distributed under the Apache License 2.0, supporting both commercial and open-source applications. We have made consistent efforts to minimize bias and ensure transparency throughout the dataset construction and evaluation process.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62473271, 62176026), the Beijing Natural Science Foundation (Grant No. QY25338), and the Fundamental Research Funds for the Beijing University of Posts and Telecommunications (Grant No. 2025AI4S03). This work is also supported by the Engineering Research Center of Information Networks, Ministry of Education, China. We would also like to thank the anonymous reviewers and area chairs for constructive discussions and feedback.

References

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, Sarina M Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, and 4 others. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19552–19564.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2025. SciAssess: Benchmarking LLM proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2335–2357, Albuquerque, New Mexico. Association for Computational Linguistics.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, and Wanxiang Che. 2024a. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 37, pages 54872–54904. Curran Associates, Inc.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024a. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13109–13125. PMLR.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024b. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7701–7719.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. 2025. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *Preprint*, arXiv:2406.09098.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and Rongrong Ji. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 9096–9105, New York, NY, USA. Association for Computing Machinery.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024a. Multi-modal latent space learning for chain-of-thought reasoning in language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18180–18187.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang.

2024b. Cmmu: A benchmark for chinese multimodal multi-type question understanding and reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 830–838. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22170–22183.

Muye Huang, Lingling Zhang, Han Lai, Wenjun Wu, Xinyu Zhang, and Jun Liu. 2025. Vprochart: Answering chart question through visual perception alignment agent and programmatic solution reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3689–3696.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. 2024. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6297–6305. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *Preprint*, arXiv:2407.10362.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024a. M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446, Miami, Florida, USA. Association for Computational Linguistics.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Wei Li, Mao Su, Shufei Zhang, Wanli Ouyang, Yuqiang Li, and Dongzhan Zhou. 2025. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):415–423.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

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

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. 2024c. MMSci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design - Vienna 2024*.

Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2025a. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jiapeng Liu, Liang Li, Shihao Rao, Xiyan Gao, Weixin Guan, Bing Li, and Can Ma. 2025b. Union is strength! unite the power of llms and mllms for chart question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5487–5495.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. 2024. Micro-bench: A microscopy benchmark for vision-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 30670–30685. Curran Associates, Inc.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What factors affect multi-modal in-context learning? an in-depth exploration. In *Advances in Neural Information Processing Systems*, volume 37, pages 123207–123236. Curran Associates, Inc.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems*, volume 37, pages 18695–18728. Curran Associates, Inc.

Jinhong Wang, Shuo Tong, Jian Liu, Dongqi Tang, Weiqiang Wang, Wentong Li, Hongxia Xu, Danny Z. Chen, Jintai Chen, and Jian Wu. 2025a. Orderchain: Towards general instruct-tuning for stimulating the ordinal understanding ability of mllm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3477–3487.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2025b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024a. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. 2024b. Stop reasoning! when multimodal LLM with chain-of-thought reasoning meets adversarial image. In *First Conference on Language Modeling*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 90277–90317. Curran Associates, Inc.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *Preprint*, arXiv:2506.12594.

Zhengzhuo Xu, Bowen Qu, Yiyan Qi, SiNan Du, Chengjin Xu, Chun Yuan, and Jian Guo. 2025. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In *International Conference on Learning Representations*, volume 2025, pages 78550–78572.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, Vienna, Austria. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 5168–5191. Curran Associates, Inc.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024a. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *Preprint*, arXiv:2411.12591.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024b. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

Yuhao Zhou, Yiheng Wang, Xuming He, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, Wenxuan Huang, Jiaqi Wei, Dan Si, YAO XIUQI, Jia Bu, Haiwen Huang, Tianfan Fu, SHIXIANG TANG, Ben Fei, Dongzhan Zhou, and 8 others. 2025. Scientists' first exam: Probing cognitive abilities of MLLM via perception, understanding, and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Examples from SPUR

To provide a comprehensive view of our dataset, this section visualizes representative examples across the predefined task hierarchy:

Panel-Level Fine-Grained Perception Cases. Figures 10–12 illustrate examples of Numerical Perception (NP), Morphological Perception (MP), and Information Localization (IL).

Cross-Panel Relation Understanding Cases. Figures 13 and 14 present examples of Trend Analysis (TA) and Heterogeneous Integration (HI).

Expert-Level Reasoning Cases. Figures 15 and 16 demonstrate Qualitative Reasoning (Qual.) and Quantitative Reasoning (Quant.) featuring a representative multi-panel configuration.

B Details of Annotation

This section provides details on the construction and annotation pipeline of SPUR. Table 6 outlines the precise quantitative breakdown of data retention and exclusion at each step.

B.1 Image Collection and Preprocessing

Source Paper Selection Criteria. Candidate papers were strictly limited to open-access PMC publications from the past 10 years with a source journal Impact Factor (IF) > 3.0, ensuring both scientific timeliness and authority. Human annotators confirmed the presence of complex experimental images, explicitly excluding review papers, meta-analyses, and theoretical studies lacking original experimental data. Figure 7 illustrates the user interface of the selection platform.

Manual Image–Text Alignment. For the retained images, annotators performed the following standardization: (1) extracted 3–5 key sentences from the main text detailing the experimental background, methods, or results; (2) standardized the original captions by correcting typos, supplementing ambiguous acronyms, and ensuring terminology consistency; and (3) categorized the images into seven predefined disciplinary categories (e.g., Cell Biology, Molecular Biology, and Oncology) based on the source journal.

B.2 Panel Recognition

The YOLO-based panel detector for the initial complexity filter was pre-trained on a proprietary dataset comprising annotated panel coordinates and categories from biomedical and materials science journals. This model automatically excluded images with ≤ 6 panels to guarantee structural complexity. Figure 8 visualizes the detection of the six recognized panel categories in the final dataset.

B.3 Expert Validation Protocols

To ensure methodological rigor and scientific validity, the retained data underwent multi-tiered expert reviews.

Raw Image Review.

- **Validation Criteria.** Reviewers verified experimental design completeness (e.g., clear control groups, experimental variables, and measurable outcomes), methodological integrity (e.g., clear sample labeling, appropriate reagents, and repeatable operations), and image representativeness.
- **Execution Rules.** Each image was independently evaluated by two experts. Images were retained upon dual approval and excluded upon dual rejection. Conflicting judgments underwent decisive arbitration by a senior expert. Initial consistency was reached on 78% of the images, with 22% requiring senior arbitration.

QA Pair Review. Retained questions were evaluated along three dimensions (scored 0–1) to eliminate factual inaccuracies or multimodal misalignments:

- **Scientific Validity.** Conformance to scientific facts, absence of logical errors in conclusions, and correct usage of professional terminology.
- **Task Alignment.** Verification that the QA pair strictly aligns with the predefined task hierarchy and that the required reasoning matches the visual data (e.g., a "quantitative trend analysis" query must correspond to quantifiable chart data).
- **Visual Reasoning Necessity.** Final confirmation that the question cannot be answered via textual shortcuts or common sense, strictly requiring visual interpretation of the image.

B.4 Fine-Grained Distribution

As illustrated in Figure 9, SPUR further categorizes staining images into four fine-grained subtypes (e.g., Cell and Tissue) within the Panel-Level Fine-Grained Perception stage, and evaluates performance across five typical experimental paradigms within the Expert-Level Reasoning stage.

Screening Step	Operation Description	Removed Quantity	Retained Quantity
Initial Image Extraction	Extracted candidate images from PMC papers	–	Images: 5,632
Complexity Filtering	Excluded images with ≤ 6 panels via YOLO detector	Images: 4,368	Images: 1,264
Raw Image Review	Excluded images lacking complete workflows or methodological rigor	Images: 180	Images: 1,084
QA Pair Generation	Generated candidate QA pairs using GPT-4o based on retained images	–	Images: 1,084; QA Pairs: 7,608
Textual Shortcut Elimination	Discarded QA pairs with ≥ 5 correct text-only responses	QA Pairs: 1,612	Images: 1,084; QA Pairs: 5,996
QA Pair Review	Discarded QA pairs with factual errors or task misalignment	QA Pairs: 1,732	Images: 1,084; QA Pairs: 4,264

Table 6: Quantitative breakdown of data retention and exclusion at each curation step of SPUR.

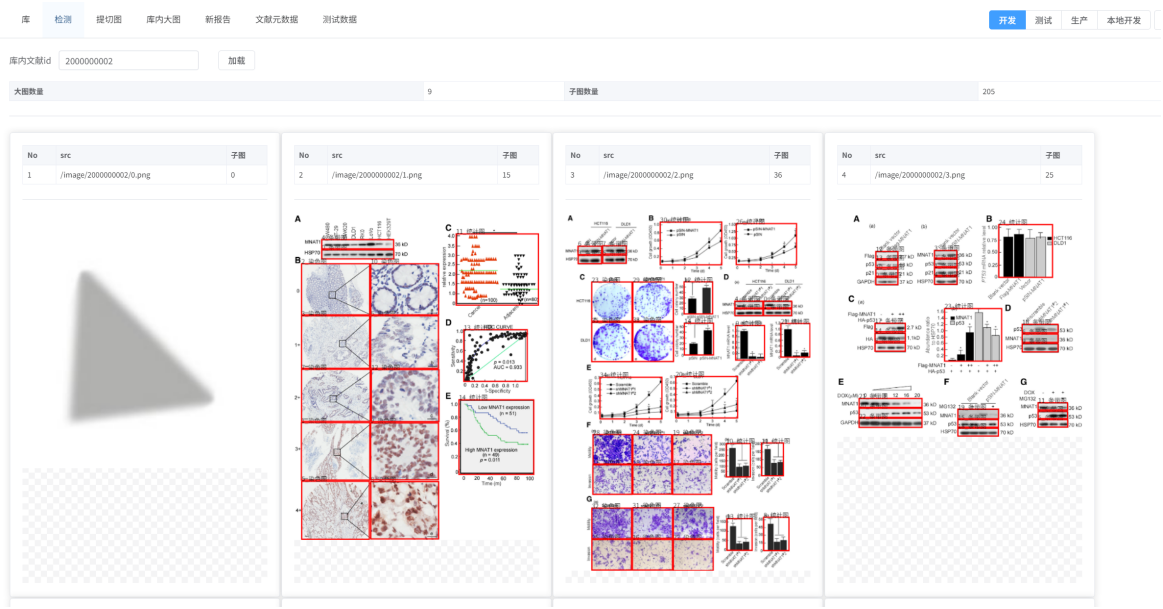


Figure 7: User interface of the academic paper selection platform utilized during the paper-level curation step.

C Experiment Details and Analysis

C.1 Evaluated Models

Since SPUR requires visual interpretation of complex scientific figures, we exclusively evaluate MLLMs with advanced vision capabilities. Table 7 details the proprietary and open-source models evaluated, including their providers, release dates, versions, and parameter sizes.

C.2 Fine-Grained Analysis

Staining Image Categories. This analysis evaluates MLLM performance on fine-grained biological structures within the Panel-Level Fine-Grained Perception stage, specifically categorizing staining images into Cell, Tissue, Microorganism, and Subcellular/Organelle. Detailed quantitative results are presented in Table 8.

Number of Relations. Within the Cross-Panel Relation Understanding stage, task complexity is quantified by the number of cross-panel relation pairs examined across all options of a given question.

Experimental Paradigms. We further analyze model capabilities within the Expert-Level Reasoning stage across five fundamental scientific experimental paradigms: Cell Experiments (CE), Animal Experiments (AE), Observational Studies (OS), Interventional Studies (IS), and Pathology & Imaging (P&I). Performance breakdown in these Expert-Level Reasoning tasks is detailed in Table 9.

D Use of AI Assistants

This research was driven entirely by the authors, who provided all core scientific insights, experimental designs, and analyses. We acknowledge the use of AI assistants during the preparation of this manuscript: Cursor was utilized to aid in code writing and data processing, while large language models (LLMs) assisted with language editing to enhance readability. Furthermore, `GPT-4o` assisted in QA pair construction for our benchmark, with detailed methodologies in Appendix B. We emphasize that all AI-assisted content was thoroughly reviewed and validated by the authors, who bear full responsibility for the scientific integrity and fundamental contributions of this work.

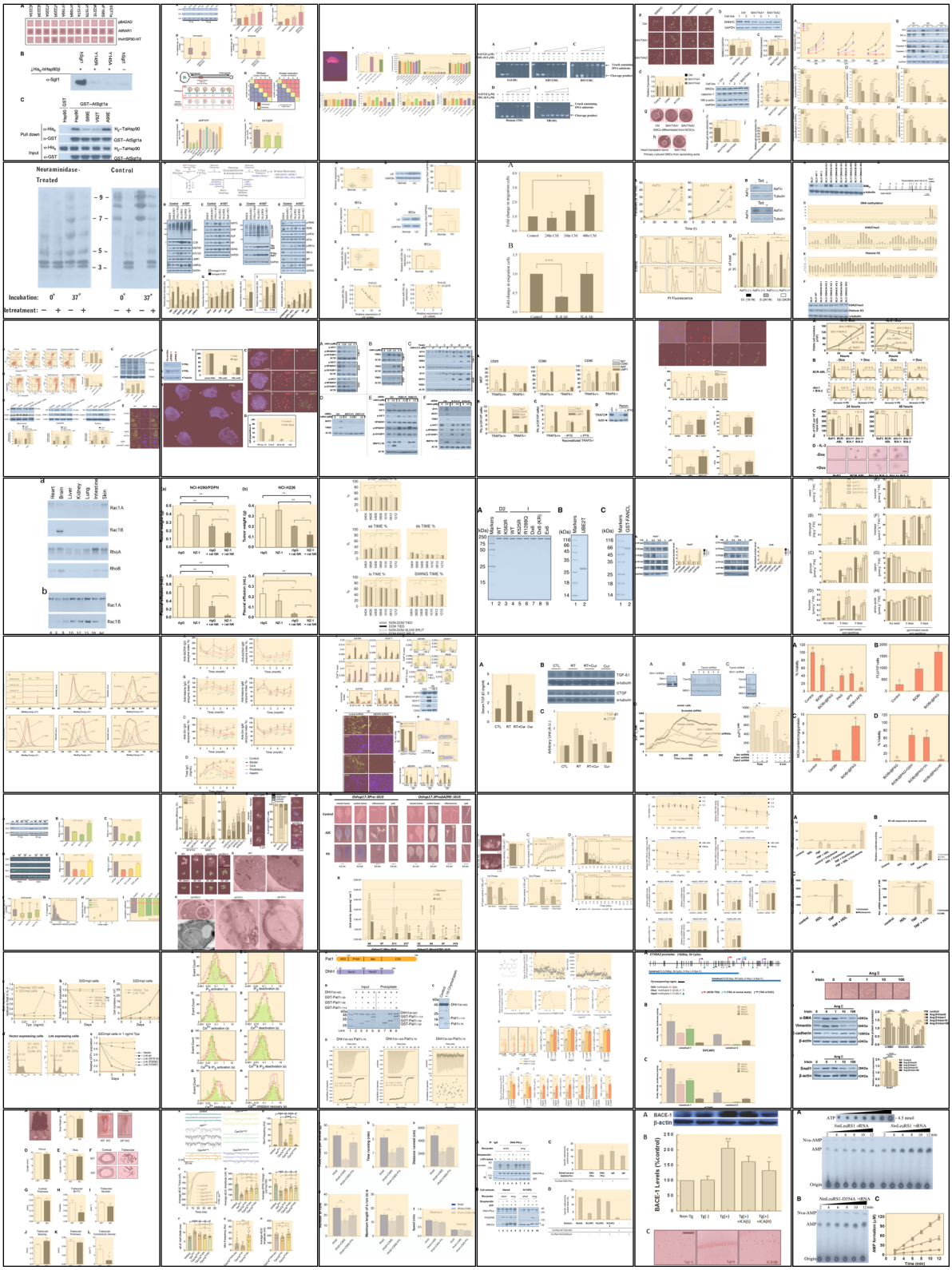


Figure 8: Visualization of panel category detections in SPUR. Red, light yellow, and light blue masks denote staining images (comprising four fine-grained subtypes), charts, and Western blots, respectively.

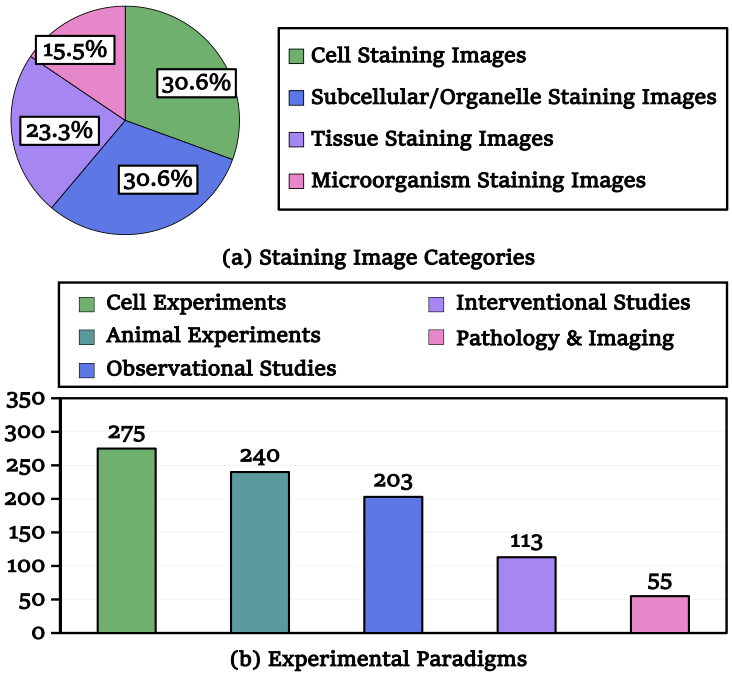


Figure 9: Distribution of **(a)** Staining Image Categories and **(b)** Experimental Paradigms in SPUR.

Provider	Model	Release	Version	Size
		Proprietary MLLMs		
OpenAI	GPT-5.1	2025-11	gpt-5.1	–
	OpenAI o4-mini-high	2025-04	o4-mini-high	–
	GPT-4o	2024-11	gpt-4o-2024-11-20	–
Google	Gemini 3 Pro Preview	2025-11	gemini-3-pro-preview	–
	Gemini 2.5 Pro Preview	2025-06	gemini-2.5-pro-preview-06-05	–
Anthropic	Claude 3.7 Sonnet (thinking)	2025-02	claude-3.7-sonnet:thinking	–
ByteDance	Doubao-Seed-1.6	2025-06	doubao-seed-1.6-250615	–
xAI	Grok 4.1 Fast	2025-11	grok-4.1-fast	–
		Open-Source MLLMs		
Meta	Llama 4 Maverick	2025-04	llama-4-maverick	400B-A17B
Google	Gemma 3 27B	2025-03	gemma-3-27b-it	27B
Mistral AI	Ministral 3 14B	2025-12	ministral-14b-2512	14B
	Ministral 3 8B	2025-12	ministral-8b-2512	8B
	Mistral Small 3.1	2025-03	mistral-small-3.1-24b-instruct-2503	24B
Alibaba	Qwen3-VL-30B-A3B-Thinking	2025-10	qwen3-vl-30b-a3b-thinking	30B-A3B
	Qwen3-VL-30B-A3B-Instruct	2025-10	qwen3-vl-30b-a3b-instruct	30B-A3B
	Qwen2.5-VL-72B	2025-01	qwen2.5-vl-72b-instruct	72B
OpenGVLab	InternVL3-78B	2025-04	internvl3-78b	78B
LLaVA Community	LLaVA-OneVision-7B	2024-08	llava-onevision-qwen2-7b-ov-hf	7B
	LLaVA-v1.5-13B	2023-09	llava-v1.5-13b-hf	13B
Z.ai	GLM-4.5V	2025-08	glm-4.5v	106B-A12B

Table 7: Detailed specifications of the proprietary and open-source MLLMs evaluated on SPUR.

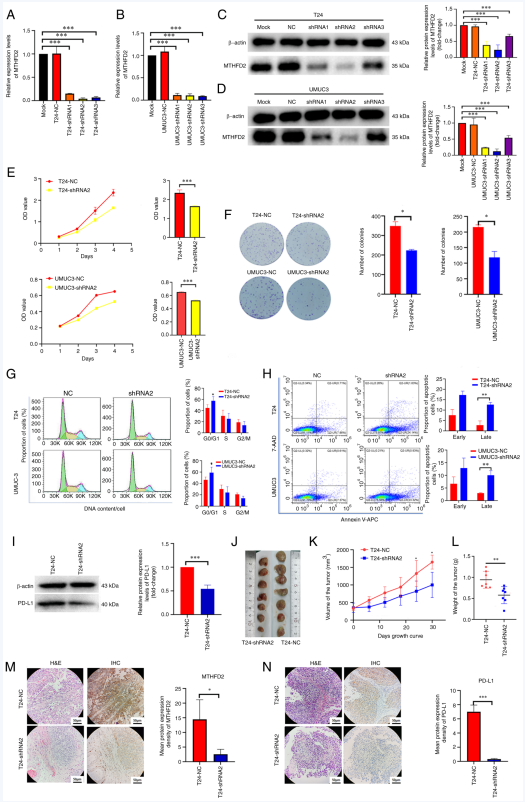
Model	Micro Avg.	Cell	Tissue	Microorganism	Subcellular/ Organelle
Proprietary MLLMs					
Gemini 3 Pro Preview	62.92	66.92 (↑ 4.00)	61.14 (↓ 1.78)	53.70 (↓ 9.22)	67.22 (↑ 4.30)
Claude 3.7 Sonnet (thinking)	60.50	59.69 (↓ 0.81)	64.16 (↑ 3.66)	51.52 (↓ 8.98)	67.11 (↑ 6.61)
Gemini 2.5 Pro Preview	58.65	57.99 (↓ 0.66)	57.70 (↓ 0.95)	55.76 (↓ 2.89)	65.07 (↑ 6.42)
GPT-5.1	58.33	61.47 (↑ 3.14)	56.69 (↓ 1.64)	51.44 (↓ 6.89)	63.69 (↑ 5.36)
OpenAI o4-mini-high	59.72	61.68 (↑ 1.96)	66.30 (↑ 6.58)	56.72 (↑ 3.00)	66.38 (↑ 6.66)
Doubao-Seed-1.6	57.81	57.70 (↓ 0.11)	56.31 (↓ 1.50)	54.84 (↓ 2.97)	62.39 (↑ 4.58)
GPT-4o	50.64	39.66 (↓ 10.98)	55.10 (↑ 4.46)	56.99 (↑ 6.35)	55.96 (↑ 5.32)
Grok 4.1 Fast	52.09	51.39 (↓ 0.70)	51.70 (↓ 0.39)	46.43 (↓ 5.66)	58.35 (↑ 6.26)
Open-Source MLLMs					
GLM-4.5V	59.12	58.81 (↓ 0.31)	59.47 (↑ 0.35)	51.79 (↓ 7.33)	64.77 (↑ 5.65)
Ministral 3 14B	56.25	51.05 (↓ 5.20)	54.36 (↓ 1.89)	42.80 (↓ 14.45)	70.52 (↑ 14.27)
Ministral 3 8B	55.19	46.20 (↓ 8.99)	53.53 (↓ 1.66)	48.93 (↓ 6.26)	68.88 (↑ 13.69)
Llama 4 Maverick	56.03	53.44 (↓ 2.59)	56.30 (↑ 0.27)	49.18 (↓ 6.85)	58.41 (↑ 2.38)
Qwen3-VL-30B-A3B-Thinking	53.99	50.56 (↓ 3.43)	58.29 (↑ 4.30)	49.28 (↓ 4.71)	62.62 (↑ 8.63)
InternVL3-78B	49.36	44.98 (↓ 4.38)	52.61 (↑ 3.25)	51.65 (↑ 2.29)	55.47 (↑ 6.11)
Mistral Small 3.1	49.36	29.56 (↓ 19.80)	56.22 (↑ 6.86)	51.33 (↑ 1.97)	62.07 (↑ 12.71)
Qwen3-VL-30B-A3B-Instruct	49.50	40.45 (↓ 9.05)	51.94 (↑ 2.44)	58.21 (↑ 8.71)	52.48 (↑ 2.98)
Gemma 3 27B	44.33	26.77 (↓ 17.56)	49.76 (↑ 5.43)	61.29 (↑ 16.96)	48.07 (↑ 3.74)
Qwen2.5-VL-72B	44.38	25.30 (↓ 19.08)	58.10 (↑ 13.72)	47.27 (↑ 2.89)	64.52 (↑ 20.14)
LLaVA-v1.5-13B	31.75	20.15 (↓ 11.60)	46.81 (↑ 15.06)	35.90 (↑ 4.15)	52.88 (↑ 20.13)
LLaVA-OneVision-7B	31.49	18.75 (↓ 12.74)	45.26 (↑ 13.77)	34.34 (↑ 2.85)	51.15 (↑ 19.66)

Table 8: Performance breakdown for fine-grained staining images within the Panel-Level Fine-Grained Perception stage. Upward/downward arrows (↑/↓) indicate the absolute deviation from the model's corresponding Micro Avg. within this stage.

Model	Micro Avg.	OS	CE	AE	IS	P&I
Proprietary MLLMs						
Gemini 3 Pro Preview	70.29	72.50 (↑ 2.21)	72.79 (↑ 2.50)	73.25 (↑ 2.96)	65.18 (↓ 5.11)	47.27 (↓ 23.02)
Claude 3.7 Sonnet (thinking)	69.93	73.71 (↑ 3.78)	67.57 (↓ 2.36)	74.25 (↑ 4.32)	66.67 (↓ 3.26)	54.90 (↓ 15.03)
Gemini 2.5 Pro Preview	68.24	71.13 (↑ 2.89)	70.04 (↑ 1.80)	71.43 (↑ 3.19)	65.77 (↓ 2.47)	37.25 (↓ 30.99)
GPT-5.1	67.23	73.76 (↑ 6.53)	65.44 (↓ 1.79)	73.97 (↑ 6.74)	57.89 (↓ 9.34)	41.82 (↓ 25.41)
OpenAI o4-mini-high	66.44	65.05 (↓ 1.39)	64.71 (↓ 1.73)	72.00 (↑ 5.56)	50.42 (↓ 16.02)	45.28 (↓ 21.16)
Doubao-Seed-1.6	63.43	71.78 (↑ 8.35)	61.40 (↓ 2.03)	65.84 (↑ 2.41)	63.16 (↓ 0.27)	32.73 (↓ 30.70)
GPT-4o	60.73	68.66 (↑ 7.93)	67.16 (↑ 6.43)	69.14 (↑ 8.41)	48.67 (↓ 12.06)	23.64 (↓ 37.09)
Grok 4.1 Fast	57.79	61.88 (↑ 4.09)	56.25 (↓ 1.54)	63.37 (↑ 5.58)	53.51 (↓ 4.28)	34.55 (↓ 23.24)
Open-Source MLLMs						
GLM-4.5V	66.59	68.81 (↑ 2.22)	68.75 (↑ 2.16)	71.19 (↑ 4.60)	57.02 (↓ 9.57)	47.27 (↓ 19.32)
Ministral 3 14B	62.49	64.85 (↑ 2.36)	67.28 (↑ 4.79)	64.20 (↑ 1.71)	53.98 (↓ 8.50)	40.00 (↓ 22.49)
Ministral 3 8B	59.77	59.90 (↑ 0.13)	61.40 (↑ 1.62)	69.42 (↑ 9.65)	51.75 (↓ 8.02)	25.45 (↓ 34.32)
Llama 4 Maverick	67.01	71.43 (↑ 4.42)	67.08 (↑ 0.07)	75.42 (↑ 8.41)	70.97 (↑ 3.96)	34.29 (↓ 33.72)
Qwen3-VL-30B-A3B-Thinking	61.75	65.84 (↑ 4.09)	63.97 (↑ 2.22)	63.49 (↑ 1.74)	58.41 (↓ 3.34)	33.96 (↓ 27.79)
InternVL3-78B	59.80	61.89 (↑ 2.09)	60.22 (↑ 0.42)	62.86 (↑ 3.06)	59.70 (↓ 0.10)	40.49 (↓ 19.31)
Mistral Small 3.1	56.84	64.95 (↑ 8.11)	49.61 (↓ 7.22)	63.76 (↑ 6.92)	55.66 (↓ 1.18)	35.19 (↓ 21.65)
Qwen3-VL-30B-A3B-Instruct	57.00	56.44 (↓ 0.56)	65.07 (↑ 8.07)	56.38 (↓ 0.62)	52.63 (↓ 4.37)	30.91 (↓ 26.09)
Gemma 3 27B	55.95	63.35 (↑ 7.40)	61.22 (↑ 5.27)	56.84 (↑ 0.89)	43.64 (↓ 12.31)	23.53 (↓ 32.42)
Qwen2.5-VL-72B	59.95	63.72 (↑ 3.77)	62.15 (↑ 2.20)	61.50 (↑ 1.55)	54.08 (↓ 5.87)	35.20 (↓ 24.75)
LLaVA-v1.5-13B	45.20	49.04 (↑ 3.84)	45.91 (↑ 0.71)	48.26 (↑ 3.06)	42.80 (↓ 2.40)	31.67 (↓ 13.53)
LLaVA-OneVision-7B	40.34	42.55 (↑ 2.21)	42.36 (↑ 2.02)	43.64 (↑ 3.30)	35.65 (↓ 4.69)	26.32 (↓ 14.02)

Table 9: Performance breakdown across typical experimental paradigms within the Expert-Level Reasoning stage. Upward/downward arrows (↑/↓) indicate the absolute deviation from the model's corresponding Micro Avg. within this stage.

Panel-Level Fine-Grained Perception: Numerical Perception



Question: **[Perception_2010073308_3_4]**

Observe Fig. E Which statement is incorrect?

A. In T24 cells, the proliferation curve of the shRNA2 group is consistently lower than that of the NC group.

B. In T24 cells, the gap between the proliferation curves of the shRNA2 group and the NC group increases over time.

C. The OD value of the UMUC3-shRNA2 group increases more slowly than that of the NC group.

D. The degrees of decrease in column height for the shRNA2 groups of both cell lines are similar.

E. Cannot be determined.

Answer: D

Explanation:

Option A: In T24 cells (Fig. E, top left), the proliferation curve of the shRNA2 group (yellow) remains below that of the NC group (red) at all time points. Therefore, this statement is correct.

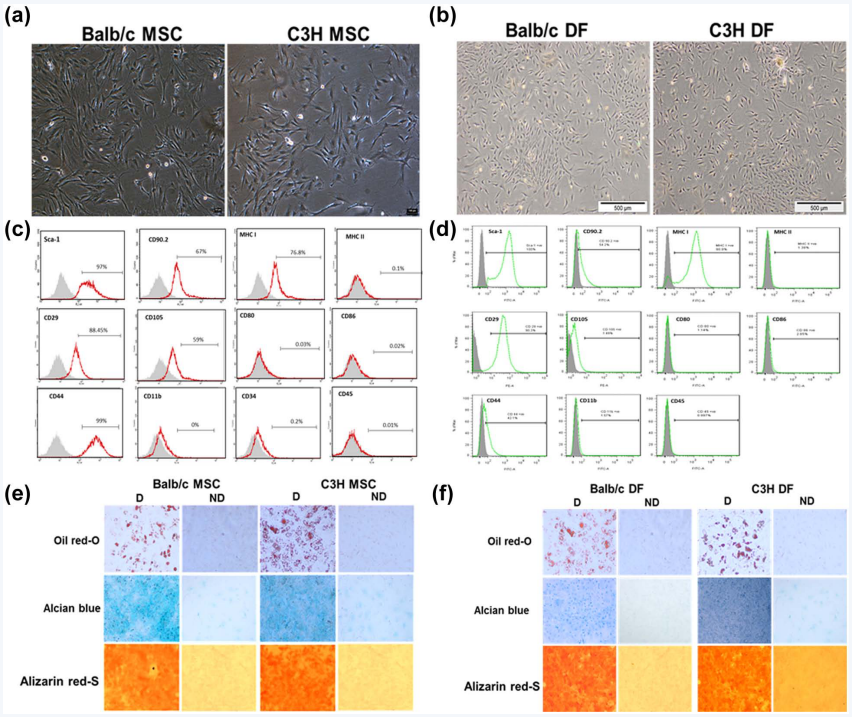
Option B: In T24 cells (Fig. E, top left), the difference between the shRNA2 group and the NC group becomes more pronounced over time. For example, by Day 5, the NC group is approximately 2.2, whereas the shRNA2 group is approximately 1.4. Therefore, this statement is correct.

Option C: In UMUC3 cells (Fig. E, bottom left), the OD value of the shRNA2 group increases more slowly than that of the NC group, indicating reduced proliferative activity. Therefore, this statement is correct.

Option D: This statement is incorrect. The reduction associated with shRNA2 is not similar between the two cell lines. Based on the plotted values, the decrease in T24 is substantially larger than that in UMUC3. Thus, the extent of reduction cannot be considered similar across the two cell lines.

Figure 10: Example of a Numerical Perception (NP) task in the Panel-Level Fine-Grained Perception stage.

Panel-Level Fine-Grained Perception: Morphological Perception



Question: **[Perception_2010002387_0_1]**

Analyze the differentiation staining results of MSC in Fig. (e) (Oil red-O, Alcian blue, Alizarin red-S; D = differentiated, ND = non-differentiated). Which statement is correct?

A. In MSC Oil red-O staining, the Balb/c ND group shows more lipid droplets than the C3H D group.

B. Comparing Alizarin red-S staining of Balb/c MSC and C3H MSC, the orange-red mineralization signals in the differentiated groups (D) are stronger than those in the non-differentiated groups (ND) in both.

C. Alizarin red-S staining of MSC shows that the Balb/c D group has more obvious orange mineralization than the C3H ND group.

D. After induced differentiation (D group), the positive signals of the three stainings (Oil red-O, Alcian blue, Alizarin red-S) in Balb/c MSC are weaker than those in the corresponding ND groups of C3H MSC.

E. Cannot be determined.

Answer: B

Explanation:

Option A: This statement is incorrect. In Oil Red O staining, differentiated groups (D) show more lipid droplets than non-differentiated groups (ND), so the Balb/c ND cannot exceed the C3H D group.

Option B: In Alizarin Red S staining, both Balb/c MSC and C3H MSC show stronger mineralization signals in D than in ND. Therefore, this statement is correct.

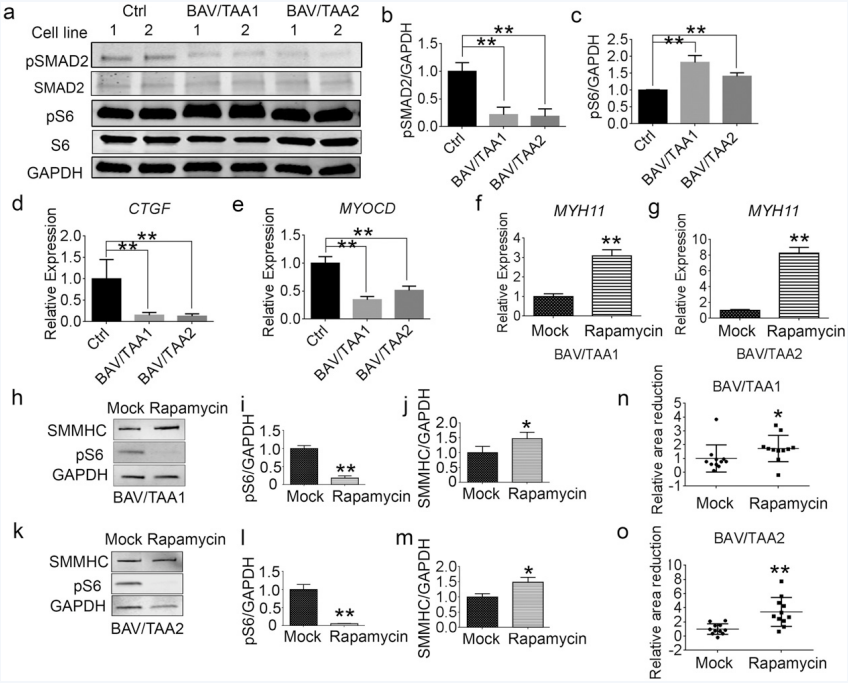
Option C: Although Balb/c D appears stronger than C3H ND, this is a cross-strain comparison and is less directly supported than the within-strain D-versus-ND comparison in Option B.

Option D: This statement is incorrect. After differentiation, Balb/c MSC D shows stronger, not weaker, staining signals than the corresponding C3H MSC ND group.

Option E: This statement is incorrect. The staining images provide enough visual information to compare differentiation-related signal intensity across groups.

Figure 11: Example of a Morphological Perception (MP) task in the Panel-Level Fine-Grained Perception stage.

Panel-Level Fine-Grained Perception: Information Localization



Question: **[Perception_2010198904_5_1]**

Analyze the WB bands in Fig. a (pSMAD2, SMAD2, pS6, S6, GAPDH). Which statement is correct?

A. In both BAV/TAA1 and BAV/TAA2 groups, the expression of pSMAD2 is significantly higher than that in the Ctrl group.

B. GAPDH is stably expressed in all groups and can be used as an internal reference protein for normalizing the expression of other target proteins.

C. The pS6 band in the BAV/TAA2 group is darker than that in the Ctrl group.

D. In the Ctrl group, the expression of S6 is significantly lower than that of SMAD2.

E. Cannot be determined.

Answer: B

Explanation:

Option A: The pSMAD2 bands in the BAV/TAA1 and BAV/TAA2 groups appear weaker than those in the Ctrl group, rather than stronger. Therefore, this statement is incorrect.

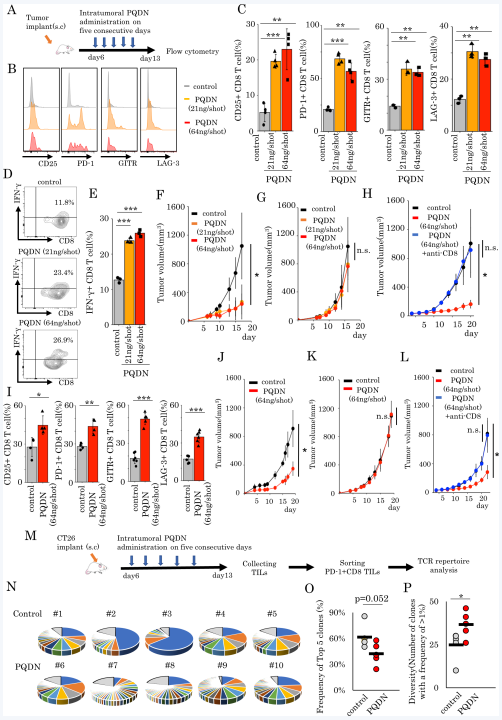
Option B: The GAPDH bands show relatively consistent intensity across the Ctrl, BAV/TAA1, and BAV/TAA2 groups, supporting the use of GAPDH as a loading control for normalization of the target proteins. Therefore, this statement is correct.

Option C: The pS6 band in the BAV/TAA2 group does not appear darker than that in the Ctrl group; it appears similar in intensity. Therefore, this statement is incorrect.

Option D: This statement is not supported by the blot. The band intensities of S6 and SMAD2 should not be directly compared within the same lane to infer which protein is expressed at a higher level, because they are different proteins detected with different antibodies and potentially different exposure conditions. Therefore, this statement is incorrect.

Figure 12: Example of an Information Localization (IL) task in the Panel-Level Fine-Grained Perception stage.

Cross-Panel Relation Understanding: Trend Analysis



Question: **[Understanding_2010105083_3_4]**

By comparing experimental images in Fig. C, Fig. E, and Fig. F, which visual perception statement is incorrect?

A. In Fig. C, PQDN groups visually show a trend of higher molecular expression with increasing dosage.

B. In Fig. E, the PQDN (64ng/shot) group exhibits greater bar-height increase than the 21ng/shot group, aligning with Panel C's molecular expression trend.

C. In Fig. F, tumor volume trendlines of PQDN (64ng/shot) and 21ng/shot groups overlap due to error bars in later stages, suggesting no visual difference.

D. Across panels, PQDN (64ng/shot) visually demonstrates a coherent trend: rising molecular bars → elevated cytokine secretion bars → slowed tumor volume growth.

E. Cannot be visually confirmed.

Answer: A

Explanation:

Option A: This statement is incorrect. In Fig. C, the PQDN groups do not show a uniformly increasing trend in molecular expression with increasing dose across all markers. Although some bars in the 64 ng/shot group appear higher than those in the 21 ng/shot group, this pattern is not sufficiently consistent to support the generalized claim of a clear dose-dependent increase.

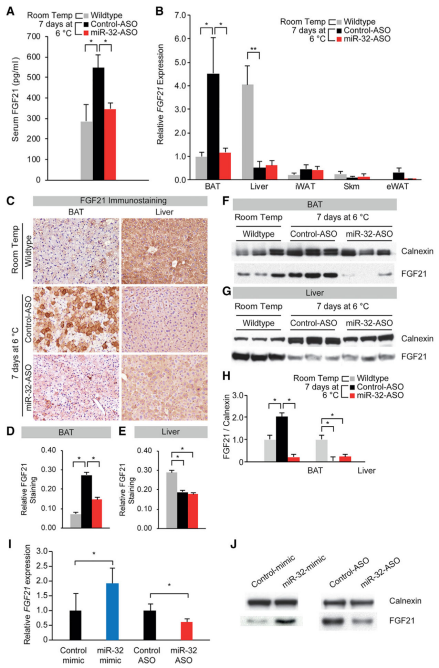
Option B: In Fig. E, the PQDN 64 ng/shot group shows a greater increase in bar height than the 21 ng/shot group, and both are higher than the control group. This is visually consistent with the stronger activity observed in the higher-dose group.

Option C: This statement is acceptable as a visual description. In Fig. F, the tumor growth curves of the PQDN 64 ng/shot and 21 ng/shot groups become visually close in the later stages, and the overlapping error bars reduce the clarity of the difference by visual inspection alone.

Option D: Across Fig. C, Fig. E, and Fig. F, the PQDN 64 ng/shot group shows a broadly consistent visual pattern: relatively higher molecular-response bars, increased cytokine secretion, and slower tumor growth. Therefore, this statement is supported at the level of overall visual trend.

Figure 13: Example of a Trend Analysis (TA) task in the Cross-Panel Relation Understanding stage.

Cross-Panel Relation Understanding: Heterogeneous Integration



Question: **[Understanding_2010143105_4_1]**

How do the visual FGF21 expression levels correlate between the staining and protein band results?

A. Both Fig. C and Fig. F visually show significant FGF21 enhancement by cold stimulation, which is then substantially suppressed by miR-32-ASO treatment in both, demonstrating a highly consistent trend.

B. Fig. C's FGF21 staining enhances with cold but visually shows no significant change after miR-32-ASO treatment, while Fig. F's FGF21 protein band shows enhancement followed by suppression. Their visual trends are not entirely consistent.

C. Under cold stimulation, Fig. C's FGF21 staining visually weakens, contrasting with the visual enhancement trend of Fig. F's FGF21 protein band.

D. miR-32-ASO treatment visually enhances FGF21 staining in Fig. C, and simultaneously, the FGF21 protein band in Fig. F also visually strengthens.

E. Cannot be determined.

Answer: A

Explanation:

Option A: Both Fig. C (immunostaining) and Fig. F (Western blot) show consistent trends. Under cold stimulation (7 days at 6 °C, Control-ASO group), Fig. C demonstrates darkened BAT/Liver FGF21 staining (enhancement), and Fig. F demonstrates thickened/brightened BAT FGF21 bands (enhancement). After miR-32-ASO treatment, Fig. C shows lighter staining (suppression), and Fig. F shows weakened/disappeared bands (suppression). Thus, trends align perfectly across methods.

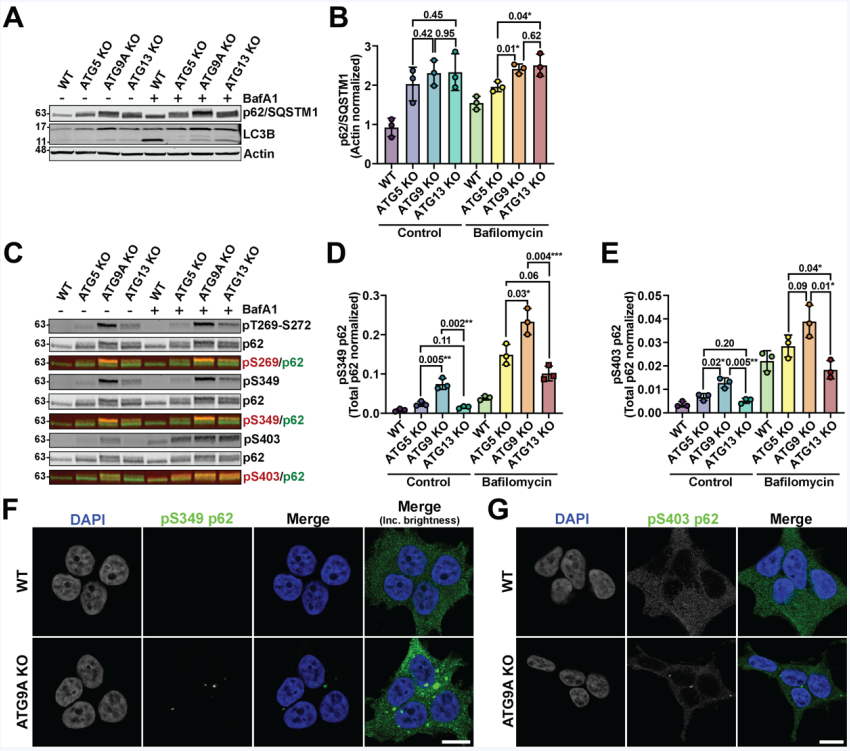
Option B: This option claims that Fig. C shows no significant change after miR-32-ASO treatment. However, Fig. C clearly exhibits lighter staining, directly contradicting this claim. Therefore, it is incorrect.

Option C: This option claims that cold weakens Fig. C staining, but Fig. C actually shows darkening (enhancement) with cold, which is consistent with Fig. F. Therefore, it is incorrect.

Option D: This option claims that miR-32-ASO enhances FGF21, but both methods clearly demonstrate weaker signals (suppression). Therefore, it is incorrect.

Figure 14: Example of a Heterogeneous Integration (HI) task in the Cross-Panel Relation Understanding stage.

Qualitative Reasoning: Chart + Staining Image + Western Blot



Question: **[Reasoning_2010143347_3_2]**

Based on the experimental images, which of the following is correct?

A. In Fig. A, ATG5 knockout leads to p62 accumulation, indicating blocked autophagic degradation independent of other ATG genes.

B. In Fig. B, p62 levels remain unchanged across ATG knockouts after BafA1 treatment, suggesting functional similarity.

C. Figs. C–E show phosphorylated p62 in ATG9A knockout, indicating its role in regulating p62 phosphorylation.

D. Figs. F and G show ATG9A knockout does not affect p62 localization, suggesting no impact on subcellular distribution.

E. Cannot be determined.

Answer: C

Explanation:

Option A: In Fig. A, ATG5 knockout leads to p62 accumulation, but it cannot indicate that the blocked autophagic degradation is independent of other ATG genes, because there may be synergistic interactions among autophagy-related genes. The claim of “independence” goes beyond the scope of the data. Thus, A is incorrect.

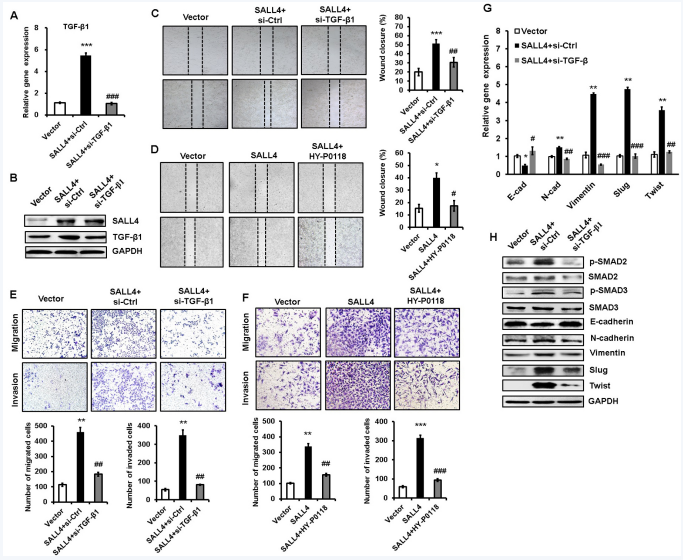
Option B: In Fig. B, the p62 levels of different ATG gene knockout groups change differently after BafA1 treatment, indicating that these ATG genes have different functions in regulating autophagy, rather than similar ones. Thus, B is incorrect.

Option C: Figs. C–E show that in the ATG9A knockout group, the phosphorylation form of p62 changes, which indicates that ATG9A is involved in regulating the phosphorylation of p62. Thus, C is correct.

Option D: Figs. F and G show that the subcellular localization of p62 changes in the ATG9A knockout group, indicating that ATG9A affects the subcellular distribution of p62. Thus, D is incorrect.

Figure 15: Example of a Qualitative Reasoning task in the Expert-Level Reasoning stage.

Quantitative Reasoning: Chart + Staining Image + Western Blot



Question: **[Reasoning_2010045119_4_2]**

In SALL4-overexpressing HGC-27 cells, which best describes key changes?

A. SALL4 raises TGF-β1 mRNA 6×, wound closure $\sim$ 60%, invasion 5×; si-TGF-β1 cuts migration $>$ 70%.

B. si-TGF-β1 lowers Slug, Vimentin, Twist mRNA 60%–80%, protein and p-SMAD2/3 decrease, showing TGF-β1/SMAD inhibition.

C. HY-P0118 reduces wound closure 60% $\rightarrow$ 20%, migration 450 $\rightarrow$ 100, p-SMAD2/3 returns to control, blocking TGF-β1.

D. si-TGF-β1 leaves N-cadherin, Vimentin mRNA unchanged, minor protein drop, E-cadherin stable; no EMT reversal.

E. Cannot be determined.

Answer: B

Explanation:

Option A: This statement is not the best description. Although SALL4 overexpression appears to increase TGF-β1 expression, wound closure, and invasion, the numerical claims (such as "6×," "$\sim$ 60%," "5×," and "$>$ 70%") are overly specific and combine results from different assays into a single summary that is not the most directly supported by the figure. Therefore, A is not the correct choice.

Option B: In Figs. G and H, si-TGF-β1 reduces the mRNA levels of Slug, Vimentin, and Twist, and is also associated with decreased protein levels and reduced p-SMAD2/3 signals. These findings support inhibition of the TGF-β1/SMAD pathway and partial reversal of the EMT-related phenotype. Therefore, B is the best-supported statement.

Option C: Although HY-P0118 appears to reduce wound closure, migration/invasion, and p-SMAD2/3 levels, the statement that p-SMAD2/3 "returns to control" is too strong and not directly established by the figure. Therefore, C is incorrect.

Option D: This statement is contradicted by the data. After si-TGF-β1 treatment, N-cadherin and Vimentin decrease rather than remain unchanged, and E-cadherin increases rather than stays stable, indicating reversal rather than persistence of the EMT-related phenotype. Therefore, D is incorrect.

Option E: The figure provides clear visual and quantitative evidence for changes in TGF-β1 expression, cell migration/invasion, and EMT-related markers. Therefore, the results can be determined from the data.

Figure 16: Example of a Quantitative Reasoning task in the Expert-Level Reasoning stage.