VISATOM: DO MULTIMODAL MODELS TRULY SEE BEFORE THEY THINK?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have achieved impressive scores on established reasoning benchmarks, yet recent evidence shows that many such benchmarks can be solved through textual shortcuts rather than genuine visual interpretation. This gap is critical because answer accuracy alone can overestimate capabilities and reward models that fail when visual conditions change. We introduce VISATOM, a methodology and benchmark for evaluating visual perception, reasoning with reduced textual cues, and generalization to modified problems. Starting from existing image-based problems, we extract *atomic facts*: elemental, directly observable statements describing content supported by an image, and use them to construct and check diagnostic tasks. These facts connect three evaluation tiers: **Perception Probing** tests basic visual observation, **Pruned Problems** remove image-redundant text, and **Edited Variants** programmatically alter problem instances to test generalization beyond public benchmark examples. Sourced from 2,711 challenging problems across 12 existing benchmarks spanning 8 domains, VISATOM provides a unified, multi-granularity diagnostic system on top of existing community benchmarks. Across ten frontier and open-weight MLLMs, VISATOM reveals persistent weaknesses in visual estimation (81-92%) and universal degradation after text pruning (4-11%). Edited variants further expose a 12.4% generalization gap for the strongest model (GPT-5.4), suggesting that strong reasoning performance does not always transfer to new problem variants.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made remarkable progress in solving problems that integrate visual and textual information. With reinforcement learning post-training, frontier models now achieve strong performance on challenging scientific and mathematical reasoning benchmarks (Yue et al., 2024; He et al., 2024; Lu et al., 2024). This progress has driven the development of increasingly demanding evaluations, ranging from problems drawn from Olympiad-level competitions (He et al., 2024; Huang et al., 2024) to expert-curated science questions (Yue et al., 2024; Zhao et al., 2025; Phan et al., 2025).

However, a growing body of work reveals that *many multimodal benchmarks do not actually require visual understanding to solve*. Models can retain surprisingly high accuracy when visual input is removed or replaced by captions (Yue et al., 2025; Zhang et al., 2024b; Hao et al., 2025), indicating that they often exploit textual shortcuts instead of interpreting the image. These findings expose a fundamental evaluation gap: answer accuracy alone does not reveal *whether and how a model uses visual information*. Scores achieved through redundant text or familiar problem instances may therefore fail to predict performance when visual conditions change.

Solving a multimodal problem requires more than producing the final answer: a model must first *perceive* the relevant visual content, ground its *reasoning* in those observations, and *transfer* that reasoning beyond previously seen problem instances. A diagnostic evaluation targeting these capabilities is needed to distinguish models that perceive and reason from those that reach correct answers through textual cues or familiarity with existing problems.

We introduce VISATOM, a reusable methodology for constructing complementary diagnostic tasks from existing image-based problems, together with a benchmark built using this methodology. The

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.

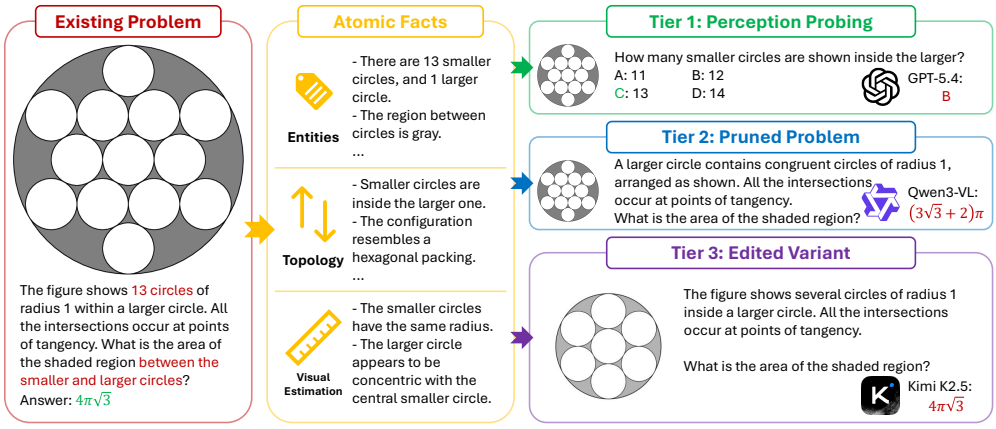


Figure 1: **Overview of VISATOM.** From each source problem, we extract *atomic facts* (entities, topology, and visual estimation) to construct three tiers: Tier 1 probes visual perception, Tier 2 removes text redundant with the image, and Tier 3 creates programmatically edited variants to test generalization. Here, GPT-5.4 miscounts the circles, Qwen3-VL fails once the shortcut "13 circles" is pruned, and Kimi K2.5 repeats the original answer on the edited variant.

benchmark is unified by **atomic facts**: elemental, directly observable statements extracted from each problem's images (Figure 1). Atomic facts provide a shared representation for constructing perception probes, identifying overlap between text and images, and guiding problem edits. Rather than simply filtering for vision-requiring problems (Yue et al., 2025; Hao et al., 2025), we use these facts to construct three evaluation tiers: targeted perception questions, problems with image-overlapping text shortcuts removed, and edited variants with changed conditions or diagrams. Each derived item remains linked to its source problem, enabling deeper reuse of community datasets through joint perception–reasoning analyses and paired comparisons of original and transformed problems. VISATOM draws from 12 benchmarks spanning eight domains, combining difficulty filtering and deduplication with model-assisted generation, checking, and human review.

Our contributions and key findings are:

- **Reusable construction methodology and benchmark.** Atomic facts connect perception probing, text pruning, and problem editing within one workflow. Applying this workflow to 2,711 source problems supports complementary diagnostic analyses of existing community datasets.
- **Fine-grained perception testing.** Tier 1 (Perception Probing) generates targeted questions from atomic facts. Visual estimation is the weakest subcategory across all models (81-92% accuracy), with up to an 11% gap between the best and worst models.
- **Diagnosing textual shortcuts through pruning.** Tier 2 (Pruned Problems) compares original problems with variants that remove text overlapping with image content. On 1,035 paired problems, pruning lowers accuracy by 4-11% across all ten models, revealing exploitable textual shortcuts. Modality ablation further shows a larger image-removal penalty after pruning, connecting shortcut removal to increased reliance on visual information.
- **Generalization testing through programmatic instance editing.** Tier 3 (Edited Variants) changes problem conditions, questions, or diagrams and compares the resulting variants with their originals. The strongest model on the original problems, GPT-5.4, shows a 12.4% generalization gap, demonstrating that strong original performance does not always transfer to modified instances.

2 RELATED WORK

Multimodal reasoning benchmarks. Multimodal reasoning benchmarks increasingly target challenging problems across diverse domains. MMMU (Yue et al., 2024) covers college-level subjects, while OlympiadBench (He et al., 2024) and OlympicArena (Huang et al., 2024) draw on competition-level problems, with the latter also providing process-level evaluation. Humanity's Last Exam (Phan et al., 2025) uses expert-curated questions to probe frontier-model limitations. MathVista (Lu et al., 2024), MATH-Vision (Wang et al., 2024), and MathVerse (Zhang et al., 2024b) evaluate visual mathematical reasoning, while EMMA (Hao et al., 2025) integrates multimodal reasoning across STEM and coding. SUPERChem (Zhao et al., 2025) and MME-Reasoning (Yuan et al., 2025) focus on chemistry and logical reasoning, respectively. These benchmarks establish demanding tests of

end-to-end problem solving. Complementary diagnostic benchmarks (Johnson et al., 2017; Thrush et al., 2022; Kembhavi et al., 2016; Masry et al., 2022; Guan et al., 2024; Tong et al., 2024b;a; Xu et al., 2025) probe specific visual capabilities. Our focus is a common construction procedure that links such diagnoses to pruning and editing of the same source problems.

Shortcuts and validity in multimodal evaluation. A growing concern is that multimodal benchmarks often admit solutions without genuine visual understanding (Goyal et al., 2017; Agrawal et al., 2018). MMMU-Pro (Yue et al., 2025) identifies substantial text-only solvability in MMMU and introduces filtering, 10-option augmentation, and vision-only settings to mitigate it. MathVerse (Zhang et al., 2024b) finds that some models perform better when diagrams are removed. Its six human-annotated versions vary the information available in text and images, providing an important precedent for comparisons aligned to the same source item. EMMA (Hao et al., 2025) filters problems solvable from text and image captions, and shows that test-time compute scaling does not resolve visual reasoning gaps. PuzzleVQA (Chia et al., 2024) decomposes perception, induction, and deduction in abstract-pattern problems using ground-truth explanations. VisuLogic (Xu et al., 2025) evaluates human-verified problems organized by visual reasoning categories.

VISATOM complements these approaches through a reusable construction workflow. Atomic facts connect perception probes, pruned problems, and edited variants, with each derived item linked to its source. This supports joint analysis of local perception, reasoning with reduced textual cues, and generalization to modified instances. The contribution is a shared procedure for constructing and relating these diagnostic tasks across community datasets.

3 THE VISATOM BENCHMARK

3.1 OVERVIEW

VISATOM is a vision-centric multimodal reasoning benchmark composed of three evaluation tiers, each constructed from a shared pool of source problems through the unifying lens of *atomic facts*. Each tier has its own generation and quality-control procedure: a source image can yield several perception probes, and a source problem can yield a pruned or edited variant.

We source problems from 12 established benchmarks (He et al., 2024; Huang et al., 2024; Zhao et al., 2025; Hao et al., 2025; Wang et al., 2024; Yue et al., 2024; Lu et al., 2021; Zhang et al., 2024b; Yuan et al., 2025; Phan et al., 2025; Lu et al., 2024; Chia et al., 2024), spanning 8 domains: mathematics, physics, chemistry, biology, geography, computer science, engineering, and puzzles. Problems include single-selection and multi-selection multiple-choice as well as free-form answer formats. Table 1 summarizes the key statistics and Figure 2 shows the domain and answer format distributions. The evaluated Tier 1 set contains 11,094 perception questions, Tier 2 contains 1,035 pruned problems, and Tier 3 contains 283 human-verified variants. Table 1 reports the source-pool size and the number of evaluated items in each tier.

Atomic facts. We decompose visual content into **atomic facts**: single-sentence statements directly verifiable from the image without solving the underlying problem. Reading labels or conventional symbols may require representational knowledge, but facts do not derive new conditions from a solution or an unstated assumption. We organize atomic facts into three categories:

- **Entities**: Labels, marks, equations, coordinates, and other symbolic content (*e.g.*, "the label at the top-left vertex is A" or "the molecule contains a hydroxyl group -OH").
- **Topology**: Spatial relationships and connectivity (*e.g.*, "point D lies on segment $\overline{BC}$").
- **Visual estimation**: Approximate quantitative properties expressed qualitatively (*e.g.*, "the angle at vertex C appears obtuse").

These atomic facts serve as the foundation for all three tiers: Tier 1 directly tests them, Tier 2 uses them to identify and remove textual redundancy, and Tier 3 uses them to guide principled problem editing (see Figure 1 for an illustrative example). We retain this distinction when using the facts: Tier 1 tests the stated observation, while Tier 2 also considers how the problem statement licenses interpretation of its diagram.

Three evaluation tiers. Each tier targets a distinct aspect of vision-centric reasoning:

Table 1: **Key statistics** of the evaluated VISATOM sets.

Statistic	Value
Source benchmarks	12
Source-pool domains	8
Total source problems	2,711
Tier 1 perception questions	11,094
Tier 2 pruned problems	1,035
Tier 3 edited problems	283
Total evaluation items	15,123

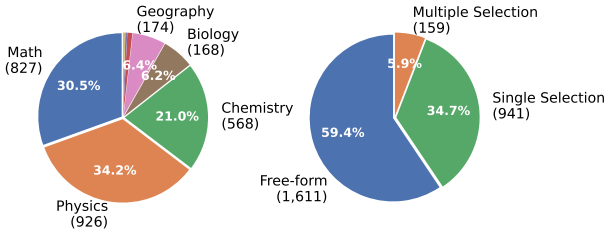


Figure 2: **Distribution of source problems** by domain (left) and answer format (right). These are source-pool distributions, not the retained distributions of each tier.

- **Tier 1 (Perception Probing):** Three multiple-choice questions per source image test entity recognition, topology, and visual estimation.
- **Tier 2 (Pruned Problems):** We remove text overlapping with image content and compare the pruned problems with their originals to assess reliance on textual shortcuts.
- **Tier 3 (Edited Variants):** We change problem text, diagrams, or both, update the solutions, and evaluate generalization to the resulting variants. Edits include text-only changes with semantically equivalent images and joint text-image changes; human annotations characterize edit type and relative difficulty.

3.2 DATASET CONSTRUCTION

The construction of VISATOM is guided by four principles: **vision-centered diagnosis**, which measures the contribution of visual evidence; **multi-granularity diagnosis**, with separate probes for perception, reasoning, and generalization; **grounded construction via atomic facts**, so that all three tiers derive from the same extracted visual content; and **instance variation**, which tests transfer to changed problems. Difficulty filtering and deduplication support the reuse of challenging problems from existing benchmarks.

3.2.1 PROBLEM PREPROCESSING

We aggregate problems from 12 established benchmarks, retaining only those with at least one image. To ensure sufficient difficulty, we filter using three lightweight models (GPT-5 nano, Gemini 3 Flash, Claude Haiku 4.5) with reasoning disabled; a problem is kept only if none produces the correct answer in a single pass. We then deduplicate across source benchmarks using embedding-based similarity with the Gemini Embedding 2 model and a cosine threshold of 0.9. Full details of the filtering and deduplication procedures are provided in the supplementary material (Appendix D).

3.2.2 ATOMIC FACT EXTRACTION

For each retained problem, we extract atomic facts from its image(s) in the context of the problem text. The text identifies referents and task context, but each fact must be supported by the diagram, its labels, or its symbols; text-only conditions remain separate from image evidence. We prompt GPT-5.4 to produce a structured fact list in three categories: entities, topology, and visual estimation. Each fact must be: (1) directly verifiable from the image, (2) expressed as a single sentence, and (3) limited to the strength of its visual support, without deriving new conditions from the answer, a theorem, or an unstated assumption. Estimates remain explicitly approximate; convention-based interpretations are not treated as explicit visual marks.

To improve completeness and accuracy while reducing dependence on a single model, we perform automatic cross-model revision. Gemini 3 and GPT-5.4 independently check GPT-5.4's initial extraction for completeness and each fact for visual accuracy, suggesting corrections for missing, incorrect, or overly inferential facts. GPT-5.4 then aggregates this feedback to revise the list, removing or revising facts that require inference beyond direct observation. The resulting atomic facts form the shared foundation for all three tiers.

3.2.3 TIER 1: PERCEPTION PROBING

For each source image, we generate three single-selection multiple-choice questions, one per atomic fact category: (1) an **entity recognition** question (*e.g.*, "What label appears at the top-right vertex?"), (2) a **topology understanding** question (*e.g.*, "Relative to Point A, where is Point B located?"), and (3) a **visual estimation** question (*e.g.*, "Approximately what is the aspect ratio of the rectangle?").

Construction follows a three-stage pipeline: (1) **Proposal.** GPT-5.4 proposes three perception multi-choice questions grounded in the original problem and its atomic facts. (2) **Verification.** Three frontier models (GPT-5.4, Gemini 3.1 Pro, Claude Opus 4.6) independently answer each question; those with unanimous agreement are retained. (3) **Human review.** Questions without consensus are revised or discarded by human reviewers. A core team of five STEM-trained annotators performs this review, with specialist consultation for questions outside the team's expertise. Appendix C describes the protocol, and Appendix C.1 reports the original independent audit and its pairwise agreement rates.

3.2.4 TIER 2: PRUNED PROBLEMS

Tier 2 addresses the text-shortcut problem directly: we rewrite each source problem's text to remove information that is already conveyed by the image, forcing models to actively perceive and extract that information visually. These overlaps include repeated labels and values, visible spatial or topological relationships, and diagram interpretations licensed by the problem statement. Each pruned problem is compared with its original.

Pruning follows the intended reading of the complete problem, including its diagram conventions. For example, retained wording such as *arranged as shown* can refer the reader to the arrangement depicted in the image. Review checks that the proposed removal is supported by this combined context, preserves necessary text-only givens, and leaves the target quantity and reference answer unchanged. Approximate visual estimates are distinguished from exact diagram conditions when assessing that support. This design targets several forms of text-image redundancy within the source tasks; modality ablations then measure how pruning changes reliance on images.

We prioritize *information completeness* over aggressive pruning: a pruned problem that is missing necessary information is unsolvable and thus useless for evaluation. The construction follows a four-step propose-and-verify pipeline with built-in safeguards:

(1) **Proposal.** GPT-5.4 identifies text spans redundant with the image and atomic facts, replaces them with `<MASK: [hint]>` tags, and then rewrites the result into a natural-language pruned problem. (2) **Recovery verification.** In a fresh context, the model reconstructs the masked content from the image and atomic facts, and a separate model instance checks whether the recovered problem faithfully matches the original. This checks recoverability with the supplied image and text-informed fact inventory. (3) **Coverage verification.** The model verifies that every atomic fact required for solving the problem remains available in at least one modality, either stated in the pruned text or visible in the image. This check assesses coverage relative to the extracted fact inventory. (4) **Human review.** Problems that fail either verification stage are revised or discarded. Reviewers check the proposed removals and retained statement together, including image references, necessary givens, answer options, fluency, and answer validity.

3.2.5 TIER 3: EDITED VARIANTS

Tier 3 evaluates whether models can generalize to novel visual configurations or exhibit reduced generalization when problem instances change. We create new problem variants by editing the problem text, diagrams, or both and adjusting the questions and solutions accordingly. The evaluated set contains 283 human-verified variants, each paired with its source problem.

A key design choice is using *code as the intermediate representation*: rather than applying image editing models directly (structured text and symbols are challenging to current generative models (Zhang et al., 2025)), we first reproduce each image programmatically using rendering tools (*e.g.*, TikZ, Matplotlib), then modify the code to create variants. The code representation makes edits inspectable and reproducible; model checks and final human review assess their semantic validity. We implement this through a multi-agent workflow (detailed in Appendix F) with five specialized roles: a **Coder** reproduces the source diagram and updates it for the edited problem; a **Reviewer** checks

the reproduction against the source; an **Editor** proposes the changed problem and answer; a **Verifier** checks the proposed solution; and an **Auditor** checks consistency between the updated drawing and the edited problem. These checks provide feedback for revision before final human review. Every retained question variant receives a human check of visual quality, image-text consistency, solvability, answer correctness, and the absence of added solution hints or revealed solution steps.

Preserving challenging problems. Relying on models to generate new problem variants introduces the risk of generating overly simplistic tasks. A naive pipeline would only produce variants simple enough for the Verifier to solve independently, biasing Tier 3 toward easy problems. Our design counters this through three mechanisms. First, the Verifier receives the original problem's step-by-step solution as a reference, enabling it to verify variants that exceed its standalone solving capability. Second, the multi-agent architecture distributes generation and verification across complementary models, reducing the chance that any single model's limitations constrain problem difficulty. Third, a final round of human verification checks correctness and quality. Post-hoc difficulty annotation (see Appendix E.5) confirms that 75.3% of edited variants maintain the same difficulty level as the original, validating this approach.

3.3 Construction statistics and human review

Human review covers flagged Tier 1 and Tier 2 candidates and every retained Tier 3 variant. Tier 2 candidates with trivial or minor pruning are excluded after quality control. Appendix C provides the construction workflow counts (Table 4), review protocol, and independent agreement results.

4 Experiments

4.1 Setup

Models. We evaluate a diverse set of frontier and open-weight MLLMs, including:

- **Proprietary:** GPT-5.4, GPT-5.4 mini, Gemini 3.1 Pro, Gemini 3.1 Flash-Lite, Claude Opus 4.6, and Claude Sonnet 4.6.
- **Open-weight:** Qwen3-VL-235B-A22B, Qwen3.5-397B-A17B, Gemma 4 31B, and Kimi K2.5.

All tiers, including Tier 1 perception, are evaluated with thinking enabled. Claude Sonnet 4.6 uses low reasoning effort; Claude Opus 4.6, GPT-5.4, and Qwen3-VL-235B use high reasoning effort. Appendix H.1 gives the answering and scoring prompts and the available inference settings.

Baselines. For Tier 2, we additionally report text-only LLM baselines (no image input) to quantify residual text solvability.

Evaluation protocol. All models are evaluated in a zero-shot setting. Tier 1 answers are scored by direct option-letter matching. For the original, Tier 2, and Tier 3 problems, Claude Opus 4.8 judges both multiple-choice and free-form answers against the reference answer using the prompt in Appendix H. For multiple-choice answers, option identification and formatting normalization are instructions to the judge, rather than a deterministic parser followed by a judge fallback. For free-form answers, the judge assesses semantic equivalence, including equivalent expressions and rounding differences.

Metrics. We report accuracy as the fraction of evaluated items judged correct, together with per-tier, per-domain, and per-atomic-fact-category breakdowns. Original and transformed outcomes are paired by source item. Original and transformed outcomes are paired by source item. Appendix E.2 reports paired significance tests, while Appendix E.3 reports variation across repeated evaluation runs.

4.2 Main Results

Tables 2 and 3 present results on the original problems and across all three tiers. We highlight several observations.

Overall performance on original problems. Table 2 reports accuracy on the full set of 2,711 source problems. GPT-5.4 achieves the highest overall accuracy (78.9%), followed by Gemini 3.1 Pro (75.2%) and Qwen3.5-397B (66.7%). A clear gap separates top-tier models (GPT-5.4, Gemini 3.1

Table 2: **Overall and per-domain accuracy (%) on the original 2,711 source problems.** Domain sizes: Physics (926), Math (827), Chemistry (568), Geography (174), Biology (168), Puzzle (26), Engineering (12), CS (10). Domains with fewer than 100 samples (Puzzle, Eng., CS) may exhibit high variance; best results are not highlighted in these columns.

Model	Overall	Phys.	Math	Chem.	Geo.	Bio.	Puzzle	Eng.	CS
Proprietary models									
GPT-5.4	**78.9**	**74**	**91**	**73**	71	**71**	81	100	100
GPT-5.4 mini	57.6	55	69	50	51	48	50	67	90
Gemini 3.1 Pro	75.2	72	86	66	**72**	**71**	85	75	90
Gemini 3.1 Flash-Lite	48.8	46	56	45	44	46	46	25	50
Claude Opus 4.6	61.0	65	65	51	56	57	54	42	80
Claude Sonnet 4.6	57.9	61	64	48	52	51	69	8	70
Open-weight models									
Qwen3-VL-235B	33.4	29	37	32	35	41	35	25	50
Qwen3.5-397B	66.7	65	80	51	66	63	77	42	70
Gemma 4 31B	55.7	48	70	48	49	54	77	33	80
Kimi K2.5	56.1	56	67	46	45	50	69	33	40

Table 3: **Results across all three tiers.** Tier 1: perception accuracy (%) on 11,094 sub-questions. Tier 2: accuracy (%) on 1,035 pruned problems vs. the same problems with original text. Tier 3: accuracy (%) on 283 edited problems vs. the same problems in original form. Δ denotes the accuracy change in %.

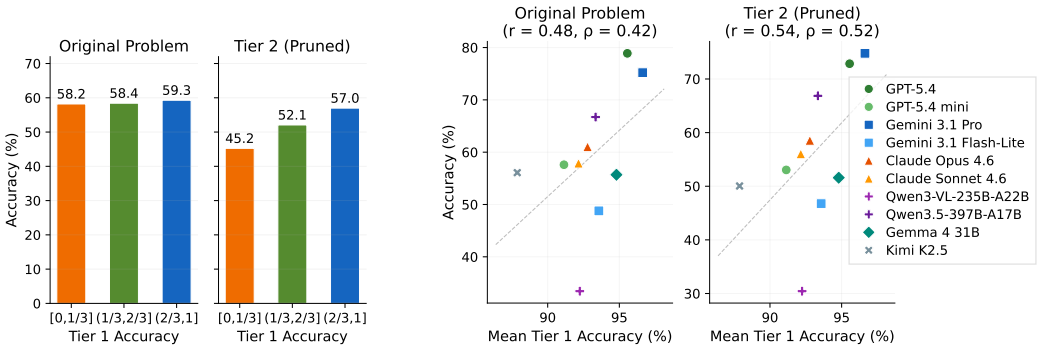
Model	Tier 1: Perception				Tier 2: Pruning			Tier 3: Editing		
	Entity	Topo.	Estim.	Overall	Orig.	Pruned	Δ	Orig.	Edited	Δ
Proprietary models										
GPT-5.4	99.1	97.9	89.6	95.6	**82.6**	72.9	−9.8	**96.5**	84.1	−12.4
GPT-5.4 mini	98.7	93.0	81.8	91.3	60.1	53.0	−7.1	74.9	70.0	−4.9
Gemini 3.1 Pro	**99.6**	**98.0**	**92.3**	**96.8**	78.7	**74.8**	−4.0	90.5	**89.0**	−1.4
Gemini 3.1 Flash-Lite	98.8	96.0	86.1	93.8	52.0	46.8	−5.2	64.0	64.3	+0.4
Claude Opus 4.6	98.4	93.4	87.0	93.0	69.6	58.6	−11.0	70.7	62.2	−8.5
Claude Sonnet 4.6	98.0	93.8	84.2	92.1	64.3	56.0	−8.2	65.7	66.1	+0.4
Open-weight models										
Qwen3-VL-235B	98.5	94.4	84.0	92.4	35.8	30.4	−5.4	41.3	48.8	+7.4
Qwen3.5-397B	98.7	96.3	85.4	93.5	72.2	66.9	−5.3	87.3	84.1	−3.2
Gemma 4 31B	99.1	96.9	88.6	95.0	57.5	51.6	−5.9	77.7	76.0	−1.8
Kimi K2.5	96.1	87.3	81.3	88.4	59.1	50.0	−9.1	67.8	69.6	+1.8

Pro) from mid-tier ones.Open-weight models span a wide range, with Qwen3.5-397B competitive at 66.7% while Qwen3-VL-235B lags at 33.4%.

Per-domain breakdown. The per-domain columns of Table 2 reveal that model performance varies substantially across disciplines. Unlike prior benchmarks that focus on a single domain (*e.g.*, mathematics), VISATOM spans 8 domains, enabling cross-domain comparison. GPT-5.4 and Gemini 3.1 Pro lead across most domains, with GPT-5.4 achieving 91% on Mathematics and Gemini 3.1 Pro reaching 86%. Chemistry proves challenging for open-weight models, with Qwen3-VL-235B achieving only 32% and Kimi K2.5 46%. Domain difficulty rankings are broadly consistent across models, with Physics, Mathematics, and Chemistry forming the core of challenging problems.

Tier 1: Perception is not yet solved. Table 3 shows that overall perception accuracy ranges from 88.4% (Kimi K2.5) to 96.8% (Gemini 3.1 Pro). Entity recognition is near-ceiling (96.1-99.6%), but the gap widens for topology (87.3-98.0%) and especially visual estimation (81.3-92.3%), which is consistently the weakest subcategory with an 11.0% spread. Since each perception error can cascade into a reasoning failure, these error rates have outsized impact on downstream problem-solving. We show in Section 4.3 that perception and reasoning accuracy are positively correlated, with a steeper problem-level trend for Tier 2 problems (Figure 3).

Tier 2: Text removal causes universal performance drops. Comparing Tier 2 accuracy to the original accuracy on the *same* 1,035 problems (Table 3) reveals that removing textual shortcuts degrades performance for *all* models without exception. The drops range from 4.0% to 11.0%, from Gemini 3.1 Pro (−4.0%) to Claude Opus 4.6 (−11.0%). The universal degradation confirms that



(a) Reasoning accuracy by perception score bin, pooled across all models.

(b) Model-level correlation between mean perception and reasoning accuracy.

Figure 3: **Tier 1 perception accuracy correlates with reasoning performance.** Left: problems with higher perception scores yield higher reasoning accuracy (binned analysis). Right: models with better perception achieve better reasoning (model-level correlation).

existing benchmarks provide textual shortcuts that inflate multimodal reasoning scores, and that genuine vision-dependent reasoning remains substantially harder than aggregate accuracy suggests.

The modality ablation below complements this paired comparison by varying the availability of images and textual descriptions. It shows a larger image-removal penalty after pruning, providing further evidence that reducing textual shortcuts shifts the information models rely on. Paired significance tests for the 1,035-problem comparison are reported in Appendix E.2.

Tier 3: Strong models show reduced generalization. The Tier 3 results (Table 3) reveal clear generalization gaps for the strongest models. GPT-5.4 exhibits the largest drop (-12.4%, from 96.5% to 84.1%), followed by Claude Opus 4.6 (-8.5%) and GPT-5.4 mini (-4.9%). These drops indicate degraded generalization when problem instances change. Gemini 3.1 Pro, the second-strongest model, shows only a modest decline (-1.4%), suggesting more robust generalization. Weaker models that performed poorly on the originals show near-zero or small positive deltas (*e.g.*, Qwen3-VL-235B $+7.4$, Kimi K2.5 $+1.8\%$), indicating that they have less room for degradation and may benefit from the cleaner programmatic renderings used in Tier 3.

4.3 Analysis

4.3.1 Perception-Reasoning Correlation

We investigate whether Tier 1 perception accuracy is correlated with downstream reasoning performance (Figure 3).

Problem-level analysis. At the problem level, we bin examples by Tier 1 perception score and measure pooled reasoning accuracy across all 10 models. Higher perception scores are associated with better reasoning performance, with a substantially steeper relationship on Tier 2 pruned problems than on the original problems. The steeper trend for Tier 2 is consistent with visual perception being more critical when textual shortcuts are removed, while the flatter trend for original problems suggests that text shortcuts can partially compensate for perception failures. These trends describe associations; the pooled bins may also differ in item difficulty, domain, and model competence.

Model-level analysis. We compute each model's mean Tier 1 accuracy and correlate it with its reasoning accuracy on the original problems and Tier 2 pruned problems. The results show a positive correlation between perception and reasoning: Pearson $r = 0.48$ for original problems and $r = 0.54$ for Tier 2. These results suggest that visual perception is associated with multimodal reasoning performance, especially on vision-dependent tasks, though it is not the sole determinant, as reasoning ability and domain knowledge also play significant roles.

4.3.2 Modality Ablation and Pruning Validation

To understand the role of each input modality, we conduct an ablation study on a randomly sampled 200-item subset of the 1,035-problem Tier 2 evaluation using four models (GPT-5.4, Gemini 3.1

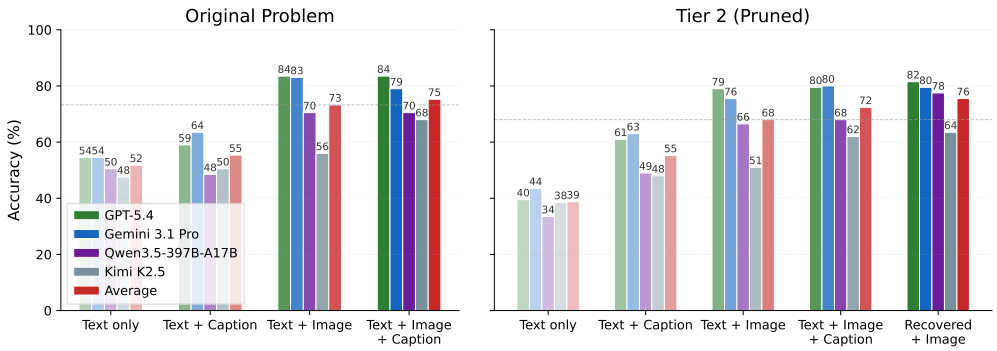


Figure 4: **Modality ablation on original and Tier 2 pruned problems** (random 200-item subset, 4 models). Each group varies the availability of images and atomic fact captions. Dashed lines mark the canonical (text + image) average.

Pro, Qwen3.5-397B, Kimi K2.5). We vary the input across five settings: (1) **text only**; (2) **text + captions** (atomic facts as text, replacing images); (3) **text + image** (canonical); (4) **text + image + captions**; and (5) **recovered question + image** (Tier 2 only: pruned text with masked content re-inserted through the recovery procedure using the image and atomic facts; Section 3.2.4). Figure 4 summarizes the results.

Image availability matters more after pruning. Removing images causes a larger accuracy drop on pruned than original problems (29.2 vs. 21.4 pp), consistent with greater reliance on image-provided information after pruning. This suggests that pruning reduces textual cues redundant with the image.

Captions partially bridge the modality gap. Replacing images with atomic fact captions improves over text-only input, especially on pruned problems (+16.4 vs. +3.6 pp), consistent with pruned text lacking more redundant visual descriptions. However, captions remain below actual images on pruned problems (55.2% vs. 68.0%), showing that the supplied atomic facts do not fully substitute for direct visual perception.

Adding captions to images provides modest benefit. Adding captions to canonical image input further improves accuracy, with a larger gain on pruned problems (4.4 vs. 2.0 pp). This suggests that explicit descriptions can help models use visual information that may otherwise be missed during reasoning.

Restoring explicit information improves performance. On Tier 2, replacing pruned text with recovered questions, where masked content is filled using image- and text-informed atomic facts (Section 3.2.4), improves accuracy from 68.0% to 75.5%. This supports the view that the removed information remains recoverable from the image and its atomic-fact descriptions under the diagram conventions licensed by the problem statement. The fact that recovered questions outperform canonical pruned questions suggests that making visual information explicit in text helps models use that information, with Kimi K2.5 (+12.5%) and Qwen3.5-397B (+11.0%) showing the largest gains. This ordering (pruned text only $<$ pruned + image $<$ pruned + image + caption $<$ recovered + image) illustrates the utility of atomic facts as a diagnostic bridge between visual and textual modalities. The paired pruning results and larger image-removal penalty after pruning provide complementary evidence of reliance on textual shortcuts.

Additional analyses on pruning ratio effects and a detailed breakdown of the Tier 3 generalization gap by difficulty level, edit type, and source dataset for the 283 pairs are in Appendix E.

5 Conclusion

We presented VisAtom, a vision-centric methodology and benchmark that uses atomic facts to decompose multimodal reasoning evaluation into three complementary tiers: perception probing, pruned problems, and edited variants. These tasks enable joint analysis of perception and transformed-task performance. Across 10 models, VisAtom reveals persistent visual estimation weaknesses, universal 4-11% degradation from text pruning, and a 12.4% generalization gap for the strongest model on edited variants, showing that aggregate accuracy can mask brittle visual reasoning.

AI USE STATEMENT

Generative AI is used in the benchmark construction and evaluation procedures described in Section 3 and the Appendix. These uses include atomic-fact extraction and revision, question generation and verification, text pruning and recovery checks, code-based diagram reproduction and editing, and model-answer evaluation. AI-generated benchmark items undergo the automated checks and tier-specific human review described in the paper: human review covers flagged Tier 1 and original Tier 2 items, every retained conservative Tier 2 item, and every retained Tier 3 variant.

We also used generative AI to assist with code implementation, manuscript revision, and literature and reference checking. The authors developed the research questions and experimental design without generative AI assistance. The authors review and revise AI-assisted manuscript content and retain responsibility for the correctness, originality, and integrity of the work, including the code, references, analyses, and conclusions.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *ACL Findings*, 2024.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *ACL Findings*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced multimodal reasoning benchmark. In *ICML*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems. In *ACL*, 2024.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. OlympicArena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI. In *NeurIPS Datasets and Benchmarks*, 2024.

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 the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL-IJCNLP*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*, 2022.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024b.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

Jiakang Yuan, Tianshuo Peng, Yilei Jiang, Yiting Lu, Renrui Zhang, Kaituo Feng, Chaoyou Fu, Tao Chen, Lei Bai, Bo Zhang, and Xiangyu Yue. MME-Reasoning: A comprehensive benchmark for logical reasoning in MLLMs. *arXiv preprint arXiv:2505.21327*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

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. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele (Mike) Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *NeurIPS Datasets and Benchmarks*, 2024a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024b.

Tianyu Zhang, Xinyu Wang, Lu Li, Zhenghan Tai, Jijun Chi, Jingrui Tian, Hailin He, and Suyuchen Wang. STRICT: Stress-test of rendering image containing text. In *EMNLP*, 2025.

Zehua Zhao, Zhixian Huang, Junren Li, Siyu Lin, Junting Zhou, Fengqi Cao, Kun Zhou, Rui Ge, Tingting Long, Yuexiang Zhu, Yan Liu, Jie Zheng, Junnian Wei, Rong Zhu, Peng Zou, Wenyu Li, Zekai Cheng, Tian Ding, Yaxuan Wang, Yizhao Yan, Tingru Wei, Haowei Ming, Weijie Mao, Chen Sun, Yiming Liu, Zichen Wang, Zuo Zhang, Tong Yang, Hao Ma, Zhen Gao, and Jian Pei. SUPERChem: A multimodal reasoning benchmark in chemistry. *arXiv preprint arXiv:2512.01274*, 2025.

A LIMITATIONS

Atomic fact completeness. Our atomic fact extraction, while systematic, may not capture all visual information relevant to solving a problem. Subtle visual cues (*e.g.*, line thickness, color difference) may be missed. We mitigate this through human review, but acknowledge that completeness is difficult to guarantee.

Tier 3 edit diversity. Current Tier 3 edits are predominantly local modifications such as changing numeric conditions or asking about different elements in the same diagram. More fundamental structural changes (*e.g.*, loosening geometric constraints or altering problem topology) remain challenging for automated pipelines to generate reliably. Future iterations could target deeper structural edits as model generation capabilities improve, or incorporate human-authored variants alongside model-generated ones.

Tier 3 validation and domain scope. Model-assisted editing and verification can share errors and can favor variants that the verifying models can solve. Final human review checks the retained problem, diagram, and answer; this does not establish the residual error rate of the retained set. Tier 3 edits can change difficulty and rendering, so model-specific performance changes are examined alongside the post-hoc analyses in Appendix E.5. The generation workflow and final-image review are described in Appendix F.

Contamination residue. Despite filtering and deduplication, some source problems may still appear in model training data in altered forms (Dong et al., 2024; Zhang et al., 2024a; White et al., 2025). Tier 3 partially addresses this by creating novel variants, but Tier 1 and Tier 2 are built on potentially seen images. Future work could incorporate private held-out test sets to bound contamination more tightly.

B ETHICS, BROADER IMPACT, AND RELEASE PLAN

Data provenance. Source problems are drawn from publicly available benchmarks. The original source of each problem is recorded in the metadata to ensure proper attribution.

Broader impact. VISATOM is designed to improve the evaluation of multimodal AI systems, contributing to a more accurate understanding of their capabilities and limitations. We do not anticipate direct negative societal impacts from the benchmark itself, as it contains only scientific and mathematical problems without sensitive content. By revealing gaps in visual reasoning, VISATOM may help prevent overconfident deployment of multimodal models in vision-critical applications (*e.g.*, medical imaging, autonomous systems).

Release plan. We plan to release: (1) the full benchmark dataset with metadata and atomic facts; (2) evaluation scripts and automated scoring tools; and (3) the construction pipeline and prompt templates.

C HUMAN REVIEW PROTOCOL

Human review is organized by the unit being validated: a Tier 1 perception question and answer, a Tier 2 pruned problem paired with its original, or a Tier 3 edited problem with its diagram and solution. All reviews are conducted by the same core team of five computer science researchers: two undergraduates and three PhD researchers. The core team reviews most questions without external input. For the few questions requiring specialist input, reviewers meet with PhD or postdoctoral researchers in the relevant domain to resolve the uncertainties.

For jointly annotated acceptance decisions, an item is accepted if and only if $n_a > n/2$, where n_a is the number of acceptance votes among n reviewers. Thus, a tie does not lead to acceptance.

Review coverage is tier- and policy-specific. Tier 1 uses human review for questions flagged by disagreement among the verifying models, and aggressive Tier 2 flags failed recovery or coverage checks. Conservative Tier 2 requires human verification of every retained item. All 283 evaluated Tier 3 variants receive final human checks of the problem, answer, and exported diagram. Review excludes variants that provide extra solution guidance by making relevant conditions or solution steps explicit in the text or diagram.

Original construction workflow. Table 4 quantifies the workflow across the three tiers. Tier 1 automatically accepts 9,808 questions and retains another 1,286 after review, yielding 11,094 perception questions. Of the 2,711 source problems, 2,615 yield proposed Tier 2 candidates; the remaining 96 do not successfully complete the workflow. Among the proposed candidates, 1,803 pass automatic checks and another 424 pass after human review, yielding 2,227 quality-control-passed candidates. We then exclude 1,192 candidates with trivial or minor pruning and select 1,035 with non-negligible character-level differences from the original questions for evaluation. Tier 3 retains 283 of 418 candidates, with final human verification of every retained variant. These counts distinguish quality control, subsequent pruning-based selection, and the population actually evaluated.

Table 4: Construction workflow statistics for the main benchmark. The human-review pool is the set routed to review; accepted-from-review counts are additional to automatic acceptance. The Tier 2 proposed pool excludes 96 source problems whose workflows did not complete. Tier 2 acceptance counts precede the exclusion of 1,192 trivial/minor-pruning candidates, leaving 1,035 evaluated problems. The table does not assign those later exclusions to either acceptance route. For Tier 3, accepted-from-review counts reflect all review rounds; the *Not retained* column includes deferred and rejected items. A dash denotes no comparable automatic-only acceptance count.

Tier / unit	Proposed	Auto accepted	Review pool	Accepted from review	Not retained
Tier 1 / question	11,529	9,808	1,721	1,286	435
Tier 2 / problem	2,615	1,803	812	424	388
Tier 3 / variant	418	-	418	283	135

Conservative Tier 2 review pool. All five reviewers work on the same 200 conservative Tier 2 candidates for shared rubric calibration, followed by non-overlapping individual assignments. The review pool contains 943 candidates. The updated record retains 387, rejects 498, and excludes 58 Chinese-language candidates without the assigned review, accounting for all 943. Thus, 556 are not retained; the 943-item review-pool size should be distinguished from the 885 items with recorded accept/reject decisions. The shared calibration and later assignments are parts of this pool, not additional items.

C.1 Independent audit of the original release

The original audit used Tier 1-2 material derived from 100 randomly sampled source problems and 100 randomly sampled Tier 3 variants from the 283-question split, with four, three, and two annotators assigned to the respective tiers. Annotators applied the tier-specific criteria described in this paper and recorded labels independently before adjudication. The original audit verdicts were retained without post-discussion changes. A source problem can produce multiple Tier 1 questions, so the number of source problems is distinct from the number of audited decisions and the overlap for an annotator pair. Within the randomly sampled source problems, the Tier 1 audit included only questions flagged by a lack of model consensus; its agreement rates do not estimate a random sample's residual error rate across the full benchmark.

Table 5: Independent agreement in the original audit, before adjudication. Tier 3 audits variants from the 283-question split. Rates are pairwise label agreement, not correctness estimates.

Audit	Decision and reported comparisons	Agreement (%)
Tier 1 (4 annotators)	Review decision; four reported pair comparisons	96.8, 92.6, 90.0, 88.8
Tier 2 (3 annotators)	Prunability; three pair comparisons	90.9, 72.7, 70.9
Tier 3 (2 annotators)	Variant validity	88.1
Tier 3 (2 annotators)	Relative difficulty: exact label match	70.5

Agreement is the fraction of common, valid decisions on which a pair assigns the same label under the stated label mapping. For Tier 1, revise and discard decisions are merged because valid revisions need not be unique; this measures agreement on the review decision, not equivalence of the revised wording. The four reported comparisons are a subset of the six possible pairs among four annotators. For Tier 2, disagreement analysis identified a rubric difference: one annotator also treated stylistic

shortening without image-overlapping content as pruning. This motivated clarification that pruning must concern information supported by the image.

Tier 3 validity agreement is separate from the easier/same/harder difficulty annotation. Difficulty labels agree exactly in 70.5% of compared cases, and no comparison assigns easier by one annotator and harder by the other. The absence of opposite judgments does not imply exact agreement. Jointly annotated acceptance decisions follow the strict-majority rule above; neither majority voting nor inter-annotator agreement establishes the residual error rate of retained items.

C.2 Conservative Tier 2 calibration and Tier 3 review

All five reviewers worked on the shared first 200 candidates in conservative Tier 2 as a calibration exercise distinct from the original independent audit. Reviewers revisited decisions that conflicted with others' annotations. When persuaded by another reviewer's reasoning, they changed their verdict and adjusted their rubric; otherwise, they recorded a justification in the notes. Remaining acceptance disagreements were resolved by the same strict-majority rule. The subsequent assignment schedule divided the remaining candidates into non-overlapping batches of approximately 150 per reviewer. Only reconciled labels were retained; the initial independent labels are unavailable. Consequently, these records cannot support a pre-discussion inter-annotator agreement estimate, and any agreement computed from them would describe post-discussion consistency.

Each Tier 3 candidate is assigned to one reviewer. This reviewer's verdict is retained, and any revisions made during review are visible to the other reviewers. Tier 3 review approves a valid item, including one approved after a completed quick fix. Ambiguity or an unfinished repair leads to deferral; fundamental invalidity or major repair leads to rejection. Final acceptance checks the exported diagram together with the problem and answer. Of 418 candidates, 283 are accepted.

D Problem Preprocessing Details

D.1 Source Selection and Problem Filtering

We aggregate problems from 12 established multimodal reasoning benchmarks, selecting only problems that include at least one image. To ensure sufficient difficulty, we filter using lightweight variants of frontier models (GPT-5 nano, Gemini 3 Flash, and Claude Haiku 4.5) with chain-of-thought reasoning disabled. A problem is retained only if *none* of these three models produces the correct answer in a single pass.

D.2 Deduplication

Problems sourced from 12 benchmarks inevitably contain overlaps, as several datasets draw from similar competition pools. We encode each multimodal problem (text and image jointly) into a dense feature vector using the Gemini Embedding 2 model. To avoid false positive detections caused by shared problem templates within a single dataset, we only consider problem pairs that span across different source benchmarks. We construct a similarity graph and iteratively remove near-duplicates: at each step, we identify the cross-benchmark pair with the highest cosine similarity above a threshold of 0.9 (selected empirically by inspecting pairs at various similarity levels). Between the two problems in that pair, we remove the one with more remaining duplicates (*i.e.*, the node with higher degree in the similarity graph), as it is more likely to be a repeated instance across datasets. This process repeats until no above-threshold pairs remain, removing the cross-benchmark near-duplicates identified by this procedure.

E Additional Analysis

E.1 Additional test with conservative pruning

We additionally test whether removing a narrower subset of textual shortcuts also reduces accuracy. This conservative variant limits pruning to information explicitly encoded in the image or supported by unambiguous visible topology, retaining conditions that rely only on an unmarked geometric

Table 6: Conservative Tier 2 on its 387 approved seed problems. Each original is paired with its retained conservative variant. Accuracies are percentages; Δ is conservative minus original in %.

Model	Original (%)	Conservative (%)	Δ (%)
GPT-5.4	77.00	72.09	-4.91
Claude Opus 4.6	63.31	59.43	-3.88
Claude Sonnet 4.6	58.14	50.90	-7.24
Qwen3-VL-235B	41.86	37.47	-4.39

appearance. The experiment complements the broader pruning design by testing model sensitivity to partial shortcut removal.

Design and construction. Conservative pruning permits explicitly encoded information and unambiguously visible topology, such as identifiable point-segment incidence or wire connectivity. It retains exact unmarked geometric conditions when their only support is that they appear true in the drawing: apparent contact, a right-looking angle, similar lengths, or parallel-looking strokes alone do not justify deleting tangency, perpendicularity, equality, or parallelism. A printed value or conventional symbol supplies stronger evidence than an unmarked estimate. Thus a length or angle written in the diagram, a right-angle mark, or an unambiguous direction symbol can support removal, while values absent from the diagram and conditions requiring mathematical derivation remain in the statement. Ambiguous incidence or connectivity is retained rather than inferred from proximity. Review checks that the pruned statement preserves image-reference tokens, the target quantity, answer options, and necessary text-only givens. A connective phrase such as *as shown* may repair wording but does not replace the conservative evidence criterion. The reference answer must remain valid for the retained problem. Conservative variants are regenerated from the original problems using the same four-stage proposal, recovery, coverage, and human-review workflow described in Section 3.2.4. Prompt changes impose the narrower evidence criterion. Recovery uses the image and atomic facts extracted with source-text context; coverage checks the available fact inventory. Every retained conservative variant receives human verification. Additional LLM checks provide advisory verdicts with supporting reasons to human reviewers in conservative mode. The checks have access to the atomic-fact descriptions. Candidates are checked first by Claude Opus 4.8; those not rejected proceed to GPT-5.6 Sol for a second check. A candidate rejected by Opus skips the GPT check but is still presented for human review. The model judgments serve only as reference feedback; human review determines the final acceptance or rejection. Eligibility requires substantive fact removal: candidates with no removable facts and paraphrase-only rewrites are excluded from Tier 2. Of 943 candidates routed to the review pool, 387 are retained, 498 rejected, and 58 Chinese-language candidates excluded without the assigned review. Thus, 556 are not retained, and the 943-candidate pool contains 885 recorded accept/reject decisions. No comparable automatic-only acceptance count is reported. Review begins with a shared 200-candidate calibration exercise and continues with individual assignments; only reconciled calibration labels remain (Appendix C).

Results on the conservative seed set. Model responses were generated anew for the conservative evaluation and the comparison on shared retained seeds, including responses to the original and aggressively pruned conditions; responses from the earlier evaluation were not reused. Table 6 compares the 387 retained variants with their paired originals. All four evaluated models lose accuracy, by 3.88-7.24%. Performance therefore remains sensitive to removing textual cues even when pruning is limited to this narrower class of image-supported information. The original/conservative correct counts are 298/279 for GPT-5.4, 245/230 for Claude Opus 4.6, 225/197 for Claude Sonnet 4.6, and 162/145 for Qwen3-VL-235B, each out of 387. The main Tier 2 evaluation separately reports the original pruning design on its own 1,035 seed problems (Table 3). These seed sets differ, so their aggregate accuracies are not a matched comparison.

Comparison on shared retained seeds. The two constructions have 232 source problems with retained variants under both designs. Table 7 evaluates the originals and both variants on this shared set, holding source membership fixed. We use A for the original, broader pruning design (termed aggressive in the construction records) and C for the conservative variant. On the shared set, conservative pruning lowers accuracy by 2.59-10.34%, compared with 9.91-20.26% for the original design. The smaller but consistent conservative drops support the same shortcut finding after partial removal. For example, GPT-5.4 changes from 79.31% on the originals to 73.28% under conservative

Table 7: Tier 2 policy comparison on the same 232 source problems retained under both modes. O = original; A = aggressive (convention-licensed); C = conservative. Accuracies are percentages, and differences are in %. All three conditions use the same source membership.

Model	O (%)	A (%)	C (%)	A-O (%)	C-O (%)	C-A (%)
GPT-5.4	79.31	59.05	73.28	-20.26	-6.03	14.22
Claude Opus 4.6	67.24	50.86	62.50	-16.38	-4.74	11.64
Claude Sonnet 4.6	64.22	49.14	53.88	-15.09	-10.34	4.74
Qwen3-VL-235B	41.81	31.90	39.22	-9.91	-2.59	7.33

pruning and 59.05% under the original design. The original/A/C correct counts are 184/137/170 for GPT-5.4, 156/118/145 for Claude Opus 4.6, 149/114/125 for Claude Sonnet 4.6, and 97/74/91 for Qwen3-VL-235B. The separately recorded old-pruned scores on all 387 conservative seeds include variants rejected under the original design and are excluded from the valid retained-set comparison. Marginal counts establish the accuracies; paired discordant outcomes are required for significance tests.

E.2 Paired accuracy comparisons

Table 9 reports the exact paired McNemar tests from the original-release evaluation, using 1,035 original/pruned pairs for Tier 2 and 283 original/edited pairs for Tier 3. Let b count pairs correct on the original and incorrect on the transformed item, and c count the reverse. The accuracy change is $100(c-b)/n\%$. Missing or failed model answers are counted as incorrect and remain in the paired evaluation denominator. The two-sided exact test conditions on the $b+c$ discordant pairs and compares their split with a binomial null probability of $1/2$. It tests the paired correctness change on this evaluated set, not its cause. The exact contingency counts from the rebuttal evaluation are given in Table 8; they also recover the original/transformed correct totals as $a+b$ and $a+c$.

Table 8: Paired correctness counts for the original aggressive Tier 2 and the 283-question Tier 3 split. a: both correct; b: original only correct; c: transformed only correct; d: both incorrect.

	Tier 2 aggressive ($n = 1{,}035$)				Tier 3 ($n = 283$)			
Model	b	c	a	d	b	c	a	d
GPT-5.4	158	57	697	123	42	7	231	3
GPT-5.4 mini	170	97	452	316	54	40	158	31
Gemini 3.1 Pro	109	68	706	152	25	21	231	6
Gemini 3.1 Flash-Lite	166	112	372	385	43	44	138	58
Claude Opus 4.6	172	58	548	257	62	38	138	45
Claude Sonnet 4.6	171	86	494	284	47	48	139	49
Qwen3-VL-235B	160	104	211	560	45	66	72	100
Qwen3.5-397B	142	87	605	201	38	29	209	7
Gemma 4 31B	162	101	433	339	43	38	177	25
Kimi K2.5	223	129	389	294	43	48	149	43

We report nominal tests and additionally apply a Holm family-wise correction at $\alpha = 0.05$ across the ten models and two tiers as a multiplicity sensitivity analysis. Under this reporting convention, all ten original-policy Tier 2 drops and the GPT-5.4 Tier 3 drop remain significant. Claude Opus 4.6's Tier 3 result is nominal ($p = 0.0210$; adjusted $p \approx 0.189$); the other Tier 3 changes are not significant even before correction. Failure to reject does not establish equivalence or robust generalization.

Table 9: Original-release paired accuracy changes (%) and nominal exact McNemar p-values. Holm-adjusted values use all 20 comparisons as one family and are computed from the exact paired counts before rounding.

	Tier 2 ($n = 1{,}035$)			Tier 3 ($n = 283$)		
Model	Δ	Nominal p	Holm p	Δ	Nominal p	Holm p
GPT-5.4	-9.8	3.63×10^{-12}	6.90×10^{-11}	-12.4	3.62×10^{-7}	6.16×10^{-6}
GPT-5.4 mini	-7.1	9.32×10^{-6}	1.40×10^{-4}	-4.9	0.1797	1.0000
Gemini 3.1 Pro	-4.0	0.0025	0.0254	-1.4	0.6587	1.0000
Gemini 3.1 Flash-Lite	-5.2	0.0014	0.0158	+0.4	1.0000	1.0000
Claude Opus 4.6	-11.0	2.68×10^{-14}	5.36×10^{-13}	-8.5	0.0210	0.1888
Claude Sonnet 4.6	-8.2	1.25×10^{-7}	2.24×10^{-6}	+0.4	1.0000	1.0000
Qwen3-VL-235B	-5.4	6.83×10^{-4}	0.0082	+7.4	0.0572	0.4573
Qwen3.5-397B	-5.3	3.38×10^{-4}	0.0044	-3.2	0.3284	1.0000
Gemma 4 31B	-5.9	2.03×10^{-4}	0.0028	-1.8	0.6570	1.0000
Kimi K2.5	-9.1	6.19×10^{-7}	9.90×10^{-6}	+1.8	0.6752	1.0000

E.3 Variation across evaluation runs

We conduct three additional evaluation runs using a separate evaluation setup from Table 3. Tier 2 uses the same 1,035 original/pruned pairs. Each run generates fresh target-model responses for all tiers, which Claude Opus 4.8 scores for the original, Tier 2, and Tier 3 problems. Tier 3 results concern the 283 pairs. Table 10 reports the four-model results as mean accuracy and standard deviation across runs, with each accuracy difference computed within a run. The variability reflects the combined generation and scoring process within this evaluation setup and is distinct from uncertainty due to benchmark-item sampling. Tier 1 uses deterministic letter scoring; its reported accuracies have zero standard deviation at the displayed precision across these three runs.

Table 10: Four-model evaluation over three additional runs: mean $\pm$ standard deviation. Accuracies are percentages; differences are paired within each run and reported in %. Tier 3 uses the 283-question set. Tier 2 uses the same 1,035 pairs as Table 3, with a separate evaluation setup.

Model	T1 overall	T2 orig.	T2 pruned	T2 Δ	T3 orig.	T3 edited	T3 Δ
Claude Opus 4.6	92.99 ± 0.00	70.14 ± 0.26	58.81 ± 0.15	-11.34 ± 0.24	71.14 ± 0.20	62.43 ± 0.20	-8.72 ± 0.20
Claude Sonnet 4.6	92.09 ± 0.00	64.38 ± 0.11	55.33 ± 0.15	-9.05 ± 0.20	66.78 ± 0.61	67.14 ± 0.00	$+0.35 \pm 0.61$
GPT-5.4	95.57 ± 0.00	78.52 ± 0.46	67.54 ± 0.19	-10.98 ± 0.64	95.17 ± 0.20	83.51 ± 0.20	-11.66 ± 0.35
Qwen3-VL-235B-A22B	92.37 ± 0.00	31.76 ± 0.11	26.41 ± 0.06	-5.35 ± 0.15	38.16 ± 0.35	46.29 ± 0.00	$+8.13 \pm 0.35$

E.4 Pruning Ratio Analysis

We define the pruning ratio as the fraction of characters removed from the original question text. On the original Tier 2 set, the mean pruning ratio is 22.3% (median 19.5%). This character-level measure describes text reduction; it does not measure the fraction of necessary information removed. We classify problems into three bins by pruning ratio and compute accuracy pooled across all 10 models (Figure 5). The observed accuracy changes are -2.6% for lightly pruned problems (0-15%), -7.7% for moderately pruned problems (15-30%), and -13.8% for heavily pruned problems ($>$30%). These are comparisons between groups of different source problems, rather than multiple pruning levels randomly assigned to the same problem. Heavier pruning is associated with larger accuracy drops. Because source difficulty, domain, wording changes, and ambiguity can also vary with pruning ratio, this is a descriptive relationship; policy-specific validity is assessed through the evidence checks and human review in Section 3.2.4.

E.5 Dissecting the Generalization Gap

The difficulty, edit-type, and source-dataset analyses below examine generalization on the 283 variants and their paired originals.

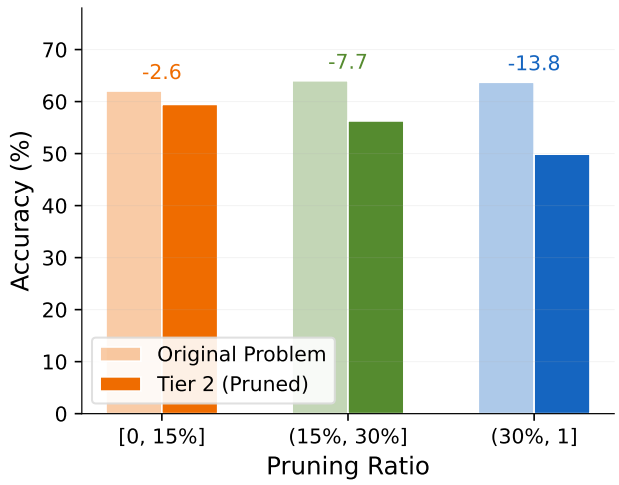


Figure 5: **Pruning ratio and accuracy.** Bars compare original and pruned accuracy within the fixed pruning-ratio bins $[0, 15\%]$, $(15\%, 30\%]$, and $(30\%, 100\%]$, pooling all ten models. Annotations give pruned minus original accuracy in %. Larger reductions in accuracy are associated with heavier pruning.

We examine how model generalization varies across human difficulty labels, edit types, and source datasets.

Difficulty and edit-type breakdown. The 283 variants were annotated post hoc as same difficulty (213), easier (30), or harder (40). Edit-type annotations distinguish text-only semantic changes (102), evaluated with semantically equivalent programmatic redraws, from changes to both text and image content (181). These labels describe the achieved edits; difficulty is not an inclusion criterion or a directional generation target. Table 11 reports edited minus original accuracy in % for each subset.

The generalization gaps persist on the 213 variants rated as the same difficulty as their originals: GPT-5.4 loses 12.2% accuracy and Claude Opus 4.6 loses 8.0%. The losses therefore extend to variants that human annotators judge equally difficult, rather than being confined to the harder group. Both models incur larger losses on the harder variants: 17.5% for GPT-5.4 and 22.5% for Claude Opus 4.6. GPT-5.4 also loses accuracy under both edit types: 7.8% for text-only semantic changes and 14.9% when both text and image are edited. Its generalization gap is thus present across difficulty and edit categories, with a larger loss for joint edits than for text-only edits. The groups contain different problems, and text-only labels describe semantic edits rather than identical image pixels. These post-hoc comparisons characterize where performance fails to transfer; they do not isolate the contributions of task difficulty, rendering, or textual shortcuts.

Per-source-dataset analysis. We report the generalization gap grouped by source dataset, focusing on the five datasets with sufficient representation in the Tier 3 subset ($\geq$29 problems): MathVision (Wang et al., 2024) (84), Geometry3k (Lu et al., 2021) (64), OlympiadBench (He et al., 2024) (45), MathVerse (Zhang et al., 2024b) (35), and EMMA (Hao et al., 2025) (29).

Averaging the reported changes across ten models, OlympiadBench has a -9.1% change, EMMA -5.2%, and MathVision -4.9%. These summarize performance on the retained source subsets.

Geometry3k has a reported mean change of $+1.1\%$ and MathVerse $+12.3\%$, illustrating the variation across source subsets. The five displayed sources account for 257 of the 283 variants; the remaining 26 come from other source benchmarks and are included in the full-set evaluation.

GPT-5.4 loses accuracy in all five displayed sources: EMMA -27.6%, OlympiadBench -15.6%, MathVision -11.9%, MathVerse -11.4%, and Geometry3k -7.8%. Its reduced generalization is therefore observed across multiple source benchmarks, rather than being confined to one dataset. Together, the model and source breakdowns show why edited variants complement original benchmark scores: models differ in their ability to sustain performance when problem instances change. The source-level comparisons are descriptive; the full-set paired tests are reported in Appendix E.2.

Table 11: **Tier 3 accuracy change (Δ, %): edited minus original**, on the 283-question split, stratified by post-hoc difficulty and edit type. Subset deltas are descriptive.

Model	By Difficulty			By Edit Type	
	Same (213)	Easier (30)	Harder (40)	Text only (102)	Text+Image (181)
Proprietary models					
GPT-5.4	−12.2	−6.7	−17.5	−7.8	−14.9
GPT-5.4 mini	−6.6	+20.0	−15.0	+2.0	−8.8
Gemini 3.1 Pro	−1.9	+6.7	−5.0	−2.9	−0.6
Gemini 3.1 Flash-Lite	−0.9	+30.0	−15.0	+2.0	−0.6
Claude Opus 4.6	−8.0	+6.7	−22.5	−12.7	−6.1
Claude Sonnet 4.6	−0.5	+13.3	−5.0	+1.0	+0.0
Open-weight models					
Qwen3-VL-235B	+6.1	+23.3	+2.5	+6.9	+7.7
Qwen3.5-397B	−3.3	+10.0	−12.5	−1.0	−4.4
Gemma 4 31B	−3.3	+20.0	−10.0	+2.0	−3.9
Kimi K2.5	−0.5	+20.0	+0.0	+2.0	+1.7

F TIER 3 MULTI-AGENT EDITING PIPELINE

The evaluated Tier 3 set consists of 283 variants generated through the reproduction-first workflow. The workflow changes problem instances through text, diagram, or joint edits. The post-hoc annotations in Appendix E.5 characterize the resulting changes in difficulty and edit type.

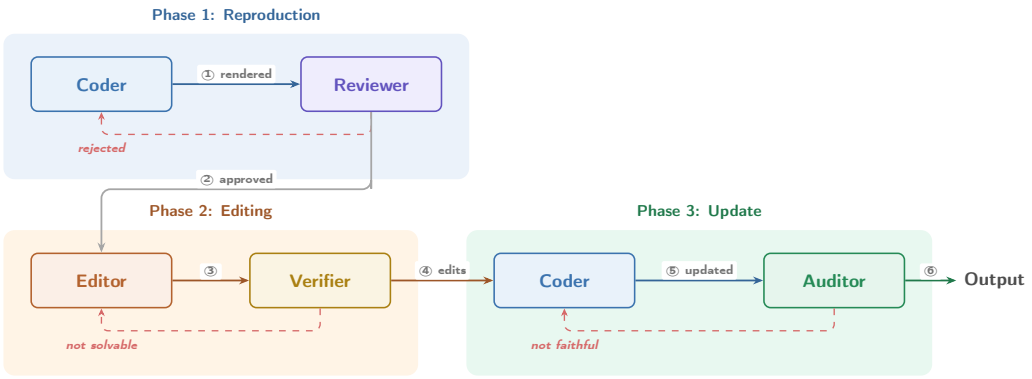


Figure 6: **Tier 3 multi-agent editing pipeline (reproduction-first).** The pipeline proceeds through three phases: reproduction of the original diagram, editing of the problem, and updating the diagram to match the edits. Dashed arrows indicate retry loops triggered by quality failures. Pipeline outputs then undergo final human review.

Reproduction and review. A Coder constructs a programmatic reproduction of the source diagram using an appropriate rendering tool. A Reviewer compares the reproduction with the source, checking labels, elements, connections, and relationships relevant to the problem. Feedback supports iterative correction of the rendered diagram. Code and rendered outputs make the changes inspectable; semantic checks establish whether the reproduced content supports the intended problem.

Problem editing and verification. An Editor proposes changes to the question, conditions, or diagram and updates the reference answer. A Verifier attempts to solve the proposed variant and checks solvability and answer consistency, using the source problem's step-by-step solution as a reference. Failed checks return feedback for revision. The role inputs and checking procedure are described below.

Diagram update and final review. The Coder updates the rendering code to match the edited problem, and an Auditor checks the updated diagram against the intended variant. Human reviewers

then inspect the final problem, image, and answer together, correcting fixable issues and excluding unresolved or invalid candidates. Human review retains 283 of 418 candidates; 135 are not retained.

F.1 Evaluated-run configuration and export provenance

For the evaluated 283 variants, Claude Opus 4.6 serves as the Coder, Editor, and Auditor, while GPT-5.4 serves as the Reviewer and Verifier.

In the documented workflow, the Coder works from problem text and image captions and inspects its own renders; it does not receive the original images. The Reviewer compares the original images with the rendered diagram and can inspect the code. The Editor and Verifier use problem text and captions without image input, while the Auditor inspects the updated diagram against the edited problem. The Verifier receives the edited problem record, including the proposed answer, and the original step-by-step solution. It starts in a fresh context without the Editor's reasoning or earlier rejection history and is instructed to solve the variant and compare answers.

The workflow uses rejection limits of five for reproduction review, three for edit verification, and five for diagram auditing, with the Auditor enabled. These limits count total rejections. Rejected reproductions and diagrams return to the Coder with feedback; rejected problem edits return to the Editor. A missing usable verdict after retries fails the run rather than approving the candidate.

Construction-only answers and internal captions are kept separate from the input presented to the evaluated model.

F.2 Final export checks

Evaluation uses the final diagrams approved by human reviewers. Reviewers inspect the exported diagram together with the problem and answer, checking legible labels, correct geometric relationships, missing or overlapping elements, and unintended duplication of question text or answer options inside the image. After a modification, reviewers inspect the final version again before it is used in evaluation. The 102 text-only variants use semantically equivalent programmatic redraws.

G Qualitative Examples

We provide qualitative examples for each of the three evaluation tiers. Figure 7 shows Tier 1 perception probing questions, where each source image is paired with three multiple-choice questions targeting entity recognition, topology, and visual estimation, with the correct answer highlighted. Figure 8 illustrates text transformations under the original pruning design, with the fraction of original text retained shown for each example. Appendix E.1 describes the additional conservative variant. Figure 9 shows Tier 3 examples with original and edited images, questions, and answers.

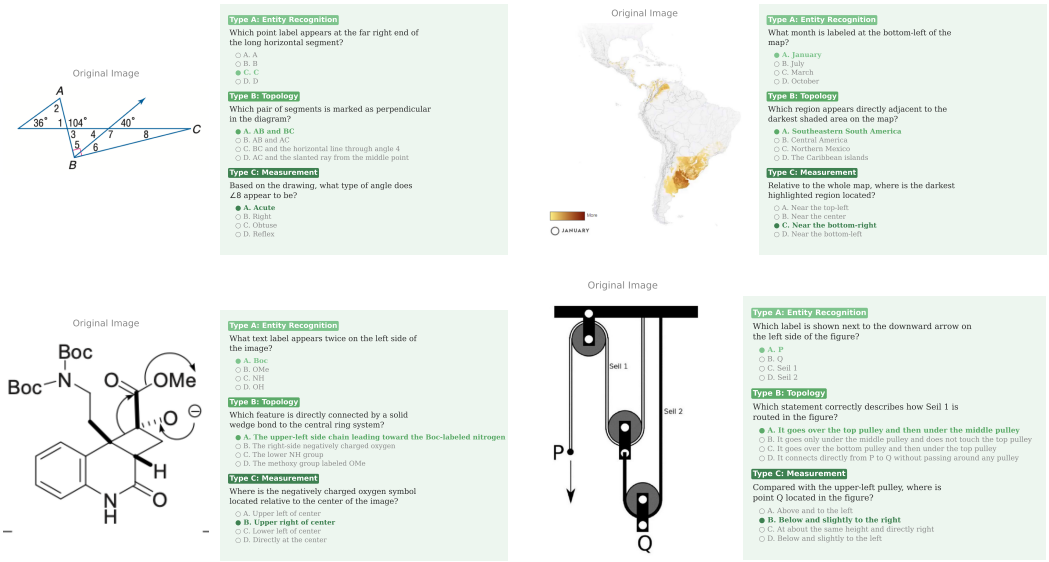


Figure 7: **Tier 1 examples.** Perception probing questions across geometry, geography, physics, and chemistry domains.

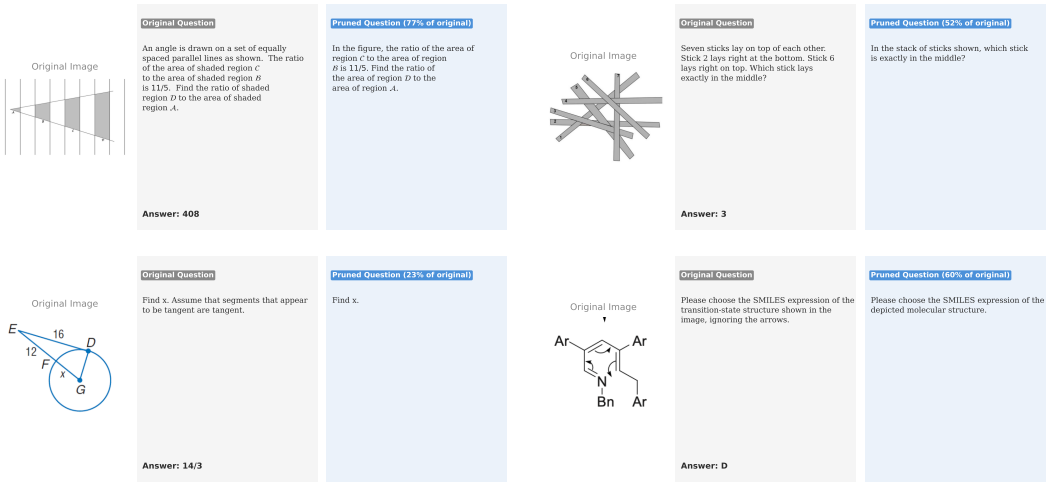


Figure 8: **Tier 2 examples.** Pruned problems with image-redundant text removed. Each percentage reports the fraction of the original text retained.

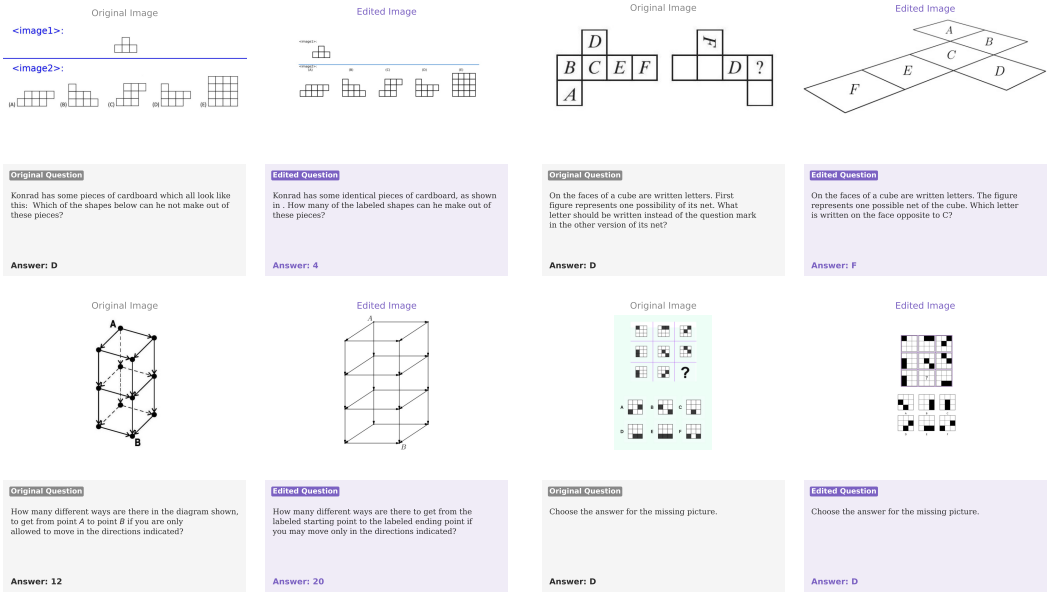


Figure 9: **Tier 3 examples.** Edited problem variants with programmatically modified images and adjusted questions.

H PROMPT TEMPLATES

We list the prompt templates used for model inference and answer judging. Prompts for benchmark construction (atomic fact extraction, Tier 1 question generation, Tier 2 text pruning, and Tier 3 multi-agent editing) are planned for release with the construction code. Placeholders in curly braces ({...}) are filled at runtime with problem-specific content.

H.1 EVALUATION CONFIGURATION AND SCORING DETAILS

The model names in Section 4.1 identify the evaluated families. Thinking is enabled in all tiers; a request for a short final answer does not imply a disabled reasoning budget. Claude Sonnet 4.6 uses low reasoning effort, while Claude Opus 4.6, GPT-5.4, and Qwen3-VL-235B use high reasoning effort in both the original and new evaluations.

Final-answer extraction takes the last boxed expression, falling back to the final 1,500 response characters when no boxed answer exists, and appends trailing solution context. In the current scoring implementation, Claude Opus 4.8 receives the question text and options, reference answer, and extracted model answer with this context, without the problem images. Malformed or failed judge responses are retried until a valid judgment is obtained. The free-form rubric accepts semantically equivalent expressions, notation, rounding, and correct rephrasings. It does not specify a fixed numerical tolerance or an executable symbolic-equivalence rule. These prompt-level allowances must therefore be distinguished from deterministic numerical or symbolic checking.

The same answering prompts are used across original, Tier 2 (pruned), and Tier 3 (edited) problems. The user message contains the question text interleaved with images; for selection questions, options are appended as a lettered list.

Answering - Single-Selection (System)

```
You are an expert problem solver.  Answer the following question by
selecting the correct option(s).
Show your full reasoning and solution process step by step.
At the very end, state your chosen option letter on its own line in
the format:
\boxed{your chosen letter}
For example:  \boxed{A}
```

Answering - Multiple-Selection (System)

```
You are an expert problem solver.  Answer the following question by
selecting the correct option(s).
Show your full reasoning and solution process step by step.
At the very end, state your chosen option letters on its own line in
the format:
\boxed{your chosen letters}
For example:  \boxed{A, C}
```

Answering - Free-Form (System)

```
You are an expert problem solver.  Answer the following question.
Show your full reasoning and solution process step by step.
At the very end, clearly state your final answer on its own line in
the format:
\boxed{your final answer}
```

For Tier 1 perception sub-questions, thinking is enabled, as in the other tiers. The prompt requests a single option letter and discourages solving the source problem; these answer-format and task instructions do not disable the model's reasoning mode. The user message includes the diagram image before the text.

Tier 1 Perception (System)

```
You are evaluating a visual perception question about a diagram.
Answer based solely on what you can directly see in the image.  Do
not solve the problem, apply theorems, or perform calculations.
```

Tier 1 Perception (User)

```
Question:  {question}

Options:
A. {A}
B. {B}
C. {C}
D. {D}

Respond with only a single letter:  A, B, C, or D.
```

A separate judge model (Claude Opus 4.8) evaluates answers for original, Tier 2, and Tier 3 problems. Tier 1 answers are evaluated by direct letter matching. The LLM judge handles both multiple-choice and free-form answers. For selection questions, identifying the selected option(s) and normalizing minor formatting differences are instructions to this judge; there is no deterministic option-matching stage with judge fallback for these problems. The model's final answer is extracted by finding the last `\boxed{...}` expression in the response. If no boxed answer is found, the last 1,500 characters of the response are used as a fallback. In both cases, trailing context from the solution is appended to help the judge resolve ambiguities.

LLM Judge (System)

```
You are a strict but fair judge evaluating whether a model's answer is
correct.

You will be given:
- A question (possibly with images)
- The ground-truth answer
- The model's final answer (extracted from a longer solution)
- The question type

Judging rules:
- Compare the model's final answer against the ground-truth answer.
- Do NOT re-derive or re-solve the problem.  Just compare the two
answers.
- For single_selection / multiple_selection:  The model must select
exactly the correct option letter(s).  Minor formatting differences
are OK (e.g., "A" vs "a" vs "(A)").
- For free_form:  Be more lenient.  The model's answer is correct if
it is semantically equivalent to the ground truth.  Accept equivalent
mathematical expressions, different notations, rounding differences,
or rephrased but correct answers.

Respond with EXACTLY this JSON format (no markdown fencing, no extra
text):
{"correct":  true/false, "reasoning":  "brief explanation"}
```

LLM Judge (User)

```
Question:  {question}
Options:  {options}
Question type:  {question_type}

Ground-truth answer:  {ground_truth_answer}
Model's final answer:  {extracted_final_answer}
```