BLAMEBENCH: DIAGNOSING ATOMIC PERCEPTION IN MULTIMODAL REASONING BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models ground reasoning in fine-grained visual evidence, but final-answer accuracy conflates errors in perception and subsequent reasoning. We introduce **BlameBench**, a task-grounded benchmark that uses visual evidence from solution trajectories with correct final answers to decide what to test. We extract visual observations, verify them against source images, and convert the resulting facts into atomic perception questions that can be answered without solving the full source problem. This construction yields a fixed 2,500-question benchmark spanning six source benchmarks and seven reporting categories, including 2,058 direct-perception questions and 442 supplementary concept probes. Across 11 frontier models, source and category profiles reveal differences concealed by overall accuracy: Qwen 3.8 Max and Claude Opus 5 are separated by only 0.64 percentage points overall but lead on different sources and categories. The same questions also provide an upstream readout: across four InternVL3.5-Pretrained checkpoints, source-centered scores before fine-tuning correlate with downstream accuracy after fine-tuning ($r = 0.834$). **BlameBench** thus supports final-model perceptual diagnosis and pretraining-stage diagnosis.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Qwen Team, 2026; Team et al., 2026; DeepSeek-AI, 2026) must ground their reasoning in fine-grained visual evidence to answer questions about charts, documents, and mathematical diagrams. A model may read the wrong value from a plot, or read the value correctly and use it in the wrong calculation. Final-answer accuracy evaluates these distinct steps together and therefore cannot locate the source of an error. What is needed is a task-grounded perception measurement that directly tests visual evidence used in successful downstream solutions and preserves its link to the originating problem. The visual and interpretive competencies that support such grounding emerge during pretraining, so the same measurement can diagnose final models and provide an early readout before supervised adaptation.

Recent benchmarks explicitly evaluate visual perception. BLINK reformulates classical visual tasks (Fu et al., 2024); VisFactor adapts cognitive tests (Huang et al., 2025); and BabyVision evaluates visual abilities beyond linguistic knowledge (Chen et al., 2026). CharXiv separates descriptive chart questions from reasoning questions (Wang et al., 2024b). PerceptionBench identifies atomic capabilities through failure analysis on existing multimodal benchmarks (Lin et al., 2026). These approaches (Yue et al., 2024; Lee et al., 2026) establish focused visual tests and useful perception-reasoning decompositions. A complementary construction is needed to determine which visual evidence to measure for each downstream reasoning problem. Such a construction should let task-relevant evidence, rather than a predefined capability taxonomy, determine the evaluation target while retaining its link to originating problem.

Solution trajectories provide a concrete starting point because they record visual observations used in successful downstream reasoning, including chart values, diagram relations, and text strings. Across multiple models and repeated runs, correct trajectories are available for 97.6% of source problems. This breadth lets the construction draw candidate evidence from nearly the full source-task distribution rather than from a narrow subset of solved examples. A correct final answer provides observations from a successful solution, while image verification establishes whether an extracted observation is a valid visual fact. A separate difficulty screen is needed because a challenging source

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.

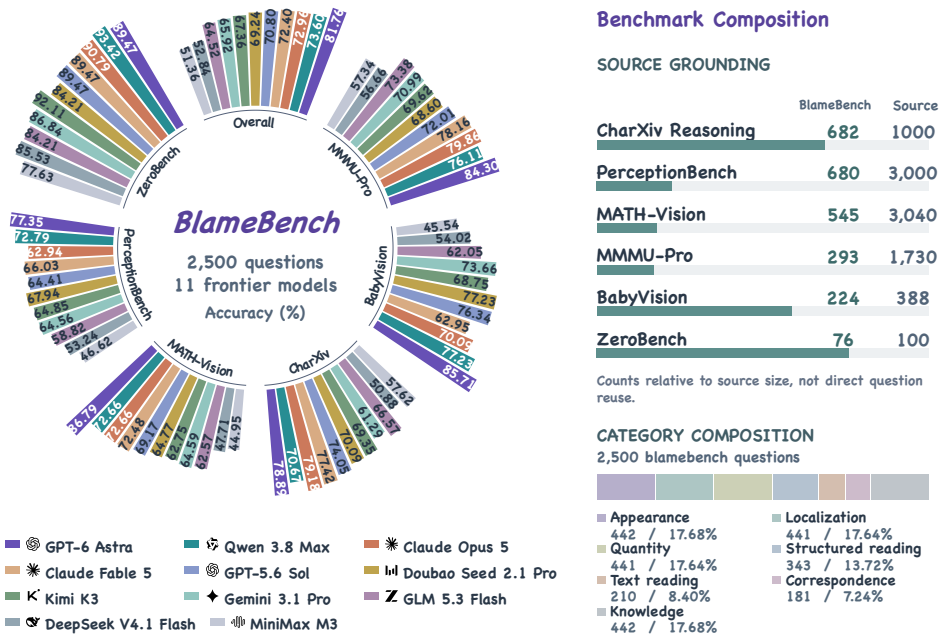


Figure 1: **Overview of BlameBench. Left:** BlameBench spans six source benchmarks across 11 models. **Right:** Composition of BlameBench and category.

problem can contain an easy visual fact. The resulting facts provide a principled basis for independently answerable perception questions that retain their links to the originating tasks.

We introduce **BlameBench**, which uses visual evidence in correct solution trajectories to decide *what to test*. We retain source problems with disagreeing model responses and extract visual statements from a successful trajectory, including values, relations, attributes, counts, and text. Each statement is checked against the image, resolved to its visual elements, and rewritten as a self-contained fact. The facts become direct-perception questions answerable from the image without solving the full source task. Repeated answering screens questions for discriminative difficulty, and model and manual review validate reference answers. The fixed set applies this procedure across heterogeneous sources, while source and category labels support complementary diagnostics. Supplementary concept probes cover concepts needed to interpret source questions and figures.

BlameBench contains 2,500 questions from six sources and seven reporting categories, including 2,058 direct-perception questions and 442 supplementary concept probes. Figure 1 shows this composition and the accuracy range across 11 frontier models. Source and category profiles complement overall accuracy. Qwen 3.8 Max and Claude Opus 5 differ by only 0.64 percentage points overall, with Qwen scoring higher on BabyVision-derived questions and Claude higher on CharXiv-derived questions. Their category profiles diverge: Qwen scores 78.45% on correspondence versus Claude's 67.40%, while Claude scores 64.48% on knowledge versus Qwen's 51.81%. These scores locate differences across downstream settings and visual operations that overall accuracy obscures. The same questions also provide an upstream readout: across four InternVL3.5-Pretrained checkpoints, source-centered BlameBench scores before fine-tuning correlate with downstream accuracy after fine-tuning (Pearson $r = 0.834$). This shared measurement connects pretraining-stage perceptual capability with final-model diagnosis. Our contributions can be summarized as:

- **Grounded construction.** BlameBench converts verified visual evidence from successful trajectories into atomic questions with image verification and answer review.
- **Two-axis diagnosis.** Source and category profiles locate errors across six downstream settings and seven visual operations, revealing differences hidden by overall accuracy.
- **Lifecycle measurement.** The same questions diagnose final models and provide a pretraining-stage readout associated with downstream performance after fine-tuning.

Table 1: **Comparison with existing perception benchmarks.** Columns summarize task grounding, construction provenance, and diagnostic use. Source QA, Correct trajectory, and Source trace indicate existing-problem construction, successful-solution extraction, and preserved source identities. ✓: reported; ○: partial; ×: not reported.

Benchmark	Target	Task-grounded construction			Diagnostic use	
	Reasoning link	Source QA	Correct trajectory	Source trace	Dual profiles	Pretrain diagnosis
MMVP (Tong et al., 2024b)	×	×	×	×	×	×
Q-Bench+ (Zhang et al., 2024)	×	×	×	×	✓	×
VisOnlyQA (Kamoi et al., 2024)	×	×	×	×	✓	○
VisFactor (Huang et al., 2025)	×	×	×	×	✓	×
Do You See Me (Kanade & Ganu, 2026)	✓	✓	×	✓	✓	×
SalBench (Dahou et al., 2025)	×	×	×	×	✓	×
PerceptionRubrics (Wei et al., 2026)	×	×	×	×	✓	×
PerceptionBench (Lin et al., 2026)	✓	✓	×	✓	✓	×
BlameBench (Ours)	✓	✓	✓	✓	✓	✓

2 Related Work

General and atomic visual perception. Recent multimodal benchmarks explicitly evaluate visual perception through complementary designs. BLINK reformulates classical vision tasks, while BabyVision and VisFactor assess basic visual skills and cognitive abilities (Chen et al., 2026; Huang et al., 2025; Saxena et al., 2026). Do You See Me uses controlled perceptual tasks; VisOnlyQA assesses geometric and numerical evidence in scientific figures; MMVP and SalBench probe visual distinctions and salient patterns (Kanade & Ganu, 2026; Kamoi et al., 2024; Tong et al., 2024b; Dahou et al., 2025). PerceptionBench is particularly close in its focus on atomic capability diagnosis: it derives a taxonomy from earliest-error analyses on existing benchmarks and constructs questions through failure-guided decomposition and task creation (Lin et al., 2026). PerceptionRubrics evaluates generated captions against instance-specific atomic visual criteria (Wei et al., 2026). Unlike these approaches, BlameBench selects targets from visual observations in solution trajectories with correct final answers, verifies them against source images, and converts them into atomic perception questions. This retains each question's link to its source problem and visual evidence.

Specialized perceptual capabilities. Specialized benchmarks examine individual perceptual dimensions in depth. Q-Bench+ evaluates low-level visual attributes and image quality, while OCR-Bench v2 and UNICBench focus on visual text and counting, respectively (Zhang et al., 2024; Fu et al., 2026; Rong et al., 2026; Liu et al., 2024). CV-Bench and SpatialScore assess spatial and geometric understanding; MMPerspective, DORI, and DisTANCE target perspective, object orientation, and visual estimation (Tong et al., 2024a; Tang et al., 2026; Li et al., 2026). These benchmarks define tasks around particular perceptual abilities. BlameBench instead applies a shared fact-extraction and question-construction procedure to visual evidence from heterogeneous downstream benchmarks. Source labels locate each item's originating task setting, while perception tags describe the specific visual operation it tests.

Separating perception from reasoning. Several benchmarks separate visual perception from the reasoning that uses it (Zhang, 2026; Sun et al., 2026; Pan et al., 2026). MathLens evaluates perception, reasoning, and their integration in textbook-style geometry, while Visual Room 2.0 uses six progressive questions per sample across low, middle, and high levels to assess perception-cognition alignment (Chung et al., 2026; Li et al., 2025). CharXiv separates descriptive questions about chart elements from reasoning questions that synthesize information across complex visual elements (Wang et al., 2024b; Garre et al., 2026; Masry et al., 2025). For its direct-perception items, BlameBench follows this diagnostic direction with questions independently answerable from the source image without solving the full source problem. Each item is traced from a solution trajectory with a correct final answer through a verified visual fact. Table 1 summarizes the construction and diagnostic distinctions among these benchmarks.

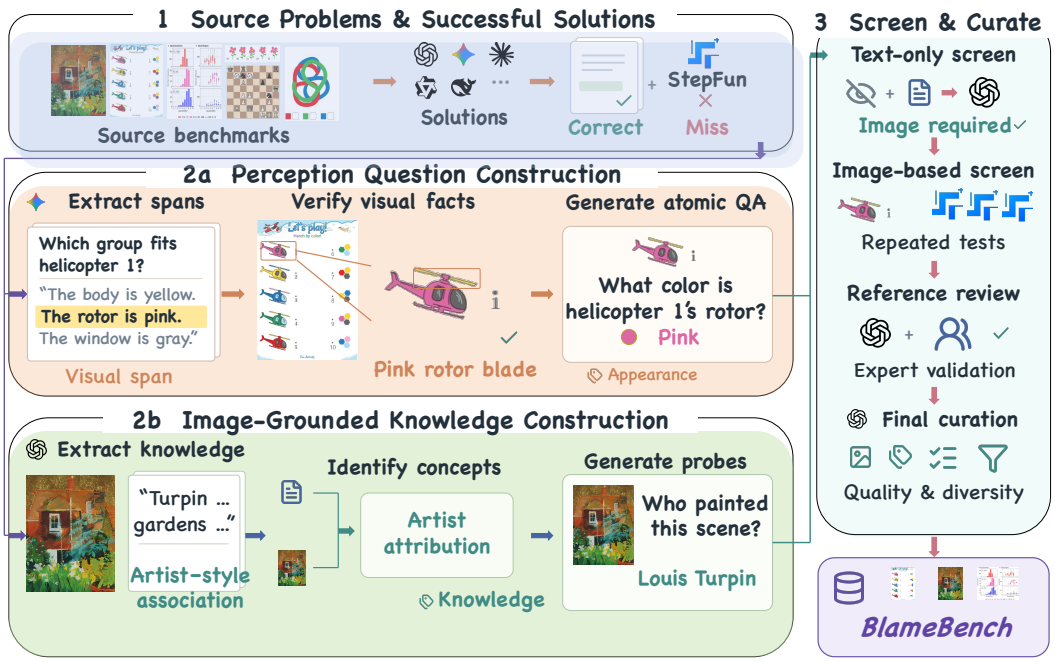


Figure 2: **From solution evidence to perception questions.** Observations extracted from correct-final-answer trajectories are verified against source images and organized into self-contained facts. Each suitable fact supports an open-ended question with a short answer. Fragment and fact indices connect the resulting measurement to its source trajectory.

3 Benchmark Construction

BlameBench uses visual observations in successful downstream trajectories to determine what to test. It turns these observations into verified facts and independently answerable perception questions, and complements them with image-grounded probes of the knowledge for interpreting source problems and figures. Image-dependence screening, repeated answering, and reference validation produce a fixed benchmark of 2,500 questions across six sources and seven reporting categories. Figure 2 illustrates the perception construction; App. A details the protocols.

3.1 Selecting Source Problems and Reference Trajectories

Source coverage. We collect 9,258 problems from MMMU-Pro (Yue et al., 2025), BabyVision (Chen et al., 2026), CharXiv Reasoning (Wang et al., 2024b), MATH-Vision (Wang et al., 2024a), PerceptionBench (Lin et al., 2026), and ZeroBench (Roberts et al., 2025). Two runs from each of 11 frontier models produce over 200,000 valid response trajectories. At least one trajectory with a correct final answer is available for 9,036 problems (97.6%), making successful solutions available across nearly the entire source collection. Correctness follows each source benchmark's answer metric.

Reference selection. We select correct reference trajectories from source problems that Step 3.7 Flash (Huang et al., 2026) answers incorrectly in a single run, yielding 16,052 trajectories. When an eligible source problem has multiple correct trajectories, we retain all of them. This selection combines two requirements: the source problem challenges the screening model, and a successful response provides candidate observations for question construction. We retain the source-problem identifier throughout the pipeline, linking each derived question to its originating task. The difficulty of the derived questions is assessed separately in Section 3.3.

3.2 Constructing Questions and Assigning Skills

Perception questions. Gemini 3.7 Flash constructs perception questions in three stages:

1. **Extract visual observations.** Given only the trajectory text, the model copies minimal continuous spans describing objects, attributes, counts, spatial relations, visible text, or chart readings. The prompt excludes observation plans, hypotheses, external knowledge, and non-visual derivations. Both an initial observation and its later correction remain available for verification. A mechanical check records each span's match to the trajectory after whitespace normalization. After this step, we obtain 151,456 extracted observation spans.
2. **Verify and organize facts.** Given the source image and indexed spans, the model resolves object references, checks observations against the image, and resolves conflicting descriptions. It expresses supported observations as self-contained facts and records their supporting span indices. The extracted spans define the content to verify; the image supplies the evidence for verification. After this step, we obtain 8,761 verified visual facts.
3. **Generate atomic QA.** From the image, indexed facts, and fixed style examples, the model generates one open-ended question per suitable fact, with a short reference answer, question type, and supporting fact indices. Each question targets one visual measurement, such as a count, attribute, position, text string, or chart value. Its wording must identify the target without revealing the answer. Facts that cannot support an unambiguous short-answer question are skipped. After this step, we obtain 8,466 candidate perception questions.

Here, atomicity refers to a single measurement target. Locating the relevant region can provide context, while the answer tests the specified visual fact without requiring the source task's full solution. App. A.2 gives the extraction and verification constraints.

Image-grounded knowledge questions. GPT-5.6 Terra constructs probes of the concepts needed to interpret source problems and figures:

1. **Extract candidate knowledge.** Given the source question, correct source answer, trajectory, and image, the model identifies transferable, self-contained knowledge propositions. Each proposition records its type, discipline, necessity, confidence, and trajectory evidence. Instance-specific calculations and directly visible facts are excluded from this knowledge extraction. After this step, we obtain 4,243 candidate knowledge propositions.
2. **Identify concepts to test.** The question-based route identifies source-question prerequisites and aligns them with extracted knowledge. The figure-based route identifies domain concepts presupposed by the image, including mechanisms, quantities, and units. Concepts retain their question or figure origin and any matching trajectory evidence. After this step, we obtain 3,915 concepts, including 1,906 question-based and 2,009 figure-based concepts.
3. **Generate image-dependent probes.** The model generates open-ended questions testing recall, recognition, distinction, application, transfer, or applicability boundaries. The image must supply a necessary referent, value, or configuration, while visual identification remains straightforward. This focuses the challenge on using the concept. Outputs include the expected answer, grading rubric, concept identifiers, probe kind, and difficulty label. Concepts that do not support a natural image-dependent question can be skipped. After this step, we obtain 11,523 candidate knowledge probes across 6 probe kinds.

App. A.3 details concept alignment and generation, including probes combining multiple concepts.

Fine-grained skills and reporting categories. For each perception question, Gemini 3.7 Flash receives the image, question, reference answer, a closed set of skill definitions, and classification examples. It assigns the single skill that directly produces the answer. Contextual steps, such as locating a region before reading its text, do not determine the label unless they are themselves the measurement target. The model separately records the visual carrier, such as a photograph, chart, document, or musical score. The assigned skill and carrier labels are subsequently reviewed by humans for consistency with the annotation definitions. After this step, 8,328 perception questions have valid skill and carrier labels.

We consolidate the 20 intermediate labels then into six perception categories. Knowledge probes receive the Knowledge category, giving seven reporting categories in total. Skill map is in App. A.4.

3.3 Screening, Validation, and Final Curation

Image dependence and question difficulty. Generation checks screen for answer leakage and copied question wording. Perception questions and knowledge probes then undergo the same validation pipeline, beginning with two answering-based screens. Together, the two construction branches provide 19,989 candidates before answering-based screening.

1. **Text-only screening.** GPT-5.6 Terra receives only the derived question, without the image, trajectory, construction facts, or reference answer. It must commit to a short answer even when uncertain. We remove candidates judged correct without the image, retaining 19,851 questions.
2. **Repeated image-based screening.** Step 3.7 Flash answers each retained question in three independent runs, receiving the original image and derived question. We retain questions answered correctly at most once: 1,572 with no correct answers and 3,714 with one, yielding 5,286 candidates.

These screens select questions that resist text-only answering and remain challenging when the image is available. App. A.6 describes the answering protocol.

Reference-answer validation. GPT-5.6 Sol reviews the references after difficulty screening. Domain experts then inspect only the image, question, and derived reference answer and assign one of three labels: *correct* when the reference is supported and answers the question; *incorrect* when it is wrong; or *unclear* when ambiguity or insufficient evidence prevents a determinate judgment. Each candidate is first labeled independently by three experts and receives a correct or incorrect label only when all three agree; all other cases enter the unclear set. Each unclear case is then labeled by a second group of three experts and is accepted as correct only when all three agree that it is correct. This process yields 4,035 correct candidates for final curation.

Five-criterion selection. We select the final 2,500 questions using five scores in $[0, 1]$:

1. **Source and category coverage:** the contribution to underrepresented sources and reporting categories.
2. **Image diversity:** preference for images used by fewer questions in the selected set.
3. **Target focus:** a clearly bounded visual measurement or conceptual test.
4. **Image grounding:** necessary image evidence and wording that preserves the intended test.
5. **Answerability:** a clear referent and a determinate answer supported by the material.

GPT-5.6 Luna evaluates the three content-quality criteria from the image, question, reference answer, measurement target, and supporting material. Coverage and image diversity are computed from selection counts. After deduplication and exclusion of demonstration images, we iteratively select the highest-scoring candidate using the mean of the five scores. Content-quality scores remain fixed, while coverage and diversity update as the selected set grows. App. A.7 specifies the scoring anchors, coverage and diversity functions, and tie-breaking rule.

3.4 Benchmark Composition and Diagnostic Views

The final benchmark contains 2,500 questions, each with one source label and one reporting category Tab. 1. Source profiles locate model differences across downstream settings, while category profiles identify the capabilities involved. Perception questions retain supporting fact and span indices; knowledge questions retain concept identifiers, origins, and available trajectory evidence. These records connect the diagnostic scores to the underlying measurement targets.

Five fixed few-shot demonstration questions are held separately as prompt examples, and evaluation questions share no images with them. All evaluated models receive the same fixed question set, and scoring uses a shared set of derived reference answers. Section 4 uses the source and category views to characterize model capabilities and pretrained checkpoints.

4 EXPERIMENTS

We evaluate BlameBench as a diagnostic of perceptual differences across 11 frontier models and of pretrained checkpoint quality before supervised fine-tuning. Source and category profiles reveal differences hidden by overall accuracy; pretraining-stage measurements are compared with downstream performance after a shared fine-tuning procedure.

4.1 EVALUATION SETUP

Benchmark and models. We evaluate 11 frontier models on the same 2,500 BlameBench questions across six sources and seven categories. We report overall, source, and category accuracy, weighting questions equally within each set. Table 2 pairs scores on source-derived BlameBench questions (BB) with performance on the corresponding full downstream benchmarks (DS), providing focused perception measurements alongside full-task outcomes.

Frontier inference and scoring. BB evaluation is zero-shot, single-turn, and tool-free. Models receive the original images in their recorded order and the derived question, with an 8,192-token output budget and one saved answer per question. We request disabled reasoning where supported; GLM uses low reasoning and Claude Fable uses adaptive reasoning with low effort. GPT-5.6 Luna grades semantic equivalence to the reference using the question and both answers as text. App. B.1 and App. B.2 specify the inference and scoring procedures.

Downstream evaluation. Frontier DS evaluation uses no external tools, the highest available reasoning effort, and a requested 256K-token generation budget, with other parameters following official API defaults. We report pass@1 for all sources except ZeroBench, which uses pass@5. BB and DS scores describe their respective question sets and evaluation protocols (App. B.1).

Pretraining-stage evaluation. We evaluate the 1B, 2B, 4B, and 8B InternVL3.5-Pretrained checkpoints with five-shot prompts on BlameBench and five downstream benchmarks: MMMU-Pro vision, BabyVision, CharXiv reasoning, MATH-Vision, and PerceptionBench. Each benchmark uses five fixed demonstrations in the same order across checkpoints. After a shared fine-tuning procedure, we evaluate the five downstream benchmarks zero-shot. The paired analysis covers 2,424 BlameBench questions from these sources. Each checkpoint uses the same 30,000-example MMFineReason subset, selected after evaluation-overlap filtering and training deduplication. Training makes one pass with incomplete-batch dropping, consuming 29,952 examples in 234 updates. App. B.3 details the evaluation settings, data selection, and training recipe.

4.2 MAIN RESULTS

Tables 2 and 3 report source and category profiles. Overall accuracy ranges from 51.36% to 81.76%. GPT-6 Astra ranks first overall and leads on four of six source subsets and five of seven categories; Doubao Seed 2.1 Pro leads on Appearance, and Qwen 3.8 Max on Text reading.

Similar totals conceal distinct capability profiles. Qwen 3.8 Max and Claude Opus 5 differ by only 0.64 percentage points overall (73.60% versus 72.96%), yet their category differences exceed 11 points in opposite directions. Qwen scores 78.45% on Correspondence versus Claude's 67.40%, while Claude scores 64.48% on Knowledge versus Qwen's 51.81%. Category profiles identify which capabilities separate models whose overall scores are nearly identical. Across models, Knowledge is the lowest-scoring category overall, whereas Correspondence shows the widest performance range.

Source profiles locate differences in downstream settings. Qwen leads Claude on BabyVision-derived questions (77.23% versus 70.09%), whereas Claude leads on CharXiv-derived questions (79.18% versus 70.67%). The CharXiv ordering also differs from full-task accuracy, where Qwen scores 89.00% and Claude 85.80%. The source-linked probes therefore expose differences in reading task-relevant visual evidence that the downstream ranking alone does not reveal. Across sources, BB and downstream rankings are generally aligned but remain nonredundant, especially on ZeroBench. ZeroBench-derived BB is easiest, while its full task remains hardest. Together, source and category profiles locate these differences by task setting and capability.

Table 2: **Main results by source.** Accuracy (%) on BlameBench questions (BB) and the corresponding full downstream benchmarks (DS). Overall is accuracy on all 2,500 BB questions. BB and DS evaluate different question sets within each source. Bold and underlined values denote the best and second-best scores in each column, including ties.

Model	Overall (BB)	MMMU-Pro		BabyVision		CharXiv		MATH-Vision		PerceptionBench		ZeroBench	
		BB	DS	BB	DS	BB	DS	BB	DS	BB	DS	BB	DS
GPT-6 Astra	**81.76**	**84.30**	**86.72**	**85.71**	**88.40**	78.89	**92.00**	**86.79**	93.52	**77.35**	**72.90**	89.47	**40.00**
Qwen 3.8 Max	73.60	76.11	82.25	77.23	82.00	70.67	89.00	72.66	**94.67**	72.79	63.50	**93.42**	26.00
Claude Opus 5	72.96	79.86	84.01	70.09	43.04	**79.18**	85.80	72.66	91.15	62.94	55.20	90.79	18.00
Claude Fable 5	72.40	78.16	84.62	62.95	42.50	77.42	86.20	72.48	92.96	66.03	57.20	89.47	20.00
GPT-5.6 Sol	70.80	72.01	83.64	76.34	65.50	74.05	87.90	69.17	93.26	64.41	59.70	89.47	22.00
Doubao Seed 2.1 Pro	69.24	68.60	81.68	77.23	73.70	70.09	86.10	64.77	91.61	67.94	55.00	84.21	26.00
Kimi K3	67.36	69.62	81.45	68.75	50.26	69.35	86.00	62.75	90.26	64.85	58.50	92.11	23.00
Gemini 3.1 Pro	65.92	70.99	80.50	73.66	55.90	61.29	84.40	64.59	87.40	64.56	56.20	86.84	16.00
GLM 5.3 Flash	64.52	73.38	79.21	62.05	53.40	66.57	79.50	62.57	83.09	58.82	49.43	84.21	11.00
DeepSeek V4.1 Flash	52.84	56.66	77.44	54.02	41.24	50.88	71.60	47.71	82.20	53.24	42.97	85.53	12.00
MiniMax M3	51.36	57.34	80.00	45.54	20.62	57.62	76.10	44.95	74.80	46.62	33.10	77.63	10.00

Table 3: **Main results by perception category.** Accuracy (%) on the same fixed 2,500-question set. Model order matches Table 2. Bold and underlined values denote the best and second-best scores in each column, including ties.

Model	Overall	Appearance	Correspondence	Knowledge	Localization	Quantity	Structured reading	Text reading
GPT-6 Astra	**81.76**	78.05	**89.50**	**67.42**	**89.80**	**85.71**	**88.05**	77.62
Qwen 3.8 Max	73.60	78.73	78.45	51.81	80.05	73.47	80.17	**80.48**
Claude Opus 5	72.96	76.02	67.40	64.48	76.19	68.48	82.51	76.19
Claude Fable 5	72.40	74.43	66.85	61.54	74.60	68.93	83.97	79.52
GPT-5.6 Sol	70.80	69.46	71.82	58.37	80.05	66.67	81.92	70.00
Doubao Seed 2.1 Pro	69.24	**78.96**	65.19	48.42	69.61	72.34	83.67	65.24
Kimi K3	67.36	74.66	65.75	52.94	74.60	61.00	74.64	70.00
Gemini 3.1 Pro	65.92	70.14	65.19	43.67	74.38	67.80	70.85	74.76
GLM 5.3 Flash	64.52	70.14	61.88	52.49	72.11	59.64	70.26	65.24
DeepSeek V4.1 Flash	52.84	61.31	54.70	34.84	60.54	48.75	60.64	50.95
MiniMax M3	51.36	62.22	44.75	39.37	50.57	42.40	67.06	54.29

4.3 Pretraining-Stage Diagnosis

We next examine whether BlameBench can diagnose *relative checkpoint quality* before fine-tuning. Within each source, do checkpoints with higher base BlameBench scores also achieve higher downstream accuracy after a shared fine-tuning procedure? We compare this signal with direct pre-SFT evaluation on the full downstream benchmarks. Table 4 reports the measurements for four checkpoints across five paired sources.

Diagnosing relative checkpoint quality. To compare checkpoints within sources, we remove source-specific mean score levels. Let $x_{m,s}$ denote base BlameBench accuracy for checkpoint m on questions from source s, and $y_{m,s}$ its post-SFT accuracy on the corresponding downstream benchmark. We center each measurement over the four checkpoints:

$$\widetilde{x}_{m,s} = x_{m,s} - \frac{1}{4}\sum_{m'=1}^{4} x_{m',s}, \qquad \widetilde{y}_{m,s} = y_{m,s} - \frac{1}{4}\sum_{m'=1}^{4} y_{m',s}. \tag{1}$$

We apply the same centering to base downstream accuracy, obtaining $\widetilde{d}_{m,s}$ as the direct-evaluation reference. Centered scores express deviations from source's checkpoint mean in percentage points.

Stronger linear association with fewer evaluation questions. Across the 20 checkpoint–source pairs, source-centered base BlameBench scores correlate with post-SFT downstream scores at Pearson $r = 0.834$, compared with $r = 0.756$ for base downstream scores (Figure 3). In this four-

Table 4: **Pretraining-stage diagnosis by source.** Accuracy (%) for four InternVL3.5-Pretrained checkpoints. BB denotes base BlameBench accuracy on questions derived from each source, Base DS denotes direct downstream evaluation before fine-tuning, and Post-SFT denotes downstream performance after the shared fine-tuning procedure. The resulting five sources × four checkpoints form the 20 pairs used for source-centered analysis.

Checkpoint	MMMU-Pro			BabyVision			CharXiv			MATH-Vision			PerceptionBench		
	BB	Base DS	Post-SFT	BB	Base DS	Post-SFT	BB	Base DS	Post-SFT	BB	Base DS	Post-SFT	BB	Base DS	Post-SFT
1B	14.68	12.14	5.38	17.86	6.96	1.80	9.53	18.70	10.90	14.86	8.55	4.34	17.35	21.10	7.80
2B	18.77	15.61	10.64	17.41	9.79	1.80	12.90	23.60	15.90	20.92	7.40	10.59	18.38	23.87	7.90
4B	29.69	17.86	17.17	29.02	14.69	3.35	18.18	27.70	26.80	28.07	10.86	23.49	26.47	22.60	13.63
8B	32.76	21.62	21.10	30.80	16.75	3.87	19.50	28.60	31.00	30.83	13.13	26.81	31.18	24.00	15.13

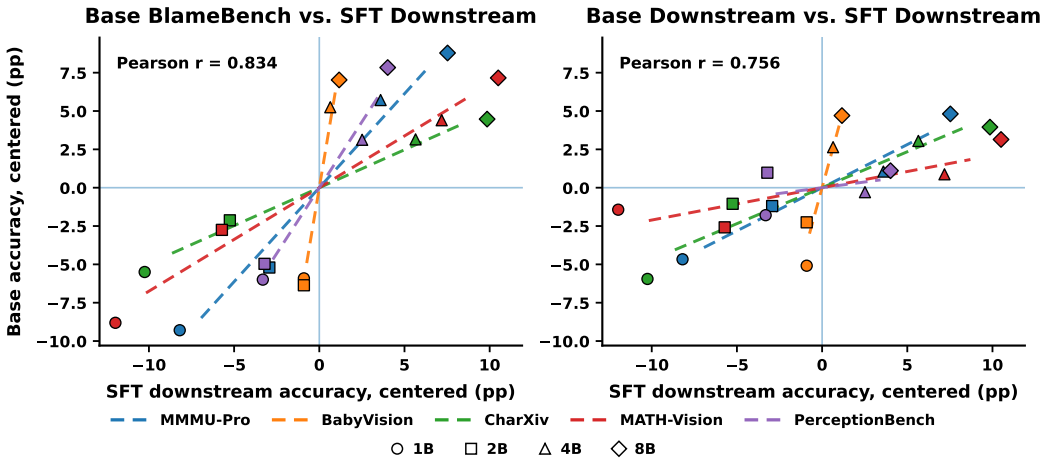


Figure 3: **Pretraining-stage diagnosis of relative checkpoint quality.** Source-centered post-SFT downstream accuracy versus base BlameBench accuracy (left) or base downstream accuracy (right) across 20 checkpoint–source pairs. BlameBench shows a stronger linear association ($r = 0.834$ vs. 0.756) using 26.5% as many questions. Panels share axis scales; colors denote sources and markers denote checkpoint sizes.

checkpoint comparison, BlameBench more closely tracks the performance differences observed after the shared fine-tuning procedure. On MATH-Vision and PerceptionBench, base BlameBench preserves the post-SFT checkpoint ordering where direct base evaluation shows reversals. The five paired BlameBench subsets contain 2,424 questions, compared with 9,158 in the corresponding five downstream benchmarks: 26.5% as many questions. This combination of stronger linear association and smaller evaluation size supports BlameBench as a compact pretraining-stage diagnostic of downstream checkpoint quality.

5 Conclusion

BlameBench uses verified visual evidence from successful solution trajectories to determine what to test in multimodal tasks. It converts evidence into independently answerable direct-perception questions and complements them with image-grounded concept probes. The benchmark contains 2,500 questions across six sources and seven reporting categories, while preserving each item's link to its source problem. Across 11 frontier models, source and category profiles reveal perceptual strengths and weaknesses concealed by similar overall scores. Across four InternVL3.5-Pretrained checkpoints, source-centered BlameBench scores achieve a stronger association with post-SFT downstream performance than base downstream evaluation ($r = 0.834$ versus 0.756), using 26.5% as many questions. BlameBench therefore provides one task-grounded measurement for diagnosing final models and assessing relative checkpoint quality before fine-tuning, while retaining links to the source problems and evidence underlying the measured capabilities.

AI USE STATEMENT

Generative AI was integral to benchmark construction and evaluation. Gemini 3.7 Flash extracted observations, verified visual facts, generated perception questions, and assigned skill labels; GPT-5.6 Terra generated knowledge probes and performed text-only screening; Step 3.7 Flash screened source problems and question difficulty; GPT-5.6 Sol reviewed reference answers; and GPT-5.6 Luna scored item quality and evaluated model responses. OpenAI Codex was additionally used for methodological feedback, quantitative checks, interpretation of results, and manuscript drafting and editing. All AI-generated benchmark content underwent automated screening and human review. Domain experts validated derived answers under the unanimity protocol described in Section 3.3, and human reviewers inspected the assigned skill labels. The authors verified the reported analyses, numerical claims, and final text and take responsibility for all AI-assisted content and research artifacts.

ETHICS STATEMENT

BlameBench is constructed from questions and images in six existing multimodal benchmarks rather than newly collected personal data. Source images remain subject to their original licenses and usage conditions, which must govern any redistribution. Domain experts participated only in validating benchmark items from the image, question, and proposed answer; the study does not analyze their personal information. The benchmark may inherit cultural, linguistic, topical, and representational biases from its source datasets and from the models used during construction and screening. Its scores characterize model behavior on the selected tasks and should not be treated as comprehensive measures of visual ability or used for high-stakes decisions. Human validation, provenance records, and the reported scope limitations are intended to reduce unsupported claims and facilitate responsible interpretation.

REPRODUCIBILITY STATEMENT

Section 3 specifies the benchmark construction pipeline, selection counts, validation procedure, and final composition. App. A documents model roles, construction records, prompt constraints, skill mappings, screening rules, human annotation, and the deterministic final-selection formula. Section 4.1 and App. B report the frontier-model inference and scoring protocols. App. B.3 provides the pretrained checkpoint identifiers, training dataset revision, overlap filters, example-selection rule, random seed, evaluation prompts, and shared fine-tuning configuration. All models are evaluated on the same fixed 2,500 questions and five demonstrations with shared reference answers. These records support recomputation of the benchmark, frontier-model, and pretraining-stage results under the stated protocols.

REFERENCES

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

Jiwan Chung, Neel Joshi, Pratyusha Sharma, Youngjae Yu, and Vibhav Vineet. What mllms learn about when they learn about multimodal reasoning, 2026. URL https://arxiv.org/abs/2510.01719.

Yasser Dahou, Ngoc Dung Huynh, Phuc H Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. Vision-language models can't see the obvious. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24159–24169. IEEE, 2025.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Suhaas Garre, Chris Mutty, Sushant Mehta, and Edwin Chen. Chartography: A benchmark for professional chart understanding. *arXiv preprint arXiv:2608.10677*, 2026.

Ailin Huang, Ang Li, Aobo Kong, Bin Wang, Binxing Jiao, Bo Dong, Bojun Wang, Boyu Chen, Brian Li, Buyun Ma, et al. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters. *arXiv preprint arXiv:2602.10604*, 2026.

Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, and Haodong Duan. human cognitive benchmarks reveal foundational visual gaps in mllms. *arXiv preprint arXiv:2502.16435*, 2025.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *arXiv preprint arXiv:2412.00947*, 2024.

Aditya Sanjiv Kanade and Tanuja Ganu. Do you see me: A multidimensional benchmark for evaluating visual perception in multimodal llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7285–7326, 2026.

Eunsoo Lee, Jeongwoo Lee, Minki Hong, Jangho Choi, and Jihie Kim. Visdot: Enhancing visual reasoning through human-like interpretation grounding and decomposition of thought. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 610–640, 2026.

Haokun Li, Yazhou Zhang, Jizhi Ding, Qiuchi Li, and Peng Zhang. Visual room 2.0: Seeing is not understanding for mllms. *arXiv preprint arXiv:2511.12928*, 2025.

Yifan Li, Zhenghao Chen, Ziheng Wu, Kun Zhou, Ruipu Luo, Can Zhang, Yufei Zhan, Xin Zhao, Minghui Qiu, et al. Unleashing perception-time scaling to multimodal reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 44407–44436, 2026.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Ningxin Pan, Hanyu Li, and Yehui Tang. Complexityworld: Benchmarking vision-language models on verifiable visual decision making. *arXiv preprint arXiv:2608.07584*, 2026.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Chenggang Rong, Tao Han, Zhiyuan Zhao, Yaowu Fan, Jia Wan, Song Guo, Yuan Yuan, and Junyu Gao. Unicbench: Unified counting benchmark for mllm. *arXiv preprint arXiv:2603.00595*, 2026.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint arXiv:2603.06148*, 2026.

Yuanfu Sun, Yuanhang Ren, Kang Li, Chuanhao Ji, Jiaxi Li, Jiajin Liu, Ninghao Liu, and Qiaoyu Tan. Graphverse: A comprehensive visual graph reasoning benchmark for multimodal large language models. *arXiv preprint arXiv:2608.06769*, 2026.

Yunlong Tang, Pinxin Liu, Mingqian Feng, Zhangyun Tan, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, et al. Mmperspective: Do mllms understand perspective? a comprehensive benchmark for perspective perception, reasoning, and robustness. *Advances in Neural Information Processing Systems*, 38, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 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, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

Yana Wei, Hongbo Peng, Yanlin Lai, Liang Zhao, Kangheng Lin, En Yu, Keyu Lv, Han Zhou, Yin Tang, Haodong Li, et al. Perceptionrubrics: Calibrating multimodal evaluation to human perception. *arXiv preprint arXiv:2606.28322*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Genpei Zhang. The visual insensitivity gap: Diagnosing when vision-language models fail to use visual evidence. *arXiv preprint arXiv:2609.00868*, 2026.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench $^{+}$: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024.

A Benchmark Construction Details

A.1 Model Roles and Construction Records

BlameBench retains the link between a downstream problem and the measurement derived from it. Perception questions are constructed from observations in successful solution trajectories and verified against the original images. Knowledge probes are constructed from concepts needed to interpret source questions or specialized figures. Table 5 specifies the construction roles. Figure 7 illustrates the resulting records using two questions in the fixed evaluation set.

Table 5: **Construction models and their roles.** Reference-answer review, difficulty screening, and final selection are separate stages.

Model	Role
Gemini 3.7 Flash	Extract observations; verify facts against images; generate perception questions; assign skill and carrier labels.
GPT-5.6 Terra	Extract knowledge and generate concept probes; answer candidates without images.
Step 3.7 Flash	Screen source problems and answer derived questions in three independent image-based runs.
GPT-5.6 Sol	Review references after the difficulty screen.
GPT-5.6 Luna	Score content quality for final curation.

A.2 Perception Question Construction

Trajectory selection and observation extraction. The source collection contains 9,258 problems; 9,036 have at least one correct-final-answer trajectory. We retain all available correct trajectories for eligible source problems that Step 3.7 Flash answers incorrectly in its single-run source screen, obtaining 16,052 reference trajectories. The extractor sees only the trajectory text. It copies minimal continuous spans describing objects, attributes, counts, spatial relations, visible text, or chart readings:

> Copy each span EXACTLY, character-for-character, as it appears in the CoT.

Observation plans, hypotheses, external knowledge, and non-visual derivations are excluded. Both an observation and a later correction remain available for verification. A mechanical check records span matches after whitespace normalization; this flag is diagnostic metadata rather than an automatic rejection rule.

Image verification. The verifier receives the original image and indexed extracted spans. Its content constraint is "The fragments are your ONLY source material." It resolves visual referents, checks the observations against the image, and reconciles conflicting descriptions. Supported observations become self-contained facts with supporting span indices. Thus, the successful solution selects the evidence to inspect, and the image establishes the fact used to construct the question.

Atomic question generation. The generator receives the image, indexed facts, and fixed style examples. Each suitable fact yields one open-ended question, a short reference answer, a question type, and supporting fact indices. The question identifies its target without revealing the answer. Atomicity means one measurement target: locating a region may provide context, but answering does not require the full source solution. Facts without an unambiguous short-answer formulation are skipped. The stages yield 151,456 extracted spans, 8,761 verified facts, and 8,466 candidate perception questions.

A.3 Image-Grounded Knowledge Probes

The supplementary Knowledge category measures concepts needed to interpret source problems and figures. Its construction preserves the distinction between a visual fact and a concept applied to an image-provided referent.

Knowledge and concept origins. Given the source question, correct answer, trajectory, and image, GPT-5.6 Terra extracts transferable, self-contained propositions with type, discipline, necessity, confidence, and trajectory evidence. Instance-specific calculations, elementary arithmetic, generic image-reading instructions, and directly visible facts are excluded. The question-based route aligns source-question prerequisites with these propositions. The figure-based route identifies mechanisms, quantities, units, and other concepts presupposed by the figure. Question and figure origins remain distinct. Figure-origin concepts can extend beyond the reference trajectory when the visual material presupposes a concept that the successful solution does not state.

Probe generation. The six probe kinds test recall, recognition, distinction, application, transfer under changed conditions, or applicability boundaries. A probe may combine multiple concepts. The instruction is:

> each probe must be impossible to answer without the image, while the looking itself stays trivial.

The image supplies a necessary referent, value, or configuration; the target is use of the concept. Each probe records its question, expected answer, grading rubric, concept identifiers, kind, and difficulty. Unsuitable concepts are skipped. The branch produces 4,243 candidate propositions, 3,915 concepts (1,906 question-based and 2,009 figure-based), and 11,523 candidate probes. Probe kinds describe how concepts are tested; all belong to the single Knowledge reporting category.

A.4 Skill Annotation and Reporting Categories

Gemini 3.7 Flash receives the image, derived question, reference, closed skill definitions, and classification examples. It assigns one of 30 leaf skills according to the operation that *produces the answer*, and separately assigns one of eight visual carriers. The carrier describes format, such as a photograph, chart, document, abstract puzzle, or musical score. Human reviewers inspect both labels for consistency with the definitions. At this stage, 8,328 perception questions have valid skill and carrier labels.

Two fixed mappings consolidate the 30 leaves into 20 intermediate labels and then six perception categories. Table 6 lists the intermediate-to-category mapping and the separate Knowledge label. Leaves are merged when they share the same evidence transformation and differ in visual subtype or answer surface. Distinct operations remain separate: reading an axis value, identifying a series, and finding an extremum have different intermediate labels. Knowledge probes receive a seventh category. The hierarchy is fixed before evaluation.

Table 6: **Intermediate skills and reporting categories.** The 20 intermediate perception labels map to six categories; Knowledge is a separate reporting category. Counts refer to the fixed set of 2,500 questions.

Category	Intermediate labels	Count
Appearance	`color_identify`, `color_complex`, `fine_grained`, `shape_geometry`, `depth_3d`	442
Localization	`spatial_relative`, `spatial_grid_coord`, `spatial_orientation`, `visual_search`	441
Quantity	`count_objects`, `count_grid`, `count_ticks_lines`	441
Correspondence	`match_correspond`, `visual_compare`	181
Text reading	`ocr_label`, `ocr_text`, `ocr_notation`	210
Structured reading	`chart_axis_value`, `chart_series_id`, `chart_extremum`	343
Knowledge	`knowledge_grounding` (supplementary concept probes)	442
Total	**2,058 perception questions + 442 concept probes**	**2,500**

The measurement target determines the label even when another operation provides context. Reading text at a specified position is Text reading; returning a coordinate is Localization. Selecting the tallest plotted bar is Structured reading, whereas comparing object heights is Correspondence. Counting musical flat signs is notation reading. Figure 4 shows why source and category are useful

together: the same source contributes several measurement types, and each category occurs across source settings.

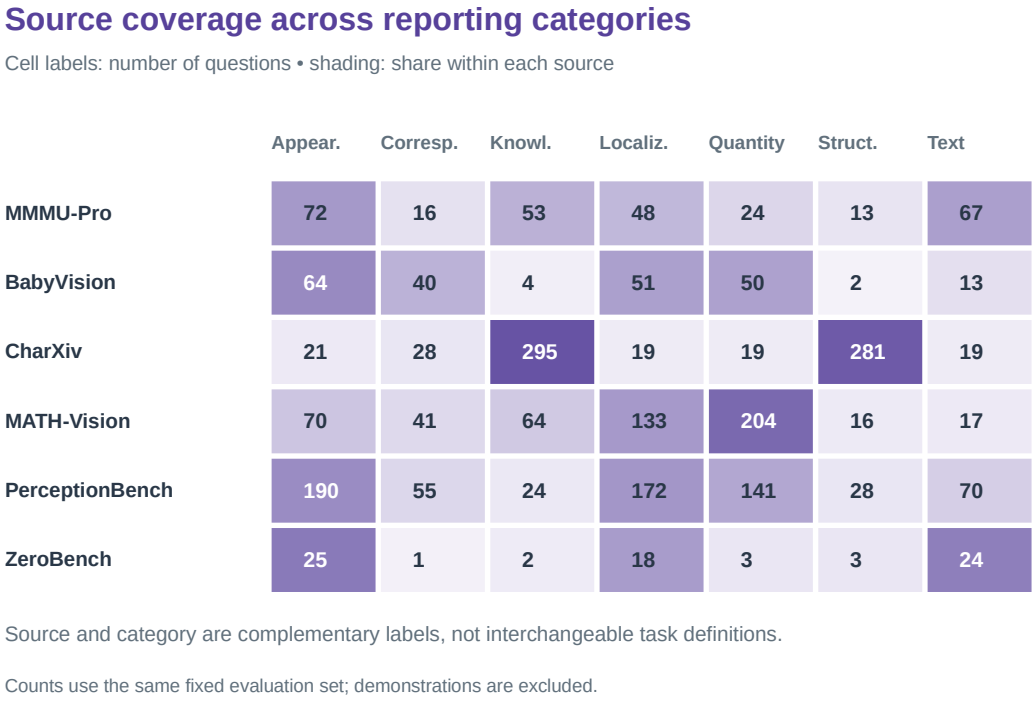


Figure 4: **Joint coverage of sources and measurement targets.** Each cell gives the number of questions; shading shows its fraction within the source. The seven columns include the supplementary Knowledge category. Source identity, visual carrier, and measurement category are distinct annotations.

A.5 Question Form and Image Reuse

The released questions use short answers while allowing enough wording to identify a precise visual target. Figure 5 summarizes question length, reference-answer length, and the number of questions associated with an image. Word counts use whitespace-delimited tokens. For these distributions, image identity uses the archived image-file hash, and repeated occurrences of the same hash are counted together. These distributions describe the fixed evaluation set and do not include prompt demonstrations.

A.6 Screening and Answer Validation

Image dependence. The two branches supply 19,989 candidates. GPT-5.6 Terra answers each derived question as text only, without images, trajectories, facts, concept statements, or references. It must commit to a short answer even when uncertain. Candidates judged correct against the normalized reference are removed, leaving 19,851. This tests text-only solvability under the answering protocol.

Difficulty with the image available. Step 3.7 Flash receives the original image and derived question in three independent runs, without construction evidence or references. For binary correctness $a_{q,r}$, the retention rule is $\sum_{r=1}^{3} a_{q,r} \leq 1$. The retained pool has 1,572 zero-correct and 3,714 one-correct candidates, totaling 5,286. This question-level screen is separate from the source-problem screen.

Reference validation. GPT-5.6 Sol reviews references after difficulty screening. Each candidate is then independently assessed by three domain experts, who see only the image, derived question, and

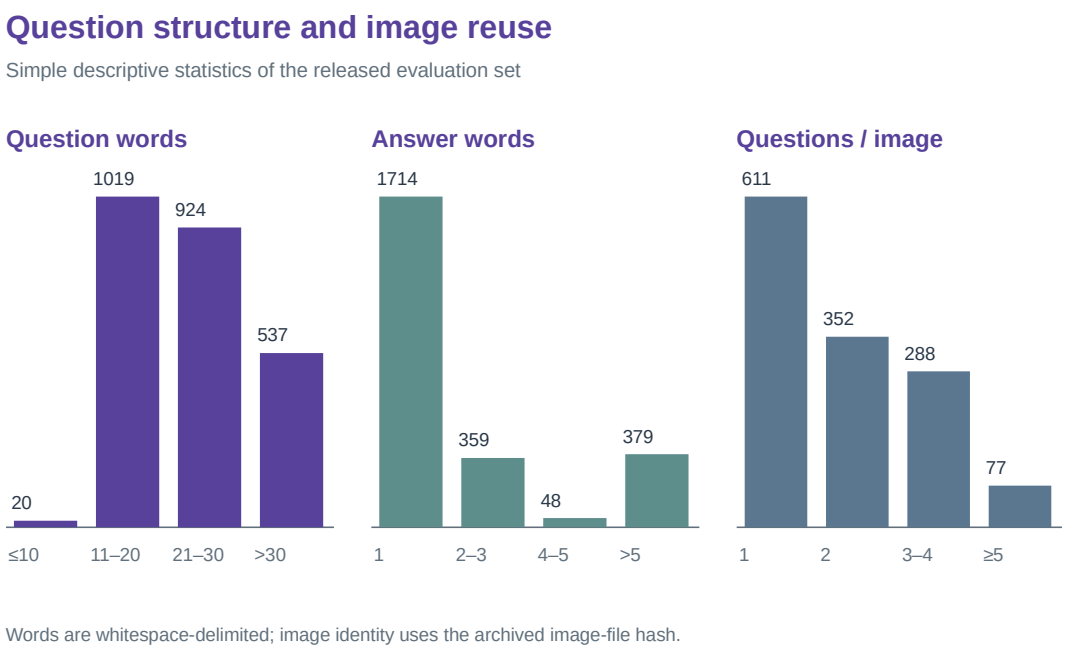


Figure 5: **Question structure and image reuse.** Most references are one word, while questions provide sufficient context to identify the requested visual measurement. The final panel reports the number of fixed-set questions associated with each image.

reference. *Correct* means that the reference is supported and answers the question; *incorrect* means it is wrong; *unclear* denotes an indeterminate judgment due to ambiguity or insufficient evidence. Only unanimous correct or incorrect labels are assigned in the first round. All other cases enter an unclear pool, reviewed by a second group of three experts; acceptance requires unanimous correct labels from this group. The protocol yields 4,035 correct candidates for final curation.

A.7 Five-Criterion Final Curation

Final selection combines coverage, image diversity, target focus, image grounding, and answerability, each in $[0, 1]$. GPT-5.6 Luna scores the last three from the image, question, reference, measurement target, and supporting material, returning an evidence-based rationale. Table 7 gives the anchors; intermediate scores are permitted.

Table 7: **Content-quality anchors for final selection.**

Criterion	**1**	**0.5**	**0**
Focus (F)	One bounded measurement or concept test.	An extra independent subtask obscures the target.	Broad, multi-step reconstruction of the source solution.
Grounding (G)	Necessary image evidence; no answer leakage.	Weak image role or partial textual shortcut.	Image unnecessary or answer disclosed.
Answerability (A)	Clear referent and determinate, supported answer.	Avoidable ambiguity in referent, precision, or scope.	No determinate, supported answer.

Let S be the selected set and $K = 2{,}500$. A candidate q has source $s(q)$, category $c(q)$, and image set $I(q)$. Write $n_s(S)$ and $n_c(S)$ for selected source and category counts, and $n_i(S)$ for the count

using image i. The coverage and diversity terms are

$$C(q \mid S) = \frac{1}{2}\left[\max\left(0, 1 - \frac{n_{s(q)}(S)}{K/6}\right) + \max\left(0, 1 - \frac{n_{c(q)}(S)}{K/7}\right)\right], \tag{2}$$

$$D(q \mid S) = \frac{1}{1 + \max_{i \in I(q)} n_i(S)}. \tag{3}$$

The equal-share reference levels are soft preferences rather than quotas. After deduplication and exclusion of demonstration images, selection repeatedly adds the candidate with the highest score

$$R(q \mid S) = \frac{C(q \mid S) + D(q \mid S) + F(q) + G(q) + A(q)}{5}. \tag{4}$$

Content scores remain fixed; coverage and diversity update after each addition. Ties use a fixed item-ID order. Selection stops at 2,500 items. For Knowledge, focus means a bounded conceptual test, and grounding requires a necessary image-provided referent or configuration with straightforward visual identification.

B Evaluation and Fine-Tuning Protocols

B.1 Frontier-Model Inference

BlameBench inputs. All 11 models receive the same 2,500 questions, one per single-turn, zero-shot request. The original images appear in their recorded order, followed by the derived question. No external tools or appended chain-of-thought instruction are used. Each run saves one answer per question with an 8,192-token output budget. Table 8 records the reasoning requests. Construction observations, facts, concept statements, rubrics, and reference answers are withheld from the evaluated model. The five BB demonstrations are used only for pretrained-checkpoint evaluation.

Table 8: **Reasoning requests in frontier BB evaluation.** These are requested inference settings; all models have the same 8,192-token output budget.

Models	Request
GPT-6 Astra; GPT-5.6 Sol	Reasoning effort set to none
Qwen 3.8 Max; Claude Opus 5; Doubao Seed 2.1 Pro; Kimi K3; DeepSeek V4.1 Flash; MiniMax M3	Thinking disabled
Gemini 3.1 Pro	Thinking budget 0
GLM 5.3 Flash	Low reasoning effort
Claude Fable 5	Adaptive thinking, low effort

Full downstream tasks. Frontier downstream evaluation uses no external tools, the highest available reasoning effort, and a requested generation budget of 256K tokens. Other parameters follow official API defaults. MMMU-Pro, BabyVision, CharXiv reasoning, MATH-Vision, and PerceptionBench use pass@1. ZeroBench uses pass@5: a problem is solved if any of five attempts is correct. BB and DS columns therefore describe their respective question sets and protocols. The pretrained-checkpoint experiment uses the separate settings below.

B.2 Answer Scoring

GPT-5.6 Luna judges semantic equivalence using exactly three text fields: the question, saved model answer, and reference answer. It receives no image, construction evidence, or per-item rubric. The judge uses medium reasoning effort, temperature 1.0, a 4,096-token completion budget, and at most three call attempts. Its saved binary verdict determines correctness. Numerical equivalence is assessed in context, with no additional fixed absolute or relative tolerance.

The fixed prompt asks the judge to explain equivalence and return `[Analysis]` followed by `[Judgment Result]` with 0 or 1. It instructs the judge to grade the supplied response rather than continue it, and to mark incomplete or repetitive answers incorrect. The user template is:

```
[QUESTION] {question}
[ANSWER] {answer}
[GROUND-TRUTH] {ground_truth}
```

Failed or ungradable requests remain in the denominator and do not count as correct. Truncated responses remain in the denominator and are judged from their saved content. Human construction-stage review establishes the references; this evaluation judge compares model outputs with those references.

B.3 Pretrained Checkpoints and Fine-Tuning

Checkpoints and prompting. We use the four public checkpoint identifiers

`OpenGVLab/InternVL3_5-{1B,2B,4B,8B}-Pretrained`.

The five paired sources are MMMU-Pro vision, BabyVision, CharXiv reasoning, MATH-Vision, and PerceptionBench. Base evaluation is five-shot on BB and each paired downstream benchmark. Each benchmark has its own five fixed demonstrations, in the same order across checkpoints. Post-SFT evaluation is zero-shot on the five downstream benchmarks. Both stages allow 8,192 new tokens within a total context limit of 32,768 tokens. Downstream scoring uses the same Luna semantic-equivalence judge through benchmark-specific adapters.

Training data and overlap filtering. The shared training subset is drawn from the public dataset `OpenDataArena/MMFineReason-SFT-123K-Qwen3-VL-235B-Thinking`, revision `cda6140bfba449bef9e57065a528b7424966e144`. Before selection, we remove matches to evaluation material using normalized question text, image-file byte hashes, decoded-pixel hashes, and 64-bit image dHash with Hamming distance at most two. We also deduplicate the training data. Eligible examples fit the 8,192-token training limit for all four checkpoints. We order eligible examples by `SHA256(seed:id)` with seed 17 and select the first 30,000. Every checkpoint uses the same subset.

Shared optimization. The language model and multimodal projector are fully trainable; the vision encoder is frozen and LoRA is not used. Table 9 gives the recipe. One pass with incomplete batch dropping yields 234 updates and 29,952 consumed examples. The same 48-example tail is omitted for every checkpoint. Evaluation uses the final update, with no early stopping.

Table 9: **Shared fine-tuning configuration.** All four checkpoints use identical data selection and optimization settings.

Parameter	Setting
Trainable / frozen modules	Language model and projector / vision encoder
Optimizer	AdamW; $\beta = (0.9, 0.999)$; $\epsilon = 10^{-8}$
Learning rate / weight decay	2×10^{-5} / 0.01
Schedule / clipping	3% linear warmup, then linear decay / max norm 1.0
Sequence length	8,192 tokens
Effective batch size	$128 = 8$ workers $\times 1$ microbatch $\times 16$ accumulation
Precision	BF16 mixed precision; FP32 optimizer states
Seed	17
Duration / checkpoint	234 updates; 29,952 examples / final update

Source-centered analysis. The paired analysis uses five sources and four checkpoints: 20 checkpoint–source pairs covering 2,424 BB questions. For each measurement, subtract the source mean over the four checkpoints, without variance rescaling. For example, MMMU-Pro's base BB scores $(14.68, 18.77, 29.69, 32.76)$ have mean 23.975 and center to $(-9.295, -5.205, 5.715, 8.785)$ percentage points. Post-SFT scores and base downstream scores each use their own source means. Pearson correlation over the resulting pairs reproduces $r = 0.834$ for BB and $r = 0.756$ for base downstream evaluation. The corresponding evaluation sizes are 2,424 and 9,158 questions, so BB

uses 26.5% as many questions. This is a comparison of relative checkpoint quality within source settings under the shared recipe.

Figure 6 expands this comparison by source. The panels use the same four checkpoint sizes and the protocols described above. Reading the curves within a source shows how the compact BlameBench measurement tracks relative checkpoint quality across the five paired downstream settings.

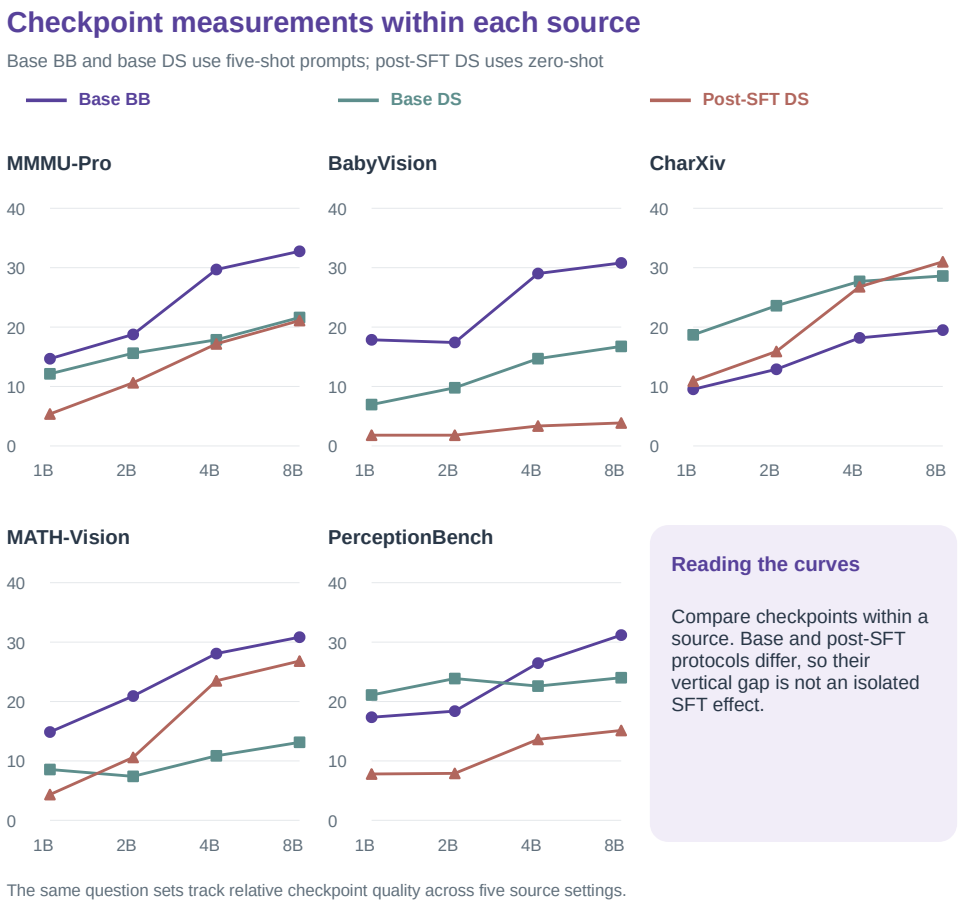


Figure 6: **Checkpoint measurements within each paired source.** Base BlameBench and base downstream evaluations use five-shot prompts; post-SFT downstream evaluation uses zero-shot prompts. The same question sets are used for every checkpoint within a source.

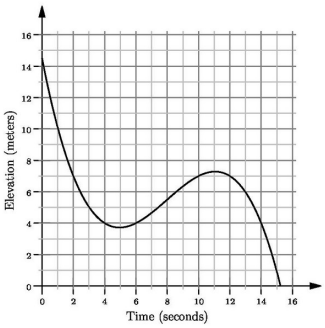
C Source-Linked Questions and Visual Evidence

C.1 From a Successful Solution to a Focused Measurement

Figure 7 illustrates the two construction branches. In the graph example, the source solution computes a total duration; the derived question isolates one elevation reading. The stored observation and verified fact make the selected evidence explicit. In the circuit example, the image supplies a current-reference arrow and the question tests the sign convention relative to it. The former measures direct perception; the latter measures a supplementary concept with a visible referent.

(a) From a successful solution to a visual measurement

SOURCE TASK
In total, how many seconds does she spend at an elevation between 4 and 7 meters?



01 Extract

At x=1 (halfway between 0 and 2), y is about 10.

02 Verify

At a time of 1 second, the curve passes through an elevation of approximately 10 meters.

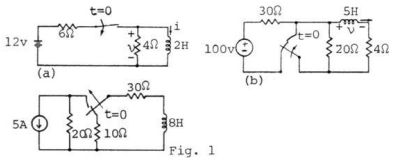
03 Ask

At a time of 1 second on the horizontal axis, what is the elevation of the curve in meters?
Answer: 10

Source answer: 8 seconds. The derived question tests one graph value.

(b) An image-grounded concept probe

SOURCE MATERIAL
A circuit and its current arrow



Knowledge

Current sign is relative to the chosen reference direction.

Concept Q6

A negative current flows opposite to the reference arrow.

Probe Q6-P3

For the current reference arrow in panel (a), compare the physical directions represented by i=+2 A and i=-2 A.
Answer: +2 A downward; -2 A upward.

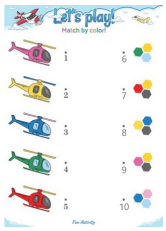
Question-origin concept; the image supplies the current-reference direction.

Figure 7: **Two source-linked construction examples.** **(a)** A trajectory observation is verified against the graph and converted into a single-value question. **(b)** The source question motivates a signed-current concept, and the image supplies the reference direction needed to apply it. Both examples belong to the fixed set. Evaluated models receive original images and derived questions; the displayed construction evidence is withheld.

C.2 Distinct Visual Operations Across Source Settings

Figure 8 covers all seven reporting categories. The first row tests an attribute, a comparison, a coordinate, and a count. The second row tests visible notation, a chart value, and a concept applied to a visible current-reference arrow. Source labels identify the originating task; category labels identify the operation that produces the answer.

Appearance



What color is the rotor blade of helicopter 1?

Answer: pink

BabyVision

Correspondence

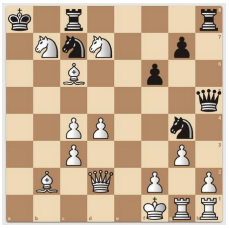


Which of the flowers labeled (B) or (D) has more green leaves on its stem?

Answer: (B)

MATH-Vision

Localization

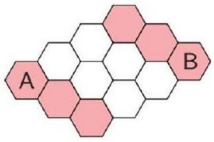


On what square is the black queen located?

Answer: h5

ZeroBench

Quantity

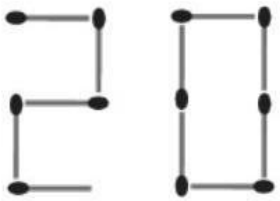


How many white hexagons are in the grid?

Answer: 6

MATH-Vision

Text reading

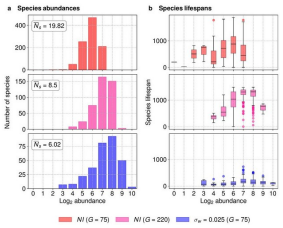


What two-digit number is formed by the matchsticks in the image?

Answer: 20

MATH-Vision

Structured reading

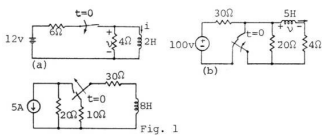


In the top-left red histogram, at what Log_2 abundance value does the tallest bar occur?

Answer: 6

CharXiv

Knowledge



For the current reference arrow in panel (a), compare the physical directions represented by i=+2 A and i=-2 A.

Answer: +2 A downward; -2 A upward.

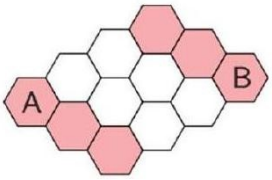
MMMU-Pro

Figure 8: **Representative questions with inspectable visual targets.** All images, questions, and references are from the fixed evaluation set. The first six categories target visible evidence; Knowledge applies a concept to an image-provided referent. Images are scaled for presentation, with their content preserved. The chart and circuit evidence also appear at larger scales in Figures 11 and 7.

C.3 Additional Focused Response Cases

Figure 9 shows four source-linked measurements: discrete counting, reading a graph value, applying a signed quantity convention, and coordinate localization. These panels illustrate that short questions can expose different answer-producing operations even when they originate from visually and conceptually rich source tasks.

Quantity / MATH-Vision

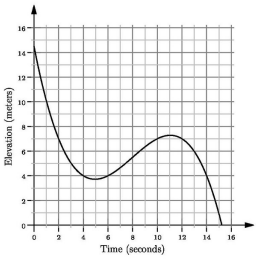


How many white hexagons are in the grid?

Reference: 6

Kimi K3	There are 7 white hexagons in the grid.	**X**
Claude Opus 5	There are 8 white hexagons in the grid.	**X**
GPT-5.6 Sol	There are 5 white hexagons.	**X**
GPT-6 Astra	There are 6 white hexagons in the grid.	**OK**

Structured reading / MATH-Vision

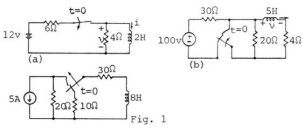


At a time of 1 second on the horizontal axis, what is the elevation of the curve in meters?

Reference: 10

Qwen 3.8 Max	Answer: 10	**OK**
Claude Opus 5	At t = 1, the curve reads approximately 10.5 meters.	**X**
GPT-5.6 Sol	At 1 second, the elevation is approximately 10 meters.	**OK**
GPT-6 Astra	At 1 second, the elevation is approximately 10 meters.	**OK**

Knowledge / MMMU-Pro



For the current reference arrow in panel (a), compare the physical directions represented by i=+2 A and i=-2 A.

Reference: +2 A: down; -2 A: up

Qwen 3.8 Max	+2 A: down / -2 A: up	**OK**
Claude Opus 5	+2 A: downward / -2 A: upward	**OK**
GPT-5.6 Sol	+2 A: downward / -2 A: upward	**OK**
GPT-6 Astra	+2 A: downward / -2 A: upward	**OK**

Localization / ZeroBench



On what square is the black queen located?

Reference: h5

Qwen 3.8 Max	the black queen is located on square h5.	**OK**
Claude Opus 5	The black queen is located on h5.	**OK**
GPT-5.6 Sol	The black queen is on h5.	**OK**
GPT-6 Astra	The black queen is on h5.	**OK**

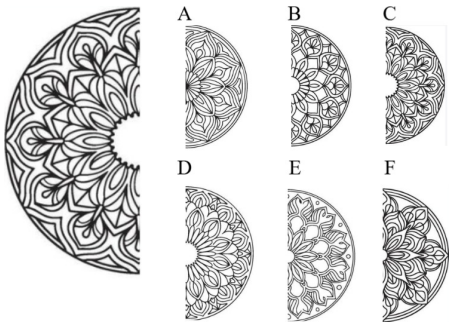
Figure 9: **Focused responses across four measurement types.** From top to bottom: Quantity, Structured reading, Knowledge, and Localization. Each displayed verdict applies to the corresponding saved response. Images are scaled for presentation; evaluation uses the originals.

C.4 Complementary Answers on the Same Visual Input

The two cases in Figure 10 make the Qwen–Claude profile differences concrete. Qwen identifies the reference option in a boundary-feature question, while Claude identifies the reference value in a histogram question. Each probe specifies one visual target that can be inspected directly. These illustrative cases connect the aggregate profiles to particular measurements; the full-set paired analysis follows in Appendix D.

(a) Identifying a local boundary feature

PerceptionBench · Correspondence



QUESTION

Which of the labeled options (A to F) features a jagged, stepped indentation on its flat left edge?

Reference: C

Qwen 3.8 Max

"Therefore, the correct option is C."

Correct

Claude Opus 5

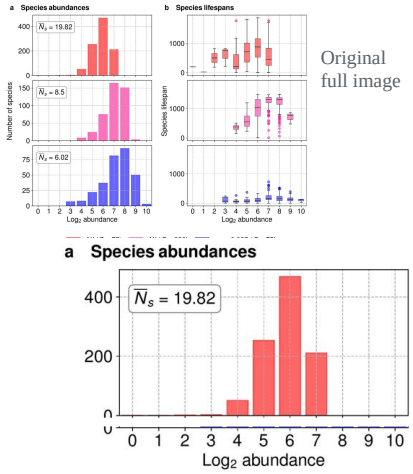
"D"

Incorrect

The requested boundary feature can be inspected independently of the complete pattern-matching task.

(b) Reading the horizontal position of a histogram peak

CharXiv · Structured reading



Original full image

Top histogram + shared axis (aligned crops)

QUESTION

In the top-left red histogram, at what Log_2 abundance value does the tallest bar occur?

Reference: 6

Qwen 3.8 Max

"the tallest bar occurs at a Log_2 abundance value of 7."

Incorrect

Claude Opus 5

"The tallest bar in the top-left red histogram occurs at Log_2 abundance = 6"

Correct

Figure 10: **Opposite outcomes at similar overall accuracy.** Each panel shows the source image, exact BB question and reference, and excerpts of saved responses with their binary verdicts. Markdown emphasis is removed for typography. **(a)** Qwen selects C and Claude selects D. **(b)** Qwen returns 7 and Claude returns 6; the tallest red bar aligns with 6 on the shared horizontal axis. The original full image is used in evaluation. These are selected examples, not estimates of category-level frequencies.

Figure 11 enlarges the histogram evidence used in one of the comparisons. The target subplot is identified in the full source image, and the crop preserves the shared horizontal axis needed to inter-

pret the tallest bar. The evaluated models receive the original full image rather than the annotated presentation view.

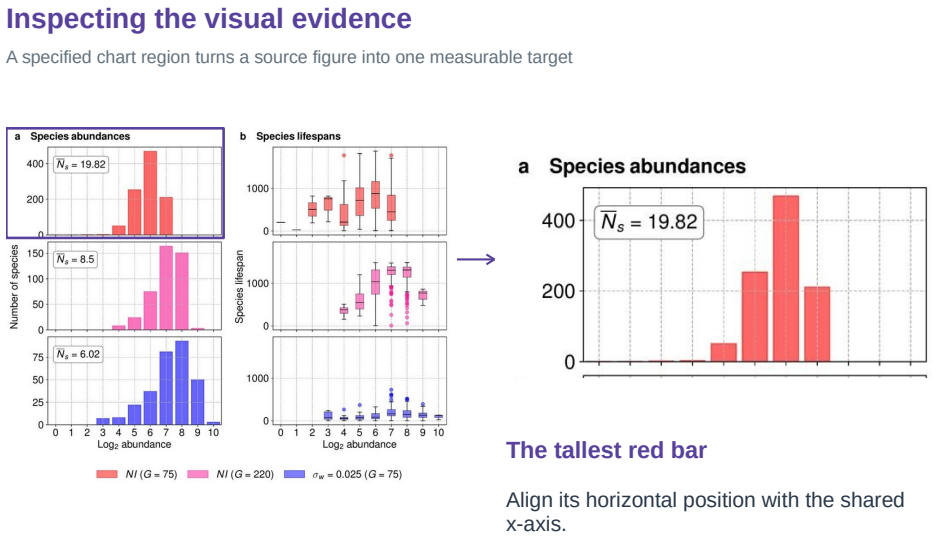


Figure 11: **Visual evidence for the histogram readout.** The target region is enlarged for inspection. The tallest red bar aligns with a Log_2 abundance value of 6 on the shared horizontal axis.

D Complementary Diagnostic Evidence

D.1 What Similar Overall Scores Leave Unresolved

Qwen 3.8 Max and Claude Opus 5 score 73.60% and 72.96% on the same 2,500 questions. Table 10 partitions their saved binary outcomes. Their overall gap is 16 correct answers, yet they differ in correctness on $344+328 = 672$ questions, or **26.88%** of the benchmark. The two directional groups nearly cancel in the total. BlameBench's item-level measurements expose this difference, and the source and category annotations identify where it occurs.

Table 10: **Paired correctness on the full fixed set.** Counts sum to 2,500; percentages use that common denominator.

Outcome	Both correct	Only Qwen	Only Claude	Neither
Questions	1,496	344	328	332
Share (%)	59.84	13.76	13.12	13.28

D.2 Locating the Complementary Measurements

Figure 12 shows the signed category differences underlying the similar overall scores. Positive values favor Qwen and negative values favor Claude Opus. The direction changes across categories: Qwen has its largest advantage on Correspondence, whereas Claude has its largest advantage on Knowledge. This profile locates the operations that contribute to the aggregate difference.

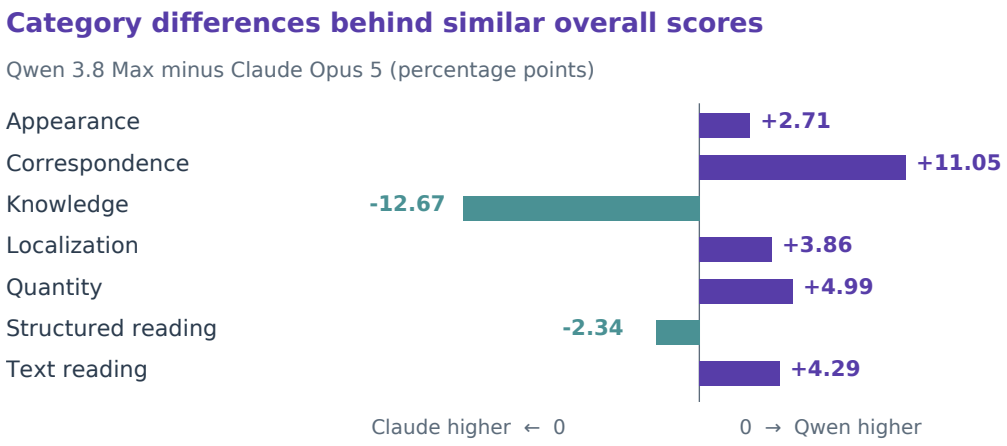


Figure 12: **Category differences behind similar overall scores.** Values are Qwen 3.8 Max minus Claude Opus 5 in percentage points, computed from the fixed-set category scores reported in the main text.

D.3 Shared and Complementary Error Sets

For model m's error set E_m, Figure 13 reports Jaccard overlap $|E_m \cap E_n|/|E_m \cup E_n|$ for the ten models on its axes; this subset excludes Claude Fable 5. Larger values indicate more similar error sets; the diagonal is one.

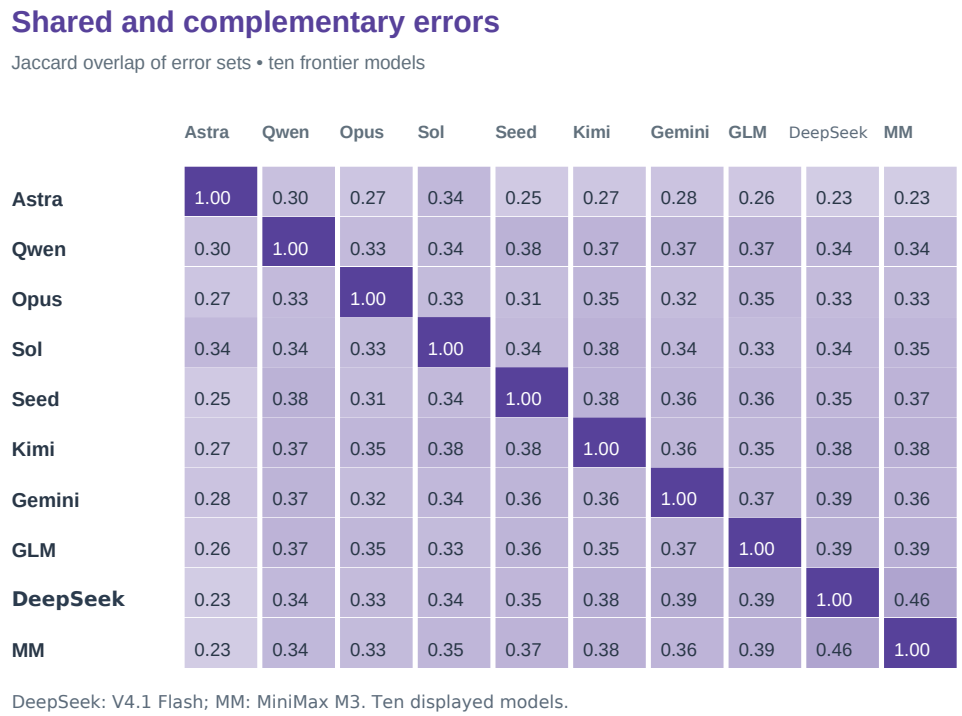


Figure 13: **Pairwise overlap of frontier-model error sets.** All outcomes are aligned on the same fixed questions before computing Jaccard overlap. The figure compares the ten models shown on its axes.