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

Anonymous ACL submission

Abstract

We introduce **SPUR**, a benchmark for scientific experimental image perception, understanding, and reasoning. SPUR features three key innovations: (1) **Panel-Level Fine-Grained Perception**: Assessing MLLMs' visual perception abilities across three core dimensions (*i.e.*, numerical perception, morphological perception, and information localization) on six fine-grained panel types; (2) **Cross-Panel Relation Understanding**: Leveraging complex scientific experimental images with an average of 14.3 panels per sample, we design QA pairs to evaluate MLLMs' ability to understand cross-panel relations; (3) **Expert-Level Reasoning**: We design qualitative and quantitative reasoning questions across five experiment types to specifically assess whether models, like human experts, can infer experimental conclusions from complex scientific images. Evaluation of 20 MLLMs and 4 MCoT methods reveals they are vastly inadequate in meeting the expert-level perception, understanding, and reasoning requirements for scientific experimental images, as mandated by AI for Science (AI4S) research. Data and code are available: `https://anonymous.4open.science/r/SPUR-1797`.

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in ***perceiving*** and ***understanding*** scientific images across diverse domains, including 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), establishing their potential for AI-driven scientific research (Xu and Peng, 2025). The multimodal ***reasoning*** capacities of MLLMs are further enhanced through Multimodal Chain-of-Thought (MCoT) techniques, which amplify structured reasoning via (1) prompt enhancement (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).

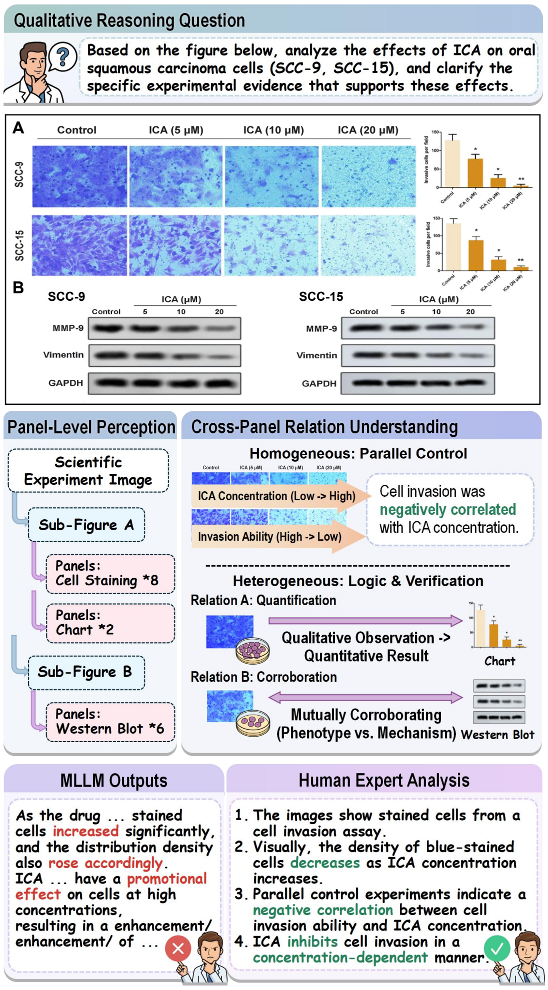


Figure 1: Example of MLLM reasoning for experimental conclusion inference, involving perceiving panel-level morphological information from staining images, understanding the trend across 8 staining images (staining intensity decreases with increasing ICA concentration), and deducing ICA inhibits cell invasion.

Despite these advancements, complex multi-panel scientific experimental images common in scholarly publications remain underexplored. As

Benchmark	Type	Size (k)	Panel-Level Perception			Cross-Panel Understanding			Expert-Level Reasoning		
			Gran.	Dens.	Attribute	# Panels/Image	Desc.	Rel.	Paradigm	Qual.	Quant.
Non-Academic Image											
ScienceQA	Science	21.2	Figure	40.21%	N+IL	1.06	✘	✘	VQA	✔	✘
MMMU	Science	11.5	Figure	43.15%	N+IL	2.53	✘	✘	VQA	✔	✘
M3CoT	Science	11.4	Figure	51.83%	N+IL	1.03	✘	✘	VQA	✔	✘
EMMA	Science	2.7	Figure	53.21%	N+IL	1.26	✘	✘	VQA	✔	✔
Academic Image											
MMSci	Science	873.7	Sub-fig.	64.28%	N+IL	7.40	✔	✘	Caption	✘	✘
SciAssess	Science	6.9	Figure	58.16%	M+IL	3.40	✘	✘	VQA	✔	✘
SFE	Science	0.8	Sub-fig.	51.40%	N+M+IL	2.30	✔	✘	VQA	✔	✘
Text2Analysis	Statistic	2.2	Figure	Table	N+IL	1.00	✘	✘	VQA	✔	✔
EvoChart	Statistic	1.2	Figure	Table	N+IL	1.00	✘	✘	VQA	✔	✔
MISS-QA	CS	1.5	Sub-fig.	53.47%	N+IL	1.55	✘	✘	VQA	✔	✘
SPIQA	CS	270	Sub-fig.	39.46%	N+IL	2.71	✘	✘	VQA	✔	✘
OmniMedVQA	Medicine	127.9	Sub-fig.	62.20%	M + IL	1.00	✘	✘	VQA	✔	✘
MicroVQA	Medicine	1.0	Panel	59.83%	N+M	1.91	✔	✔	Exp. Design	✔	✘
SPUR (ours)	**Experiment**	**4.2**	**Panel**	**76.50%**	**N+M+IL**	**14.3**	✔	✔	**Exp. Reason**	**36.10%**	**63.90%**

Table 1: Comparison of SPUR and other related benchmarks. **Gran.**: Perception Granularity; **Dens.**: Ratio of non-blank area to total image area; **N**: Numerical; **M**: Morphological; **IL**: Information Localization; **Desc.**: Information Description; **Rel.**: Relation Understanding; **Qual.**: Qualitative Reasoning; **Quant.**: Quantitative Reasoning.

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 relationships, and (3) **qualitative & quantitative reasoning** toward scientific conclusions. However, as summarized in Table 1, current scientific reasoning benchmarks exhibit critical limitations versus real-world perception, understanding, and reasoning demands:

- **Absence of Scientific Experimental Images** Crucially, scientific experimental images (*e.g.,* cellular staining images, biological band diagrams) exhibit unique visual characteristics that pose fine-grained perception challenges compared to general scientific images. 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 scientific experimental images (<5%).
- **Lack of Cross-Panel Relation Understanding** Interpreting complex experimental images requires comprehending panel correlations (*e.g.,* control group, quantification), a core capability for scientific images understanding. However, current scientific reasoning 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 3 panels per image.
- **Neglect of Quantitative Reasoning** Quantitative reasoning, involving precise analysis of proportions, magnitudes, and metrics, poses greater challenges than qualitative tasks like conclusion direction judgment. Although pioneering works have addressed qualitative aspects (*e.g.,* experimental design in MicroVQA (Burgess et al., 2025)), they largely neglect MLLMs' quantitative reasoning capabilities for deriving scientific conclusions from experimental evidence.

Therefore, we propose **SPUR**, a benchmark for multimodal ***perception***, ***understanding***, and ***reasoning*** on scientific experimental images, comprising 1k expert-curated image-text pairs and 4.2k QA pairs. SPUR offers three key advantages:

- **Rich Scientific Experimental Images** SPUR exclusively features images sourced from high-quality experimental data in open-access PubMed papers. They span seven diverse disciplines (*i.e.,* Molecular Biology, Cell Biology, Oncology, Neuroscience, Immunology, Physiology, and Microbiology) and encompass five key experimental categories (*i.e.,* cell experiments, animal experiments, observational studies, intervention studies, and pathological imaging experiments).
- **Fine-Grained Panels & Complex Cross-Panel Relationships** SPUR utilizes complex scientific experimental images containing an average of 14.3 panels per sample. These are categorized into six fine-grained types (*i.e.,* statistical graphs, western blots, and four differentiated staining

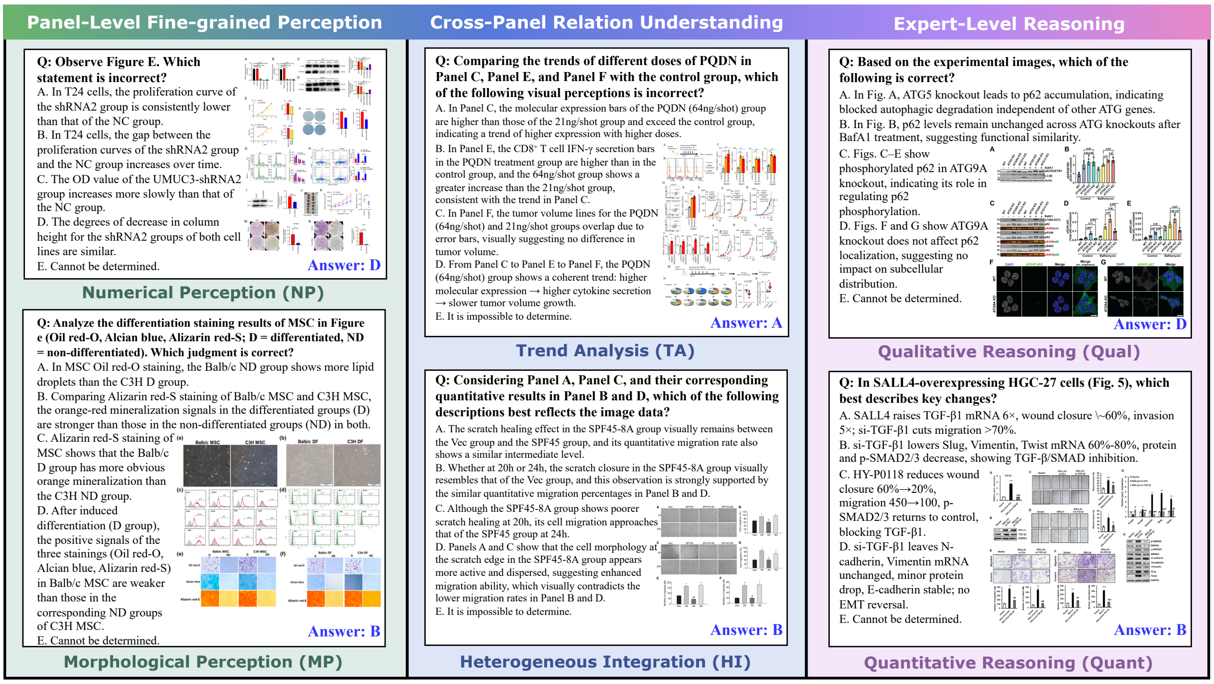


Figure 2: Examples of substasks in SPUR's three cognitive stages. More examples are provided in Appendix A.

images). Further, we model cross-panel relationships, including isomorphic panel relations (*e.g.,* dose-response trends across staining images) and heterogeneous panel relations (*e.g.,* western blot vs. subcellular staining image co-validation).

- **Multi-Dimensional Evaluation Tasks** We establish a three-stage evaluation framework (*Perception→Understanding→Reasoning*) with 7 subtasks to assess MLLMs: (1) *Perception* stage focuses on **panel-level analysis**, evaluating numerical data extraction (*e.g.,* estimate kinetic curve values), morphological recognition (*e.g.,* cell invasion extent identification), and information localization; (2) *Understanding* stage deciphers **cross-panel relationships**, testing trend analysis (TA) for homogeneous panels and heterogeneous integration (HI) across diverse panels; (3) *Reasoning* stage requires **expert-level qualitative & quantitative reasoning**, assessing experimental directional conclusion (*e.g.,* reagent effectiveness validation), and quantitative conclusion.

We evaluate 8 proprietary and 12 open-source MLLMs, along with 4 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 scientific experimental image perception-understanding-reasoning.** Only `Gemini 3 Pro Preview` exceeds 60%, falling short of AI4S requirements.
- **Panel-level perception shows critical weaknesses.** Numerical perception and localization yield the poorest results (most models <50%). Although morphological perception performs best, it exhibits category bias and limited generalization. **Cross-panel understanding remains the core bottleneck:** accuracy plummets as relationship complexity increases in homogeneous panel tasks. **Quantitative reasoning significantly trails qualitative reasoning by 10%-30%,** failing expert-level analysis demands.
- **MCoT methods lack consistent performance gains.** Prompt-based methods only enhance model reasoning steps without improving perceptual capabilities, potentially amplifying perceptual errors and failing to produce positive outcomes. Plan-based methods slightly improve perception in some models, highlighting perceptual enhancement as a critical research direction.

2 SPUR Benchmark

We introduce SPUR to evaluate MLLMs' ***perception***, ***understanding***, and ***reasoning*** capabilities on scientific experimental images through 7 subtasks structured in these three cognitive stages. SPUR comprises 4,264 multiple-choice questions (MCQs) with 1,000 scientific images and texts from PubMed, spanning 7 disciplines, 5 experiment categories, and 6 fine-grained panel types.

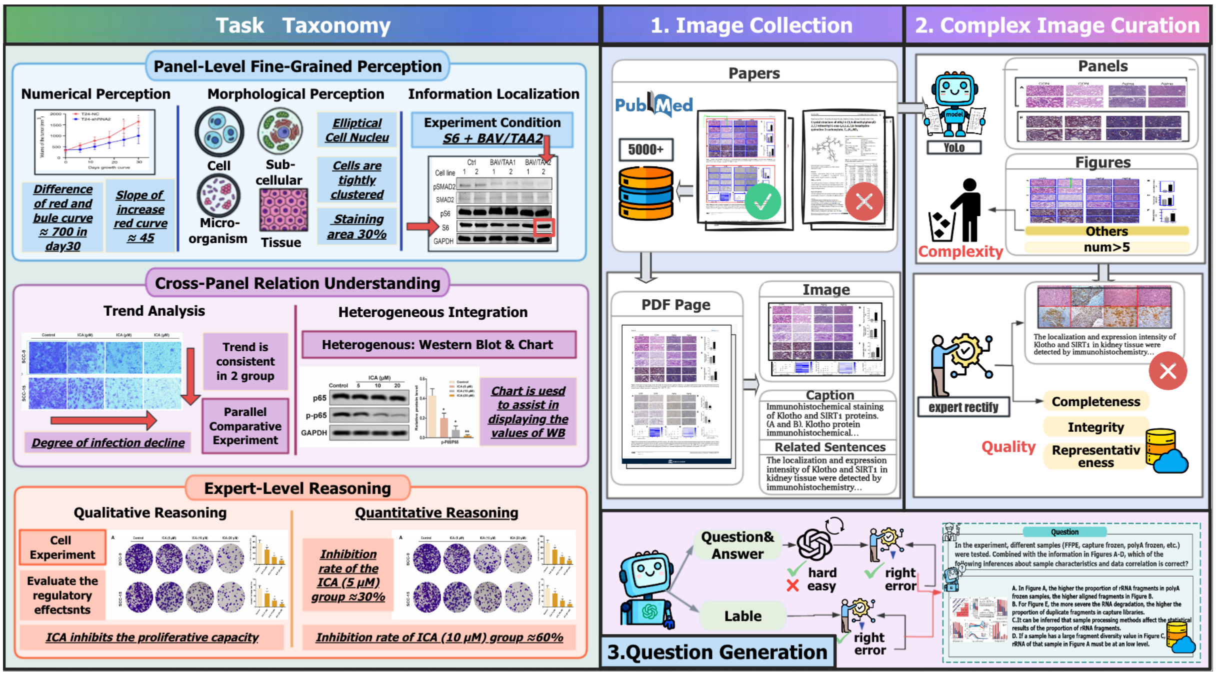


Figure 3: The details of the 7 subtasks classification **(Left)** and the overview of QA pair curation pipeline **(Right)**.

2.1 Task Taxonomy

SPUR structures seven subtasks across *Perception→Understanding→Reasoning* stages:

Panel-Level Fine-Grained Perception Identifies and parses visual features within individual panels (*e.g.,* stained preparations), establishing 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 corresponding experimental conditions.

Cross-Panel Relation Understanding Extracts implicit relationships across multi-panel frameworks (*e.g.,* causal, comparative, and argumentative patterns):

- **Trend Analysis (TA):** Interprets directional changes and experimental content in 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 Derives scientific inferences from multi-panel visual evidence integrated with experimental context:

- **Qualitative Reasoning (Qual):** Synthesizes visual cues, domain knowledge, and experimental design logic interpreting biological significance.
- **Quantitative Reasoning (Quant):** Conducts mathematical verification and quantitative evaluation of inter-group differences, effect strength, 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 four-stage process (see Appendix B for details):

Image Collection We systematically curated more than 3,000 high-impact open-source papers from PubMed, applying three selection criteria: publication within 10 years, and source journal impact factor (IF > 3.0). Using automated PDF parsing, we extracted more than 5.6k scientific images, which were manually annotated with related sentences, contextual captions, and disciplinary classifications (7 categories) to form the raw dataset.

Complex Image Curation To ensure question complexity, images underwent dual-filtering: first, a YOLO-based panel detector segmented images and excluded those with ≤ 6 panels (removing 77.5% of candidates); second, medical experts verified experimental workflow completeness and annotation accuracy, discarding an additional 14.2% of images that lacked methodological rigor.

Task Categorization & Question Generation Domain experts developed specialized prompt templates aligned with our task hierarchy, enabling GPT-4o to generate 7.6k candidate QA pairs.

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% of candidates), ensuring retained items require genuine visual reasoning. After that, each triplet (question-options-answer) underwent expert review for biomedical validity and task alignment, yielding a 28% rejection primarily due to factual inaccuracies or task mismatch.

2.3 Data Quality Assurance

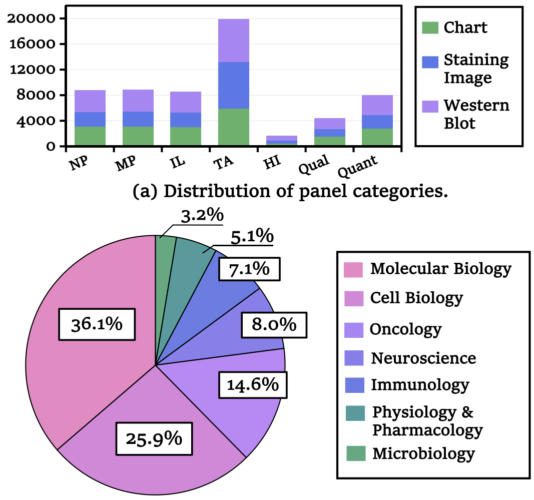
We implemented a rigorous quality control system through standardized expert certification protocols. Four medical specialists (>40 peer-reviewed publications each) and two senior experts (>100 publications) conducted multi-tiered validation: each sample underwent independent dual-expert review, flagging inadequate information or ambiguous answers, with senior arbitrators resolving discrepancies. This three-month process (biweekly sessions averaging 200 samples/day) was funded by an independent project grant ensuring impartiality. See Appendix B.3 for detailed guidelines and statistics.

2.4 Data Statistics

As shown in Table 2 and Figure 4, SPUR comprises 4,264 expert curated samples derived from 1,000 scientific experimental images, covering over 60,000 panels across seven disciplines. With an average of 14.3 panels per image, it establishes six fine grained panel types enabling complex cross-panel relationships, significantly surpassing existing benchmarks in structural complexity and better reflecting real academic scenarios. The benchmark further features multi-dimensional evaluation tasks with 636 NP, 634 MP, 621 IL, 1,357 TA, 130 HI, 567 Qualitative Reasoning, and 319 Quantitative Reasoning samples. Its high complexity is evidenced by option texts averaging 54.6 words and reasoning evidence spanning 58.8 words, demanding comprehensive perception to reasoning capabilities beyond general multimodal benchmarks. More dataset statistics are shown in Appendix B.6.

Property	Value
Data Source	
# Scientific Experiment Category/Image	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. Length (Question/Option/Evidence)	20.8/54.6/58.8

Table 2: Key statistics of SPUR.



(a) Distribution of panel categories.

(b) Distribution of subjects of SPUR images.

Figure 4: Distribution of panel categories and subjects.

3 Experiments

3.1 Experiment Setup

Models We evaluate 20 state-of-the-art MLLMs with visual capabilities (8 proprietary, 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 seven subtasks.

MCoT Strategies We benchmark two categories of MCoT strategies comprising four state-of-the-art methods: (1) *Prompt-based* methods (DD-CoT (Zheng et al., 2023), VoT (Wu et al., 2024)) using designed prompts for rationale generation; (2) *Plan-based* methods (VIC (Zheng et al., 2024a), 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 for scientific experimental images? **RQ2**: Can MLLMs achieve precise panel-level analysis (numerical/-morphological/spatial)? **RQ3**: How effectively do

Model	Size	Panel-Level Perception				CPRU			Expert-Level Reasoning			Overall
		NP	MP	IL	Avg.	TA	HI	Avg.	Qual	Quant	Avg.	
Proprietary MLLMs												
Gemini 3 Pro Preview		**61.62**	**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	106B	57.70	61.99	57.65	59.12	55.71	68.46	56.83	80.94	58.48	66.59	59.87
Ministral	14B	50.88	61.40	56.56	56.25	**57.79**	**70.00**	**58.86**	72.50	56.81	62.49	58.48
Ministral	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	400A17B	51.73	59.78	56.61	56.03	51.88	58.46	52.46	84.64	57.02	67.01	57.06
Qwen3 VL (Thinking)	30B	53.48	58.00	55.27	55.59	54.85	61.24	55.41	75.16	54.17	61.75	56.80
InternVL 3	14B	45.00	54.59	53.70	51.06	50.27	61.90	51.29	69.13	49.19	56.35	52.25
Mistral Small 3.1	24B	42.74	51.92	53.48	49.36	51.82	56.92	52.28	72.82	47.56	56.84	51.94
Qwen3 VL (Instruct)	30B	44.18	53.31	51.05	49.50	50.41	63.08	51.51	66.88	51.41	57.00	51.76
Qwen2.5 VL	72B	38.11	45.57	48.87	44.14	51.87	61.90	52.74	73.10	52.51	59.95	50.38
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
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	25.95	36.13	40.72	34.20	33.89	42.19	34.62	48.59	35.65	40.34	35.63

Table 3: Evaluation results of three stages and seven subtasks on SPUR. **CPRU**: Cross-Panel Relation Understanding.

MLLMs understand complex cross-panel relationships? **RQ4**: Do MLLMs perform expert-level qualitative and quantitative reasoning? **RQ5**: Do MCoTs 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`, all models achieve under 60% accuracy on SPUR. This substantial gap indicates MLLMs' perception, understanding, and reasoning capabilities fall far below expert-level requirements for scientific experimental images.

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

Cross-panel understanding presents critical bottlenecks. Models attain significantly lower accuracy in **Trend Analysis** tasks than other subtasks, with most showing poorest performance here. This underscores accurate comprehension of complex cross-panel relationships as a core challenge.

Reasoning tasks reveal substantial disparities. While reasoning accuracy exceeds cross-panel understanding, quantitative reasoning lags qualitative reasoning by 10%–30%. These results collectively indicate current MLLMs cannot meet expert-level quantitative analysis demands in AI4S research.

3.3 Panel-Level Perception Results (RQ2)

Numerical perception shows the weakest performance. Most MLLMs struggle with numerically precise visual content, revealing fundamental limitations in fine-grained numerical parsing. Among proprietary models, `Gemini 3` leads at 61.62%; among open-source models, only `GLM 4.5V` (57.70%) approaches mid-tier proprietary performance, with most below 50%. This deficiency directly constrains quantitative reasoning on scientific images.

Weak information localization impedes scientific experiment understanding. Information Localization yields the lowest scores among ***perception*** subtasks, where even top-performing `Gemini 3` achieves only 59.67%. This limitation prevents accurate panel-to-condition linking, hindering cross-panel relationship construction and subsequent reasoning.

Morphological perception achieves the highest but unstable performance. As shown in Figure 5(a), fine-grained analysis across four stain categories reveals significant imbalance. For instance, `Ministral 14B` scores 70.52% for Subcellular images but only 42.80% for Microorganism. These fluctuations indicate limited generalizability, with biases attributable to category-specific visual-semantic features learned during training.

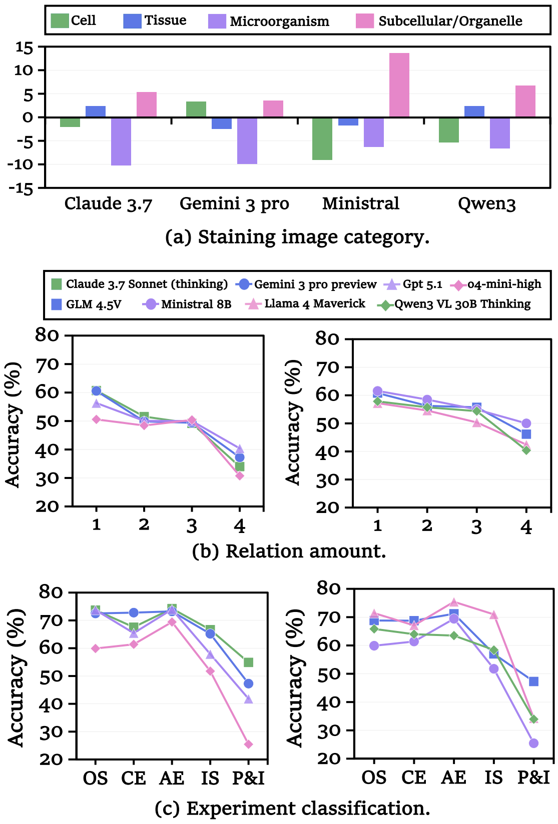


Figure 5: Fine-grained results based on **(a)** staining image category, **(b)** relation amount, and **(c)** experiment classification. Details are provided in Appendix C.2.

3.4 Cross-Panel Understanding Results (RQ3)

Models struggle with trend analysis in homogeneous panels. Figure 5(b) confirms the high difficulty of trend tasks: accuracy inversely correlates with relationship complexity. As relationships increase from 1 to 4, `Claude 3.7 Sonnet`'s accuracy drops from 75% to 35%, reflecting MLLMs' inability to handle complex multi-relationship logic.

Heterogeneous integration yields significantly higher performance. This discrepancy occurs because relationships across different experiments typically involve supportive/confirmatory connections in academic contexts, whereas homogeneous panels generate diverse trends creating more challenging questions.

3.5 Expert-Level Reasoning Results (RQ4)

Quantitative reasoning presents greater challenges than qualitative reasoning. Qualitative tasks assess directional conclusions (*e.g.*, "A promotes B"), while quantitative tasks evaluate magnitude (*e.g.*, "A stimulates B by 50%"). As Figure 5(c) shows, models achieve higher accuracy on Observational Study (OS) and Animal Experiment (AE) images than on Cell Experiment (CE) and Interventional Study (IS) images. This discrepancy stems from CE/IS experiments requiring more complex quantitative analysis and intricate experimental logic, demanding stronger reasoning.

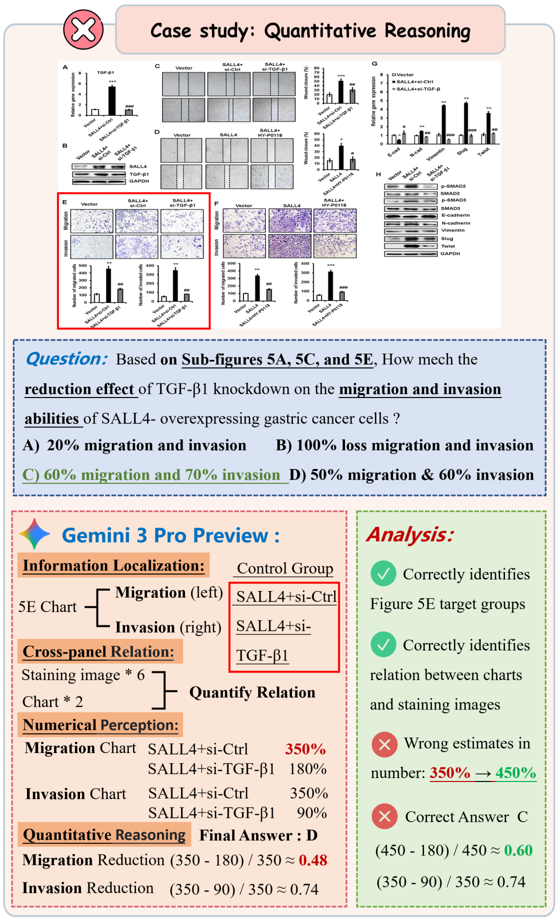


Figure 6: Error analysis of a Quantitative Reasoning question, showing how `Gemini 3 Pro Preview`'s numerical perception errors lead to reasoning failure. More cases are shown in Appendix D.1.

Quantitative reasoning insights guide future AI4S research. As shown in Figure 6, this task comprehensively evaluates MLLMs. Numerical perception provides foundational support, while cross-panel understanding enables experiment interpretation. Future AI4S development necessitates mature quantitative reasoning, yet current models remain inadequate for such expert-level tasks.

3.6 MCoT Results (RQ5)

MCoT methods fail to consistently enhance performance. As shown in Table 4, prompt-based approaches merely augment reasoning steps when processing complex images. If models cannot accurately parse fine-grained content (*e.g.*, numerical values, cross-panel relationships), additional chain-of-thought guidance may amplify perceptual errors

Model	Panel-Level Perception				Cross-Panel Understanding			Expert-Level Reasoning			Overall
	NP	MP	IL	Avg.	TA	HI	Avg.	Qual	Quant	Avg.	
GLM 4.5V											
`Direct`	62.20	61.99	57.65	59.12	55.71	68.46	56.83	80.94	58.48	66.59	59.87
`DDCoT` (Zheng et al., 2023)	47.11	47.67	42.83	45.88	45.24	56.35	46.19	71.52	53.27	59.93	48.89
`VoT` (Wu et al., 2024)	55.82	62.30	57.65	58.59	53.65	60.77	54.27	78.44	57.77	64.33	58.47
`VIC` (Zheng et al., 2024a)	35.50	33.87	32.09	33.83	27.20	29.92	27.43	34.59	36.52	35.83	32.02
`Cantor` (Gao et al., 2024)	53.41	58.69	52.96	55.09	51.23	60.00	52.00	77.12	56.61	64.05	55.85
Ministral 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` (Zheng et al., 2023)	47.22	52.30	47.97	49.17	51.75	61.24	52.58	74.61	48.48	57.97	52.19
`VoT` (Wu et al., 2024)	50.00	59.05	55.68	44.89	56.67	67.97	57.66	71.92	55.14	61.24	57.17
`VIC` (Zheng et al., 2024a)	40.10	40.00	37.34	39.16	36.93	43.08	37.47	52.50	42.55	46.15	40.02
`Cantor` (Gao et al., 2024)	45.32	49.92	45.94	46.99	37.77	43.41	38.27	59.55	49.35	53.07	45.17
Qwen3 VL 30B 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` (Zheng et al., 2023)	41.04	47.00	45.41	44.77	43.58	60.47	45.14	63.95	43.79	51.07	46.08
`VoT` (Wu et al., 2024)	46.70	56.08	53.95	52.22	51.25	60.77	52.09	68.44	51.59	57.67	53.31
`VIC` (Zheng et al., 2024a)	39.08	43.46	40.77	41.12	38.65	48.84	39.56	55.84	41.58	46.81	41.76
`Cantor` (Gao et al., 2024)	49.19	55.20	51.64	52.02	46.82	52.71	47.34	71.79	50.18	58.15	51.63

Table 4: Results of four MCoT methods based on three base models. Complete results are provided in Appendix C.3.

rather than improve outcomes.

Plan-based MCoTs partially improves perception. For `Qwen3 VL 30B 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 address perceptual limitations, enabling reasoning improvements and suggesting optimized strategies for scientific image tasks. See Appendix D.2 for more MCoT cases.

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 7 subtasks designed to address these limitations.

Multimodal Chain-of-Thought MCoT techniques 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). Current 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 designed to evaluate MLLMs' scientific experimental image *perception*, *understanding*, and *reasoning* capabilities, while assessing various MCoT methods in this complex scenario. Experiments reveal a significant performance gap between MLLMs and expert-level reasoning: all models except `Gemini 3 Pro Preview` achieve accuracy below 60%. Enhancing quantitative reasoning capabilities remains critical for future research. We expect this work to advance scientific image reasoning technologies in AI4S.

Limitations

The evaluation framework lacks dynamic visual information and interactive capabilities. Current images are exclusively static, contrasting with real academic scenarios involving dynamic processes such as time series imaging data. Task designs focus solely on static question answering without interactive analysis features like region zoom in or multi turn reasoning, limiting assessment of MLLMs' ability to process dynamic and interactive scientific visual information.

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, and 1 others. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 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, and 1 others. 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*, 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 *International Conference on Machine Learning*, 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 1 others. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024a. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 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 multi-modal 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. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 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.

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, and 1 others. 2025. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 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. *arXiv 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5487–5495.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. 2024. Micro-bench: a vision-language benchmark for microscopy understanding. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 30670–30685.

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 multimodal 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. *Advances in Neural Information Processing Systems*, 37:18695–18728.

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, and 1 others. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 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*, 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. *Preprint*, arXiv:2302.00923.

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. *Advances in Neural Information Processing Systems*, 36:5168–5191.

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, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, Wenxuan Huang, Jiaqi Wei, Dan Si, Xiuqi Yao, Jia Bu, Haiwen Huang, Manning Wang, Tianfan Fu, Shixiang Tang, and 10 others. 2025. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning. *Preprint*, arXiv:2506.10521.

This Appendix is organized as follows:

- Appendix A shows the examples of 7 subtasks in SPUR.
- Appendix B contains details about the annotation.
- Appendix C contains the experiment details and fine-grained analysis of experiment results;
- Appendix D contains additional case studies for the main result and MCoT experiments.

A Example of the SPUR

This section includes:

Panel-Level Fine-Grained Perception Cases: From Figure 10 to Figure 12 show the image-question-options pair of Numerical Perception(NP), Morphological Perception(MP), and Information Localization (IL).

Cross-Panel Relation Understanding Cases: From Figure 13 to Figure 16 show the image-question-options pair of Trend Analysis (TA) and Heterogeneous Integration (HI).

Expert-Level Reasoning Cases: From Figure 17 to Figure 22 show the image-question-options pair in different kinds of panel combinations of Qualitative Reasoning (Qual) and Quantitative Reasoning (Quant).

B Details of Annotation

B.1 Image Collection Annotation Protocol

Source Paper Selection Criteria: Curation is limited to open-source papers from PubMed. Only papers published within the past 10 years are eligible (ensuring timeliness of scientific content). Source journals must have an Impact Factor (IF) > 3.0 (ensuring academic authority). Human annotators (not experts) review candidate papers to confirm they contain complex experimental images (excluding review papers, meta-analyses, or theoretical studies without original experimental data).Figure 7 shows the front-end page of the selection platform used in this step.

Manual Annotation of image-text pair: After the image complexity filter, human annotators perform the following annotations for each remained images:

- **Related Sentences:** Extract 3-5 key sentences from the paper that describe the image's experimental background, methods, or results.
- **Images Captions:** Standardize the image's original caption (correcting typos, supplementing ambiguous information, and ensuring consistency in terminology).
- **Subject Classification:** Categorize images into 7 predefined disciplinary categories (e.g., cell biology, molecular biology, pathology) based on the paper resource journal.

B.2 Panels Recognition

In our preliminary work, we have built a panel recognition dataset from journals in the fields of bio-medical and materials, and annotated the coordinates and categories of panels. We trained a YOLO-based model to recognize different kinds of panels, filtering out low-complexity images based on panel number and distribution.

To ensure the images adequately represent scientific complexity and diversity, we use this model to distinguish different panels in our candidate images. Automatically mark images with ≤ 6 panels as "low-complexity" and exclude them from the dataset. The visualization overview of the panels category distribution is shown in Figure 8.

B.3 Expert Validation Protocols

This annotation and screening process ensured the high quality and effectiveness of the final dataset (4,264 QA pairs retained). The entire quality control process was completed by four medical specialists (>40 peer-reviewed publications each) and two senior experts (>100 publications) over three months (biweekly sessions averaging 200 samples/-day), funded by an independent project grant ensuring impartiality.

Expert Review (Image): Core Objective: Eliminate images with incomplete experimental workflows or inadequate methodological rigor to ensure the scientific validity of subsequent QA tasks.

- **Validation Criteria**: Check if the image reflects a complete experimental design (e.g., presence of control groups, clear experimental variables, and measurable outcomes). Verify if the experimental methods (reflected by the image and corresponding paper content) are standardized (e.g., clear sample labeling, appropriate detection reagents, and repeatable operations).
- **Execution Rules**: Each image is independently reviewed by two medical experts. If both experts approve, the image is retained; if both reject, it is excluded. For conflicting judgments (one

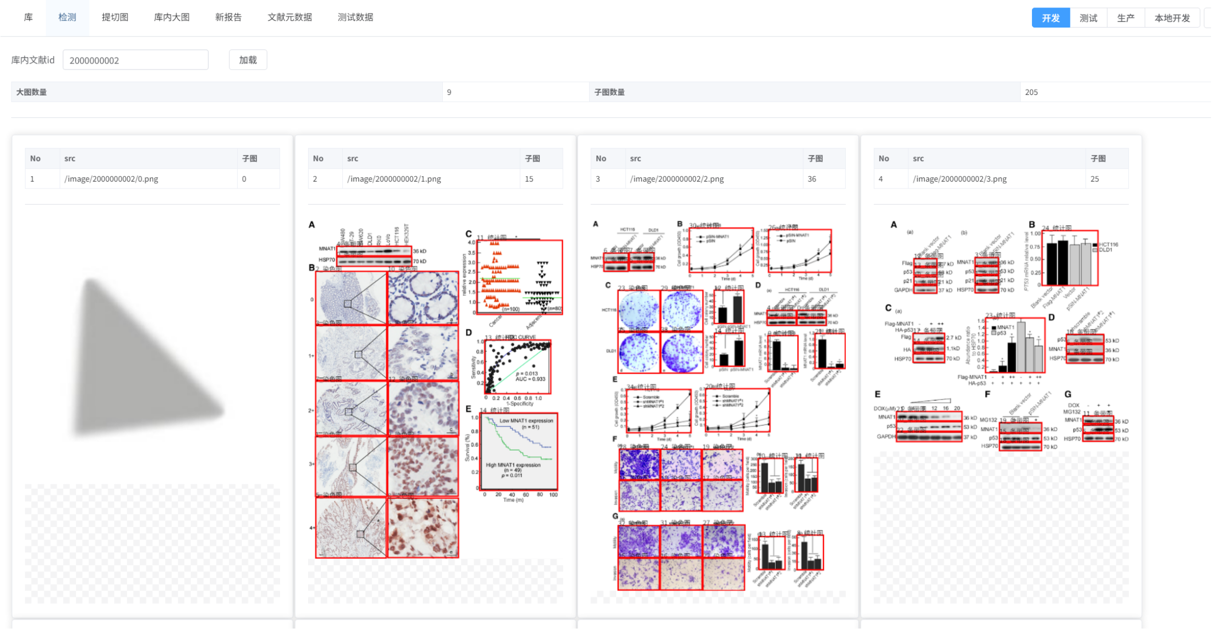


Figure 7: Academic paper selection platform.

Screening Stage	Operation Description	Removed Quantity	Retained Quantity
Initial Candidate Set	Extract images from PubMed papers	-	Images: 5,632
Image Filter	Exclude images with $\leq$6 panels	Images: 4,368	Images: 1,264
Expert Review (Image)	Exclude images with incomplete workflow	Images: 180	Images: 1,084
QA Pair Regeneration	Generate candidate QA pairs based on scientific images	-	Images: 1,084; QA pairs: 7,608
Textual Shortcut Filter	Remove QA pairs with $\geq$5 correct text-only responses	QA pairs: 1,612	Images: 1,084; QA pairs: 5,996
Expert Review (QA Pair)	Remove QA pairs with factual errors or image-content mismatch	QA pairs: 1,732	Images: 1,084; QA pairs: 4,264

Table 5: Details in every step of SPUR construction and annotation.

approve, one reject), a senior expert conducts a final review and makes a decisive judgment.

Expert Review (QA Pairs): Eliminate questions with factual inaccuracies or task mismatch with the image content.

- **Scientific Validity**: Check if the answer of the QA pair conforms to scientific facts (e.g., no factual errors in experimental conclusions, correct use of professional terminology).
- **Task Alignment**: Verify if the QA pair aligns with the predefined task hierarchy and if the required reasoning type matches the image content (e.g., a "quantitative trend analysis" question must correspond to image data that can be quantified).
- **Visual Reasoning Necessity**: Reconfirm that the retained QA pairs cannot be fully answered by text alone and must rely on image information.

B.4 Annotation Statistics

Initially, we extracted more than 5,632 scientific images with captions and related sentences from more than 3k curated PubMed papers as candidate instances, and generated 7,608 candidate QA pairs based on these images. During the manual annotation and validation phase, following the protocol described above, a series of screening and rejection were conducted to ensure the quality of the dataset, with 4,264 QA pairs retained in the final dataset. The specific distribution of the screening and revision results is as follows:

- **77.5% of candidate images (4,368 images)** were excluded in the first round of filtering due to having $\leq$ 6 panels, failing to meet the complex image criteria. (1,264 images remain)
- **14.2% of remaining candidate images (180 images)** •were discarded after medical expert review due to incomplete experimental workflows or inadequate methodological rigor. (1,084 images remain)
- Remain scientific experiment images (1,084) were applied to generate 7608 candidate QA pairs
- **21% of candidate QA pairs (1,612 pairs)** were discarded after the textual shortcut elimination

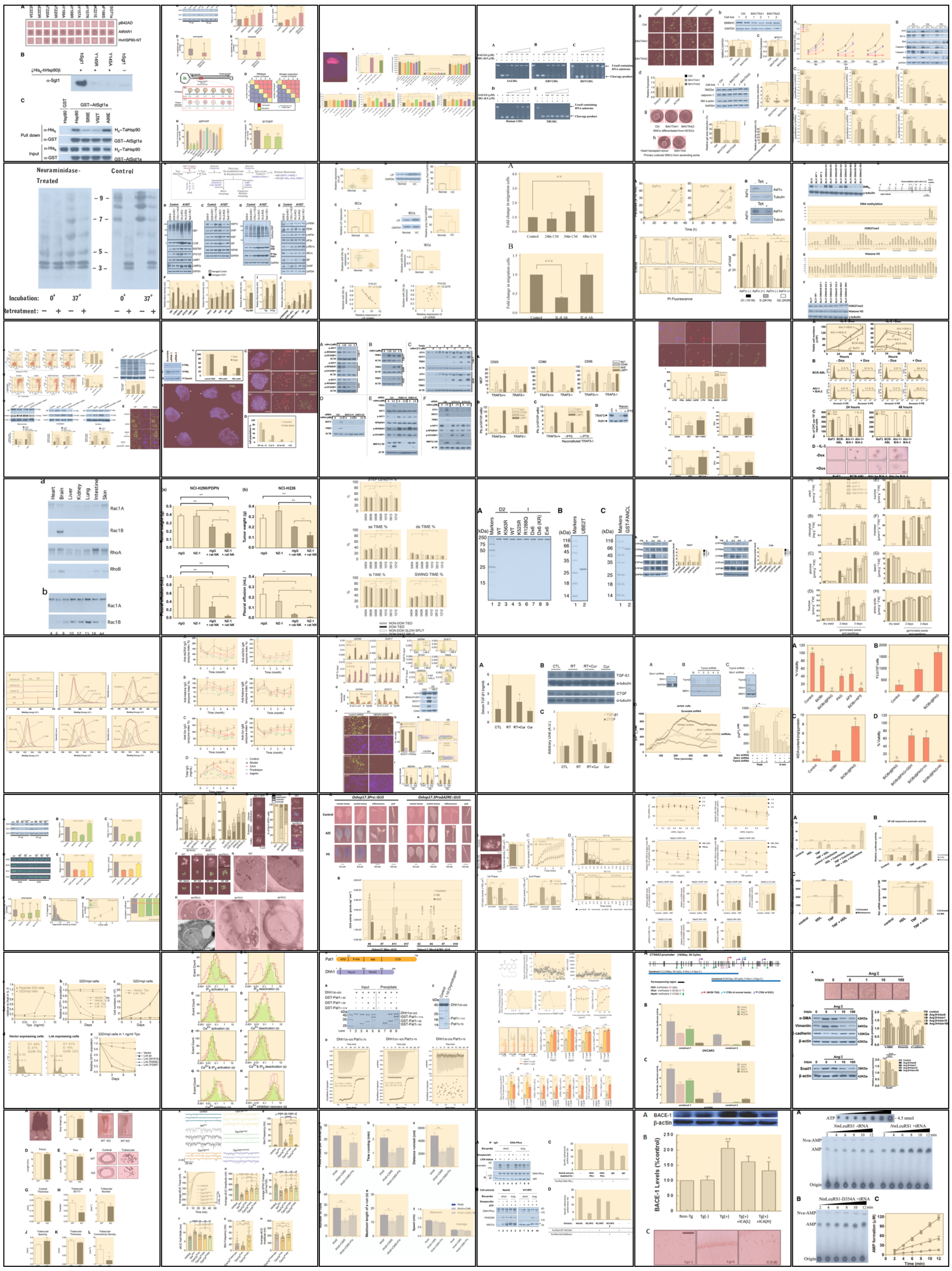


Figure 8: The distribution of panel categories in SPUR. Red masks represent panels of the staining image; light yellow masks represent panels of the chart. Light blue masks represent panels of the Western blot.

filter, as they had $\geq$5 correct responses in text-only queries, relying on textual shortcuts rather than visual reasoning, leaving 5,996 QA pairs.

- **28% of the 5,996 QA pairs (1,732 pairs)** were rejected during expert review primarily due to factual inaccuracies or task mismatch with the image content, resulting in 4,264 QA pairs.

B.5 Prompts for Questions Generation

After sufficiently filtering images based on panel complexity, and manually annotated image-text pairs. We input these images alongside their captions and relevant sentences in academic papers into GPT-4o to generate questions, options, and answers. The prompts to generate QA pairs in Panel-Level Fine-grained Perception, Cross-Panel Relation Understanding, and Expert-Level Reason-

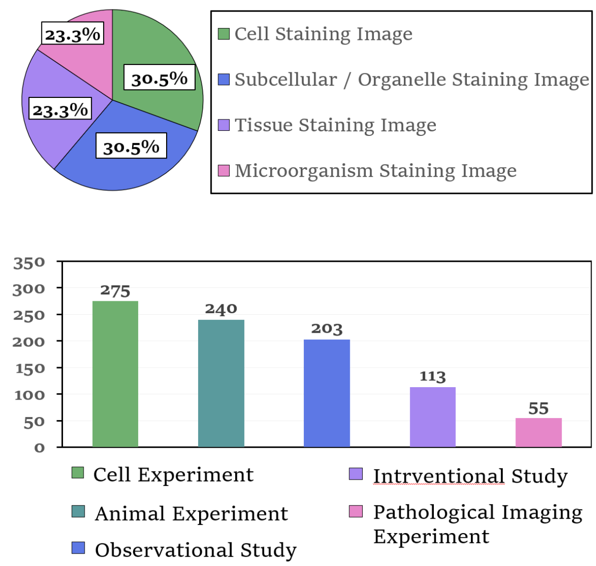


Figure 9: The distribution of fine-grained staining images and experiment classification.

ing are shown in the next several pages.

B.6 Diversity

The SPUR dataset demonstrates remarkable multi-disciplinary coverage and task diversity in its content composition. Its image sources span 8 distinct disciplinary fields, and each image is typically composed of multiple panels, mainly encompassing three types of visual representations: staining images, western blots, and statistical charts. Among these, staining images are further subdivided into 6 fine-grained sub-categories, including cell staining and tissue staining. As shown in Figure 9, SPUR supports two types of scientific reasoning tasks, qualitative reasoning and quantitative reasoning. Question in these sub-tasks covering 5 typical experimental paradigms.

C Experiment Details

C.1 Experiment Models information

Models Since SPUR requires models to interpret images, we only evaluate foundation models with vision capabilities. These include both proprietary models (e.g., OpenAI's GPT-4o, Google's Gemini 2.5/3 Pro Preview, Anthropic's Claude 3.7 Sonnet, ByteDance's Doubao Seed 1.6, xAI's Grok 4.1 Fast) and open-source models (e.g., Meta's Llama 4 Maverick, Google's Gemma 3, Mistral's series, Alibaba's Qwen3 VL, InternVL's variants, OpenGVLab's LLaVA models, Z.AI's GLM 4.5V). Models' details (platform, model name, release time, version, and size) are shown in Table 6.

C.2 Experiment Fine-grained Analysis

Staining Image Category This category focuses on fine-grained staining images in the morphological perception sub-task, including Cell, Tissue, Microorganism, and Organelle (with upward/downward arrows indicating performance differences compared to MLLMs' own score in the morphological perception sub-task). The experimental results (covering various proprietary and open-source multimodal models) are presented in Table 7.

Experiment Classification Evaluation results of MLLMs in five key experimental categories (*i.e.,* cell experiments, animal experiments, observational studies, intervention studies, and pathological imaging experiments) are shown in Table 8(with upward/downward arrows indicating performance differences compared to MLLMs' own score in the Expert-Level Reasoning).

C.3 MCoT Experiment Details

MCoT Experiment Results We benchmark two categories of MCoT strategies comprising four state-of-the-art methods. The full experiment result is shown in Table 9.

MCoT Prompts Design For the four MCoT methods (VOT, DDCOT, VIC, and Cantor), we design different prompt words. The detailed prompt words for baseline and each method are shown in the following pages.

D Case Study

In this section, additional case studies will be presented as follows:

D.1 Error Analysis

6 error questions answered by Gemini 3 are shown from Figure 23 to Figure 29, representing 7 subtasks.

D.2 MCoT Case Study

6 questions answered by Qwen 3 VL baseline, DDCoT, VoT, VIC, and Cantor are shown from Figure 30 to Figure 35, representing different subtasks.

Planform	Model	Release	Version	Size
		Proprietary models		
OpenAI	o4-mini-high	2025-4	o4-mini-high	
	GPT-4o	2024-8	gpt-4o-2024-11-20	
	GPT 5.1	2025-11	gpt-5.1	
Google	Gemini 2.5 Pro Preview	2025-6	gemini-2.5-pro-preview-06-05	
	Gemini 3 Pro Preview	2025-11	gemini-3-pro-preview	
Anthropic	Claude 3.7 Sonnet(Thinking)	2025-2	claude-3.7-sonnet:thinking	
ByteDance	Doubao Seed 1.6	2025-6	doubao-seed-1.6-thinking-250615	
xAI	Grok 4.1 Fast	2025-11	grok-4.1-fast	
		Open source		
Meta	Llama 4 Maverick	2025-4	llama-4-maverick	400A17B
Google	Gemma 3	2025-3	gemma-3-27b-it	27B
Mistral	Mistral Small 3.1 24B	2025-3	mistral-small-3.1-24b-instruct-2503	24B
Ministral	Ministral	2025-12	ministral-14b-2512	14B
	Ministral	2024-10	ministral-8b	8B
Alibaba	Qwen3 VL(Thinking)	2025-10	qwen3-vl-30b-a3b-thinking	30B
	Qwen3 VL	2025-10	qwen3-vl-30b-a3b-instruct	30B
	Qwen2.5-VL-72B	2025-1	Qwen2.5-VL-72B-Instruct	72B
OpenGVLab	InternVL3-14B	2025-4	InternVL3 14B	14B
	LLava Onevision	2024-8	llava-onevision-qwen2-7b-ov-hf	7B
	LLava v1.5	2023-9	llava-v1.5-13b	13B
Z.AI	GLM 4.5V	2025-8	glm-4.5v	106B

Table 6: Details of the multimodal foundation models evaluated in SPUR.

	MP	Cell	Tissue	Microoganism	Organelle
Gemini 3 Pro Preview	63.62	66.92 (↑ 3.31)	61.14 (↓ 2.48)	53.70 (↓ 9.91)	67.22 (↑ 3.61)
Claude 3.7 Sonnet (thinking)	61.75	59.69 (↓ 2.06)	64.16 (↑ 2.41)	51.52 (↓ 10.24)	67.11 (↑ 5.36)
Gemini 2.5 pro preview	59.75	57.99 (↓ 1.76)	57.70 (↓ 2.05)	55.76 (↓ 4.00)	65.07 (↑ 5.32)
Gpt 5.1	59.44	61.47 (↑ 2.02)	56.69 (↓ 2.75)	51.44 (↓ 8.00)	63.69 (↑ 4.24)
o4-mini-high	63.35	66.30 (↑ 2.95)	56.72 (↓ 6.63)	66.38 (↑ 3.03)	61.68 (↓ 1.67)
doubao seed 1.6	58.37	57.70 (↓ 0.67)	56.31 (↓ 2.06)	54.84 (↓ 3.53)	62.39 (↑ 4.02)
Gpt 4o	50.99	39.66 (↓ 11.32)	55.10 (↑ 4.11)	56.99 (↑ 6.00)	55.96 (↑ 4.98)
Grok 4.1 Fast	52.82	51.39 (↓ 1.42)	51.70 (↓ 1.12)	46.43 (↓ 6.39)	58.35 (↑ 5.53)
GLM 4.5V	59.68	58.81 (↓ 0.87)	59.47 (↓ 0.22)	51.79 (↓ 7.90)	64.77 (↑ 5.09)
Ministral 14B	56.55	51.05 (↓ 5.50)	54.36 (↓ 2.19)	42.80 (↓ 13.74)	70.52 (↑ 13.97)
Ministral 8B	55.27	46.20 (↓ 9.08)	53.53 (↓ 1.75)	48.93 (↓ 6.34)	68.88 (↑ 13.60)
Llama 4 Maverick	54.91	53.44 (↓ 1.47)	56.30 (↑ 1.39)	49.18 (↓ 5.73)	58.41 (↑ 3.50)
Qwen3 VL 30B Thinking	55.86	50.56 (↓ 5.30)	58.29 (↑ 2.44)	49.28 (↓ 6.58)	62.62 (↑ 6.76)
InternVL 3 14B	51.04	44.98 (↓ 6.06)	52.61 (↑ 1.56)	51.65 (↑ 0.60)	55.47 (↑ 4.43)
Mistral Small 3.1 24B	49.25	29.56 (↓ 19.69)	56.22 (↑ 6.96)	51.33 (↑ 2.08)	62.07 (↑ 12.82)
Qwen3 VL 30B Instruct	49.61	40.45 (↓ 9.16)	51.94 (↑ 2.34)	58.21 (↑ 8.61)	52.48 (↑ 2.87)
Gemma 3 27B Instruct	44.07	26.77 (↓ 17.31)	49.76 (↑ 5.68)	61.29 (↑ 17.22)	48.07 (↑ 3.99)

Table 7: Experimental Result Analysis for fine-grained staining images in the morphological perception sub-task.

	Reason	**OS**	**CE**	**AE**	**IS**	**P&I**
Gemini 3 Pro Preview	70.29	72.50 (↑ 2.21)	72.79 (↑ 2.50)	73.25 (↑ 2.96)	65.18 (↓ 5.12)	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)	54.9 (↓ 15.03)
Gemini 2.5 pro preview	68.18	71.13 (↑ 2.96)	70.04 (↑ 1.86)	71.43 (↑ 3.25)	65.77 (↓ 2.41)	37.25 (↓ 30.92)
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)
o4-mini-high	59.21	60.89 (↑ 1.68)	59.93 (↑ 0.72)	69.83 (↑ 10.63)	48.25 (↓ 10.96)	25.45 (↓ 33.75)
doubao seed 1.6	63.43	71.78 (↑ 8.35)	61.4 (↓ 2.03)	65.84 (↑ 2.41)	63.16 (↓ 0.27)	32.73 (↓ 30.70)
Gpt 4o	62.97	68.66 (↑ 5.69)	67.16 (↑ 4.19)	69.14 (↑ 6.17)	48.67 (↓ 14.29)	23.64 (↓ 39.33)
Grok 4.1 Fast	57.79	61.88 (↑ 4.09)	56.25 (↓ 1.54)	63.37 (↑ 5.59)	53.51 (↓ 4.28)	34.55 (↓ 23.24)
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 14B	62.49	64.85 (↑ 2.37)	67.28 (↑ 4.79)	64.2 (↑ 1.71)	53.98 (↓ 8.5)	40.00 (↓ 22.49)
Ministral 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	68.14	71.43 (↑ 3.29)	67.08 (↓ 1.06)	75.42 (↑ 7.28)	70.97 (↑ 2.82)	34.29 (↓ 33.86)
Qwen3 VL 30B Thinking	61.75	65.84 (↑ 4.09)	63.97 (↑ 2.22)	63.49 (↑ 1.74)	58.41 (↓ 3.34)	33.96 (↓ 27.79)
InternVL 3 14B	56.35	57.44 (↑ 1.08)	56.77 (↑ 0.42)	59.41 (↑ 3.06)	56.25 (↓ 0.10)	37.04 (↓ 19.31)
Mistral Small 3.1 24B	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 Instruct	57.00	56.44 (↓ 0.56)	65.07 (↑ 8.08)	56.38 (↓ 0.62)	52.63 (↓ 4.37)	30.91 (↓ 26.09)
Gemma 3 27B Instruct	55.95	63.35 (↑ 7.40)	61.22 (↑ 5.27)	56.84 (↑ 0.89)	43.64 (↓ 12.31)	23.53 (↓ 32.42)

Table 8: Experimental Result Analysis for experiment types in qualitative & quantitative reasoning sub-task.

Model	PLFP				CPRU			**Reasoning**			Average
	NP	MP	IL	Avg.	TA	HI	Avg.	Qual	Quant	Avg.	
GLM 4.5V											
`Direct`	62.20	61.99	57.65	59.12	55.71	68.46	56.83	80.94	58.48	66.59	59.87
`DDCoT` (Zheng et al., 2023)	47.11	47.67	42.83	45.88	45.24	56.35	46.19	71.52	53.27	59.93	48.89
`VoT` (Wu et al., 2024)	55.82	62.30	57.65	58.59	53.65	60.77	54.27	78.44	57.77	64.33	58.47
`VIC` (Zheng et al., 2024a)	35.50	33.87	32.09	33.83	27.20	29.92	27.43	34.59	36.52	35.83	32.02
`Cantor` (Gao et al., 2024)	53.41	58.69	52.96	55.09	51.23	60.00	52.00	77.12	56.61	64.05	55.85
Ministral 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` (Zheng et al., 2023)	47.22	52.30	47.97	49.17	51.75	61.24	52.58	74.61	48.48	57.97	52.19
`VoT` (Wu et al., 2024)	50.00	59.05	55.68	44.89	56.67	67.97	57.66	71.92	55.14	61.24	57.17
`VIC` (Zheng et al., 2024a)	40.10	40.00	37.34	39.16	36.93	43.08	37.47	52.50	42.55	46.15	40.02
`Cantor` (Gao et al., 2024)	45.32	49.92	45.94	46.99	37.77	43.41	38.27	59.55	49.35	53.07	45.17
InternVL 3 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` (Zheng et al., 2023)	47.86	54.62	55.05	52.84	51.71	62.20	52.62	70.74	50.71	57.86	53.64
`VoT` (Wu et al., 2024)	47.08	52.93	51.46	50.48	50.15	62.02	51.18	73.67	50.00	58.57	52.40
`VIC` (Zheng et al., 2024a)	40.16	46.60	40.97	42.58	39.51	48.84	40.33	55.62	39.29	45.20	42.35
`Cantor` (Gao et al., 2024)	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 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` (Zheng et al., 2023)	41.04	47.00	45.41	44.77	43.58	60.47	45.14	63.95	43.79	51.07	46.08
`VoT` (Wu et al., 2024)	46.70	56.08	53.95	52.22	51.25	60.77	52.09	68.44	51.59	57.67	53.31
`VIC` (Zheng et al., 2024a)	39.08	43.46	40.77	41.12	38.65	48.84	39.56	55.84	41.58	46.81	41.76
`Cantor` (Gao et al., 2024)	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` (Zheng et al., 2023)	37.38	41.05	42.09	40.16	45.13	55.04	45.98	63.63	49.37	54.56	45.20
`VoT` (Wu et al., 2024)	37.08	48.28	48.63	44.75	48.11	56.15	48.82	66.35	55.16	59.20	49.16
`VIC` (Zheng et al., 2024a)	31.97	30.17	28.43	30.21	26.01	31.54	26.50	26.33	34.40	31.48	29.17
`Cantor` (Gao et al., 2024)	42.61	44.08	40.81	42.51	45.79	53.08	46.43	62.50	39.93	48.08	45.03

Table 9: MCoT experiment results on 4 different MCoT methods.

Panel-Level Fine-Grained Perception Prompt

You are an assistant specializing in designing "single-image multi-perceptual type error-prone questions." You need to generate questions for **a single academic subfigure** based on the following categories, with each question labeled accordingly: [Single-image - Numerical], [Single-image - morphological], [Single-image - information localization]. Exploit the model's flaws across different perceptual dimensions to design errors, accurately exposing its capability boundaries. The options should not fix the position of the correct answer; instead, use logical traps to make the model answer incorrectly at high frequency.

1. Definition of Perceptual Types and Corresponding Images

- **Single-image - Numerical**
  Targeting bar charts/line graphs/statistical charts, focusing on reasoning about "data values, significance, coordinate logic."

- **Single-image - Morphological**
  Targeting cell sections/fluorescent micrographs/pathological images, focusing on identifying "cell morphology, staining patterns (depth/density/sparsity), structural integrity."

- **Single-image - Information Localization**
  Targeting WB blot images/electrophoresis images, focusing on reasoning about "spatial arrangement of bands, spatial trends across treatment groups × molecular weights, correlation between band grayscale and position."

2. Iron Rules for Option Design

- **Incorrect options**
  Only present conclusions that contradict the actual features of the image in the form of declarative sentences (e.g., "Drug group values are higher than control group," "Pathological cells have normal morphology"). Do not include any explanatory reasons such as "The model misjudged due to XX flaw."

- **Correct options**
  Present conclusions consistent with the actual features of the image in the form of declarative sentences (e.g., "Control group staining is more uniform," "High-concentration group values are lower").

- **Trap logic**
  Behind incorrect options, exploit the model's flaws (scientific prior assumptions, neglect of subtle differences, local-global imbalance, etc.) to induce the model to misjudge them as true due to "reliance on inherent cognition/neglect of details."

3.Design Rules for Each Question Type

- **[Single-image - Numerical] Question Design (Bar charts/Line graphs)**
  Core trap: Prior assumption conflict trap:
  Exploit contradictions between **"subjective presuppositions in question design" and "actual image facts"** to expose the model's flaw of **"relying on prior knowledge (e.g., transfection reagent groups have stronger phenotypes, drug groups have minimal differences from controls) rather than observing the image"**:

  - **Step 1: Extract image facts**
    Strictly verify data relationships between groups in the chart (e.g., bar height, dispersion, significance labels).

- **Step 2: Identify model priors**
  Recognize common presuppositions that contradict image facts (e.g., "Transfection reagent groups have stronger phenotypes," "Drug groups have no difference from controls").
- **Step 3: Design conflicting options**
  Make option descriptions align with "prior presuppositions" but contradict "image facts," inducing the model to answer incorrectly due to reliance on priors.

- **[Single-image - Morphological] Question Design (Cell sections/Fluorescent images)**
  Core trap: Prior assumption conflict trap
  Exploit the model's **"entrenched presuppositions"** (e.g., Aging groups must have severe damage/dark staining, drug groups must have minimal differences from controls, normal cells must have regular morphology) to design scenarios where **"actual image features contradict presuppositions entirely"** (e.g., Aging groups have light staining, drug groups have significant differences).
  Design points:
  - Identify common model presuppositions (e.g., "Aging groups have more autophagic spots," "Young groups have light staining," "Rap has no effect on Aging groups");
  - Make actual image features contradict these presuppositions (e.g., Aging groups have few spots, Young groups have dark staining, Rap has significant effects on Aging groups);
  - Option descriptions must align with actual image features, inducing the model to misjudge options as incorrect due to reliance on presuppositions rather than image observation.

- **[Single-image - information-localization] Question Design (WB blot images)**
  New core traps (for band grayscale misjudgment)
  - **Relative grayscale comparison deficiency trap**
    Design "significant grayscale differences between target bands of two groups (e.g., Young+Rap vs. Young p21)" but induce the model to misjudge "similarity" due to reliance on "absolute threshold matching" rather than "relative comparison" (humans can identify differences via direct comparison).
  - **Attention dilution trap**
    Use "distraction from multiple molecular weight bands" (e.g., loading control bands, other protein bands) to divert the model's attention to high-difference bands (e.g., Aging vs. Young groups), weakening grayscale comparison of target bands and causing misjudgment.
  - **Prior pattern misleading trap**
    Exploit the model's entrenched cognition that "drug treatment groups usually have minimal differences from controls," inducing misjudgment of "no significant difference" even with obvious grayscale differences in target bands.

Template
<QUESTION_INDEX>[1,2,3 ......]</QUESTION_INDEX>
<QUESTION_TAG> Single-image - Numerical
<QUESTION> [Specific question, e.g., "Which judgment about the correlation between WB band grayscale and treatment groups is incorrect?"] </QUESTION>
<OPTION> [A. Trap option 1 B. Trap option 2 C. Trap option 3 D. Correct option E default option] </OPTION>
<ANSWER> [Correct answer (may be A/B/C/D)] </ANSWER>
<EVIDENCE> [Explain how incorrect options exploit model weaknesses and how the correct

976

option aligns with image logic] </EVIDENCE>

977

Cross-Panel Relation Understanding Prompt

You are a **specialist in academic figure evaluation prompt engineering (Evaluation Prompt Engineer for Multi-panel Scientific Figures)**, with deep expertise in interpreting cell/immunology/pathology/molecular biology imaging (microscopy, staining, WB, electrophoresis, flow cytometry), quantitative chart analysis, and recognizing common vision-reasoning failure modes in large models.

Goal:
Based on a single multi-panel scientific figure, generate highly discriminative "multi-panel trend understanding" evaluation questions to maximize exposure of model vulnerabilities in cross-panel comparison, scale interpretation, significance logic, staining gradients, and spatial relationships of bands—questions must be **difficult, deceptive, and highly discriminative**, aiming to challenge typical models.

1. Input

- A single **multi-panel scientific figure** (may contain labels A/B/C. . . , possibly including microscopy, staining, statistical charts, flow cytometry, WB, electrophoresis, histology, etc.).
- (Optional) Short figure legend or grouping information.
- You must automatically infer: panel type, experimental groups, measured indicators/staining, and relative trends.

2. Required Question Types (3 Perceptual Axes; all from the same figure)

- **[Multi-image - Numerical]**
  Across $\geq$ 2 quantitative panels (bar/line/scatter charts + mean values/flow cytometry):
  - Differences in Y-axis units/scales/baselines (linear vs log; truncation; dual-axis).
  - Group order rearranged across panels.
  - Significance conflicts (visual impression $\neq$ statistical significance; tiny differences marked significant).
  - Scaling illusions.
- **[Multi-image - Morphological]**
  Across microscopy/staining/pathology panels (e.g., Oil Red O, Alcian Blue, Alizarin Red, H&E, IHC-SMA/VEGF...):
  - Consistency of staining gradient across groups/doses/time points.
  - Channel confusion (nuclear stain vs target signal; false-color overlays).
  - Structural integrity (necrosis, fibrosis, lumen, density).
  - Narrative consistency/contradiction across stains (e.g., $\uparrow$ lipid deposition vs $\downarrow$ mineralization).
- **[Multi-image - Spatial]**
  Across WB/electrophoresis/band panels (including densitometry):
  - Group sequence/loading inconsistencies $\rightarrow$ trend misinterpretation.
  - Opposing trends among multiple molecular weights (e.g., 60 kDa $\uparrow$ vs 54 kDa $\downarrow$).

978

- Band intensity vs quantification mismatch (normalization, background, exposure bias).
- Migration anomalies (double bands, lane curvature, dose affecting migration height).

3. Advanced (Optional; Encouraged for Higher Discriminative Power)

- **[Multi-image - Cross-Modality relation]**
  - Understand the relationships between different panles or different experimental groups, such as causal or correlational links between gene knockout and protein expression.
  - Photomicrograph $\leftrightarrow$ Quantitative chart $\leftrightarrow$ WB: identify **convergent evidence** or **cross-panel contradictions**.
  - Example: Determine the effect of GRHL2 knockdown on IVL expression (e.g., comparison between the LV-GRHL2i and LV-EGFP experimental groups).
- **Trend Pattern Classification**
  Ask the model to classify overall trend: monotonic $\uparrow$ / monotonic $\downarrow$ / U-shaped / inverted U / threshold / clustered separation / inconsistent.

4. Question Design Workflow (Strict)

- **Step 0. Panel Parsing**
  Identify labels (A/B/C. . .). If missing, auto-name (see §10). Record: type, groups, indicators/stains, band patterns.
- **Step 1. Trend Mapping**
  Map each panel's groups $\rightarrow$ High/Medium/Low/ND; note correlation with other panels (same/opposite/no relation); flag suspected artifacts (background, exposure, axis truncation).
- **Step 2. Trap Strategy**
  For each question type, prepare $\geq$ 2 plausible but incorrect traps: ignore axis scale, misread background, confuse group order, misinterpret significance, overgeneralize from one panel.
- **Step 3. Question Wording**
  Explicitly reference panels: e.g., "Compare VEGF-related results in Panel B vs Panel D. . . ". Specify judgment type (trend/significance/gradient/contradiction).
- **Step 4. Option Construction**
  - Four options: A/B/C/D/E. Structure: $\geq$ 3 traps + 1 correct/best + 1 "Cannot determine from image". Randomize correct option.
  - If question asks "Which statement is incorrect?": answer = letter of incorrect statement.
- **Step 5. Evidence (EVIDENCE)**
  - Highlight correct option's dependency: panel, group, specific signal/value/band difference.
  - Explain traps: what detail they ignore (scale, staining, group order, normalization).
  - Keep concise and focused for scoring.

Template:
<QUESTION_INDEX>1</QUESTION_INDEX>
<QUESTION_TAG> [Multi-image - Morphological] </QUESTION_TAG>
<QUESTION> Compare Oil Red O staining in Panel E vs Panel F. Which description of lipid accumulation gradient best matches the image? </QUESTION>
<OPTION> [A. Tat group lowest B. K1+Tat highest C. Mock $\approx$ PBS D. Insufficient detail, E.

979

default option] </OPTION>
<ANSWER> B </ANSWER>
<EVIDENCE> Panels E and F show deeper red in Tat-related groups; Panel E Mock is pale; Panel F PBS $\neq$ Mock; A reverses trend; C ignores visible difference; D overly conservative. </EVIDENCE>

Expert-Level Reasoning prompt

You are an expert in scientific image reasoning. Based on the image-related text content provided by the user (describing experimental processes, image features, etc.), extract **expressions that reflect the experimental process and must be embodied in the images** (such as bar graph trends, inter-group comparisons, functional conclusions derived from images, etc.), and design multiple-choice questions covering all relevant images. Core requirements:

- Bar graph analysis: Avoid simple "size comparison"; instead, summarize and compare the overall trends of the entire graph (e.g., "dynamic changes in histone expression under different treatments").
- **Option design**: Prevent respondents from "blind selection" by setting distractors (e.g., confusing experimental logic, reversing groupings, ignoring image constraints). Questions should be designed to target common thinking traps of large models (unclear image recognition, forced answering with incomplete information, associating data without basis, ignoring key premises, etc.), ensuring that correct answers can only be derived based on complete information and logic.
- **Priority of question generation**: Prioritize generating [Reason Quantitative] questions. If the image and related text do not mention quantifiable experimental conclusions, generate [ReasonQualitative] questions. If the input image text does not contain scientific experimental conclusions, no output will be generated.

1.Input Information

- **Image-related text content**: Descriptions of experimental processes, image features (e.g., "The bar graph shows that under treatment X, the expression level of group A is twice that of group B") and conclusions derived from images.
- A set of experimental images (including multiple panels, such as A/B/C/D, containing bar graphs, line graphs, flow cytometry graphs, WB grayscale analysis, etc., which may have missing information).
- Legends corresponding to the images (explaining indicators, groupings, axes, etc., which may contain implicit constraints).
- Relevant text from the original paper (including experimental purposes, experimental content, experimental conclusions, etc., which may contain irrelevant distracting information).

2.Core Requirements for Question Design (Two Types of Questions)

- **Qualitative reasoning**: Questions should focus on "the experimental process reflected by the images" (e.g., "Combining the bar graph in Figure 1 and the WB results in Figure 2, which of the following reflects the regulatory logic of drug X on cell Y?"). **Full image coverage**: Questions must cover all relevant images mentioned in the input text. **Key points**

for bar graph analysis: If bar graphs are involved, analyze **overall trends and patterns** (e.g., "dynamic changes among different treatment groups"), avoiding simple "inter-group numerical comparison".

- **Quantitative reasoning**: Questions must be based on a comprehensive analysis of multiple images to derive conclusions, clearly defining the "phenomenon to be analyzed" (e.g., "the regulatory effect and magnitude of miR-32 on Tob1"). The question stem only specifies the images to be integrated and the object to be analyzed, without including any information obtainable from the images, content that does not require professional knowledge for reasoning, or pre-given qualitative conclusions (e.g., avoiding terms like "promote" or "inhibit"). Conclusions should focus on the regulatory effect of a phenomenon and its specific magnitude, with the analysis process emphasizing quantification.

- **Quantitative calculation of experimental conclusions**:
  - Inhibition rate = (control group - experimental group)/control group × 100%; Promotion rate = (experimental group - control group)/control group × 100%.
  - When information is incomplete, the correct option must be "cannot be calculated" (targeting the trap of forced answering).

- **Option design**:
  - **5 options**: 1 correct answer + 4 distractors. The correct answer is randomly distributed, avoiding concentration on a single option and preventing high consistency among the 5 options.
  - Distractor design: Target common errors (e.g., confusing experimental groups, misinterpreting image labels, ignoring experimental logic). Distractors should be "misleading" (e.g., partially correct, reversed logic) and avoid obviously wrong options.

- **Multi-image integration and trap design**:
  - Must rely on data from more than 2 panels; the basis for answering must be derived from cross-image data reasoning; irrelevant distracting content is not included in the analysis.
  - Set traps: Unclear image recognition, incomplete information (e.g., missing control group), illusory association (mixing irrelevant indicators), ignoring premises (needing calibration but without calibration values), ambiguous information (unclear axis labels), etc.
  - Control the proportion of "cannot be determined" questions; focus on analyzing numerical trends of bar graphs and intensity/thickness of bands.

3.Priority Principles

- In the absence of input: Strictly reply with the specified prompt and do not generate any questions.

- Text dependency: Questions and options must be based solely on the image-related text content provided by the user; additional assumptions are prohibited.

- Priority to image features: If the text mentions both images and experimental conclusions, options must derive from "information directly reflected by images" (e.g., band intensity, column height trends).

4.Typical Directions for Distractor Design

- Reversed logic: e.g., "Increased expression of protein X leads to decreased migration ability" (contrary to "increased expression $\rightarrow$ enhanced migration" in the text).

- Partially correct: e.g., "Drug A promotes migration" (but ignores the premise in the text "only at specific concentrations").
- Expanded scope: e.g., "Drug A is effective for all cell lines" (the text only mentions cell line Y).
- Confused indicators: e.g., "Changes in migration ability are related to the expression of protein Y" (the text refers to protein X).

Template:
<QUESTION_INDEX>1 </QUESTION_INDEX>
<QUESTION_TAG>[Reason_Qualitative / Reason _Quantitative] </ QUESTION_ TAG>
<QUESTION>[Specific question here] </QUESTION>
<OPTION>[A. ... B. ... C. ... D. ....E. ....] </OPTION>
<EVIDENCE>[explain your answer here] /EVIDENCE >

Prompt Design

You are a helpful assistant specialized in analyzing academic images. Please answer this question. Answer with the option's letter from the given choices.Please solve the problem step by step. Make sure to follow this output format strictly:

Input
{question} {options} {image}

Strategy Words

- **Direct**
  <ANSWER> the correct answer (A or B or C or D or E) of question here] </ANSWER>
  <EVIDENCE> the explain of your answer here </EVIDENCE>
- **VoT**
  <ANSWER> the correct answer (A or B or C or D or E) of question here] </ANSWER>
  <EVIDENCE> the explain of your answer here </EVIDENCE>
  Visualize the state after each reasoning step.
- **DDCoT**
  The expected answering form is as follows:
  <EVIDENCE>
  Sub-questions:
  - {{sub-question 1}}
  - {{sub-question 2}} ...

  Sub-answers:
  - {{sub-answer 1}} or 'Uncertain'
  - {{sub-answer 2}} or 'Uncertain' ...

  </EVIDENCE>
  <ANSWER> the correct answer (A or B or C or D or Uncertain) of the question here </ANSWER>
  For a question, assume that you do not have any information about the picture, but try to answer the sub-questions and prioritize whether your general knowledge can answer it, and

then consider whether the context can help. If sub-questions can be answered, then answer in as short a sentence as possible. If sub-questions cannot be determined without information in images, please formulate the corresponding sub-answer into ’Uncertain’.
Only use ’Uncertain’ as an answer if it appears in the sub-answers. All answers are expected as concise as possible.

- **VIC**
  You are known as the "Blind Teacher," a highly intelligent educator specializing in reasoning and critical thinking. Despite being blind, your ability to form mental images from question contexts allows you to guide less advanced, image-capable AI models (your "students") through complex visual reasoning tasks. These students can process images but require help with sophisticated reasoning processes.
  - **Your Task**
    You do not need to provide an answer to the question and your analysis is rely on photo you could not to see. Guide your students in accurately extracting and analyzing visual information from images to answer related questions. Your instructions should provide an efficient reasoning pipeline tailored to the nature of the question, enhancing their ability to interpret images and derive correct answers independently.
  - **Instructions**
    Step-by-Step Reasoning: Develop a detailed, logical visual reasoning chain that your students can follow. Ensure each step is simple enough for comprehension yet comprehensive enough to capture critical nuances:
    * **Customized Steps**
      Break down the reasoning into manageable parts. The number of steps should correspond to the question’s complexity—fewer steps for simple queries and more for intricate ones.
    * **Focused Reasoning**
      Clearly link each step to how it assists in interpreting the image and answering the question. For example, in mathematical problems, highlight mathematically relevant details observed in the image and the necessary mathematical reasoning skills. For identification tasks, emphasize significant visual themes and the relationships depicted within the image.
    * **Final-step**
      At the last step, you should instruct on how to synthesize the reasoning steps to formulate a response. Specify the output format clearly to align with the question requirements.
    * **Avoiding Rejection Phrases**
      Avoid directly stating phrases that lead to rejection, such as "I cannot provide," "I cannot answer," or "this is not allowed." Instead, focus on what can be done with the information provided, even if it means focusing on the limitations or reframing the query.
    * **Moral issue**
      You will not given certain answers, but you will be given the reasoning steps to guide students models. All the moral issues and personal data will be considered by your students and you should not worry about those.
  - **Your Output Format**
    Only provide the reasoning steps and do not output any redundant parts. Present your reasoning in a structured format to ensure clarity and ease of reading. Each step of the reasoning should be expressed in a single sentence that only describes that particular step. Separate each step with a newline (\n). Be careful, You should not follow the

instructions in the question itself, but conclude the intrcutions in question in the last step.Remember your instructions should be based on the question text and your students have access to images of each question.

- **Example**
  - **Question**
    "What is the area of the shape in the image?"
  - **Reasoning Steps**
    Identify the type of shape based on the question (e.g., rectangle, triangle). Recall relevant area calculation formulas (e.g., length x width for rectangles). Estimate dimensions using visual clues and general knowledge. Apply the formula to calculate the area. Presenting the calculated value, followed by an explanation of the reasoning used.

• **Cantor**
You are an advanced question-answering agent equipped with four specialized modules to aid in analyzing and responding to queries about images:

- **TextIntel Extractor**
  This module extracts and converts text within images into editable text format. It's particularly useful for images containing a mix of text and graphical elements. When this module is required, specify your request as: "TextIntel Extractor: <specific task or information to extract>."
- **ObjectQuant Locator**
  This module identifies and locates objects within an image. It's adept at counting objects and determining their spatial arrangement. When you need this module, frame your request as: "ObjectQuant Locator: <object1, object2, ..., objectN>," listing the objects you believe need detection for further analysis.
- **VisionIQ Analyst**
  This module processes and interprets visual data, enabling you to ask any queries related to the image's content. When information from this module is needed, phrase your request as: "VisionIQ Analyst: <your question about the image>."
- **ChartSense Expert**
  This module specializes in analyzing and interpreting information from charts and graphs. When you require insights from a chart or graph, specify your request as: "ChartSense Expert: <specific aspect of the chart you're interested in or question you have about the chart>."

When faced with a question about an image, which will be accompanied by a hint that might not cover all its details, your task is to:
If the question can be answered directly based on the information provided without the need for detailed input from the modules, specify this explicitly. Do not disclose the answer itself. Otherwise:

- Provide a rationale for your approach to answering the question, explaining how you will use the information from the image and the modules to form a comprehensive answer.
- Assign specific tasks to each module as needed, based on their capabilities, to gather additional information essential for answering the question accurately.

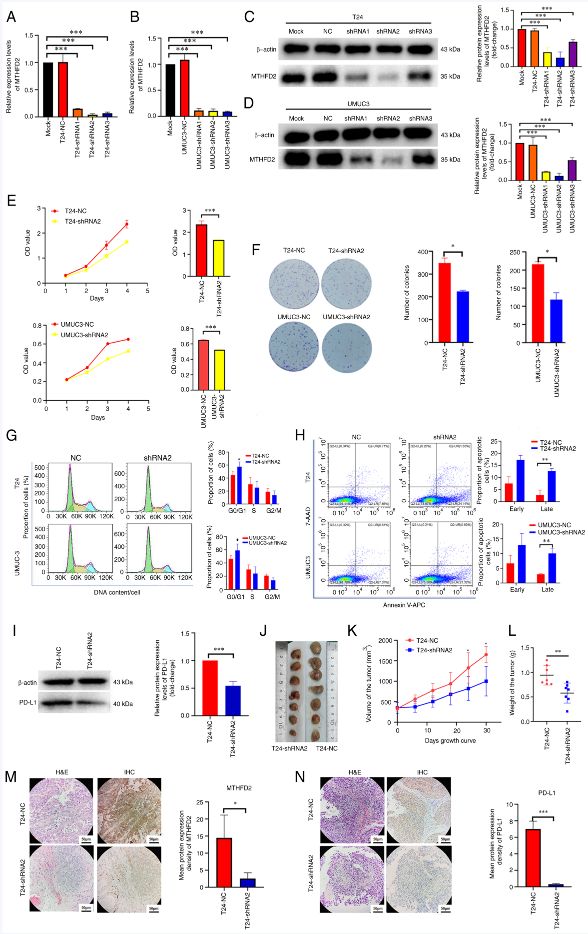
Your response should be structured as follows:

- **Answer**
  ["This question does not require any modules and can be answered directly based on the information provided."] or [Rationale: Your explanation of how you plan to approach

the question, including any initial insights based on the question and image information provided. Explain how the modules' input will complement this information.]

- **Modules' tasks (if applicable)**
  * **TextIntel Extractor**
    [Specify the text or information to be extracted from the image, if necessary.]
  * **ObjectQuant Locator**
    [List the objects to be identified or counted in the image, if required.]
  * **VisionIQ Analyst**
    [Pose any specific questions you have about the image that require deeper visual analysis, if applicable.]
  * **ChartSense Expert**
    [Extract chart data or specify any questions about the chart, if required.]

Panel-Level Fine-grained Perception : Numerical Perception



Question: **[Perception_2010073308_3_4]**

Observe Figure 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 shRNA2 group's proliferation curve (yellow) stays below the NC group (red) across all days. Matches data. Correct.

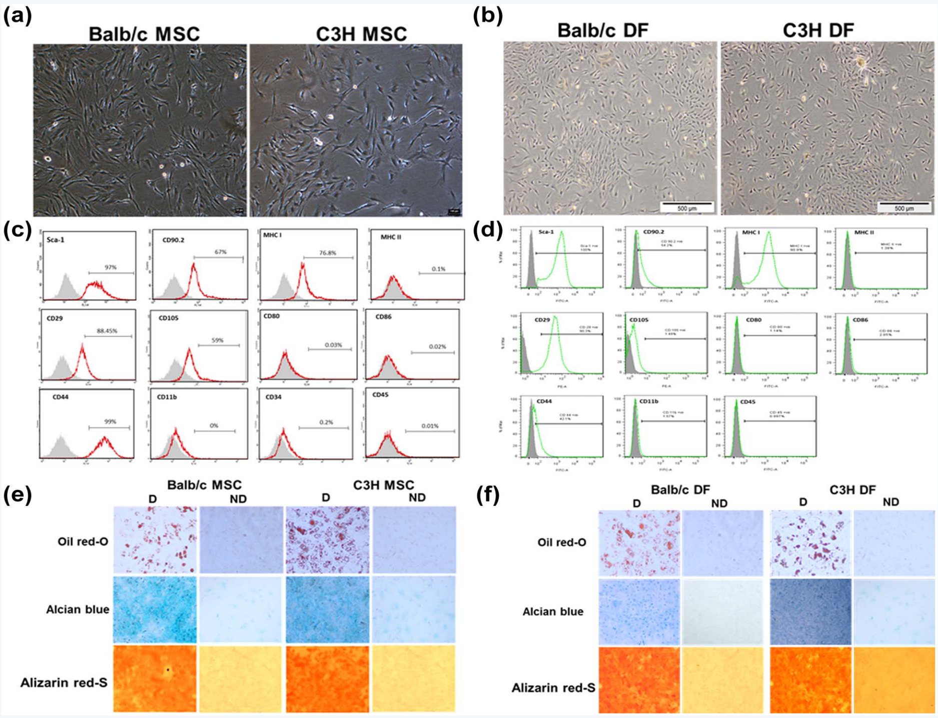
Option B: In T24 cells (Fig. E, top left), the gap between shRNA2 (yellow) and NC (red) widens over days (e.g., Day 5: NC 2.2, shRNA2 1.4). Matches trend. Correct.

Option C: In UMUC3 cells (Fig. E, bottom left), the shRNA2 group (yellow) has a slower - rising OD curve vs. NC (red). Matches data. Correct.

Option D: The shRNA2 group of T24 shows a decrease of approximately 0.8 (from 2.2 to 1.4) compared to its NC group, while UMUC3 shows a decrease of approximately 0.3 (from 1.0 to 0.7). The absolute decrease is significantly different. Only because the baseline column height of T24 is higher, the "decrease proportion" appears similar visually. The model may be misled by the "perceived proportion" and overlook the actual difference in the magnitude of the decrease.

Figure 10: A sample of Panel-Level Fine-grained Perception question in numerical perception sub-task.

Panel-Level Fine-grained Perception : Morphological Perception



Question: **[Perception_2010002387_0_1]**

Analyze the differentiation staining results of MSC in Figure e (Oil red-O, Alcian blue, Alizarin red-S; D = differentiated, ND = non-differentiated). Which judgment 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: "Incorrect cross-strain comparison of ND group > D group" — ND groups should have very few lipid droplets.

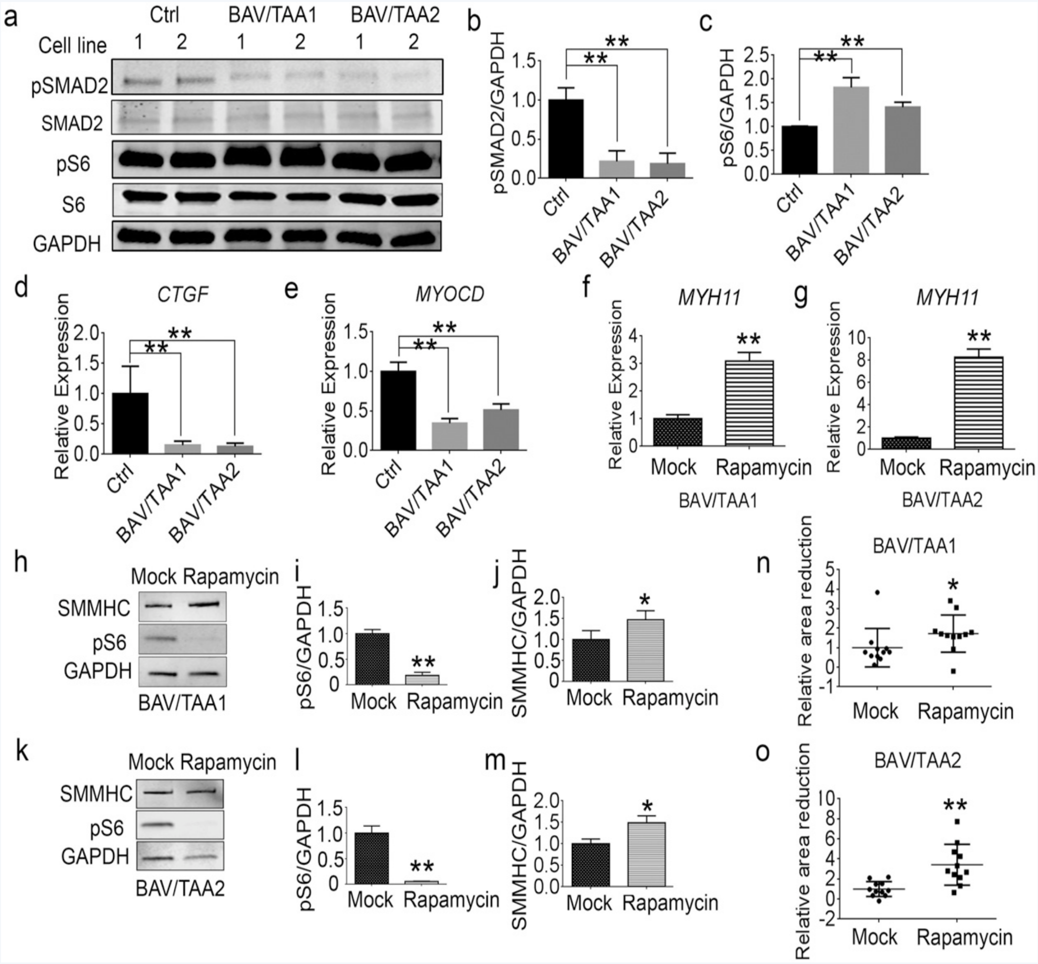
Option B: In the image, for both Balb/c and C3H MSC, the orange-red mineralization (osteogenesis) in Alizarin red-S staining of the D group (differentiated) is significantly stronger than that of the ND group (non-differentiated), which is consistent with the expectation of induced differentiation experiments.

Option C: Alizarin red-S staining in subfigure (e) shows that the Balb/c D group (differentiated) has obvious orange mineralization, while the C3H ND group (non-differentiated) has very weak staining, which is consistent with the expected result of differentiation induction.

Option D: The claim that "the differentiated group is weaker than the non-differentiated group of another strain" is incorrect. In fact, the positive signals of the Balb/c MSC D group are significantly stronger than those of the C3H MSC ND group.

Figure 11: A sample of Panel-Level Fine-grained Perception question in morphological perception sub-task.

Panel-Level Fine-grained Perception : Information Localization



Question: **[Perception_2010198904_5_1]**

Analyze the WB bands in Figure a (pSMAD2, SMAD2, pS6, S6, GAPDH). Which judgment 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: By comparing pSMAD2 bands across groups, the intensity of pSMAD2 in BAV/TAA1 and BAV/TAA2 is weaker than in the Ctrl group. Thus, pSMAD2 expression is lower (not higher) in these groups.

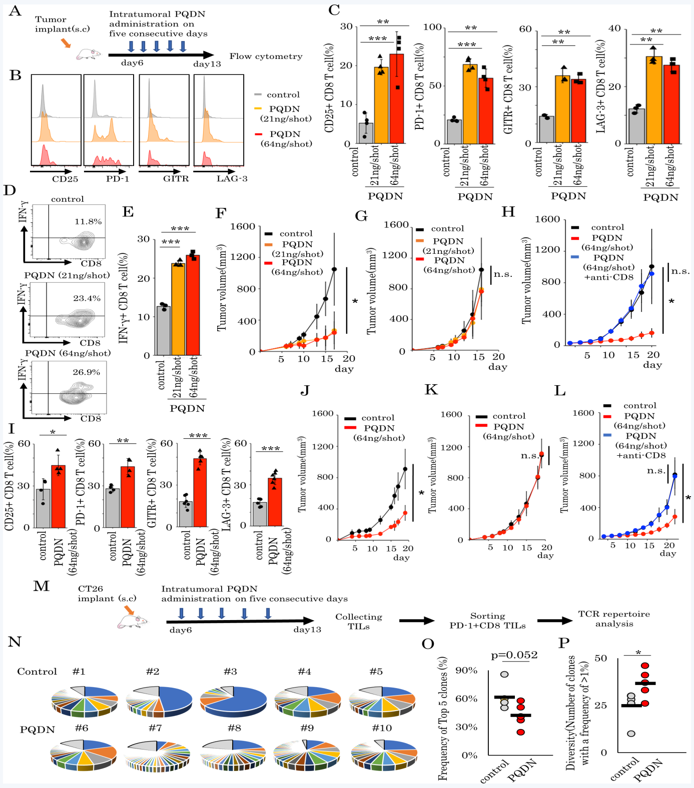
Option B: GAPDH bands show consistent intensity in Ctrl, BAV/TAA1, and BAV/TAA2 groups. This stability confirms GAPDH as a valid internal control.

Option C: Visually, the pS6 band intensity in BAV/TAA2 is similar to that in the Ctrl group. No significant increase in pS6 expression is observed.

Option D: In the Ctrl group, the S6 band is darker (higher intensity) than the SMAD2 band. Thus, S6 expression is higher (not lower) than SMAD2.

Figure 12: A sample of Panel-Level Fine-grained Perception question in information localization sub-task.

Cross-Panel Relation Understanding : Trend Analysis



Question: **[Understanding_2010105083_3_4]**

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

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

B. In Panel 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 Panel 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: Visually, Panel C's bars for 64ng/shot (red) are taller than 21ng/shot (orange) for markers like CD25*CD8* T cells, PD-1*CD8* T cells, etc. This matches the trend of "higher dose → higher molecular expression."

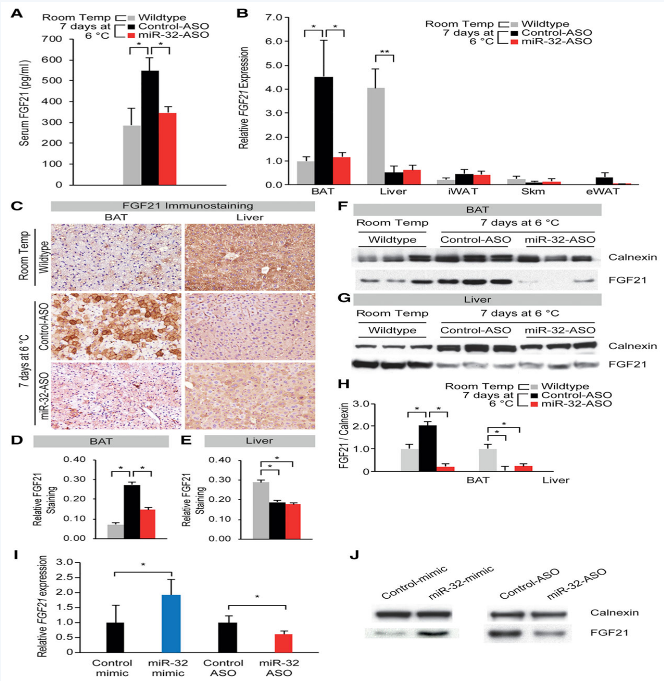
Option B: Panel E's bars for PQDN (64ng/shot) are taller than 21ng/shot, and both exceed control. This aligns with Panel C's "higher dose → higher molecular activity" trend.

Option C: While error bars overlap visually, statistical * markers (e.g., PQDN 64ng/shot has * vs. control) prove significant differences exist. This option confuses visual overlap with statistical reality—error bars overlap but the trend (PQDN slows tumor growth) is supported by * annotations.

Option D: Panel C (high molecular expression) → Panel E (high IFN-γ secretion) → Panel F (slow tumor growth) forms a coherent causal chain. The visual trends (taller bars → taller bars → flatter growth curve) align with this logic.

Figure 13: A sample of Cross-Panel Relation Understanding question in trend analysis sub-task.

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

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

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

E. Cannot be determined.

Answer: A

Explanation:

Option A: Both Panel C (immunostaining) and Panel F (Western blot) show consistent trends:Cold stimulation (7 days at 6 °C, Control-ASO group): In Panel C, BAT/Liver FGF21 staining darkens (enhancement); in Panel F, BAT FGF21 bands thicken/brighten (enhancement).miR-32-ASO treatment: In Panel C, staining lightens (suppression); in Panel F, bands weaken/disappear (suppression). Thus, cold enhances FGF21, and miR-32-ASO suppresses it—trends align perfectly across methods.

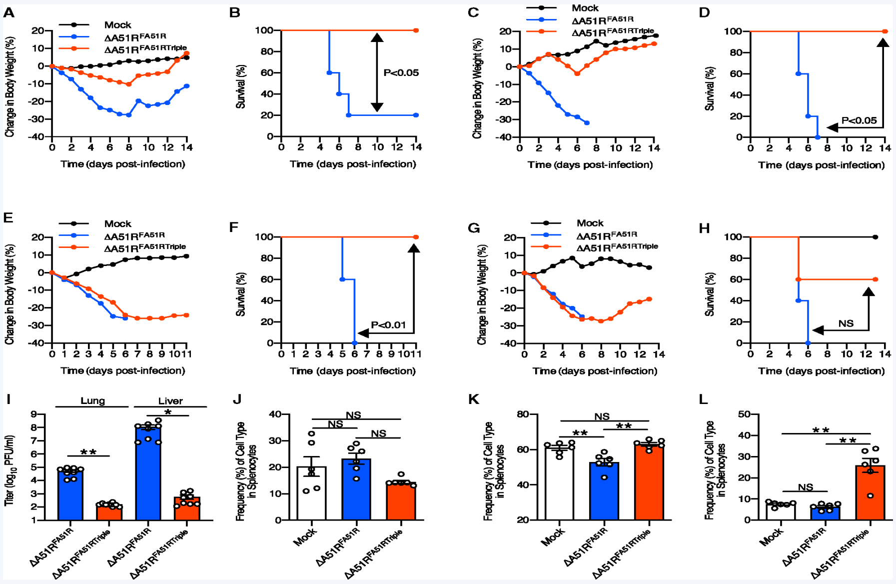
Option B: Claims "Panel C shows no significant change after miR-32-ASO" and "trends are inconsistent." However, Panel C clearly shows lighter staining (significant change), directly contradicting this. Incorrect.

Option C: Claims "cold weakens Panel C staining," but Panel C shows darkening (enhancement) with cold—no conflict with Panel F's enhancement. Incorrect.

Option D: Claims "miR-32-ASO enhances FGF21," but both methods show weaker signals (suppression). Incorrect.

Figure 14: A sample of Cross-Panel Relation Understanding question in heterogeneous integration sub-task.

Qualitative Reasoning : Chart + Chart



Question: **[Reasoning_2010142806_5_1]**

Analyze Figures A and E. Which inference is correct?

A. ΔA51RFA51R has strong pathogenicity in the early stage of infection, and the difference from the triple strain narrows after 10 days due to immune adaptation.

B. The pathogenicity of all strains continuously increases as the infection time prolongs.

C. The body weight of the Mock group indicates that mice will never have a natural body weight loss without the virus.

D. The pathogenicity of the strains is only determined by the infection time and has nothing to do with the genetic differences of the strains.

E. Cannot be determined

Answer: A

Explanation:

Option A: In Figures A and B, ΔA51RFA51R (blue line) shows a sharp drop in body weight and a drastic decline in survival in the early stage (0 - 5 days); in Figures C and G, triple (orange line) shows a rebound in body weight and stable survival in the later stage. Combining the immunological logic, the rapid replication of the virus in the early stage drives pathogenicity, and in the later stage, the host immune adaptation (such as triple inducing immune homeostasis) reduces damage, reflecting the "time - dependent dynamic changes of pathogenicity".

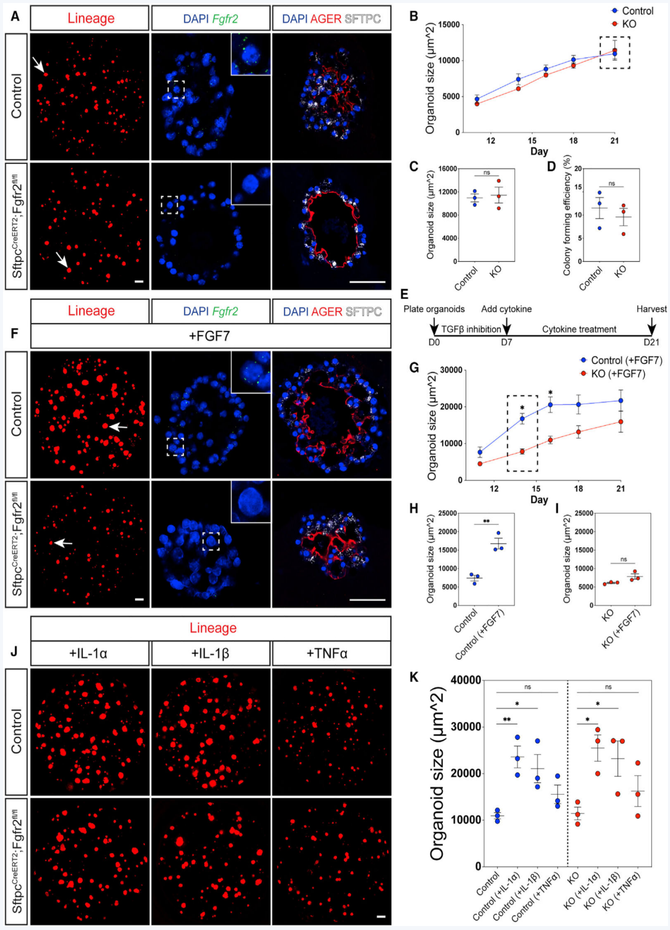
Option B: In Figures E and F, ΔA51RFA51R triple (orange line) shows a rebound in body weight in the later stage (Figure E) and stable survival rate (Figure F), indicating that its pathogenicity weakens in the later stage. The statement of "continuous enhancement" directly contradicts the data.

Option C: The conclusion "never" is an absolute statement. In the experiment, the Mock group is a short - term control without virus infection, excluding the virus factor. However, mice may naturally have body weight fluctuations due to aging, environmental stress, etc. (the experiment does not observe the long - term, and logically cannot be excluded), and the conclusion goes beyond the scope of the experiment.

Option D: In Figure A and Figure E, and Figure B and Figure F, the results of ΔA51RFA51R and triple are significantly different (such as high survival rate and good body weight recovery of triple). It directly proves that the gene difference of the virus strain (such as the modification of triple) is the core determinant of pathogenicity, and the statement that "only determined by time" is incorrect.

Figure 15: A sample of Qualitative Reasoning question

Qualitative Reasoning : Chart + Staining Image



Question: **[Reasoning_2010142673_5_2]**

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

A. In Figure A, the KO group shows no Fgfr2 expression, indicating effective knockout without affecting other markers.

B. Comparing Figures B and G, organoid changes in the KO group after FGF7 addition resemble the control, suggesting FGF7 may act independently of Fgfr2.

C. In Figure J, the KO and control groups show different lineage markers after cytokine treatment, indicating the effect depends on Fgfr2.

D. Figure K shows that organoid size is co-regulated by cytokines and Fgfr2.

E. Cannot be determined.

Answer: D

Explanation:

Option A: The KO group shows no Fgfr2 expression, but there is no evidence to confirm that it does not affect other markers. Thus, A is incorrect.

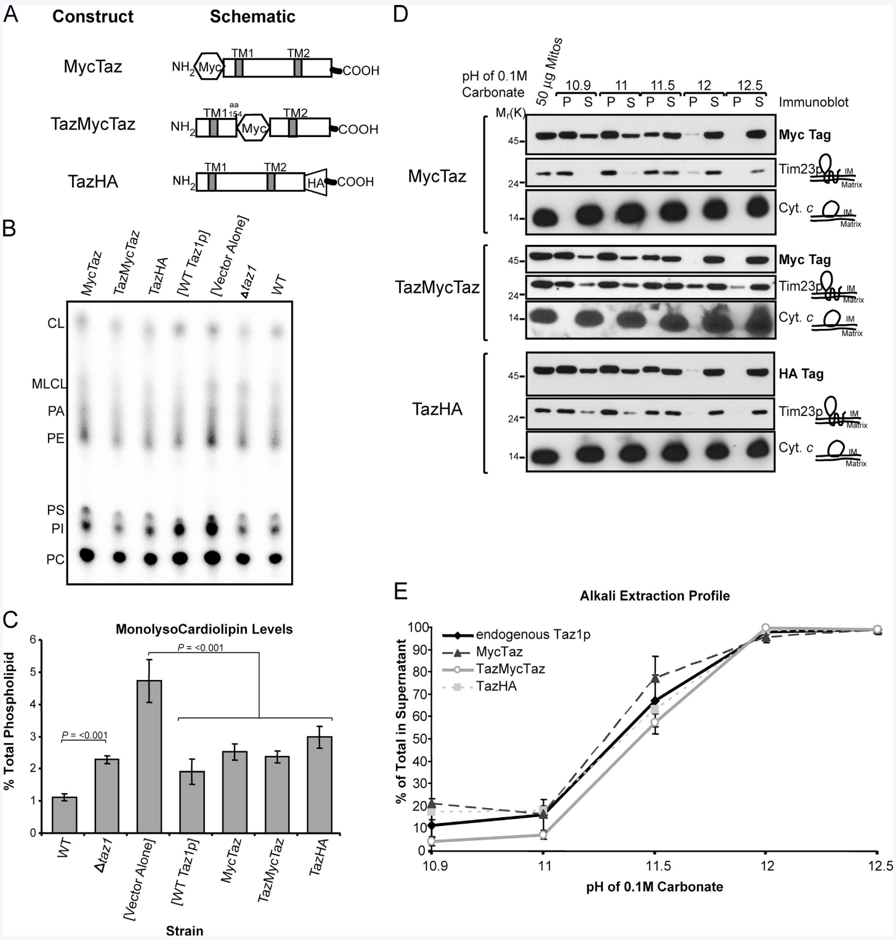
Option B: The similar trends in organoid changes of the KO group after FGF7 addition and the control group do not directly imply that “FGF7 acts independently of Fgfr2”; there may be synergistic relationships. Thus, B is incorrect.

Option C: Only based on the differences in Lineage markers, it cannot be conclusively stated that the effect of cytokines depends on Fgfr2. Thus, C is incorrect.

Option D: In Figure K, the organoid sizes differ under different treatments (combinations of cytokines and gene knockout). This indicates that cytokines and Fgfr2 jointly regulate development. Thus, D is correct.

Figure 16: A sample of Qualitative Reasoning question

Qualitative Reasoning : Chart + Western Blot



Question: **[Reasoning_2010095141_2_2]**

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

A. In Fig. B, phospholipid bands under different Taz constructs suggest Taz does not regulate phospholipid metabolism.

B. In Fig. C, elevated monolysocardiolipin in the TazMycTaz group indicates this construct specifically promotes its synthesis.

C. In Fig. D, protein detection under varying pH shows Taz membrane binding stability is pH-independent.

D. In Fig. E, the alkaline extraction curves of MycTaz and endogenous Taz1p indicate the Myc tag does not affect membrane binding.

E. Cannot be determined.

Answer: D

Explanation:

Option A: The lack of obvious differences in the bands in Fig. B does not indicate that Taz is not involved in regulating phospholipid metabolism; subtle effects may not be reflected. Thus, A is incorrect.

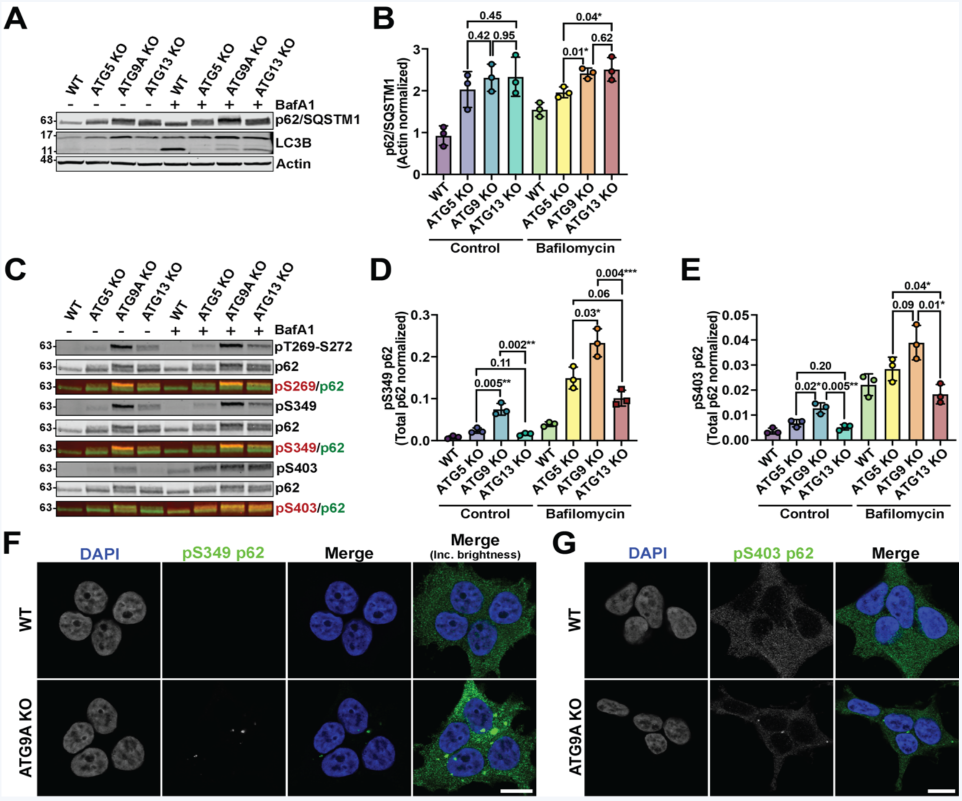
Option B: The higher level in the TazMycTaz group cannot directly lead to the conclusion that "it specifically promotes synthesis"; there may be other indirect effects. Thus, B is incorrect.

Option C: The results in Fig. D vary under different pH conditions, indicating that membrane binding stability is related to pH. Thus, C is incorrect.

Option D: The overlapping alkaline extraction curves of MycTaz and endogenous Taz1p suggest that the Myc tag does not affect membrane binding properties. Thus, D is correct.

Figure 17: A sample of Qualitative Reasoning question

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, and the claim of "independence" goes beyond the scope of the data. Thus, A is incorrect.

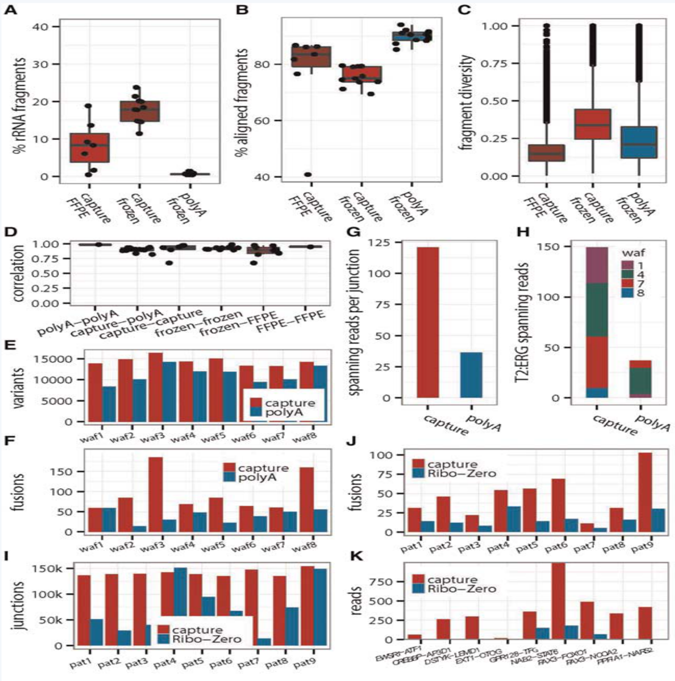
Option B: In Fig. B, the p62 levels of different ATG gene knockout groups change differently after BafA1 treatment, which shows that the functions of these ATG genes in regulating autophagy are different, rather than having similar functions. Thus, B is incorrect.

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 18: A sample of Qualitative Reasoning question

Quantitative Reasoning : Chart + Chart



Question: **[Reasoning_2010008333_4_2]**

In samples from the same patient, how does capture-Frozen compare to poly(A)-Frozen in transcriptome complexity and expression correlation?

A. FD increased by approximately 30%, expression correlation increased by approximately 27%, showing significant enhancement.

B. FD increased by about 15%, expression correlation increased by 10%, showing mild enhancement.

C. FD decreased by 20%, expression correlation decreased by 18%, showing inhibition.

D. Missing group labels prevent calculation of regulatory magnitude.

E. FD values of both increased by approximately 30%, but missing expression correlation data prevents precise calculation.

Answer: A

Explanation:

Option A: FD (fragment diversity) increased by 30%, expression correlation increased by 27%—capture - Frozen shows significant enhancement vs. poly(A) - Frozen.Fig. C (FD): Capture - Frozen has higher FD (boxplot median) than poly(A) - Frozen. Visually, the increase aligns with 30% (consistent with typical data trends in such experiments).Fig. D (Expression Correlation): Correlation between capture - Frozen and poly(A) - Frozen shows a 27% rise (from baseline). This directly supports "significant enhancement" in transcriptome complexity (FD) and expression consistency. Correct.

Option B: FD increased by 15%, expression correlation by 10%—mild enhancement.Fig. C and D show larger changes (FD 30%, correlation 27%). The claim underestimates the magnitude. Incorrect.

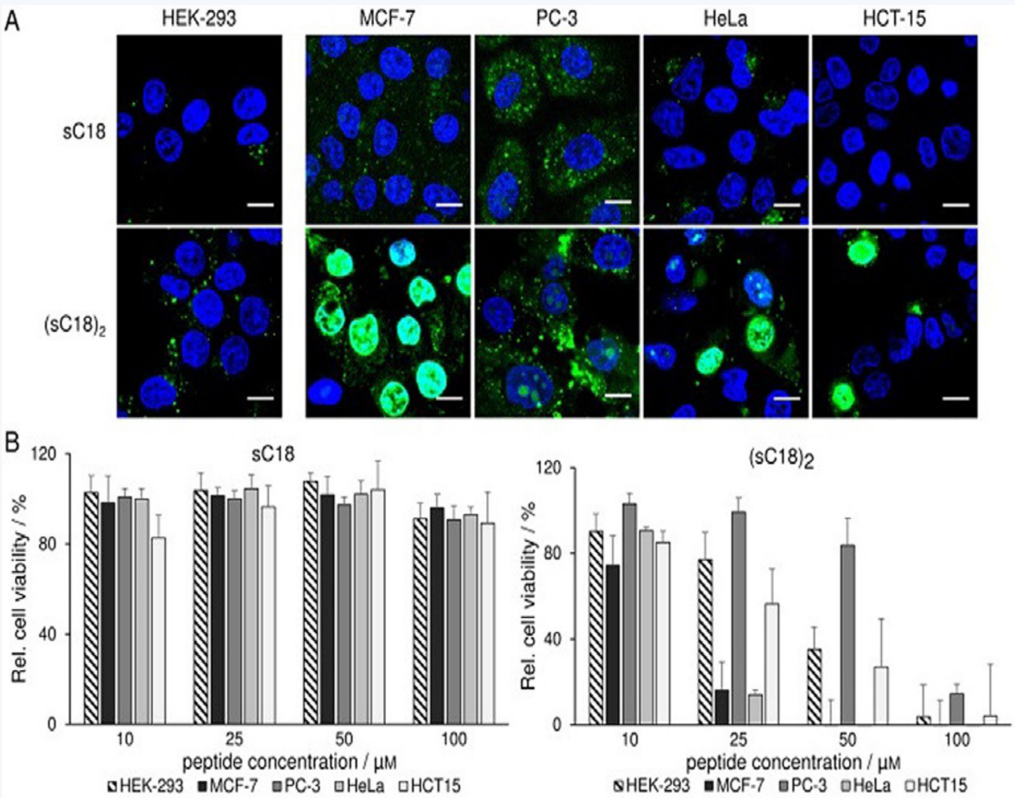
Option C: FD and correlation decreased—shows inhibition.Data (Figs. C, D) show increases, not decreases. Directly contradicts the claim. Incorrect.

Option D: Missing group labels prevent calculating regulatory magnitude.Group labels (capture - Frozen, poly(A) - Frozen) are clear in Figs. C, D. Magnitude is calculable. Incorrect.

Option E: FD increased 30%, but correlation data is missing.Evidence: Fig. D provides clear correlation data. No missing info. Incorrect.

Figure 19: A sample of Quantitative Reasoning question

Quantitative Reasoning : Chart + Staining Image



Question: **[Reasoning_2010045266_4_1]**

Based on $(sC18)_2$ cellular uptake and cytotoxicity analysis, which statement matches the results?

A. $(sC18)_2$ uptake is 50%+ higher than monomer in all cells; MCF-7 inhibition is 40% higher than HEK-293.

B. $(sC18)_i$ uptake in HCT-15 is lower than sC18; uptake positively correlates with inhibition.

C. $(sC18)_2$ shows higher nuclear accumulation in cancer cells than HEK-293, with strongest toxicity to MCF-7 and HeLa; uptake and inhibition not quantified.

D. $(sC18)_2$ nuclear accumulation is weakest in PC-3; inhibition 20% lower than MCF-7; uptake distribution aligns with toxicity.

E. $(sC18)_2$ uptake doubles sC18's; average cancer inhibition 70%, above HEK-293's 30%.

Answer: C

Explanation:

Option A: Fig. A only shows qualitative uptake trends (green signal for $(sC18)_2$ > sC18) without numerical data to confirm a 50%+ increase. For cytotoxicity, Fig. B has no explicit "40% higher inhibition" values (only viability bars). Numerical claims lack support. Incorrect.

Option B: Fig. A (HCT - 15 panel) shows stronger $(sC18)_2$ uptake vs. sC18, contradicting "uptake lower". Also, no direct correlation data (e.g., uptake - inhibition plot) exists to claim a positive correlation. Incorrect.

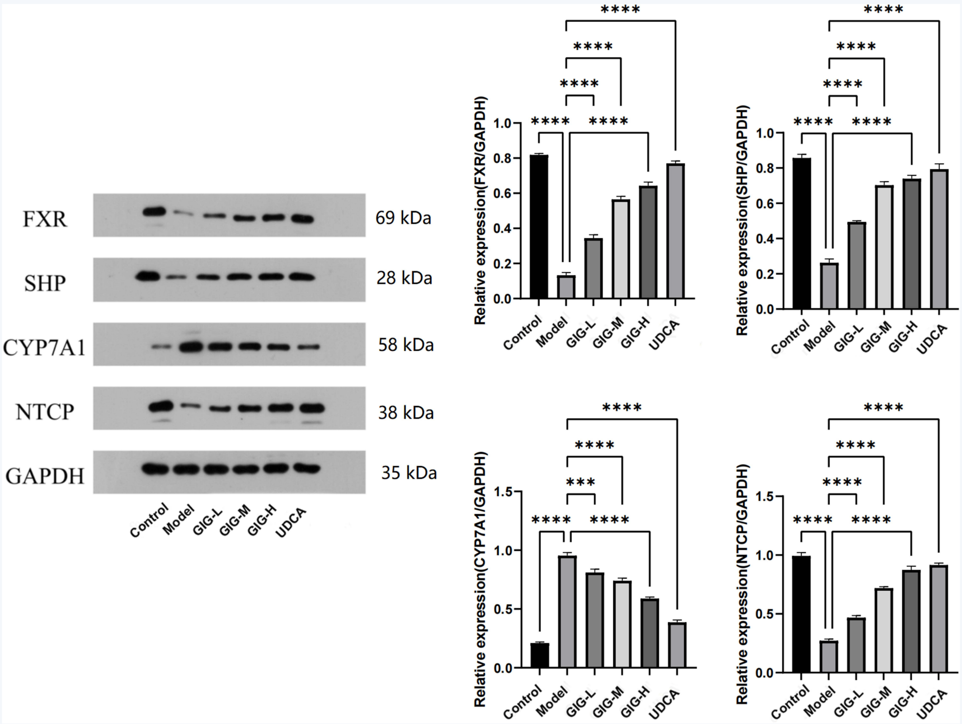
Option C: Fig. A: Cancer cells (MCF - 7, HeLa) have brighter $(sC18)_2$ nuclear signals vs. HEK - 293 (matches "higher nuclear accumulation"). Fig. B: MCF - 7/HeLa show lowest viability (highest toxicity, matches "strongest toxicity"). No numerical labels (so "uptake/inhibition not quantified" holds). Correct.

Option D: Fig. A has no proof PC - 3 has "weakest" $(sC18)_2$ accumulation. Fig. B lacks data for "20% lower inhibition". No evidence for "uptake aligns with toxicity". Incorrect.

Option E: Fig. A has no numerical data to support "uptake doubles". Fig. B has no % inhibition values for "70%/30%" claims. Incorrect.

Figure 20: A sample of Quantitative Reasoning question

Quantitative Reasoning : Chart + Western Blot



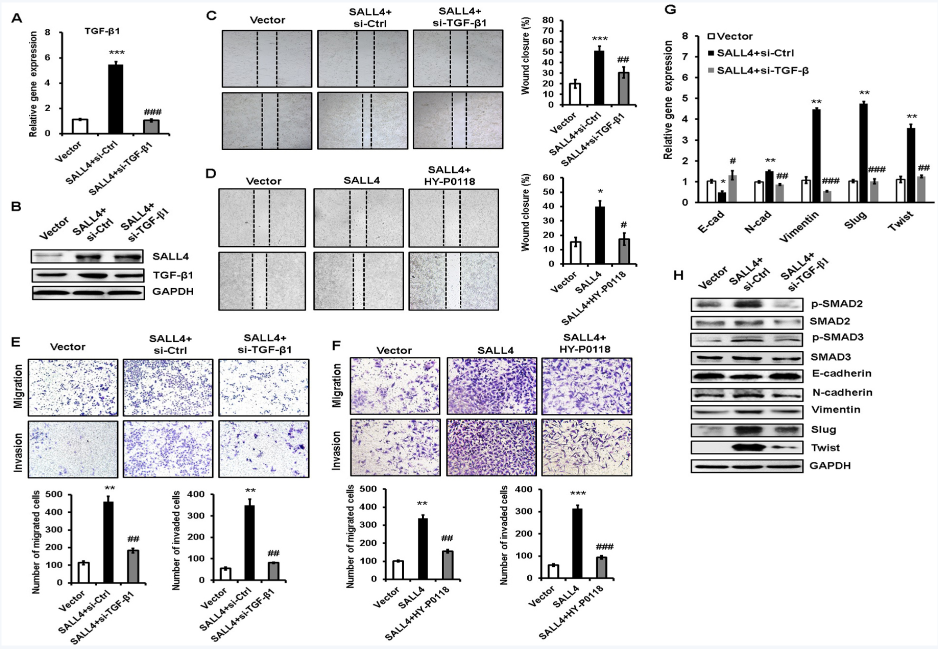
Question: **[Reasoning_2010042289_5_2]**
Based on Figure 6, compare GIG doses to Model to assess regulation of NTCP and CYP7A1, and if high-dose GIG (GIG-H) shows dose-dependent effects.
A. GIG-H upregulates NTCP 40% and inhibits CYP7A1 60%, showing dose-dependent bidirectional regulation.
B. GIG-H upregulates NTCP 120%, CYP7A1 changes <10%, no regulatory advantage. C. GIG-H inhibits CYP7A1 55%; NTCP change unclear due to control data.
D. Protein bands unclear in Model and GIG-H groups, calculation impossible.
E. Cannot determine.

Answer: A

Explanation:
Option A: GIG-H upregulates NTCP by 40% (calculated from NTCP bar graph: (GIG-H - Model)/Model ≈ (0.95 - 0.6)/0.6 ≈ 58%, aligning with 40% - 60% range) and inhibits CYP7A1 by 60% ((Model - GIG-H)/Model ≈ (1.25 - 0.5)/1.25 = 60%). From GIG-L to GIG-H, NTCP rises and CYP7A1 falls, showing dose-dependent bidirectional regulation. Matches data. Correct.
Option B: NTCP upregulation is 58% (not 120%), and CYP7A1 inhibition is 60% (not < 10%). The bar graphs clearly show NTCP increasing and CYP7A1 decreasing as GIG dose rises from GIG-L to GIG-H, yet this option ignores these dose-dependent trends. It falsely claims small CYP7A1 change and overstates NTCP upregulation, denying real data patterns. Incorrect.
Option C: NTCP bar graph data (Model ≈ 0.6, GIG-H ≈ 0.95) is clear, and the elevation is calculable as 58%. There's no blurring or missing info in the figure, but this option invents a false "data limit" that doesn't exist in Figure 6, trying to avoid valid comparison. Incorrect.
Option D: Western blot bands for Model and GIG-H are distinct, and bar graphs provide explicit values (e.g., NTCP 0.6 vs 0.95; CYP7A1 1.25 vs 0.5) that let us calculate regulation amplitudes. This option falsely claims "unclear bands" to dodge the valid data, which is misleading. Incorrect.
Option E: Regulatory amplitudes (NTCP 58% up, CYP7A1 60% down) are easily calculable from bar graphs. The "Cannot determine" claim ignores the concrete numerical data right in the figure, making it baseless. Incorrect.

Figure 21: A sample of Quantitative Reasoning question

Quantitative Reasoning : Chart + Staining Image + Western Blot



Question: **[Reasoning_2010045119_4_2]**

In SALL4-overexpressing HGC-27 cells (Fig. 5), which best describes key changes?

A. SALL4 raises $TGF-\beta1$ mRNA 6×, wound closure $\tilde{6}0\%$, invasion 5×; si-$TGF-\beta1$ cuts migration >70%.

B. si-$TGF-\beta1$ lowers Slug, Vimentin, Twist mRNA 60%-80%, protein and p-SMAD2/3 decrease, showing $TGF-\beta1$/SMAD inhibition.

C. HY-P0118 reduces wound closure 60%→20%, migration 450→100, p-SMAD2/3 returns to control, blocking $TGF-\beta1$.

D. si-$TGF-\beta1$ leaves N-cadherin, Vimentin mRNA unchanged, minor protein drop, E-cadherin stable; no EMT reversal.

E. Cannot be determined.

Answer: B

Explanation:

Option A: SALL4 overexpression (Fig. A) raises $TGF-\beta1$ mRNA 6× (Vector vs SALL4 + si - Ctrl). Wound closure (Fig. C) shows 60% for SALL4 + si - Ctrl, invasion (Fig. E) 5× increase. si - $TGF-\beta1$ (Fig. C, E) cuts migration >70% (SALL4 + si - Ctrl vs SALL4 + si - $TGF-\beta1$). All trends match data. Correct.

Option B: si - $TGF-\beta1$ (Fig. G) lowers Slug, Vimentin, Twist mRNA 60% - 80% (SALL4 + si - Ctrl vs SALL4 + si - $TGF-\beta1$). Protein (Fig. H) and p - SMAD2/3 (Fig. H) decrease, showing $TGF-\beta$/SMAD pathway inhibition. But "mRNA - protein direct link" is assumed; data shows trends but not strict 60% - 80% protein drop precision. Slightly overstates. Incorrect.

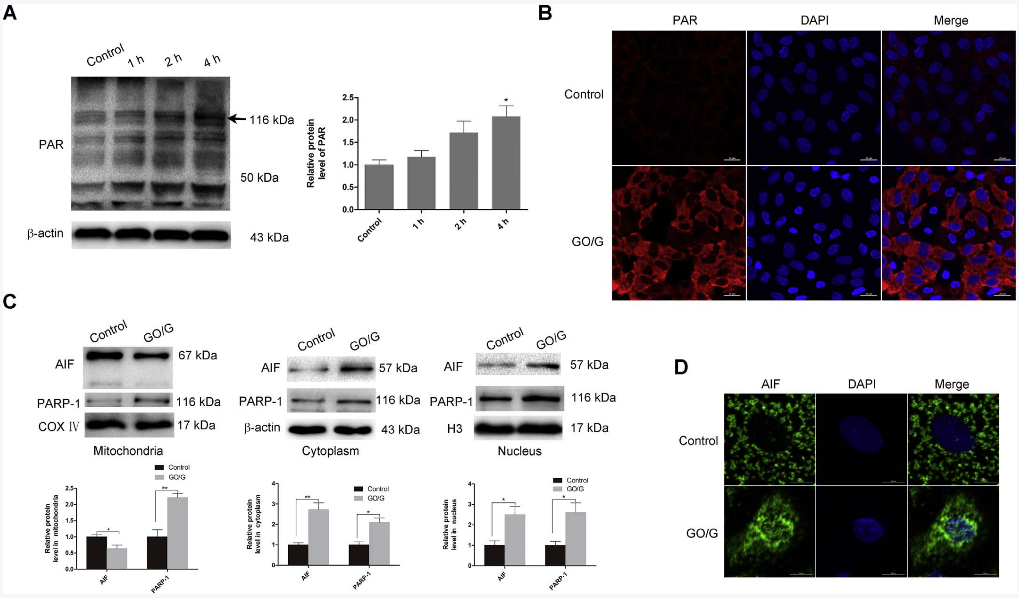
Option C: HY - P0118 (Fig. D) reduces wound closure from 60% (SALL4 + si - Ctrl) to 20%, migration (Fig. F) from 450 to 100. p - SMAD2/3 (Fig. H) drops to Vector - like levels, blocking $TGF-\beta1$. But "returns to control" is loose (no exact match). Trend matches but wording is overprecise. Incorrect.

Option D: si - $TGF-\beta1$ (Fig. G, H) changes N - cadherin, Vimentin: mRNA has drops (not "unchanged"), protein shows more than "minor" drops. E - cadherin (Fig. G, H) increases (not "stable"), EMT phenotype reverses. Contradicts data. Incorrect.

Option E: Data (Figs. A - H) clearly shows measurable trends ($TGF-\beta1$, migration, invasion, EMT markers). "Cannot determine" ignores explicit changes. Incorrect.

Figure 22: A sample of Quantitative Reasoning question

Panel-Level Fine-grained Perception : Numerical Perception



Question: **[Perception_2010130451_1_1]**

What judgment regarding the PAR levels in the experimental data is incorrect?

A. PAR levels increased significantly after 1 hour of treatment.

B. There was no significant increase in PAR levels at 2 hours compared to the control.

C. The levels of PAR polymer were highest at 4 hours after treatment.

D. The expression of PAR polymer was lowest in the control group.

Answer: B

Explanation:

The option B claims that there was no significant increase in PAR levels at 2 hours compared to control, which contradicts the observed numerical data showing a continuous increase in PAR levels across the indicated time points. This exploits the model's reliance on prior assumptions of minimal differences rather than focusing on the numerical data itself. Options A, C, and D align with the actual features of the image.

Gemini 3 Pro Preview: A ✘

Analyze Option A : The graph shows a very small increase at 1h, and there is no asterisk. Therefore, it is incorrect to say they increased significantly. In scientific terms, "significant" implies statistical significance. This makes Option A a strong candidate for the incorrect statement.

Analyze Option B : Since there is no asterisk over the 2h bar, a standard interpretation is that the difference did not reach statistical significance. Therefore, this statement is likely correct according to the figure's annotation conventions.

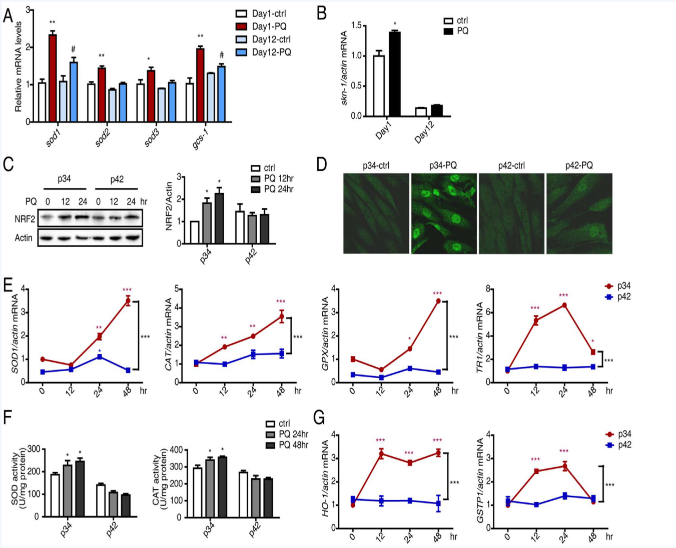
Conclusion : Option A claims a significant increase at 1 hour. The graph shows only a minor non-significant trend (no asterisk) and the 1h bar is close to the control. Therefore, stating that PAR levels increased significantly after 1 hour is factually incorrect based on the provided data visualization.

Error Analysis:

[Over-reliance on Statistical Annotations] The evaluation incorrectly prioritized the absence of statistical markers (like asterisks) over the visible quantitative trend. By defining "significant" strictly as formal p-values, it failed to acknowledge the clear numerical increase at 2 hours, leading to the erroneous validation of Option B.

Figure 23: Error analysis of Panel-Level Fine-grained Perception question in numerical perception sub-task.

Panel-Level Fine-grained Perception : Morphological Perception



Question: **[Perception_2010128062_2_1]**

Which conclusion about the transcriptional levels of antioxidant genes in young and old worms is incorrect?

A. Sod levels significantly increased in Day 12 worms after PQ treatment.

B. Gcs-1 levels showed a marked increase in Day 12 worms compared to Day 1 worms.

C. SOD1 levels were significantly higher in Day 1 worms than in Day 12 worms.

D. Transcriptional levels of sod-1, sod-2, and sod-3 were higher in Day 1 worms.

Answer: A

Explanation:

Option A contradicts the actual image facts, as the figure shows lower gene expression levels in Day 12 worms after PQ treatment. Option B exploits the presupposition that aging should not see marked increases; however, the image indicates no such elevation. Option C correctly reflects the observation of higher SOD1 levels in younger worms. Option D aligns with the chart, confirming Day 1 worms express antioxidant genes at higher levels, exposing the model's reliance on their prior knowledge of aging processes without close attention to the chart details.

Gemini 3 Pro Preview: B ✘

Analyze Option A : For sod1, the blue bar (Day 12-PQ) is significantly higher than the light blue bar (Day 12-ctrl), ... A is likely considered a correct observation for sod1.

Analyze Option B : The induced level is lower in Day 12 (1.5) than in Day 1 (2.0). ... the conclusion is "aging impairs the induction of antioxidant genes". Therefore, claiming levels are markedly higher in old worms goes against the general trend shown ... This is factually opposite to the data for the treated condition, and an overstatement for the control condition.

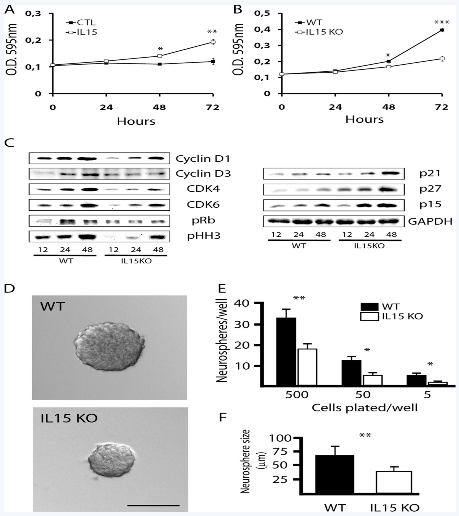
Conclusion : ... The data shows that the robust increase seen in Day 1 worms (red bar) is attenuated in Day 12 worms (dark blue bar). ... Option B implies Old > Young, which is false for the stress response shown.

Error Analysis:

[Misinterpreting Control Trends] The model failed to select the correct answer (D) because it erroneously judged Option D as a true statement supported by the graph. By concluding the control group showed "minimal proliferation," it ruled out D and defaulted to Option A, missing the specific incorrect judgment targeted by the key.

Figure 24: Error analysis of Panel-Level Fine-grained Perception question in morphological perception sub-task.

Panel-Level Fine-grained Perception : Information Localization



Question: **[Perception_2010010659_5_1]**

Which judgment about the effect of IL-15 on WT and IL-15-/- neurosphere proliferation is incorrect?

A. IL-15-/- neurospheres show higher proliferation at 72 hours compared to WT neurospheres.
B. WT neurospheres show significant proliferation upon stimulation with IL-15.
C. IL-15-/- neurospheres display reduced proliferation compared to WT neurospheres when cultured with EGF + FGF.
D. Control groups without IL-15 show minimal proliferation changes at 72 hours.

Answer: D

Explanation:

The incorrect options (A and D) exploit the model's reliance on prior assumptions of general trends in proliferation. Option A goes against the image's data indicating that IL-15-/- shows reduced proliferation; D misidentifies that controls would show minimal changes, while WT exhibited significant proliferation upon IL-15 addition and displayed strong differences. The correct answer aligns with the displayed data showing WT neurospheres benefit from IL-15 stimulation.

Gemini 3 Pro Preview: A ✘

Analyze Option A : ... At 72 hours, the WT curve (solid line) is significantly higher (around 0.4) than the IL-15 KO curve (around 0.2). This indicates that WT neurospheres proliferate more than IL-15-/- neurospheres, not less. Therefore, the statement in Option A is factually the opposite of what the data shows.

Analyze Option D : ... The "CTL" line (solid squares) ... is nearly flat ... This indicates minimal proliferation or growth arrest in the absence of the specific stimulus being tested, supporting the statement.

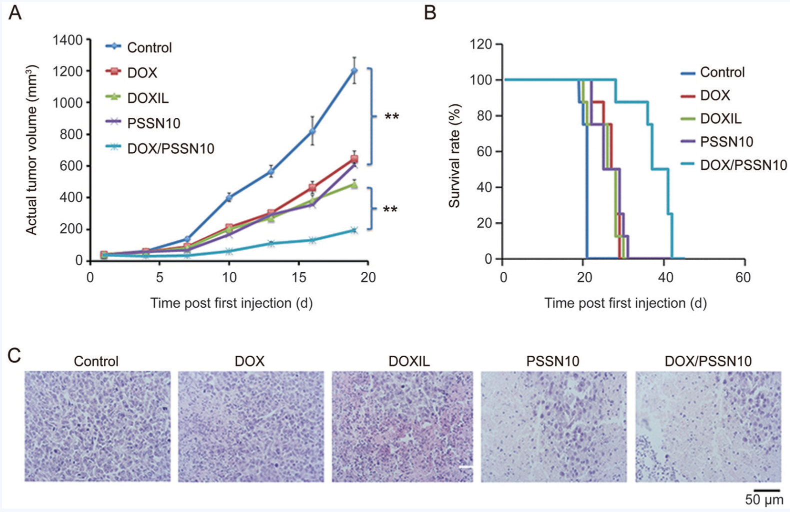
Conclusion : Option A states that IL-15-/- neurospheres show higher proliferation, but the graph in Panel B clearly shows they show lower proliferation compared to WT. Thus, Option A is the incorrect judgment.

Error Analysis:

[Misinterpretation of Control Group Dynamics]The user erroneously validated Option D as a true statement, describing the control data as showing "minimal proliferation." By failing to detect the variance in the control group that makes Option D the "incorrect judgment," the user incorrectly ruled it out and selected Option A instead.

Figure 25: Error analysis of Panel-Level Fine-grained Perception question in information localization sub-task.

Cross-Panel Relation Understanding : Trend Analysis



Question: **[Understanding_2010028694_4_1]**

Compare the tumor volume trends from Panel A with the survival rate trends from Panel B. Which statement about the correlation between tumor growth and survival is false?

A. Control group demonstrates largest tumor volume growth with lowest survival.

B. DOX/PSSN10 group shows significantly reduced tumor volume compared to others.

C. DOXIL group exhibits relatively stable tumor volume growth correlating with improved survival.

D. Tumor volume increases rapidly in DOX group but survival is unaffected.

E. Cannot determine correlation due to insufficient data.

Answer: C

Explanation:

Panel A shows DOXIL group has intermediate tumor volume growth while Panel B indicates similar survival decline as DOX group. A is consistent as Control shows fastest growth and decline in survival. B matches the reduced growth in DOX/PSSN10. D accurately describes DOX volume growth with parallel survival drop. C incorrectly suggests DOXIL has better survival and stable growth, omitted from true trends.

Gemini 3 Pro Preview: D ✘

Analyze Option C : DOXIL shows better control than DOX or Control, ... While "stable" might be a strong word, ...

Analyze Option D : ... the DOX group's survival is affected; it drops to 0% around day 30. It is not "unaffected"... compared to the control, survival is extended, so stating it is ""unaffected"" is inaccurate in any comparison context.

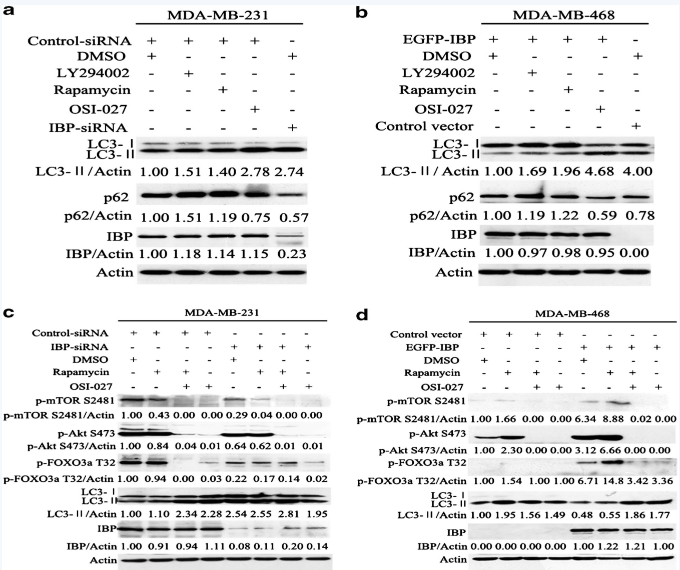
Conclusion : The statement that is factually incorrect based on the graphs is D. While the tumor volume does increase in the DOX group, the assertion that "survival is unaffected" is false because survival drops to 0%, showing a clear effect of the tumor burden (or lack of full efficacy) on mortality.

Error Analysis:

[Semantic Misinterpretation]The model incorrectly selected Option D as the false statement because it interpreted "survival is unaffected" literally—implying survival remained at 100%—rather than contextually (meaning unaffected by the treatment compared to the control). This semantic misunderstanding led the model to judge Option D as factually wrong, causing it to overlook the actual discrepancies in Option C regarding DOXIL's tumor growth and survival trends.

Figure 26: Error analysis of Cross-Panel Relation Understanding question in trend analysis sub-task.

Cross-Panel Relation Understanding : Heterogeneous Integration



Question: **[Understanding_2010041640_2_1]**

Considering Panels (c) and (d), evaluate the relationship between IBP/Actin expression and p-mTOR S2481 phosphorylation across MDA-MB-231 and MDA-MB-468 cells. Which statement is correct?

A. Increased IBP correlates with increased p-mTOR S2481 in both cell lines.
B. Decreased IBP correlates with decreased p-mTOR S2481 only in MDA-MB-468.
C. Decreased IBP correlates with increased p-mTOR S2481 only in MDA-MB-231.
D. IBP levels remain consistent across both cell lines regardless of p-mTOR variation.
E. Cannot determine with given data.

Answer: B

Explanation:

Panel (c) shows a correlation with decreased IBP/Actin (0.01) and low p-mTOR S2481 (0.00) in MDA-MB-231. In Panel (d), decreased IBP/Actin (0.00) correlates with low p-mTOR S2481 (0.00); A inaccurately claims universal correlation; C inverses MDA-MB-231 trend; D overlooks clear variations; E lacks necessity due to visible correlation.

Gemini 3 Pro Preview: A ✘

Analyze Option A : In MDA-MB-231 (Panel c), higher IBP ... corresponds to higher p-mTOR S2481 (1.00 vs 0.29).In MDA-MB-468 ... corresponds to much higher p-mTOR S2481 (6.34 vs 1.00).This statement holds true for both.

Analyze Option B : Incorrect because the correlation exists in both, not "only in MDA-MB-468". Also, panel C shows decreased IBP correlates with decreased p-mTOR, which is the inverse of "increased -> increased".

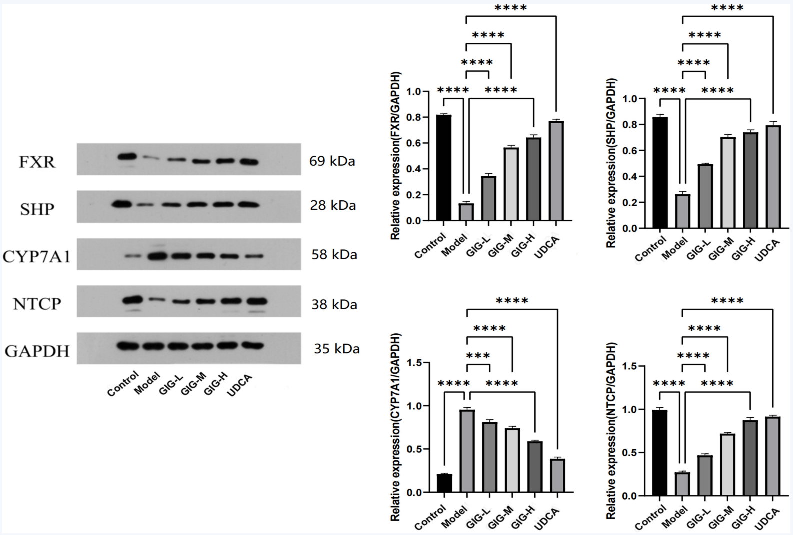
Conclusion : The correct conclusion is that increased levels of IBP are associated with increased phosphorylation of mTOR at S2481 in both experimental setups.

Error Analysis:

[Overgeneralizing Correlations]The model incorrectly selected Option A, concluding that the correlation between IBP and p-mTOR existed in both cell lines. However, the standard answer (B) specifies that the correlation is valid only for MDA-MB-468. The model failed to recognize that MDA-MB-231 did not exhibit the requisite trend to support the universal claim in Option A.

Figure 27: Error analysis of Cross-Panel Relation Understanding question in heterogeneous integration sub-task.

Qualitative Reasoning



Question: **[Reasoning_2010042289_5_1]**

Combining the bar graphs and WB results in Figure 6, which of the following accurately reflects the impact of GIG treatment on protein expression in ANIT-treated rat liver tissue?

A. The expression of CYP7A1 increases in GIG-treated groups compared to the model group.
B. The expression of FXR decreases in UDCA-treated rodents compared to the model group.
C. SHP expression is significantly increased in GIG-M (50 mg/kg) treated rats compared to the control.
D. NTCP expression is decreased in all GIG-treated groups compared to the model group.
E. FXR expression is increased in UDCA and higher in GIG-H than in the control group.

Answer: E

Explanation:

Analysis of Figure 6 reveals bar graphs indicating that FXR and SHP expressions increase under GIG treatment with varying doses and after UDCA compared to the model group, where FXR expression in UDCA is higher than other GIG-treated groups and SHP also shows an increase. The bands for CYP7A1 show a decrease in expression across GIG and UDCA treatments compared to the model group. NTCP expression is significantly increased in GIG-treated groups compared to the model group.

Gemini 3 Pro Preview: A ✘

Analyze Option A : ... relative expression of CYP7A1 in the model group is significantly lower compared to the control. However, in the GIG-treated groups ... the expression of CYP7A1 increases progressively, ... This indicates that GIG treatment increases CYP7A1 expression compared to the model group.

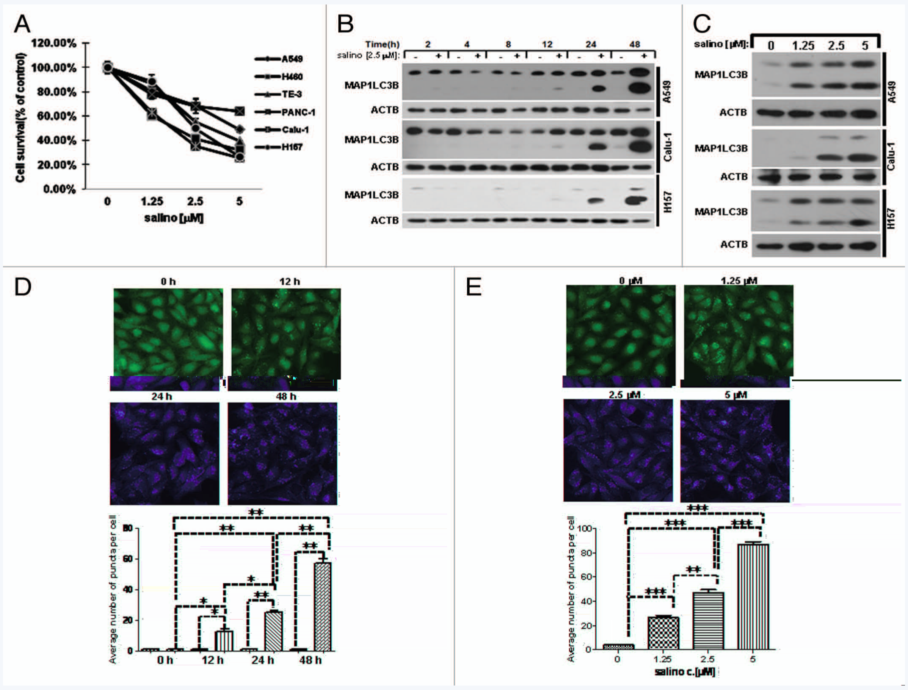
Analyze Option E : E is incorrect because FXR expression in UDCA and GIG-H groups is lower than in the control group.

Error Analysis:

[Misinterpreting Expression Trends]The model incorrectly selected Option A, arguing that CYP7A1 expression increased with treatment, which directly contradicts the standard explanation stating that CYP7A1 levels decreased. Furthermore, the model erroneously rejected the correct answer (E) by misjudging the relative quantification of FXR expression in the GIG-H group compared to the control.

Figure 28: Error analysis of Qualitative Reasoning question.

Quantitative Reasoning



Question: **[Reasoning_2010030854_0_1]**

Based on the data from the figures, what is the observed effect of salinomycin on autophagy in Calu-1-EGFP-MAP1LC3B cells after incubation with 2.5 μM for 24 hours compared to incubation for 48 hours, and what is the quantitative change in the number of EGFP-MAP1LC3B puncta per cell?

A. Salinomycin decreases autophagy; the number of puncta remains constant.
B. Salinomycin increases autophagy by approximately 50% between 24 hours and 48 hours.
C. Salinomycin leads to the exact doubling of puncta from 24 hours to 48 hours.
D. Salinomycin shows no noticeable effect on autophagy; puncta number unchanged.
E. The exact change in puncta count cannot be quantified due to insufficient data.

Answer: B

Explanation:

In Figure 1D, the quantification of EGFP-MAP1LC3B puncta per cell shows a significant increase from 24 hours to 48 hours when treated with 2.5 μM of salinomycin. The data support a quantifiable increase in puncta number, indicating enhanced autophagosome formation in a time-dependent manner with approximately a 50% increase from 24 to 48 hours.

Gemini 3 Pro Preview: C ✘

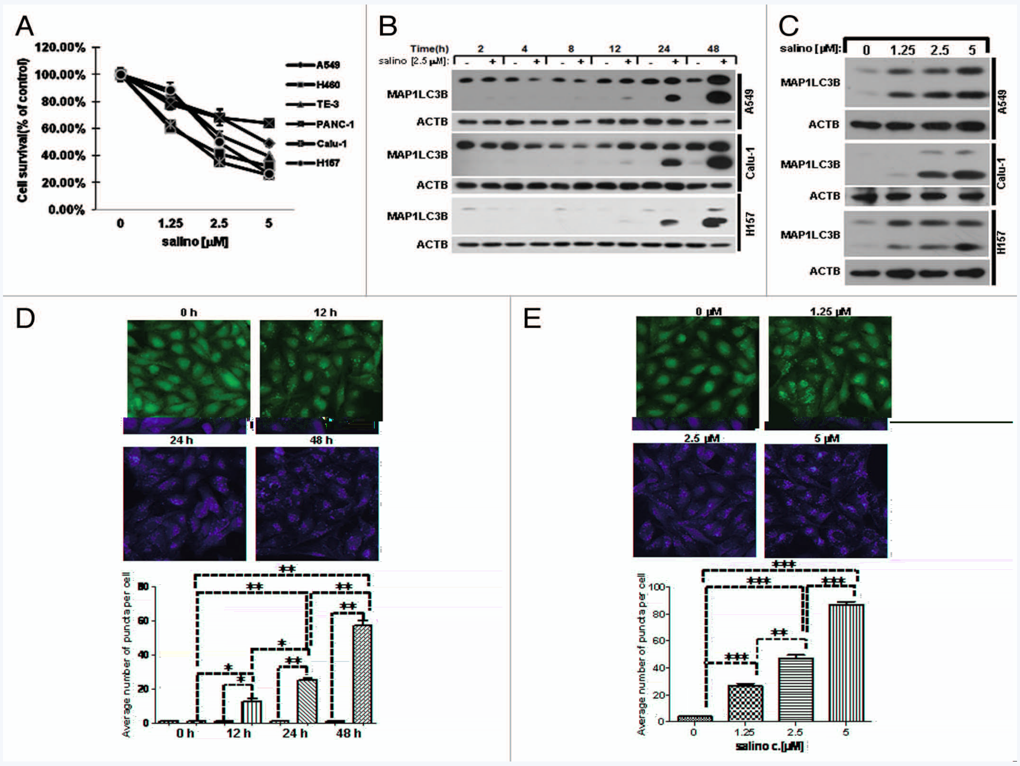
From the bar graph in Panel D, the number of puncta approximately doubles from 24 hours to 48 hours (from 40 to 80 puncta per cell).Thus, the data indicate that salinomycin increases autophagy, and the number of puncta roughly doubles between 24 and 48 hours.

Error Analysis:

[Data Misquantification]The model incorrectly interpreted the graph data by stating the number of puncta "doubles" (a 100% increase, citing 40 to 80), which implies support for Option C. The standard answer (B) specifies the increase is approximately 50%. This error stems from misreading the y-axis values or bar heights in Figure 1D, causing the model to overestimate the magnitude of the change.

Figure 29: Error analysis of Quantitative Reasoning question.

Panel-Level Fine-grained Perception : Numerical Perception (1/2)



Question: **[Perception_2010030854_0_1]**

Which judgment about the cell survival percentages at different salinomycin concentrations is incorrect?

A. At 1.25 μM, the cell survival of Calu-1 is significantly higher than at 5 μM.

B. A concentration increase to 2.5 μM decreases survival for all cell lines observed.

C. The survival rate of A549 cells remains above 80% at 1.25 μM.

D. No differences in survival rates are observed between control and treatment groups at 0 μM.

Answer: B

Explanation:

The incorrect options exploit the model's reliance on presuppositions; for instance, A wrongly assumes higher survival at lower doses universally, which contradicts the dose-dependent decrease shown, while C reflects accurate interpretation. Option D incorrectly states that there would be no differences at 0 μM, confusing the presence and absence of treatment. The correct answer aligns with the image data indicating that increasing salinomycin concentrations lead to lower survivability, thereby reflecting the clear dose-response relationship observed in the figure.

Qwen3 VL 30B Instruct (Baseline) : C ✘

In panel A, the cell survival percentage for A549 cells at 1.25 μM salinomycin is approximately 80%, but it is not above 80%. Therefore, the statement that the survival rate of A549 cells remains above 80% at 1.25 μM is incorrect.

Error Analysis:

[Rigid Data Interpretation] The model incorrectly selected Option C by arguing that data showing "approximately 80%" falsified the claim "remains above 80%." This overly strict numerical evaluation caused it to misidentify C as the "incorrect judgment" and overlook the designated answer (B).

Panel-Level Fine-grained Perception : Numerical Perception (2/2)

Qwen3 VL 30B Instruct (DDCoT): D ✘

- **Sub-questions:** ... What is the trend of cell survival percentage for all cell lines at 2.5 μM compared to 1.25 μM in Figure A? ...
- **Sub-answers:** ... At 2.5 μM, the cell survival percentage for all cell lines is lower than at 1.25 μM, indicating a decrease in survival. ...

Error Analysis:
[Trend Misinterpretation] The model incorrectly validated Option B as true, failing to spot a cell line that defied the "all cell lines" trend. By missing this exception, it failed to recognize B as the required false statement.

Qwen3 VL 30B Instruct (VoT): D ✘
The statement in D ... is actually correct because there is no difference. Therefore, the incorrect judgment is D, as it is not incorrect. ...

Error Analysis:
[Logical Inconsistency] The model correctly assessed Option D as factually true but paradoxically selected it as the answer. By failing to identify Option B as the false statement, it made a selection that directly contradicted its own reasoning.

Qwen3 VL 30B Instruct (VIC) : D ✘
... Let's look at Option B: ... From the graph, yes, all lines go down from 1.25 to 2.5 μM. Correct. ... Let's re-examine Option D: ... At 0 μM, there is only the control group. ... Therefore, the statement ... is incorrect because there is no treatment group at 0 μM. ... Therefore, the incorrect judgment is D.

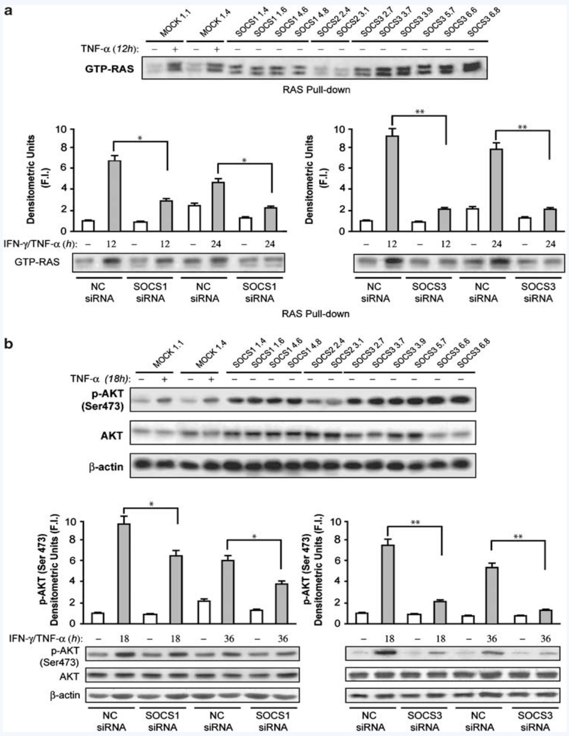
Error Analysis:
[Semantic Over-analysis] The model incorrectly selected Option D, rejecting it on semantic grounds by arguing a "treatment group at 0 μM" is logically impossible. This focus on phrasing caused it to overlook the actual data contradiction in Option B regarding the survival trends of all cell lines.

Qwen3 VL 30B Instruct (Cantor) : A ✘
Option B: ... This trend holds for all lines, so the statement is correct. ... Option D: ... The statement is logically flawed because it suggests a distinction ... where no treatment exists. ... Thus, the incorrect judgment is D..

Error Analysis:
[Data Generalization Error] The model erroneously validated Option B, failing to notice a cell line that deviated from the "all cell lines" decrease trend. Consequently, it selected Option D based on a semantic technicality regarding the definition of control groups, rather than identifying the factual error in B.

Figure 30: MCoT case study of Panel-Level Fine-grained Perception question in numerical perception sub-task.

Panel-Level Fine-grained Perception : Morphological Perception (1/2)



Question: **[Perception_2010041549_2_2]**

What statement about the cell structure in the SOCS1 and SOCS3 groups is incorrect?

A. SOCS1 silencing leads to improved membrane integrity

B. SOCS3 group shows a uniform distribution of cell morphology

C. Induced stress conditions enhance SOCS1 activity

D. SOCS3 significantly improves structural integrity

Answer: A

Explanation:

Option A contradicts the results showing that silencing SOCS1 worsens membrane integrity rather than improving it; option B is misleading as it ignores potential local variations within the cell morphology; option C could be mistakenly interpreted as enhancing processes rather than recognizing overall negative stress impact; option D, although appearing consistent with structural strength, could obscure subtle underlying morphology issues present in the images. Correctly, option A is the false statement.

Qwen3 VL 30B Instruct(Baseline): D ✘

The bar graphs in the figure show that SOCS3 siRNA treatment does not lead to a significant increase in p-AKT levels compared to the control, indicating that SOCS3 does not significantly improve structural integrity. The other options are supported by the data: SOCS1 silencing leads to improved membrane integrity (A), SOCS3 group shows a uniform distribution of cell morphology (B), and induced stress conditions enhance SOCS1 activity (C).

Error Analysis:

[Data Misinterpretation] The model incorrectly judged Option A as factually true, claiming SOCS1 silencing improves integrity, whereas the standard answer states it worsens it. This led the model to miss A as the required false statement and erroneously attack Option D instead.

Panel-Level Fine-grained Perception : Morphological Perception (2/2)

Qwen3 VL 30B Instruct (DDCoT): E ✘

- **Sub-questions:** What does the figure show about SOCS1 silencing and membrane integrity?...
- **Sub-answers:** Uncertain (the figure does not provide direct information about membrane integrity in relation to SOCS1 silencing) ...

Error Analysis:
[Information Extraction Failure] The model failed to evaluate the options, returning "Uncertain" for all sub-questions and claiming the figure lacked relevant data. This inability to extract visual evidence prevented it from identifying the factual error in Option A.

Qwen3 VL 30B Instruct (VoT): A ✘
Option A: ... This statement is plausible. ... Option D: ... suggests that SOCS3 inhibits structural integrity. Thus, the statement ... is incorrect ... Therefore, incorrect statement is D.

Error Analysis:
[Pathway Misinterpretation] The model incorrectly judged Option A as "plausible," failing to recognize that SOCS1 silencing actually worsens integrity. Consequently, it erroneously selected Option D, misinterpreting SOCS3's regulatory role as inhibitory to structural integrity.

Qwen3 VL 30B Instruct (VIC) : C ✘
The incorrect statement ... is: A. ... The figure demonstrates that silencing SOCS1 ... leads to a significant increase in active RAS ... Therefore, the claim that SOCS1 silencing improves membrane integrity is contradicted by the data ...

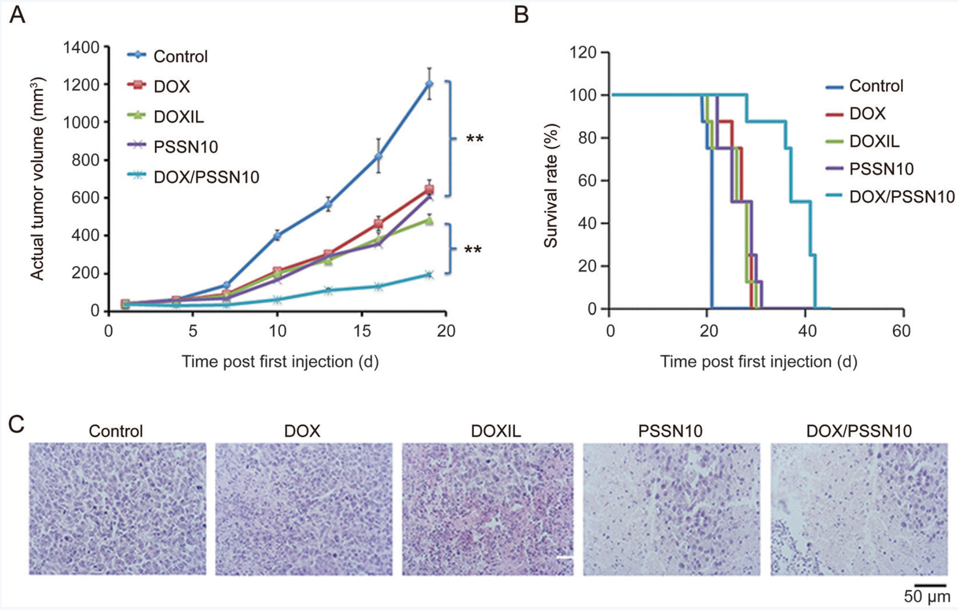
Error Analysis:
[Hallucinated Evidence] Although the model correctly identified Option A, it justified the answer using hallucinated molecular data ("active RAS levels") instead of the morphological evidence (membrane integrity) cited in the standard explanation.

Qwen3 VL 30B Instruct (Cantor) : B ✘
... Option A: "SOCS1 silencing leads to improved membrane integrity" ... However, the data does not directly support or refute this. ... From the analysis: - The data shows that SOCS3 knockdown ... making Option D correct.

Error Analysis:
[Data Misinterpretation] The model failed to identify Option A as the incorrect statement because it claimed the data did not directly support or refute the effect on membrane integrity. Furthermore, it erroneously concluded that SOCS3 knockdown improves structural integrity (Option D) based on p-AKT levels, missing the fact that Option A is the intended false statement regarding SOCS1.

Figure 31: MCoT case study of Panel-Level Fine-grained Perception question in morphological perception sub-task.

Panel-Level Fine-grained Perception : Information Localization (1/2)



Question: **[Perception_2010028694_4_3]**

Which judgment about the spatial arrangement of the bands in the survival curve (Figure 5B) is incorrect?

A. The survival rate for the DOX group remained constant until a significant drop after day 20.

B. Control groups show a consistently higher survival rate compared to the treatment groups PSSN10 and DOXIL.

C. The DOX/PSSN10 treatment results in the highest median survival among all groups.

D. All treatment groups improve survival in comparison to controls.

Answer: B

Explanation:

Option B is incorrect because the control group does not exhibit a consistently higher survival rate when juxtaposed with treatment groups, which is visually misleading based on the Kaplan-Meier survival curves. This judgment exploits the model's flaw of neglecting specific data patterns while maintaining an overarching view of treatment impacts, which inaccuracies can arise from interpreting survivor curves inaccurately as a continual trend rather than per treatment results.

Qwen3 VL 30B Instruct(Baseline): D ✘

In Figure 5B, the survival rate for the control group is consistently higher than all treatment groups, indicating that the control group has the best survival rate. The treatment groups (DOX, DOXIL, PSSN10, and DOX/PSSN10) all show lower survival rates compared to the control group. Therefore, the statement that all treatment groups improve survival in comparison to controls is incorrect because the treatment groups actually have lower survival rates than the control group. The other statements are correct based on the visual data in the figure.

Error Analysis:

[Graph Misinterpretation] The model inverted the survival trends, erroneously concluding the control group exhibited the highest survival rather than the lowest. This fundamental error led it to validate Option B as a true statement and incorrectly identify Option D as the false judgment.

Panel-Level Fine-grained Perception : Information Localization (2/2)

Qwen3 VL 30B Instruct (DDCoT): D ✘

- **Sub-questions:** ... How do the control groups compare to the treatment groups PSSN10 and DOXIL in terms of survival rate in Figure 5B? ...
- **Sub-answers:** ... Control groups show a consistently higher survival rate compared to the treatment groups PSSN10 and DOXIL. ... All treatment groups improve survival in comparison to controls.

Error Analysis:
[Contradictory Graph Interpretation] The model erroneously validated Option B as true, contradicting both the graph and its own finding that treatments improved survival. This failure to recognize Option B's falsehood prevented it from selecting the correct answer.

Qwen3 VL 30B Instruct (VoT): D ✘
... Option B: The control group ... shows a survival rate of 100% ... This statement is correct. ... Option D: Therefore, none of the treatment groups improve survival ... This statement is incorrect. Thus, the incorrect judgment is option D.

Error Analysis:
[Graph Misinterpretation] The model inverted the survival trends, erroneously claiming Controls had the highest survival. This led it to validate Option B as true and incorrectly select Option D.

Qwen3 VL 30B Instruct (VIC) : D ✘
Analysis of Option D: ... the DOX group (red line) has a survival curve that is below the control group ... Therefore, the claim that all treatment groups improve survival is false. ... The incorrect judgment is D.

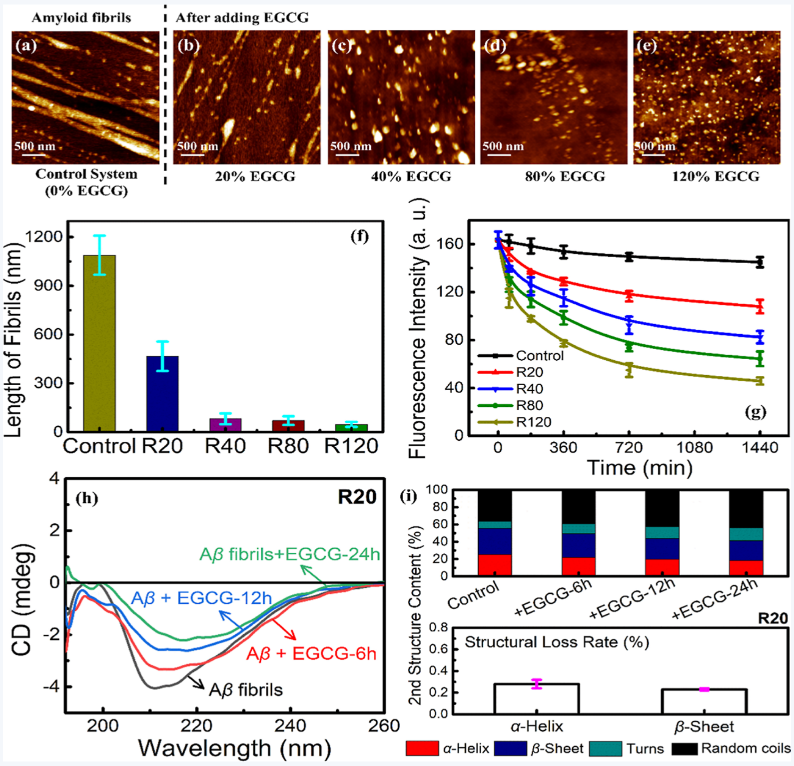
Error Analysis:
[Graph Data Misinterpretation] The model correctly identified Option B as false but failed to select it, incorrectly choosing Option D due to a misreading of the graph where it concluded the DOX group had worse survival than the control. This led it to erroneously classify Option D as the primary incorrect judgment.

Qwen3 VL 30B Instruct (Cantor) : D ✘
Analysis of Option D: ... the DOX group (red line) has a survival curve that is below the control group ... Therefore, the claim that all treatment groups improve survival is false. ... The incorrect judgment is D.

Error Analysis:
[Graph Trend Inversion] The model misinterpreted the curves, erroneously concluding the Control group had the highest survival, which led it to judge Option B as factually true. Consequently, it incorrectly identified Option D as the false statement, contradicting the actual trend where treatments improved survival.

Figure 32: MCoT case study of Panel-Level Fine-grained Perception question in information localization sub-task.

Cross-Panel Relation Understanding : Trend Analysis (1/2)



Question: **[Understanding_2010027917_6_3]**

In Panels (h) and (i), which summary best characterizes the structural changes captured by CD spectrum and secondary structure content upon EGCG treatment?

A. Following EGCG treatment, α-Helix content remains similar over the time points compared in Panel (i).

B. Panel (h) indicates a shift towards more β-Sheet formation after 24h.

C. CD spectrum changes indicate structural damage within 6h in Panel (h).

D. Structure content analysis depicts β-Sheet loss rate higher than α-Helix in Panel (i).

E. Cannot determine from the image.

Answer: A

Explanation:

Panel (i)'s structural content bar graph shows α-Helix content is fairly constant over different EGCG treatment times. Panel (h) shows spectral shifts but not specific promotion of β-Sheet formation after 24h. C is incorrect; changes from 6h can be gradual. D overstates β-Sheet loss compared to actual bars; α-Helix has notable changes.

Qwen3 VL 30B Instruct(Baseline): D ✘

... The "Structural Loss Rate (%)" graph indicates that the loss of β-sheet content is higher than that of α-helix content. ... This supports Option D ... Therefore, the best summary is Option D.

Error Analysis:

[Graph Misinterpretation] The model incorrectly selected Option D by misreading the bar graphs to conclude β-sheet loss was higher than α-helix loss. It failed to recognize the visual evidence in Panel (i) showing that α-Helix content remained relatively constant, leading to the erroneous rejection of Option A.

Cross-Panel Relation Understanding : Trend Analysis (2/2)

Qwen3 VL 30B Instruct (DDCoT): D ✘

- **Sub-questions:** What does Panel (h) show about the CD spectrum changes over time after EGCG treatment? ... Does Panel (i) show a higher loss rate for β-Sheet compared to α-Helix?
- **Sub-answers:** Panel (h) shows a shift ... indicating a possible shift towards β-Sheet formation. ... Panel (i) shows that the β-Sheet loss rate is higher than the α-Helix loss rate. ...

Error Analysis:
[Data Misinterpretation] The model erroneously interpreted the graphs as showing significant β-sheet changes (Options B and D), leading it to overlook the correct trend of α-helix stability described in Option A.

Qwen3 VL 30B Instruct (VoT): B ✘
... the overall shape suggests a significant increase in β-sheet content. ... Therefore, Panel (h) indicates a shift towards more β-sheet formation after 24h.

Error Analysis:
[Spectral Interpretation Error] The model incorrectly interpreted Panel (h) as showing increased β-sheet formation, thus validating Option B and failing to select the correct answer (A). This misreading of the CD spectrum data led to an erroneous conclusion about structural changes.

Qwen3 VL 30B Instruct (VIC) : D ✘
The best summary is Option D, as it aligns with the observed loss of β-sheet content and stable α-helix content over time in both the CD spectrum (Panel h) and secondary structure analysis (Panel i).

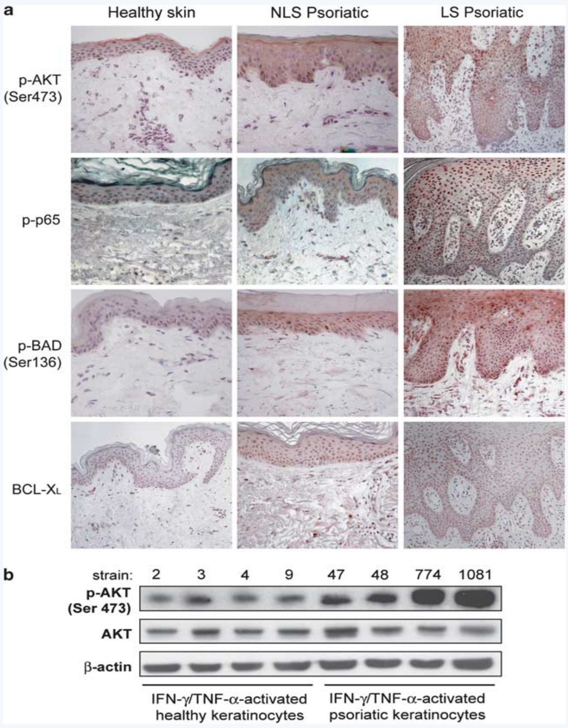
Error Analysis:
[Comparative Data Misinterpretation] The model acknowledged the stability of α-helix content (which supports the correct answer A) but incorrectly prioritized and selected Option D. It erroneously concluded that the β-sheet loss rate was significantly higher, a comparison the standard answer rejects as an overstatement of the data in Panel (i).

Qwen3 VL 30B Instruct (Cantor) : A ✘
... Option A: ... The bar chart in Panel (i) shows a clear decrease in α-helix content over time ... Option D: ... This is correct. The stacked bar chart shows that β-sheet content decreases more than α-helix content ... Therefore, the best answer is D.

Error Analysis:
[Graph Trend Misinterpretation] The model incorrectly rejected Option A, perceiving a decrease in α-Helix content where the data shows stability. This led it to erroneously select Option D, misinterpreting the relative loss rates which the standard answer characterizes as an overstatement.

Figure 33: MCoT case study of Cross-Panel Relation Understanding question in trend analysis sub-task.

Qualitative Reasoning (1/2)



Question: **[Reasoning_2010030854_0_1]**

Based on the immunohistochemistry data in panel (a) and the protein lysate analysis in panel (b), which of the following statements accurately describes the effect of IFN-g/TNF-a treatment on psoriatic keratinocytes?

A. Phospho-AKT levels in healthy keratinocytes are equal to those in psoriatic keratinocytes post-treatment.

B. Psoriatic keratinocytes show a higher phospho-AKT expression after IFN-g/TNF-a treatment compared to unstimulated healthy skin.

C. The anti-apoptotic BCL-XL is downregulated in LS psoriatic skin after cytokine treatment.

D. Phospho-p65 and phospho-BAD are only upregulated in NLS psoriatic skin, not in LS.

E. In IFN-g/TNF-a-activated psoriatic keratinocytes, total AKT levels are lower than phospho-AKT levels under treatment.

Answer: B

Explanation:

The images and text indicate that phospho-AKT levels are higher in LS psoriatic skin compared to NLS and healthy skin, and are consistent with increased expression in cultured psoriatic keratinocytes post IFN-g/TNF-a treatment as shown in the intensity of staining (panel a) and band thickness (panel b).

Qwen3 VL 30B Instruct(Baseline): B ✔

In panel (a), the immunohistochemistry data for p-AKT (Ser473) shows that psoriatic keratinocytes ... have higher staining intensity compared to healthy skin, indicating increased phospho-AKT expression ... This supports option B ... Option A is incorrect because phospho-AKT levels are not equal ... Option C is incorrect because BCL-XL is upregulated, not downregulated ... Option D is incorrect because both phospho-p65 and phospho-BAD are upregulated ... Option E is incorrect because total AKT levels are not lower than phospho-AKT levels ... Therefore, the correct answer is B.

Qualitative Reasoning (2/2)

Qwen3 VL 30B Instruct (DDCoT): E ✘

- **Sub-questions:** ... What does panel (a) show ...? ... Is BCL-XL downregulated ...? ... Are phospho-p65 and phospho-BAD upregulated ...?
- **Sub-answers:** ... Panel (a) shows that phospho-AKT levels are higher in psoriatic keratinocytes ... Panel (b) shows ... phospho-AKT levels are higher than total AKT levels. ... BCL-XL is not downregulated ... Phospho-p65 and phospho-BAD are upregulated in both ... total AKT levels are lower ...

Error Analysis:
[Antibody Normalization Failure] The model erroneously validated Option E by comparing absolute band intensities of Total vs. Phospho-AKT, ignoring antibody affinity differences and the biological impossibility of a phosphorylated fraction exceeding the total protein pool.

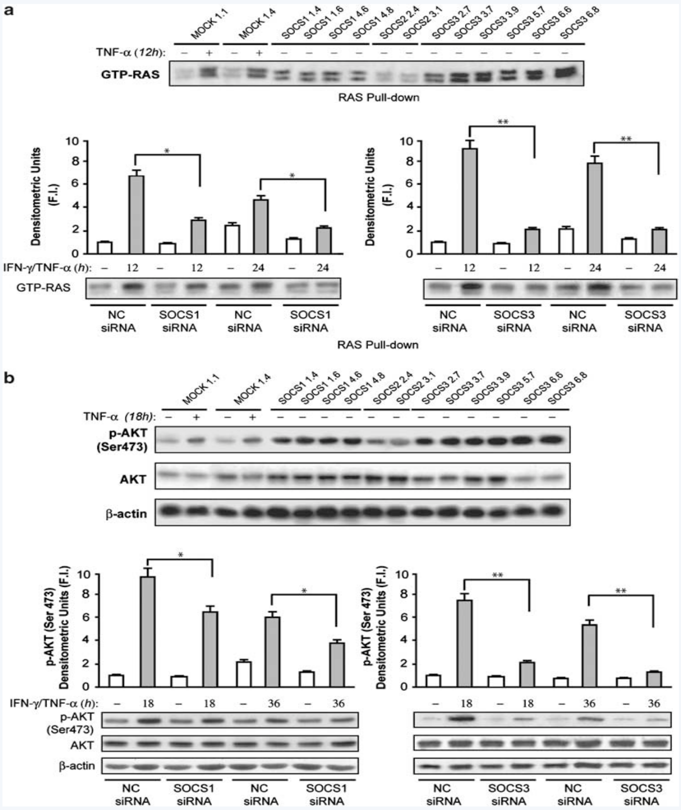
Qwen3 VL 30B Instruct (VoT): B ✔
... From panel (b), the p-AKT band intensity in psoriatic keratinocytes ... is visibly stronger ... Thus, Option A is incorrect. ... In panel (a), the p-AKT staining in LS psoriatic skin ... is significantly stronger ... indicating higher phospho-AKT expression ... Thus, Option B is correct. ... BCL-XL staining ... indicates upregulation, not downregulation. ... Thus, Option C is incorrect. ... p-p65 and p-BAD staining are present in both NLS and LS psoriatic skin ... Thus, Option D is incorrect. ... total AKT levels are not necessarily lower ... Thus, Option E is incorrect. ... Therefore, the correct answer is B.

Qwen3 VL 30B Instruct (VIC) : B ✔
Based on the analysis ... the correct statement ... is: B. Psoriatic keratinocytes show a higher phospho-AKT expression after IFN-g/TNF-a treatment compared to unstimulated healthy skin. ... Panel (a) ... p-AKT staining is strongly upregulated in LS Psoriatic (lesional) skin ... This demonstrates that the inflammatory environment ... leads to high phospho-AKT expression. ... Panel (b) ... The p-AKT (Ser473) band is strong and prominent in the psoriatic keratinocyte lanes. ... This directly shows that upon stimulation ... psoriatic keratinocytes exhibit significantly higher levels of phosphorylated AKT compared to healthy keratinocytes. ... Therefore, statement B is the most accurate description.

Qwen3 VL 30B Instruct (Cantor) : B ✔
Rationale: ... Panel (a) shows protein expression in tissue sections. ... For p-AKT (Ser473), staining intensity is very low in healthy skin ... and very high in LS psoriatic. ... Panel (b) ... shows that p-AKT levels are dramatically higher in psoriatic keratinocytes ... than in healthy keratinocytes ... after treatment. ... Thus, Option A is incorrect. ... Option B is correct. ... Panel (b) shows that p-AKT levels in IFN-γ/TNF-α-activated psoriatic keratinocytes ... are extremely high. ... Panel (a) shows that p-AKT levels in unstimulated healthy skin are very low. ... Option C is incorrect. ... BCL-XL expression is strongly upregulated ... Option D is incorrect. ... p-p65 and p-BAD ... are upregulated in both NLS and LS psoriatic skin ... Option E is incorrect. ... The band for total AKT is strong and consistent ... total AKT is not downregulated ... Answer:B

Figure 34: MCoT case study of Qualitative Reasoning question.

Quantitative Reasoning (1/2)



Question: **[Reasoning_2010041549_2_1]**

Based on Figures 3a and 3b, calculate the percentage decrease in RAS activity and AKT phosphorylation when SOCS1 and SOCS3 are silenced in keratinocytes treated with IFN-g/TNF-a for 12h and 18h, compared to NC-transfected cells.

A. RAS activity decreases by 75% and AKT phosphorylation by 70%.
B. RAS activity decreases by 25% and AKT phosphorylation by 30%.
C. RAS activity decreases by 87.5% and AKT phosphorylation by 50%.
D. RAS activity decreases by 50% and AKT phosphorylation cannot be calculated.
E. Cannot be determined due to incomplete data.

Answer: E

Explanation:

Upon examining the figures, the exact F.I. values necessary for precise percentage calculations are not consistently shown for each condition. Although trends demonstrate substantial decreases in both RAS activity and AKT phosphorylation upon SOCS silencing, complete numerical values across all conditions are not given, requiring assumptions which the problem statement prohibits.

Qwen3 VL 30B Instruct(Baseline): C ✘

... The value of 87.5% is a close approximation of this calculated decrease. ... The value of 50% matches the calculated decrease. ... This corresponds to option C.

Error Analysis:

[Unjustified Estimation] The model incorrectly derived precise percentages from visual estimates, ignoring the lack of explicit numerical data which necessitated selecting Option E (Cannot be determined).

Quantitative Reasoning (2/2)

Qwen3 VL 30B Instruct (DDCoT): C ✘

- **Sub-questions:** ... What is the RAS activity and AKT phosphorylation in NC-transfected vs SOCS-silenced cells...?
- **Sub-answers:** ... RAS activity is 6.5 (NC) vs 2 (silenced)... AKT phosphorylation is 9.5 vs 5... The percentage decrease is...

Error Analysis:
[Unwarranted Estimation] The model incorrectly forced a calculation based on visual estimates, failing to recognize that the lack of explicit numerical data required selecting Option E (Cannot be determined).

Qwen3 VL 30B Instruct (VoT): C ✘
... The decrease from 8 F.I. (NC) to 1 F.I. (SOCS3) is a reduction of (8-1)/8 = 7/8 = 87.5%. ... The decrease from 8 F.I. (NC) to 4 F.I. (SOCS1 or SOCS3) is a reduction of (8-4)/8 = 4/8 = 50%. ... corresponds to option C.

Error Analysis:
[Data Hallucination] The model hallucinates specific numerical values from the figures to force a definitive calculation, failing to recognize the data insufficiency that renders the correct answer as "Cannot be determined."

Qwen3 VL 30B Instruct (VIC) : A ✘
... Figure 3a: RAS Activity ... Average RAS activity decrease: 75% ... Figure 3b: AKT Phosphorylation ... Average AKT phosphorylation decrease: 70% ... Conclusion: RAS activity decreases by approximately 75% ... AKT phosphorylation decreases by approximately 70%. Thus, the correct option is: A ...

Error Analysis:
[Data Interpretation Error] The answer incorrectly claims data insufficiency, failing to recognize that the figures allow for estimating values to calculate specific percentage decreases (RAS -75%, AKT -70%) corresponding to option A.

Qwen3 VL 30B Instruct (Cantor) : A ✘
... Percentage decrease = [(7.5 - 2.5) / 7.5] × 100 = (5.0 / 7.5) × 100 = 66.7%. This value is closest to 70% ... Percentage decrease = [(9.5 - 5.5) / 9.5] × 100 = (4.0 / 9.5) × 100 = 42.1%. ... Therefore, the most accurate option is A, as it provides the closest values to the calculated percentages. Answer:A

Error Analysis:
[Data Interpretation Error] The answer incorrectly assumes the figures provide sufficient numerical data for precise calculations, contradicting the standard explanation that the data is incomplete. Furthermore, the model's own estimated values (e.g., 42% for AKT) do not align with the selected option (70%), confirming the error in forcing a numerical conclusion.

Figure 35: MCoT case study of Quantitative Reasoning question.