OMNISCIENCE++: A HIERARCHICAL CORPUS FOR EVIDENCE-GROUNDED SCIENTIFIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific conclusions often depend on observations spread across multiple panels, measurements, and experimental conditions. Understanding these figures requires identifying the relevant evidence and determining how the observations support a conclusion. We introduce OmniScience++, a hierarchical corpus for evidence-grounded scientific reasoning, organizing 1.52 million figures into a figure–panel–subfigure structure with 10.02 million localized atomic-subfigure records. An evidence-first data engine turns this corpus into questions with explicit visual support. The engine selects observations and composes a reasoning program before generating the question and answer. Constraints on entities, conditions, and measurements guide evidence composition, while source regions and intermediate operations remain traceable. This construction supports scientific tasks such as comparing experimental outcomes and using an intermediate finding to select evidence for a subsequent inference. We evaluate these capabilities with a 600-question human-reviewed benchmark covering quantitative analysis, visual structure, and experimental relations. Across eight multimodal models, answer accuracy and evidence localization produce different rankings. Providing reference evidence crops improves answer accuracy by 2.36-17.70 percentage points across four models. By enabling the construction and evaluation of scientific reasoning tasks with explicit visual support, OmniScience++ provides a foundation for future AI for scientific research that connects conclusions to their underlying experimental evidence.

1 INTRODUCTION

Scientific figures are compact arguments. A microscopy image may show a phenotypic change, a chart may quantify it, and another assay may test the same contrast. Recovering their support requires a resource that preserves local observations and their experimental context. OmniScience++ provides this foundation across diverse scientific visual types (Figure 1), then uses it to construct evidence-grounded questions. A correct final answer alone does not reveal whether the supporting observations were identified and composed correctly.

In compound figures, similar-looking regions may represent different samples, conditions, assays, or views. A model could answer from question regularities or domain knowledge without recovering this support, or locate evidence but fail to combine it. Joint evaluation distinguishes these outputs, without identifying internal reasoning.

Existing resources provide figures, captions, paper context, and fine-grained regions (Subramanian et al., 2020; Hsu et al., 2021; Li et al., 2024; Tao et al., 2026; Yuan et al., 2026). QA and grounding benchmarks test scientific interpretation and support retrieval (Pramanick et al., 2024; Ding et al., 2026; Yu et al., 2025). A construction challenge remains: converting these collections into questions with explicit, compositional support. Generating a question, answer, and rationale jointly leaves their correspondence dependent on one generative process, without an independently specified evidence composition.

We address this challenge through *evidence-first construction*. OmniScience++ extracts atomic visual subfigures from OmniScience (Tao et al., 2026) with source coordinates and semantics. The engine forms localized observations, aligns entities and conditions, and samples compatible evidence and a program before generating QA. For numerical families, it computes orderings, extrema,

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

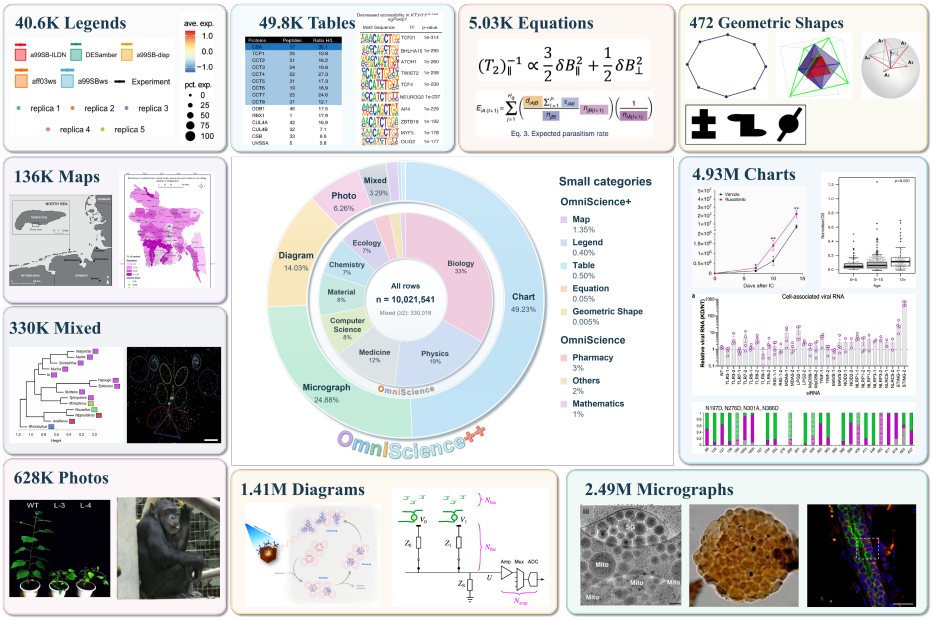

Figure 1: **A broad, fine-grained scientific evidence resource.** OmniScience++ extends OmniScience's source coverage with 10.02M atomic records across charts, microscopy, diagrams, photographs, and other visual types. Rounded counts and the Mixed display group describe the resource inventory; the 600-question benchmark is a separate curated application.

or ranking agreement before verbalization. Generated text preserves the selected identifiers and program structure. This separates evidence composition from language generation while remaining dependent on extracted observations.

We instantiate the engine in 600 human-reviewed questions across three scientific task axes. Eight-model evaluation compares answers with localization; a visual oracle tests whether reference regions help answer the same questions.

Our contributions are threefold. **(1) A fine-grained scientific resource:** 10,021,541 atomic-subfigure records connected to source figures and paper metadata, built with a detector, semantic recursion gating, and residual-region recovery. **(2) An evidence-first data engine:** a reverse sampler that composes localized observations under experimental compatibility constraints and records the selected evidence, operations, and generated QA. **(3) A diagnostic benchmark and analysis:** 600 human-reviewed questions, eight-model answer and grounding results, and visual-oracle results for four models. Category-level analysis distinguishes large relational-task gains for some models from persistent visual-structure difficulty. Extraction ablations and annotation audits characterize the supporting resource.

2 Related Work

Scientific image resources. MedICaT, SciCap, and Multimodal ArXiv align scientific visuals with captions and document context (Subramanian et al., 2020; Hsu et al., 2021; Li et al., 2024). EXSCLAIM! and FigEx study extracting scientific images and corresponding descriptions (Schwenker et al., 2023; Song et al., 2025). OmniScience supplies broad figure and panel coverage (Tao et al., 2026), while Panel2Patch constructs supervision at finer visual scales (Yuan et al., 2026). OmniScience++ builds on this line of work. Its distinguishing emphasis is the connection from atomic regions to condition-aware evidence compositions and grounded QA, rather than region granularity alone.

Scientific reasoning and evidence grounding. SPIQA addresses scientific-paper QA; CharXiv tests realistic charts; SciVQA and AnaFig examine broader scientific figures; and SPUR studies

experimental-image reasoning (Pramanick et al., 2024; Wang et al., 2024; Borisova et al., 2025; Yue et al., 2025b; Ding et al., 2026). RefChartQA and SciEGQA connect answers to visual support (Vogel et al., 2025; Yu et al., 2025). Our contribution connects these evaluation needs to evidence-first construction over a large atomic-figure resource. Table 1 compares the task settings and reference structures.

Structured generation. Program-structured visual reasoning includes CLEVR (Johnson et al., 2017); SaGe samples scene-graph reasoning traces for multimodal post-training (Yang et al., 2026). Our engine adapts structured composition to experimental figures, where valid comparisons require aligning entity, condition, measurement, and assay or sample context. Hierarchical containment locates observations, while experimental compatibility constrains which observations can be composed. Our evaluation concerns data construction and diagnostics, rather than post-training.

Table 1: Comparison with scientific QA benchmarks, visual resources, and evidence-grounding benchmarks.

Benchmark	Scale	Task		Grounding			Relations		Human review
		F+Q	Open	Region	Multi	Hier.	Cross	Path	
Scientific question answering									
FigureQA (2018)	>1M QA / >100K F	✓	–	△	△	–	–	–	–
CharXiv (2024)	2,323 F	✓	✓	–	–	–	–	–	✓
SPIQA (2024)	∼270K QA	–	✓	–	△	–	△	–	△
SciVQA (2025)	21K QA / 3K F	✓	✓	–	–	–	–	–	△
AnaFig (2025b)	2K samples	△	✓	–	–	–	△	–	✓
SPUR (2026)	4,264 MCQ / 1,084 F	✓	–	–	△	△	✓	–	✓
Scientific visual resources									
ACL-Fig (2023)	112K F / 1.7K manual	*n/a*	*n/a*	*n/a*	*n/a*	–	*n/a*	*n/a*	△
OmniScience (2026)	1.52M F / 4.46M P	*n/a*	*n/a*	*n/a*	*n/a*	✓	*n/a*	*n/a*	△
Panel2Patch (2026)	364K F / 1.30M P / 619K R	*n/a*	*n/a*	*n/a*	*n/a*	✓	*n/a*	*n/a*	–
Evidence-grounding benchmarks									
RefChartQA (2025)	73.7K pairs	✓	✓	✓	✓	–	–	–	△
SciEGQA (2025)	1.6K human / >30K auto	–	✓	✓	✓	△	–	–	✓
GroundingDocQA (2026)	2M QA / 200K D / 5K test	–	✓	✓	✓	✓	–	–	✓
OmniScience++(Ours)	1.5M F / 10M R / 600 QA	✓	✓	✓	✓	✓	✓	✓	✓

Task: F+Q = figure and question without paper/context input; Open = free-form answers beyond binary or multiple-choice labels. **Grounding:** Region = answer-region annotation; Multi = multi-region support; Hier. = figure or document hierarchy. **Relations:** Cross = cross-panel relations; Path = typed relational paths. **Scale:** F/P/R/D = figures/panels/region pairs/documents. ✓ explicit target or annotation; △ partial or indirect support; – not designated as an explicit target; *n/a* not applicable. Human review may cover only a subset.

3 OmniScience++: A Fine-Grained Evidence Resource

3.1 Atomic regions and source organization

OmniScience++ extends the figure and panel records of OmniScience with an atomic-subfigure inventory. Here, *atomic* denotes a meaningful terminal visual unit retained by the parsing pipeline. It is an operational granularity, not a claim that every crop is semantically indivisible. A panel can contain multiple such units, for example images at different magnifications or condition-specific elements arranged in a grid.

Each atomic record stores a bounding box in source-figure coordinates, source-figure and source-paper identifiers, one or more visual types, two to five semantic tags, and a generated local caption. The controlled visual vocabulary contains Chart, Micrograph, Diagram, Photo, Map, Legend, Table, Equation, and Geometric Shape. Tags describe content, scientific domain, or imaging technique. The database establishes figure–subfigure mappings that connect observations to their source figure.

This mapping preserves the correspondence between each local region and its source figure, allowing every decomposed subfigure to be traced back to the complete compound figure. It describes only source provenance and does not encode experimental or semantic relations among observations; those relations are constructed separately by the data engine.

Table 2: **Resource inventory and distinct evaluation populations.** Counts refer to records, not independently validated examples. Eval-1K and the QA benchmark serve different purposes.

Population	Count	Role
Source figures	1,520,179	Full scientific figures
Normalized DOIs	251,181	Source-paper associations
Inherited panel records	4,464,199	Intermediate visual context
Atomic-subfigure records	10,021,541	Localized visual units
Nonempty atomic captions	9,967,257	Available generated descriptions
Detector refinement figures	5,000	Human-in-the-loop training
Eval-1K figures / reference boxes	1,000 / 12,785	Extraction evaluation
Human-reviewed QA items	600	Evidence-grounded reasoning

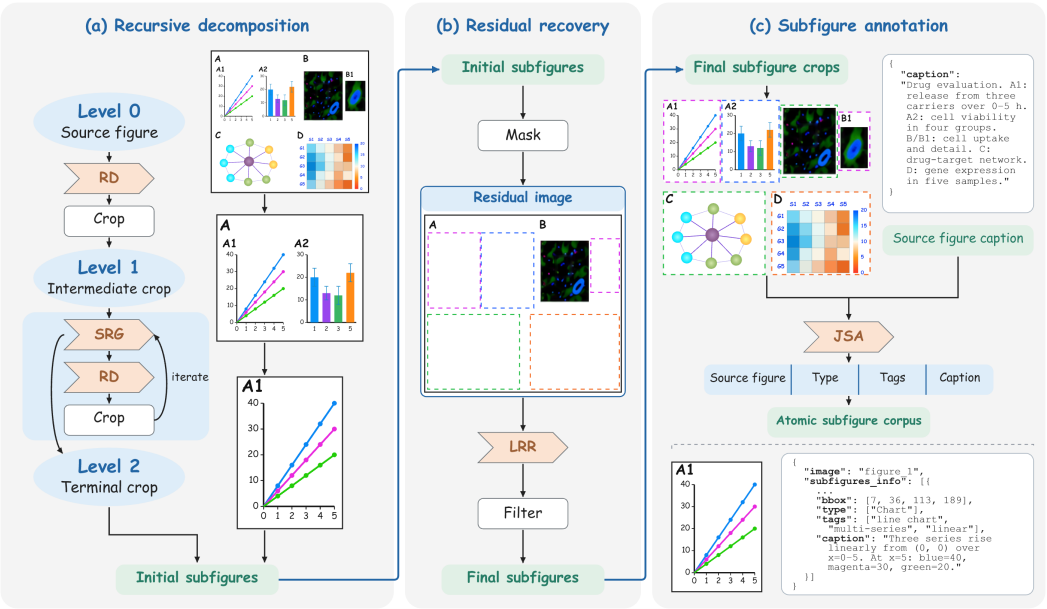

Figure 2: **Atomic-subfigure construction pipeline(ASCP).** Recursive decomposer and semantic recursion gating produce retained units; residual recovery supplements missed regions; joint semantic annotation supplies local descriptions. The displayed levels illustrate recursive processing, not a universal limit of two splits. This resource-construction pipeline precedes the QA engine in Figure 3.

3.2 Atomic-subfigure construction pipeline

The atomic-subfigure construction pipeline (ASCP) identifies small, meaningful visual units while avoiding excessive fragmentation. As shown in Figure 2, it comprises three stages—recursive decomposition, residual recovery, and semantic annotation—implemented by four components: a recursive decomposer (RD), semantic recursion gating (SRG), local residual recovery (LRR), and joint semantic annotation (JSA).

During recursive decomposition, the **RD**, implemented with AS-YOLO, extracts candidate regions from the source figure and eligible crops. AS-YOLO is initialized from inherited panel annotations and refined through human correction on 5,000 figures. **SRG** determines whether each region is

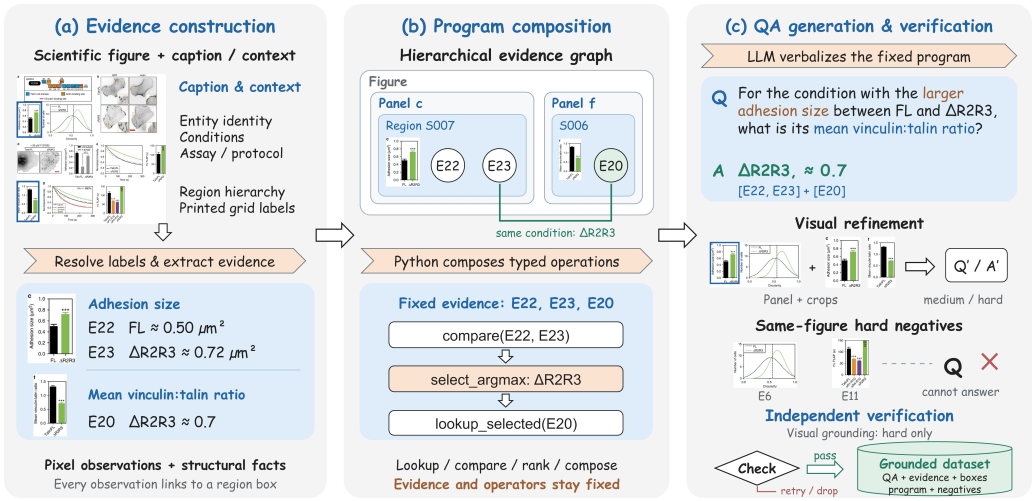

Figure 3: **Evidence-first construction.** The engine selects evidence and a program before generating QA. Adhesion-size observations select a condition; linked evidence supplies its vinculin:talin ratio. Identifiers preserve source-region provenance. This example illustrates construction, not measured solver reasoning.

meaningful and requires further subdivision. A parent is replaced only when RD produces at least two valid children, and recursion proceeds for at most ten rounds.

During residual recovery, **LRR** masks the retained regions and identifies meaningful content missed by recursive decomposition. Its proposals are mapped to source-figure coordinates and filtered before being added to the retained set.

During semantic annotation, **JSA** assigns visual types, semantic tags, and local captions to the final regions.

The resulting atomic-subfigure records provide localized visual and semantic information for the subsequent data engine. Additional implementation details and resource analyses are provided in Appendices A and G.

4 Evidence-First Data Engine

The engine constructs examples in the order

$$\text{regions} \longrightarrow \text{observations and relations} \longrightarrow (S, \pi) \longrightarrow (q, a, \mathcal{B}), \tag{1}$$

where S is a selected observation set, π a reasoning program, q a question, a its reference answer, and $\mathcal{B}$ the supporting source-coordinate boxes. The reverse sampler is the organizing mechanism: the evidence composition is chosen before the language model formulates the question.

4.1 Localized observations and complementary graphs

An atomic observation is represented as

$$e_i = (\mathrm{id}_i, b_i, s_i, p_i, t_i, \mathbf{z}_i, v_i), \tag{2}$$

where b_i is visual support, s_i, p_i identify the subfigure and parent panel, t_i is the observation type, and v_i is its value or description. Context $\mathbf{z}_i$ includes entity, condition, attribute, unit, series, sample, assay, and view. A vision-language model extracts observations using crops and caption/context hints; grid reading resolves condition labels in qualifying panels.

The **hierarchy graph** records source containment. A **semantic relation graph** connects experimentally compatible observations: the strict contrast predicate requires different conditions with

compatible entities, observation types, and attributes. Further checks constrain views, series, control/treatment, samples, and assays; other candidates align conditions across measurements. Shared figure membership alone is insufficient. These rules operate on extracted metadata and do not certify causal relations.

4.2 Reverse sampling under program constraints

The sampler chooses observations and an admissible program family before QA generation:

$$(S, \pi) \sim \mathrm{Sampler}(\mathcal{E}, G_H, G_R) \quad \text{subject to} \quad C_{\text{type}} \wedge C_{\text{condition}} \wedge C_{\text{relation}} \wedge C_{\text{program}}. \tag{3}$$

These predicates summarize code-level eligibility checks. They constrain which evidence may be combined, rather than certify the correctness of the extracted observations. Direct questions use a lookup; comparative questions use multiple lookups followed by comparison, trend analysis, or ranking. More compositional families include:

- **Contrast agreement:** compare two shared conditions within each of two measurements, then determine whether the resulting orderings agree.
- **Rank agreement:** rank three or four shared conditions within each measurement and compare the rankings.
- **Select then lookup:** identify an extremal condition in one measurement and retrieve a different attribute for the same condition from another evidence source.
- **Cross-assay composition:** combine aligned contrasts, or use an explicitly represented structural correspondence to connect observations.

For eligible numerical observations, extrema, contrast directions, and rankings are computed before verbalization, providing the generator with a selected target or intermediate result. Qualitative outputs remain model-dependent. Operation counts include lookups and do not measure independent scientific deductions.

Why generate from evidence? The benefit is controllable composition and inspectable provenance: evidence selection and intended operations are specified before verbalization. Structured generation can reduce dependence on a generator's preferred question patterns, and record the reasoning steps and evidence supporting the conclusions.

4.3 Verbalization, refinement, and verification

Given (S, π), the generator produces QA and reference steps with tags linked to stored evidence boxes. Refinement can repair wording, answers, and step descriptions while retaining the selected evidence and program inputs. Validation checks tag completeness, agreement between refined plain and grounded answers, and family-specific numerical consistency. Coordinates are copied from evidence records.

Verification combines schema/program checks, a separate model judge for answerability and evidence sufficiency, and default visual inspection for hard items. Same-figure distractor observations screen alternative support. The benchmark selects 600 human-reviewed items after construction; Appendix A details the implementation.

Construction trace. Figure 3 compares adhesion sizes for FL and ΔR2R3, selects the larger condition, and retrieves its vinculin:talin ratio from another panel. The trace identifies support for both the intermediate decision and final lookup. It is inspectable reference supervision, not a solver's internal chain of thought; training effects remain unevaluated.

5 A Benchmark for Evidence-Grounded Scientific QA

5.1 Task and coverage

The benchmark contains 600 human-reviewed questions: 200 per axis, each with 100 Basic and 100 Advanced items. Basic items emphasize local interpretation; Advanced items emphasize evidence

composition, without asserting calibrated difficulty. Given a figure and question, models produce an answer and supporting source-image regions. References include QA, evidence, and the construction trace. This curated diagnostic set is distinct from the full resource. Source counts and examples appear in Appendix C and Appendix D.

The three coverage axes describe different uses of evidence in experimental papers. **Quantitative analysis** includes reading values or ranges, comparing measurements, and ranking conditions. **Visual structure** concerns visible patterns, morphology, bands, and spatial organization. **Experimental relations** connects observations across conditions, panels, or assays. An advanced family first compares two observations to obtain an intermediate conclusion c, then combines that conclusion with a third observation:

$$c = f(e_1, e_2), \qquad a = g(c, e_3). \tag{4}$$

For example, a comparison selects a condition whose second property must be retrieved from another panel. The intended dependency is encoded by the program; it is not implied by merely presenting several regions. Three observations can occupy two unique boxes, so observation count and crop count are distinct. These content axes are separate from program families and operation counts.

Table 1 compares all task and resource families using common dimensions. Our combination provides open answers, localized multi-region evidence, hierarchical provenance, and typed paths linking intermediate decisions to subsequent observations. Figure-mining resources are distinguished from QA benchmarks, and indirect support from explicit evidence supervision.

5.2 Answer, grounding, and visual-oracle evaluation

We use five answer types with three scoring mechanisms: normalized exact match for yes/no, comparison/trend, and entity/label answers; numerical matching with frozen tolerances; and binary model judging for free text. Every required item must pass and the final answer must be consistent. Appendix C.2 specifies the rules. The evaluation excludes refusals and unresolved responses. For retained items $R_{m,c}$ of model m in condition c, *conditional accuracy* is

$$\mathrm{Acc}_{m,c} = \frac{\sum_{i \in R_{m,c}} \mathbf{1}[\hat{a}_i \text{ is correct}]}{|R_{m,c}|} \times 100\%. \tag{5}$$

Questions have equal weight; Overall pools retained items across categories. Model-specific retained sets differ from a fixed 600-item denominator. Grounding F1 averages per-question box-set F1 at IoU 0.50, 0.75, and 0.90. Empty predictions score zero; invalid and duplicate boxes can count as unmatched predictions. Appendix C specifies matching and averaging. Accuracy minus grounding F1 is not a joint error rate.

The **visual oracle** supplies the unmarked full figure, all deduplicated reference-evidence crops, and the question. Crops have no padding or annotations; the prompt identifies relevant evidence but provides no reference answer, coordinates, caption, or context. Models still predict full-figure coordinates. The same model is compared on matched questions. Crops may aid both selection and visual reading, so the intervention does not isolate localization. Appendix C records crop counts and image-integrity checks.

6 Experiments

We evaluate resource construction on Eval-1K, then answer/grounding performance and crop assistance on the separate 600-question benchmark. QA results use the response filtering in Equation 5.

6.1 Resource construction quality

We construct Eval-1K by sampling source figures not used for detector refinement and evaluate region extraction against manually calibrated reference boxes. We compare a panel-trained YOLO26x baseline with AS-YOLO and examine the contributions of recursive decomposer (RD), semantic recursion gating (SRG), local residual recovery (LRR), and global set regeneration (GSR). Missing rate measures extraction completeness, while macro F1 summarizes per-figure detection quality.

Table 3: **Extraction ablation on Eval-1K.** Missing rate (Miss) and macro F1 are percentages; subscripts denote IoU thresholds. API costs cover this 1,000-figure evaluation, excluding detector training, local compute, and semantic annotation.

Detector	Configuration	$\text{Miss}_{.50}\downarrow$	$\text{Miss}_{.90}\downarrow$	$\text{F1}_{.50}\uparrow$	$\text{F1}_{.90}\uparrow$	API cost ($)
YOLO	RD	40.1	46.7	77.7	69.5	0.000
	+ SRG	38.1	45.4	79.0	70.4	0.187
	+ SRG + LRR	37.6	45.2	79.0	70.2	20.693
	+ SRG + GSR	22.3	36.0	87.5	73.5	106.723
AS-YOLO	RD	12.7	19.9	91.4	83.6	0.000
	+ SRG	9.9	17.9	92.8	**84.5**	0.175
	+ SRG + LRR	**9.2**	**17.6**	92.8	84.2	22.997
	+ SRG + GSR	9.3	20.5	**92.9**	81.0	102.805

Table 3 shows that the additional atomic-region supervision provides the largest performance improvement, with AS-YOLO consistently outperforming the baseline detector across pipeline configurations. On top of AS-YOLO, SRG improves extraction completeness, while LRR further reduces omissions at the cost of a small decrease in localization precision under the strict IoU criterion. Because resource construction prioritizes recovering as many meaningful atomic subfigures as possible, this limited loss in precision is acceptable in our setting. Compared with GSR, LRR also provides a more favorable balance between strict localization quality and API cost. Together, AS-YOLO, SRG, and LRR provide a construction configuration that balances region completeness, localization precision, and computational cost.

We further conduct a quality audit of the automatically extracted regions using Qwen and GPT evaluators. The audit considers unit validity, boundary quality, factual reliability, and information coverage. Both evaluators indicate that the generated descriptions have generally strong factual reliability but comparatively lower information coverage. This suggests that local descriptions may not capture all visual information within a region; constructing QA data therefore still requires verification against the original image and its corresponding regions. Appendix B reports the complete score distributions, and Appendix E presents the caption-conditioned localization diagnostic.

6.2 Answer and grounding rankings differ

We compare answer accuracy and region-level F1 overall and within each scientific task axis.

Table 4: **Overall and category-level answer and grounding results.** Acc is conditional answer accuracy; F1 is mean per-question grounding F1. Overall F1 uses IoU .50, .75, and .90; category F1 uses IoU .75. All values are percentages on each model's retained responses after excluding refusals and unresolved cases. Bold marks the column maximum.

	Overall				Quantitative		Visual structure		Exp. relations	
Model	Acc	$\text{F1}_{.50}$	$\text{F1}_{.75}$	$\text{F1}_{.90}$	Acc	F1	Acc	F1	Acc	F1
Claude Opus 5.5	**88.52**	63.85	41.31	13.51	**95.89**	51.69	**69.79**	25.45	94.78	41.36
Gemini 3.8 Flash	82.81	44.73	40.05	26.23	91.07	45.63	52.25	23.33	**95.09**	45.68
GPT 6 Sol	80.91	**84.29**	**75.02**	**51.60**	93.53	**80.24**	47.50	**59.47**	91.81	**80.74**
Kimi-k3	74.25	57.69	32.23	7.38	84.62	42.60	43.65	17.86	86.55	32.56
GLM-5.3-flash	71.46	42.72	17.51	2.05	73.21	18.55	47.58	11.85	87.43	20.66
Qwen3.8-max	68.61	13.32	9.75	2.24	78.31	14.66	51.22	4.63	71.97	8.58
DeepSeek-V4.1-Flash	50.91	27.19	4.70	0.10	63.48	6.74	22.46	1.93	60.45	4.80
GPT-4o	32.62	9.42	0.29	0.00	44.21	0.00	16.43	0.00	32.97	0.82

Table 4 shows that answer accuracy and evidence grounding yield different model rankings on each model's response-filtered set of retained examples. Claude Opus 5.5 attains the highest overall conditional answer accuracy, whereas GPT 6 Sol achieves the strongest grounding performance across all IoU thresholds. Moreover, grounding F1 decreases as the IoU criterion becomes stricter, indi-

cating that identifying broadly relevant evidence regions does not necessarily translate into precise localization.

This divergence between answer accuracy and evidence grounding is further reflected in two complementary category-level patterns. First, visual-structure questions have the lowest answer accuracy and grounding performance across all models, indicating that understanding visual patterns, morphology, and spatial organization remains a major challenge for current models. Second, some models produce relatively accurate answers to experimental-relations questions without successfully localizing the corresponding supporting evidence, further demonstrating that answer generation and evidence grounding reflect distinct capabilities. Reporting both answer accuracy and grounding F1 therefore provides a more complete characterization of model performance.

Table 5: **Overall and category-level accuracy with oracle evidence crops.** The same model is compared on matched questions. Base and Oracle are conditional accuracies (%); Δ is the reported change in percentage points. Category baseline values appear in Table 4. Rounding can alter a displayed difference by 0.01 pp.

	Overall			Quantitative		Visual structure		Exp. relations	
Model	Base	Oracle	Δ	Oracle	Δ	Oracle	Δ	Oracle	Δ
Claude Opus 5.5*	90.54	92.91	+2.36	96.75	-0.81	81.16	+8.70	96.15	+1.92
GPT-4o	32.62	45.85	+13.23	56.52	+12.31	20.29	+3.86	54.35	+21.38
Qwen3.8-max	68.61	86.31	+17.70	90.68	+12.37	68.29	+17.07	95.27	+23.29
Kimi-k3	74.25	76.94	+2.69	88.41	+3.80	47.97	+4.32	86.44	-0.11

* Due to Anthropic's policy, some questions refused in either setting are excluded from both.

6.3 Performance with reference visual crops

Providing reference evidence crops improves overall answer accuracy for all four evaluated models (Table 5). The models answer the same questions with the full figure still available, supplemented by crops of the supporting regions. These results show that making the relevant visual evidence explicit can help models arrive at the correct answer. Evidence grounding therefore has value for the answering process itself, by connecting a scientific question to the observations needed to answer it.

7 Limitations and Conclusion

Limitations. Support remains limited for more open-ended scientific processes, such as formulating hypotheses, explaining conflicting observations, and designing follow-up experiments. Expanding the reasoning programs and their possible compositions is an important direction for supporting a broader range of scientific reasoning tasks.

Conclusion. We introduced OmniScience++, a hierarchical corpus of 10.02 million atomic-subfigure records derived from 1.52 million scientific figures. By linking localized regions, semantic annotations, and source metadata across figure–panel–subfigure hierarchies, it provides a reusable resource for fine-grained scientific image understanding. Our evidence-first engine connects this corpus to question generation through compatible observations and explicit reasoning programs. A 600-question human-reviewed benchmark reveals different answer and grounding rankings across eight models. Together, these contributions support research on how models identify, combine, and reason over scientific visual evidence.

AI USE STATEMENT

Data construction and evaluation. We have used Qwen for SRG and JSA and GPT for LRR (Section 3.2). These uses have produced semantic refinement decisions, additional region proposals, and visual types, tags, and local captions from existing scientific figures. OmniScience redescriptions have provided context for annotation and caption assessment. Qwen and GPT have provided automated region and caption judgments (Appendix B.2); Qwen has also generated grounding questions and screened caption uniqueness (Appendix E). Grounding reference boxes have been taken from existing annotations, not generated as answers. The models in Table 9 have been evaluated as grounding systems, separately from manuscript-writing assistance.

Research and manuscript assistance. We have used ChatGPT to assist with developing the methodological framing and research questions, proposing and refining evaluation criteria and experimental controls, and interpreting reported results. It has also assisted with English drafting and revision, translation of Chinese research notes and scoring guidelines, manuscript organization, literature search and summarization, LaTeX and supporting-script edits, table formatting, and adjustments to figure annotations and layout. This assistance has extended beyond language polishing.

Checks and responsibility. We have reviewed and revised the AI-assisted manuscript and supporting preparation artifacts. Numerical summaries have been cross-checked against the experimental reports, within-dimension score means have been recomputed from the reported distributions, and the manuscript has been compiled and visually inspected. These checks do not constitute independent human validation of generated annotations or automated quality scores; the corresponding limitations are described in Section 7. We have not used generative AI for theorem proving or proof writing, which were outside this empirical study's scope. We take full responsibility for the final text, code, data artifacts, references, and scientific claims, including material produced with the assistance of generative AI.

REFERENCES

Negin Baghbanzadeh, Mohammed Saidul Islam, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. Open-PMC-18M: A high-fidelity large scale medical dataset for multimodal representation learning. *arXiv preprint arXiv:2506.02738*, 2025. URL https://arxiv.org/abs/2506.02738v3.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing*, pp. 182–210, 2025. doi: 10.18653/v1/2025.sdp-1.18.

Junpeng Ding, Zichen Tang, Haihong E, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, Zijie Xi, Ruomeng Jiang, Peizhi Zhao, Rongjin Li, Yuanze Li, Jiacheng Liu, Zhongjun Yang, Jintong Chen, and Siying Lin. Decoding scientific experimental images: The SPUR benchmark for perception, understanding, and reasoning. In *Proceedings of ACL*, pp. 12859–12882, 2026. doi: 10.18653/v1/2026.acl-long.585.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In *Findings of EMNLP*, pp. 3258–3264, 2021. doi: 10.18653/v1/2021.findings-emnlp.277.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, Clyde Lee Giles, and Ting-Hao Huang. GPT-4 as an effective zero-shot evaluator for scientific figure captions. In *Findings of EMNLP*, pp. 5464–5474, 2023. doi: 10.18653/v1/2023.findings-emnlp.363.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of CVPR*, 2017. URL https://arxiv.org/abs/1612.06890.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *International Conference on Learning Representations Workshop Track*, 2018. URL https://arxiv.org/abs/1710.07300.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. ACL-Fig: A dataset for scientific figure classification. In *Workshop on Scientific Document Understanding*, volume 3656 of *CEUR Workshop Proceedings*, 2023. URL https://ceur-ws.org/Vol-3656/paper2.pdf.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of ACL*, pp. 14369–14387, 2024. doi: 10.18653/v1/2024.acl-long.775.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3773.

Eric Schwenker, Weixin Jiang, Trevor Spreadbury, Nicola J. Ferrier, Oliver Cossairt, and Maria K. Y. Chan. EXSCLAIM!: Harnessing materials science literature for self-labeled microscopy datasets. *Patterns*, 4(11):100843, 2023. doi: 10.1016/j.patter.2023.100843.

Jifeng Song, Arun Das, Ge Cui, and Yufei Huang. FigEx: Aligned extraction of scientific figures and captions. In *Findings of EMNLP*, pp. 16558–16571, 2025. doi: 10.18653/v1/2025.findings-emnlp.899.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of EMNLP*, pp. 2112–2120, 2020. doi: 10.18653/v1/2020.findings-emnlp.191.

Haoyi Tao, Chaozheng Huang, Nan Wang, Han Lyu, Linfeng Zhang, Guolin Ke, and Xi Fang. OmniScience: A large-scale multi-modal dataset for scientific image understanding. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9860–9870, 2026. doi: 10.1145/3770855.3817510.

Venkata Kesav Venna, Sai Madhusudan Gunda, Jyothi Swaroopa Jinka, Hrithik Sagar Rachakonda, Anirudh Srinivasan, and Ravi Kiran Sarvadevabhatla. M3Grounder: Mask-based multi-span and multi-granular grounding for document QA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23685–23695, 2026.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. RefChartQA: Grounding visual answer on chart images through instruction tuning. *arXiv preprint arXiv:2503.23131*, 2025. URL https://arxiv.org/abs/2503.23131.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3609.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, Shiyang Feng, Bin Wang, Chao Xu, Conghui He, Pinlong Cai, Min Dou, Botian Shi, Sheng Zhou, Yongwei Wang, Bin Wang, Junchi Yan, Fei Wu, and Yu Qiao. DocGenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024. URL https://arxiv.org/abs/2406.11633.

Zhiwei Yang, Yuanchen Wu, Nan Zhang, Yucong Meng, Ke Yan, and Shouhong Ding. Scene graph thinking: Reinforcing structured visual reasoning for multimodal large language models. *arXiv preprint arXiv:2607.05716*, 2026. URL https://arxiv.org/abs/2607.05716.

Wenhan Yu, Zhaoxi Zhang, Wang Chen, Guanqiang Qi, Weikang Li, Lei Sha, Deguo Xia, and Jizhou Huang. SciEGQA: A dataset for scientific evidence-grounded question answering and reasoning. *arXiv preprint arXiv:2511.15090*, 2025. URL https://arxiv.org/abs/2511.15090.

Kun Yuan, Min Woo Sun, Zhen Chen, Alejandro Lozano, Xiangteng He, Shi Li, Nassir Navab, Xiaoxiao Sun, Nicolas Padoy, and Serena Yeung-Levy. From panel to pixel: Zoom-in vision-language pretraining from biomedical scientific literature. In *Proceedings of CVPR*, pp. 42649–42658, 2026.

Tan Yue, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. F2TEval: Human-aligned multi-dimensional evaluation for figure-to-text task. In *Proceedings of EMNLP*, pp. 3932–3948, 2025a. doi: 10.18653/v1/2025.emnlp-main.195.

Tan Yue, Xuzhao Shi, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. AnaFig: A human-aligned dataset for scientific figure analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12837–12843, 2025b. doi: 10.1145/3746027.3758226.

A Resource Construction and Engine Details

A.1 Region geometry and recursion

For a candidate parent crop $I[b]$, the detector proposes child boxes. Children smaller than 10^{-4} of the crop area are discarded. A descending-area suppression pass removes a smaller box if a retained larger box covers at least 0.7 of its area. This uses intersection over the smaller box's area, not the detector's internal IoU suppression. A parent is replaced only if at least two valid children remain. Local coordinates are translated into the source image after accounting for crop position and any inference resizing. The documented limit is ten frontier-processing rounds after initial detection.

An optional augmentation branch handles a region judged subdividable when ordinary detection returns fewer than two valid children. It searches horizontal and vertical flips, rotations by 90, 180, and 270 degrees, and two diagonal transpositions. Predictions are mapped back to crop coordinates. An augmented partition requires at least two children and union coverage of at least 0.8; pooled views use an overlap threshold of 0.85 and favor smaller regions. This branch is disabled in the documented default configuration.

For residual recovery, let B_0 denote retained boxes before LRR and U_0 their union. GPT predicts supplemental proposals G_{res} from a full figure with U_0 whitened. A proposal g is retained only if

$$\frac{|g \cap U_0|}{|g|} \leq 0.33. \tag{6}$$

The reference union is fixed during this filtering pass. It is not updated after every accepted proposal. The condition measures redundancy; it does not gate whether the LRR call is made. Proposals are rescaled, clipped, merged with B_0, and passed through the common overlap, minimum-area, and duplicate-coordinate filters. The final minimum-area threshold is relative to the complete image, unlike the crop-local child filter. No surviving proposal leaves the retained set unchanged.

A.2 Detector training and construction services

AS-YOLO uses YOLO26x with a single subfigure class and 1024×1024 input. It is initialized from a detector trained on inherited panel annotations and refined through human correction on 5,000 figures. Each documented training round uses 100 epochs, no early stopping, three warmup epochs, and cosine learning-rate decay with final ratio 0.01. AdamW uses initial learning rate 10^{-3}, momentum 0.937, and weight decay 5×10^{-4}. The global batch size is 16 over four GPUs, with nominal batch size 64 for gradient accumulation. Automatic mixed precision is disabled; the recorded seed is zero with deterministic execution. The detector-refinement figures are excluded from Eval-1K at the source-figure level.

Construction uses `qwen3.7-flash` for SRG and JSA and an Azure OpenAI `gpt-5.5` deployment for LRR, with API version `2024-10-21`. Recorded temperatures are zero for SRG and 0.3 for JSA; GPT uses the service default. Long image sides are limited to 1,024 pixels for Qwen and 1,536 pixels for GPT using isotropic Lanczos resizing. GPT coordinates are mapped back to the source resolution. Calls request structured JSON. Parsing first strips Markdown fences; a fallback extracts the substring between the first and last braces. Parsing failures receive task-specific conservative handling. SDK retries are disabled, unsupported optional fields can be removed for a further attempt, and sample-level timeouts permit up to three reruns. These settings describe the resource construction pipeline and must not be confused with the separate QA-generation or solver configurations.

A.3 Engine implementation and retained supervision

The data engine integrates observation extraction, evidence organization, program sampling, and QA generation into a unified pipeline. Using localized observations and their hierarchical and semantic relations, the engine selects supporting evidence and specifies the reasoning operations and their dependencies. The generator then follows this program to produce a question, an answer, and reference reasoning steps linked to the evidence. Evidence selection and the intended reasoning process are thus fixed before QA generation. Quality control combines program consistency checks, model-based review, and visual verification of difficult examples. These checks assess whether the

question is answerable, whether the evidence is sufficient, and whether the answer is consistent with the selected observations and reasoning operations. A refinement stage revises question wording, answers, and step descriptions while preserving the selected evidence and program structure. Samples included in the benchmark also undergo human review. Alongside each QA pair, the engine retains the supporting observations, visual regions, relations among evidence sources, and reference reasoning steps. These records document how the answer is constructed from visual evidence, allowing intermediate judgments and final conclusions to be traced to specific image regions. They also provide supervision for downstream scientific reasoning and evidence grounding tasks.

B Resource Statistics and Automated Quality

B.1 Counting conventions

The resource statistics refer to frozen record versions: figure v2, panel v3, subfigure v368, and paper v127. Of 4,464,199 panel records, 4,193,429 have valid geometry; the remaining 270,770 remain in the inventory count. Of 10,021,541 atomic records, 10,021,537 have valid geometry. There are 10,352,838 visual-type assignments because some regions have multiple types. A "Mixed" display label summarizes multi-label records and is not a tenth controlled visual type. Caption availability is reported separately from caption quality.

Table 6: Caption availability and length. Counts use the same frozen snapshot. Word counts use Unicode whitespace.

Level	Records	Nonempty captions	Median words	P90 words
Figure	1,520,179	1,520,178	334	578
Panel	4,464,199	4,463,084	126	173
Atomic subfigure	10,021,541	9,967,257	40	61

The source metadata covers nine named disciplines and an Other group. Visual-type distributions are multi-label and need not sum to 100% within a discipline. Source-paper and figure grouping are relevant for any future train/test split because numerous local records can share one source. Shorter local captions indicate narrower textual scope, not verified factuality or better training outcomes.

B.2 Automated quality rubric

We assess unit validity (B1), boundary quality (B2), factual reliability (C1), and information coverage (C2), each on a 0–4 scale. Box and caption quality use separate calls: the box judge does not receive the generated caption, and the caption judge does not receive the box scores. The overlay identifies the target. Necessary axes, units, legends, and scale bars are relevant to boundary quality. Source captions supply context only when applicable to the target region. C1 concerns correctness and local attribution; C2 concerns omitted local content. An incorrect value can therefore count as addressed for coverage while being penalized for factual reliability. We do not combine dimensions into one score. This follows the broader motivation for multidimensional scientific-caption assessment (Hsu et al., 2023; Yue et al., 2025a).

Table 7: Automated quality-score distributions on 12,198 regions from Eval-1K. Bin entries are percentages; means use the 0–4 scale. Unresolved service errors are not assigned score zero.

Judge	Dimension	0	1	2	3	4	Mean
Qwen	B1: Unit validity	1.9	1.4	2.9	7.7	86.1	3.8
	B2: Boundary quality	0.2	0.2	2.4	18.3	78.9	3.8
	C1: Factual reliability	2.5	6.9	8.0	6.2	76.4	3.5
	C2: Information coverage	1.7	2.3	13.5	26.2	56.3	3.3
GPT	B1: Unit validity	0.5	0.3	1.4	1.4	96.4	3.9
	B2: Boundary quality	0.2	0.2	3.6	25.2	70.8	3.7
	C1: Factual reliability	3.2	1.8	11.0	12.7	71.3	3.5
	C2: Information coverage	1.4	2.3	17.4	21.8	57.2	3.3

The archived Qwen quality configuration uses `qwen3.8-max`, crop-and-source-caption input, and a maximum image side of 2,048 pixels. Its latest-event summary has 24,398 task slots, comprising 12,199 box and 12,199 caption assessments: 24,312 successful model events, 84 programmatic prechecks, and two unresolved API errors affecting one region. Numeric precheck results are included in the completed-score population. Marginal score distributions do not establish item-level agreement between judges or accuracy against human judgments. The human-reviewed QA benchmark is a different population from this automated caption audit.

C Benchmark Composition and Scoring

C.1 Coverage and difficulty terminology

The benchmark contains 200 questions per axis, with 100 basic and 100 advanced items in each.

The category-level inventory associates quantitative analysis, visual structure, and experimental relations with approximately 185, 183, and 184 distinct figures, respectively, and 161, 156, and 156 distinct source papers.

C.2 Answer judging and retained-response evaluation

Five answer types share a binary outcome: correct is one and incorrect is zero. Three scoring mechanisms implement these outcomes (Table 8). The same rules apply to quantitative, visual-structure, and experimental-relations questions at both benchmark levels.

Table 8: **Answer types and frozen scoring rules.** A question may contain multiple required items or mix answer types. All required items must pass, with a consistent final answer, to receive one point; there is no partial credit.

Answer type	Mechanism	Acceptance rule
Yes/no	Normalized exact match	Map explicit affirmative/negative statements to canonical labels. Equivalent wording is allowed; ambiguity or contradiction fails.
Trend / comparison	Normalized exact match	Match normalized directions, comparisons, or same/opposite labels. Rankings must preserve order and ties; all requested comparisons must be correct.
Numerical	Numerical match	Convert to the reference unit and apply frozen tolerance. All requested values and range endpoints must match, without missing or surplus values.
Entity/label	Normalized exact match	Accept predefined aliases. Compare deduplicated entity sets unless order is required; missing required entities or adding incorrect entities fails.
Free text	Binary model judgment	Cover every required fact in the frozen rubric, allowing semantic paraphrases but no substantive contradiction. Do not grade verbosity, length, or style.

For yes/no, trend/comparison, and entity/label answers, a language model first extracts and normalizes the candidate's statement without receiving the reference answer or image. The extraction must cite the candidate's own text; a program then compares it with the reference. Numerical answers also use extraction followed by programmatic checking. The judge cannot independently relax a tolerance. All answer types must preserve the requested entity, condition, unit, and experimental correspondence. Only requested information is required: a trend question does not require an incidental numerical value merely because that value appears in the reference answer.

For a reference value y and extracted value $\hat{y}$, numerical acceptance requires

$$|\hat{y} - y| \leq \max(\text{abs_tol},\ \text{rel_tol}\,|y|)\,. \tag{7}$$

Exact printed values and exact counts use zero absolute and relative tolerance. Visual readings from a linear axis use an absolute tolerance of 0.1 times the smallest visible tick interval and a relative tolerance of 5%. Logarithmic axes, axes without visible ticks, and derived quantities require item-specific tolerances and a recorded rationale fixed before evaluation; tolerances are not adjusted after inspecting model answers.

The stated judging configuration uses Sol at medium reasoning effort for primary assessment and Gemini 3.8 Flash at medium effort for review. Reference disputes and disagreements between the primary judge and reviewer enter the unresolved state rather than receiving an automatic correctness label. This hybrid procedure constrains subjective judging, while exact prompts and versioned rubrics remain necessary for full reproducibility.

The current independent view excludes each model's refusals and unresolved cases from answer and grounding summaries. Refusals were observed for Claude in the reported runs. Answer accuracy

is the fraction of retained questions marked correct. Categories are averaged only through their retained question counts when forming the overall result; equal source-category sizes do not imply equal retained sizes. This protocol reports performance conditional on retained responses rather than fixed-denominator accuracy over all 600 items. It should be preserved when interpreting the model comparison and distinguished from the same-model matched-question oracle protocol.

C.3 GROUNDING METRICS

For threshold τ, predicted regions are matched to reference regions using bounding-box IoU. Invalid predictions count as unmatched predicted boxes, and duplicate predictions are not deduplicated; surplus predictions therefore reduce precision. For a question with P predicted regions, $G > 0$ reference regions, and T_τ successful matches,

$$\text{Precision}_\tau = \frac{T_\tau}{P}, \qquad \text{Recall}_\tau = \frac{T_\tau}{G}, \qquad \text{F1}_{\text{ground}\,\tau} = \frac{2T_\tau}{P+G}, \tag{8}$$

with precision and F1 zero when $P = 0$. Reported grounding F1 is the arithmetic mean over retained questions multiplied by 100. It is neither pooled-box micro F1 nor an equal-weight mean over the three categories. Grounding and answer correctness are scored separately. Computing their joint success requires per-item outcomes, which are not reconstructed from marginal summary tables.

C.4 ORACLE INPUTS AND INTERPRETATION

The oracle input is the unmarked full figure followed by every distinct reference-evidence crop and the question. There are 317 questions with one crop, 280 with two, and three with three, totaling 886 crop occurrences. Crops have no padding or added labels. Hash checks were performed on 471 full-image files against the frozen sources. The prompt identifies the crops as relevant evidence but withholds reference answers, coordinates, captions, and paper context. Predicted evidence boxes must still refer to the full original figure.

The reported oracle protocol compares the same model on matched questions. We preserve the supplied aggregate values and reported changes. Statistical uncertainty, answer/evidence transition counts, and the fraction of initially wrong answers corrected by crops require per-item outputs and are not inferred from the aggregate differences. Visual crops may improve both target selection and effective resolution; they do not isolate one mechanism or establish faithful use of evidence before the intervention.

D BENCHMARK EXAMPLES

Figure 4 illustrates the three axes and their basic/advanced composition strata. These examples show reference answers and their supporting regions.

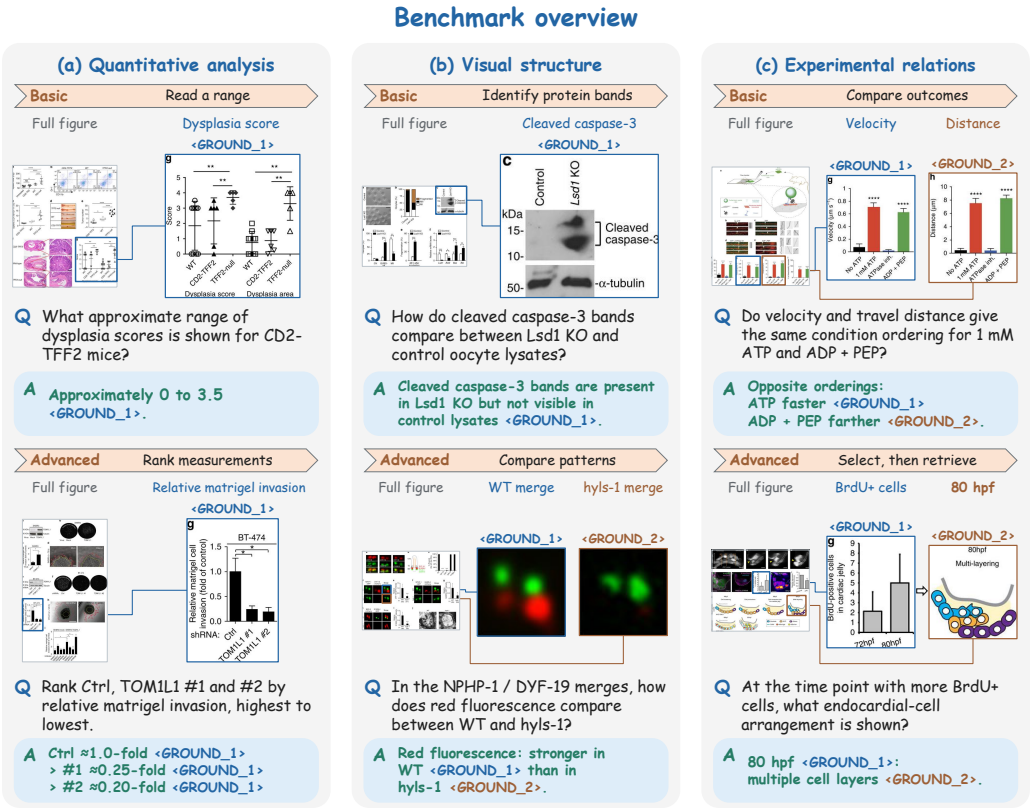

Figure 4: **Benchmark examples.** Quantitative questions read or compare measurements; visual-structure questions interpret appearance and spatial patterns; experimental-relations questions connect observations across experimental conditions or panels. Colored regions indicate reference support. Advanced examples require more involved composition within an axis.

E SEPARATE CAPTION-CONDITIONED LOCALIZATION DIAGNOSTIC

This study asks a model to locate a single described subfigure in the full figure. It is distinct from the 600-question joint reasoning benchmark: it has its own question construction, denominator, and metrics. It uses calibrated Eval-1K regions rather than automatically extracted regions from the quality audit.

Caption screening begins with 12,785 reference captions and 142,052 within-figure pairs. An embedding model (`text-embedding-v4`, 1,024 dimensions) selects 21,036 pairs at cosine similarity at least 0.85 for a two-caption matching check with `qwen3.8-flash`. Of these, 11,641 pass and 9,395 fail, yielding 4,078 distinct excluded subfigure identifiers. Failed matching can reflect errors or ambiguity; it is not proof that the underlying regions are intrinsically indistinguishable. Below-threshold pairs are not tested.

Using seed 42, the generator samples one or two reference regions per figure and produces 1,519 questions. Applying the exclusion index removes 268, leaving 1,251 questions from 876 figures. The generator sees the selected caption and reference overlay; solvers receive the question and clean full figure. Generated wording is not independently checked for uniqueness against every possible region. Filtering changes the evaluation population rather than improving the solver.

For each retained question, let u_i be IoU with its single reference box, assigning zero when no valid prediction is available. Mean IoU averages u_i over all 1,251 items. Recall at threshold τ is the fraction with $u_i \geq \tau$. Average recall averages these fractions over $\tau \in \{0.50, 0.55, \ldots, 0.95\}$. AP uses model-provided confidence, a stable question-identifier tie-break, and a 101-point interpolated precision envelope; unavailable recall levels contribute zero. Missing and abstained predictions remain in the denominator. Thus, AP50 is not the fraction of correctly localized questions.

Table 9: Caption-conditioned localization on 1,251 retained questions. All values are on the 0–1 scale. This table must not be merged with the joint QA benchmark.

Model	AP	AP50	AP75	AR	AR50	AR75	mIoU
DeepSeek-V4.1-Flash	.365	.657	.360	.506	.772	.528	.682
Kimi-k3	.642	.873	.701	.751	.922	.808	.835
MiniMax-M3	.257	.590	.193	.383	.702	.362	.604
Qwen3.8-27b	.593	.824	.674	.712	.895	.794	.807
GLM-5.3-flash	.733	.888	.809	.812	.937	.872	.867
Qwen3.8-max	.478	.764	.492	.622	.859	.673	.757
GPT 6 Sol (medium)	.762	.862	.794	.829	.906	.853	.871
Gemini 3.8 Flash	.144	.305	.115	.307	.495	.293	.472
Claude Opus 5.5 (medium)	.796	.894	.845	.851	.929	.887	.888

Recorded initial output limits are 8,192 tokens for GLM, DeepSeek, MiniMax, and the two Qwen variants, and 16,384 for Kimi with low reasoning effort. The implementation supports bounded truncation retries up to 32,768 tokens; these limits do not imply equal realized compute. Qwen predictions use normalized-1000 coordinates; the other documented adapters use transmitted-image pixels before rescaling. MiniMax has one abstention, which remains in the 1,251-item denominator. These solver settings are specific to this diagnostic and do not define the separate joint-QA inference protocol.

F Additional Category-Level Analysis

This section analyzes the existing category aggregates; it does not introduce new inference runs. Figure 5 makes two complementary patterns visible. First, the largest oracle accuracy increase occurs on experimental relations for Qwen3.8-max and GPT-4o. Second, visual structure remains the lowest-accuracy category for every reported oracle model. The latter pattern also holds for all eight baseline models, for both answer accuracy and grounding F1 at IoU 0.75 (allowing a tie for grounding).

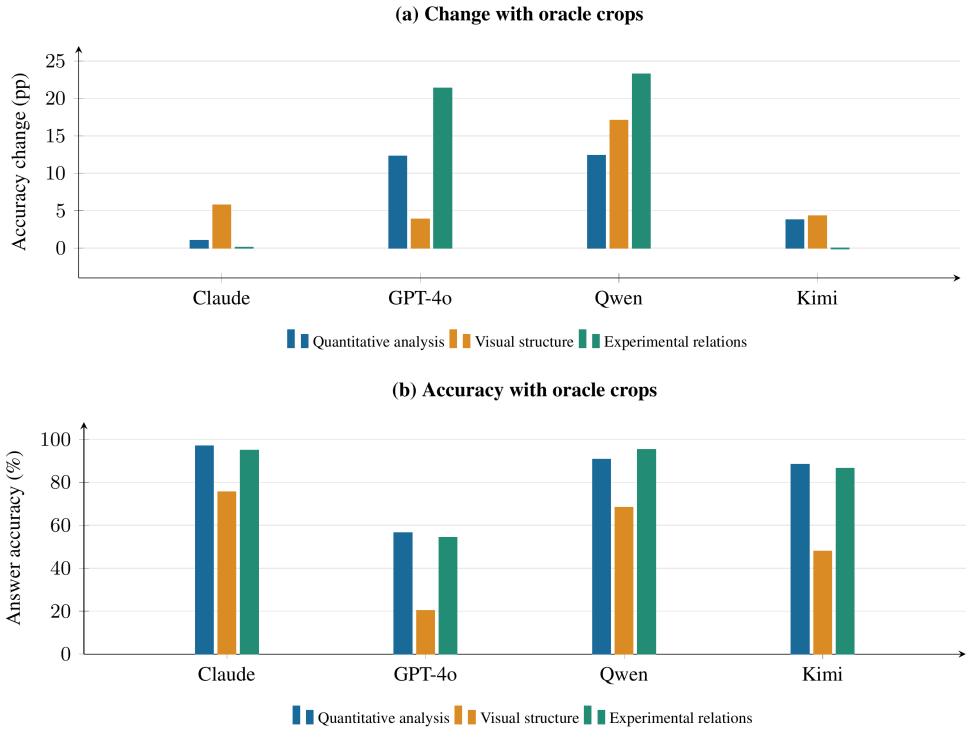

Figure 5: **Category-dependent response to oracle evidence crops.** Top: reported changes in answer accuracy. Bottom: accuracy with oracle crops. Experimental relations has the largest positive change for Qwen3.8-max and GPT-4o, while Claude and Kimi show almost no relational change. Visual structure remains the weakest oracle category in every model. These aggregates do not isolate the cause of individual errors.

Accounting descriptively for available error headroom. Let A be baseline category accuracy and ΔA the reported oracle change, both in percentage points. The descriptive relative reduction in error is $100\Delta A/(100 - A)$. Table 10 reports this transformation for all four models. It is a net aggregate error reduction, not the fraction of originally incorrect questions individually corrected by the oracle. Rounding follows the supplied category accuracies and reported changes; no significance claim is made.

Table 10: Descriptive relative reduction in error with oracle crops (%). Computed from the reported baseline accuracies and reported deltas; negative values indicate increased error.

Model	Quantitative	Visual structure	Experimental relations
Claude Opus 5.5	25.1	19.1	1.7
GPT-4o	22.1	4.6	31.9
Qwen3.8-max	57.0	35.0	83.1
Kimi-k3	24.7	7.7	-0.8

For Qwen3.8-max, the relational gain corresponds to an approximately 83.1% reduction in the baseline error rate, versus 57.0% for quantitative analysis and 35.0% for visual structure. For GPT-4o, the corresponding values are 31.9%, 22.1%, and 4.6%. Thus, those two models' larger relational gains are not solely an artifact of a larger baseline error denominator. The result remains model-dependent: Claude and Kimi show negligible or slightly negative relational changes, and a high initial relational score alone does not identify why.

An intended dependency, not a measured causal mechanism. In the select-then-lookup family, two observations identify a condition and a third supplies an attribute of that condition. This is the dependency in Equation 4; the second operation must bind to the condition selected by the first. Evidence can therefore be relevant to an intermediate decision even when it does not directly contain the final answer. The program records this intended relationship. However, the present tables aggregate the complete experimental-relations category, not only this family or its Advanced stratum. They cannot attribute the category gain specifically to a three-observation path, establish that each observation is necessary to a solver, or distinguish condition binding from local visibility. Several observations may also share one source crop, so a three-observation program is compatible with two distinct evidence crops.

Implication for benchmark design. The categories probe complementary capabilities. A model may answer many relational questions while localizing their support weakly, and may benefit from being shown the correct regions. A model may also remain inaccurate on visual-structure questions after those regions are supplied. A benchmark combining open answers, region predictions, and typed reference compositions can expose these patterns together. Joint answer–grounding success and within-family interventions would be needed for stronger causal claims; they are not reconstructed from the available category marginals.

Table 11: **Representative scientific visual resources.** Slashes separate distinct counted units rather than interchangeable dataset sizes. These resources are not all QA benchmarks.

Resource	Scale and unit	Region/context supervision
ACL-Fig (Karishma et al., 2023)	0.112M figures / 1.67K type-labeled	Figure captions and inline references
SciCap (Hsu et al., 2021)	2.17M figures / 0.417M graph plots	Captions, metadata, subfigure flag
Multimodal ArXiv (Li et al., 2024)	6.4M images / 3.9M main captions	Subfigure hierarchy and source subcaptions
DocGenome (Xia et al., 2024)	0.50M documents / 6.8M pages	Full documents, boxes, and relations
MedICaT (Subramanian et al., 2020)	0.217M figures / 7.5K manually aligned subfigures	Manual subfigure boxes and subcaptions for a subset
BioSci-Fig (Song et al., 2025)	7.17K compound figures / 43.18K boxes	Manual boxes and verified subcaptions
Open-PMC-18M (Baghbanzadeh et al., 2025)	18M medical images	Extracted crops paired with full captions
Panel2Patch (Yuan et al., 2026)	0.364M figure / 1.30M panel / 0.619M region pairs	Biomedical figure/panel/region supervision
OmniScience (Tao et al., 2026)	1.52M figures / 4.46M panels	Cross-disciplinary figures, panels, and paper context
OmniScience++	1.52M figures / 4.46M panels / 10,021,541 atomic records	Figure–panel–subfigure mappings; local boxes, types, tags, captions; downstream evidence programs

G SUPPLEMENTARY RESOURCE FIGURES AND EXPERIMENTS

Figure 1 summarizes the resource's scale and visual composition. The following figures provide further resource characterization and construction diagnostics. They concern the large figure resource and Eval-1K, whereas the joint QA benchmark contains 600 separately reviewed questions. Source-paper subject groups and generated visual types characterize different properties; the Mixed display group is not an additional controlled visual type.

G.1 EXTRACTION ABLATIONS AND TOKEN ACCOUNTING

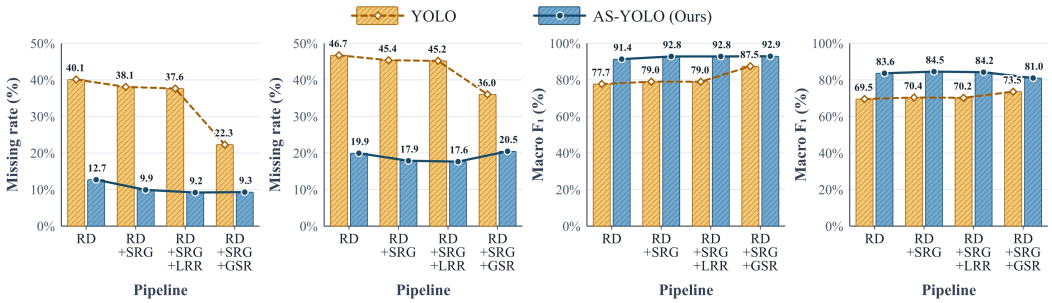

Figure 6: **Extraction ablation in graphical form.** The four panels show missing rate at IoU 0.50 and 0.90, followed by macro F1 at the same thresholds. Values correspond to Table 3; lower missing rate and higher F1 are preferable.

Table 12: **Token consumption and API cost on Eval-1K.** Token counts are millions; costs are USD. Stage-level accounting corresponds to Table 3. Totals may differ slightly from sums of displayed rounded components. Detector training, local compute, and JSA are not included.

Detector	Configuration	Qwen tokens Input	Qwen tokens Output	GPT tokens Input	GPT tokens Output	Total tokens	Qwen cost	GPT cost
YOLO	RD	.000	.000	.000	.000	.000	.000	.000
	+SRG	5.259	.260	.000	.000	5.518	.187	.000
	+SRG+LRR	5.259	.260	1.962	.357	7.837	.187	20.506
	+SRG+GSR	5.259	.260	2.525	3.130	11.174	.187	106.536
AS-YOLO	RD	.000	.000	.000	.000	.000	.000	.000
	+SRG	4.520	.346	.000	.000	4.866	.175	.000
	+SRG+LRR	4.520	.346	1.962	.434	7.261	.175	22.822
	+SRG+GSR	4.520	.346	2.620	2.984	10.470	.175	102.630

The greatest extraction improvement comes from additional atomic-region supervision for AS-YOLO. LRR then offers a different cost/localization tradeoff from globally regenerating all boxes. Its lower cost in this experiment is an API-cost comparison on Eval-1K, not the full cost of constructing ten million atomic records.

G.2 Discipline-conditioned visual composition

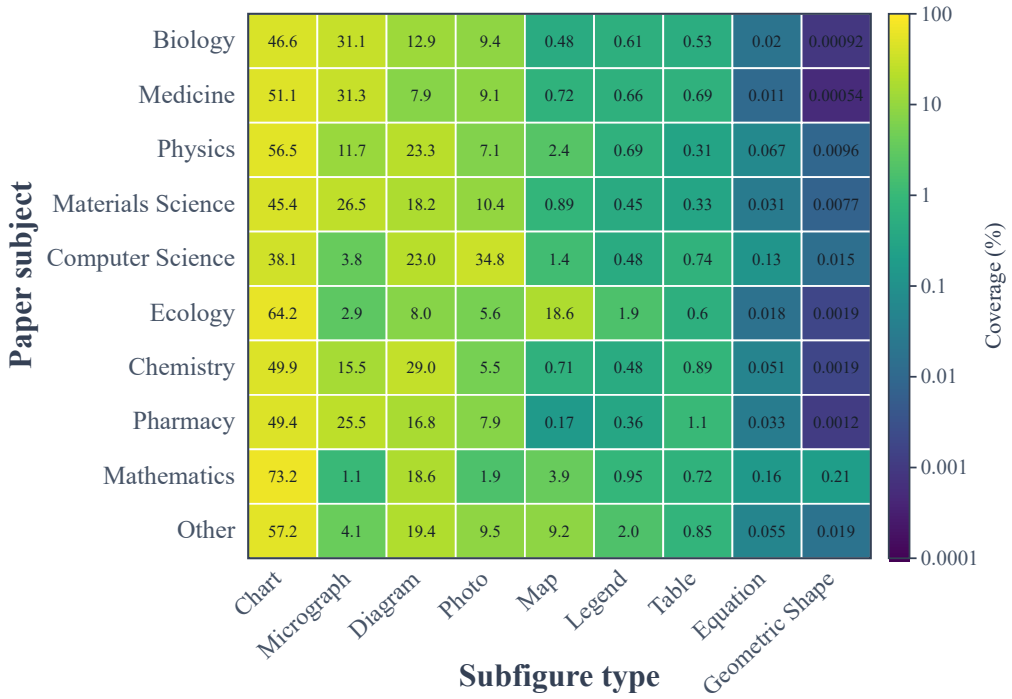

Figure 7: **Atomic visual types by source-paper subject.** Each entry is the within-subject percentage for that type, shown with a logarithmic color scale. Labels are multi-label, so rows need not sum to 100%. Other follows source metadata; unknown subjects are excluded. The plot is descriptive and does not establish balanced domain coverage.

G.3 Region multiplicity and geometry

The discipline-conditioned multiplicity summary includes 1,425,373 of 1,520,179 figures. Biology has median atomic count 6 and P90 22, while Medicine has median 5 and P90 20. Physics, Materials Science, and Pharmacy have median 3; the remaining reported medians are 2–3. These are terminal-region counts, not reasoning depth or the number of experimental relations.

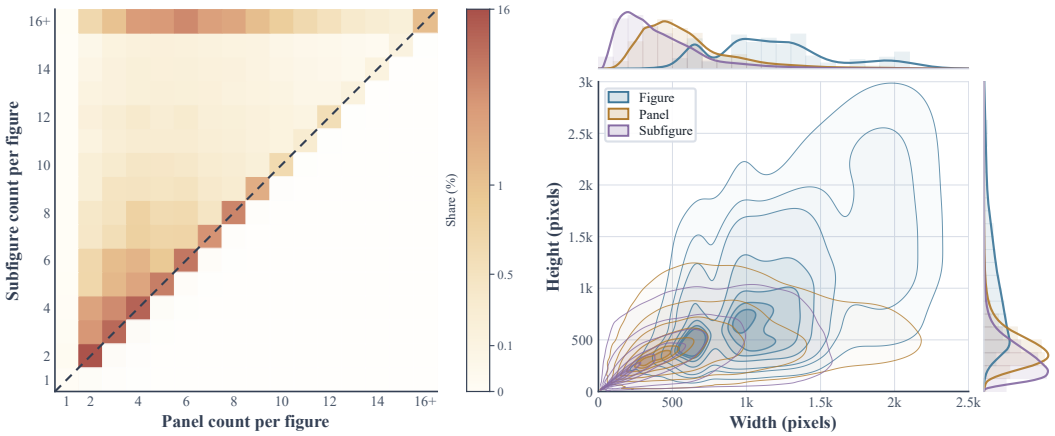

Figure 8: **Region multiplicity and spatial scale.** Left: joint panel and atomic-subfigure counts for 917,409 source figures represented in both inventories; counts of 16 or more are pooled, and color shows share on a logarithmic scale. Right: width–height distributions and marginal densities for full figures, panels, and atomic subfigures, in pixels. The right plot is a geometry distribution, not a cumulative relative-area curve.

The comparison uses the declared geometry populations: 4,193,429 valid panel boxes and 10,021,537 valid atomic boxes. Its purpose is to characterize the scale of the retained units. It does not impose a fixed number of children per panel or demonstrate that each region is independently meaningful.

G.4 Caption-length distributions

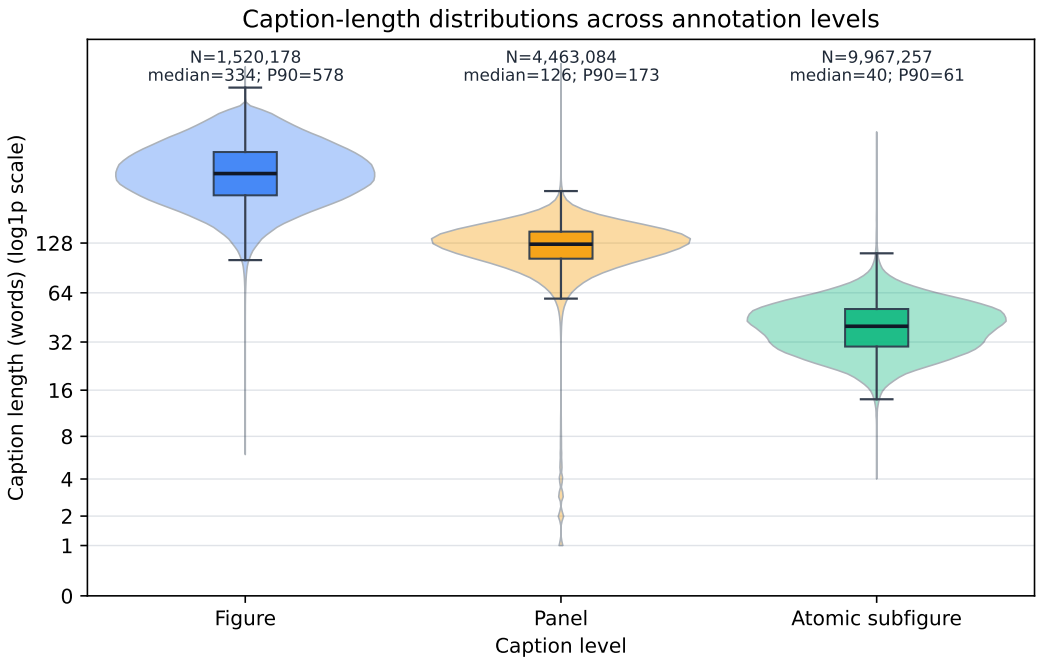

Figure 9: **Caption lengths at three annotation levels.** Nonempty captions use Unicode-whitespace word counting, with the displayed log1p scale. Figure, panel, and atomic medians are 334, 126, and 40 words, respectively; P90 values are 578, 173, and 61. These describe textual scope and availability, not factual reliability.

G.5 CAPTION-LOCALIZATION SUBSET SENSITIVITY

Table 13: **Sensitivity to caption-question screening.** Metrics reuse unchanged predictions on the original 1,519 questions and filtered 1,251-question set of Appendix E. Scores are on the 0–1 scale. Different evaluation populations explain these changes; they are not model-training improvements.

	AP		Mean IoU	
Model	Original	Filtered	Original	Filtered
GLM-5.3-flash	.6700	.7334	.8301	.8670
DeepSeek-V4.1-Flash	.3110	.3651	.6290	.6820
Kimi-k3	.6054	.6420	.8084	.8354
MiniMax-M3	.2177	.2565	.5577	.6044
Qwen3.8-max	.4041	.4782	.7022	.7569
Qwen3.8-27b	.5438	.5928	.7711	.8066
GPT 6 Sol (medium)	.6961	.7624	.8288	.8712
Claude Opus 5.5 (medium)	.7347	.7963	.8401	.8878
Gemini 3.8 Flash	.1276	.1444	.4374	.4717

Screening raises AP and mean IoU for all listed configurations on the retained population, but may alter difficulty and composition. This sensitivity is relevant when using caption-generated localization tasks as a resource diagnostic. It is separate from filtering model responses in the 600-question QA benchmark and from the visual-oracle intervention.