# PERCEPTIONBENCH: EVALUATING ATOMIC VISUAL PERCEPTION IN MULTIMODAL LARGE LANGUAGE MODELS

**Moonshot AI**

We introduce PerceptionBench, a benchmark specifically designed to evaluate the atomic visual perception capabilities of Multimodal Large Language Models (MLLMs). Existing benchmarks often fail to isolate perception: holistic evaluations conflate perceptual errors with failures in reasoning or domain knowledge, while application-driven benchmarks only cover narrow, fragmented domains shaped by heuristic designs. To address these limitations, PerceptionBench adopts a bottom-up approach: by diagnosing the earliest failure points in the responses of frontier MLLMs across 42 existing benchmarks, we construct an error taxonomy whose perception branch defines ten atomic perceptual capabilities. Guided by this taxonomy, we construct 3,000 verified questions with short, unambiguous answers, each isolating a single capability, with difficulty stemming from perception rather than reasoning or knowledge. Benchmark results across sixteen frontier MLLMs reveal that atomic perception remains largely unsolved—no model reaches 60% accuracy, perception-related hallucination is the weakest capability on average, and similar overall scores conceal sharply divergent capability profiles. PerceptionBench thus provides a capability-level standard for measuring and diagnosing the visual perception boundaries of MLLMs.

## 1 Introduction

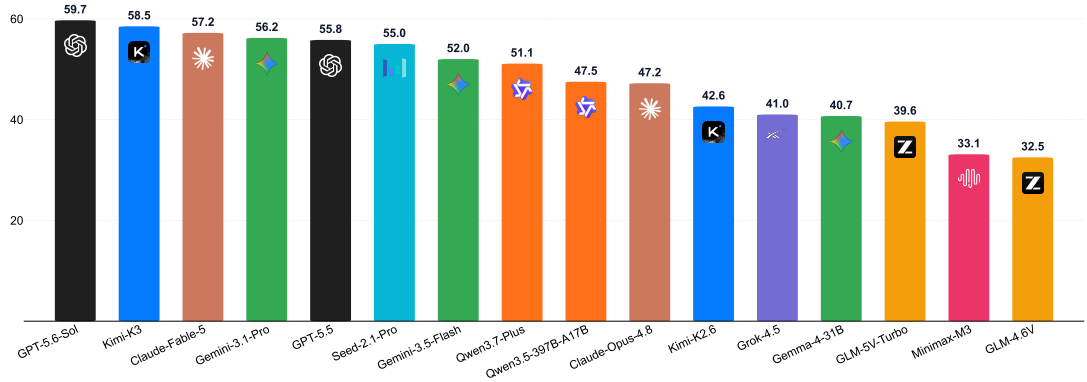



Figure 1: **Overall model accuracy on PerceptionBench.** While GPT-5.6-Sol (59.7%) and Kimi K3 (58.5%) currently lead the field, no evaluated model surpasses the 60% accuracy threshold, underscoring the significant challenge of atomic visual perception.

Accurate visual perception serves as a foundational prerequisite for Multimodal Large Language Models (MLLMs) to perform downstream reasoning and environmental interaction (Xingyu Fu et al. 2024; Rahmanzadehgervi et al. 2024; J.-T. Huang et al. 2025). However, visual perception is not a single, monolithic ability; it is composed of distinct capabilities—attribute recognition, counting, localization, and text reading—which we term atomic capabilities. Since downstream reasoning cannot reliably recover visual content that was missed, misread, or hallucinated, uncorrected perceptual errors can invalidate subsequent reasoning (Buschoff et al. 2025; Gao, Zihao Huang, et al. 2025). Consequently, mastering atomic perception represents a key leverage point for advancing MLLMs.

Blog: `https://www.kimi.com/blog/perception-bench`.
Code and data: `https://github.com/MoonshotAI/PerceptionBench`.

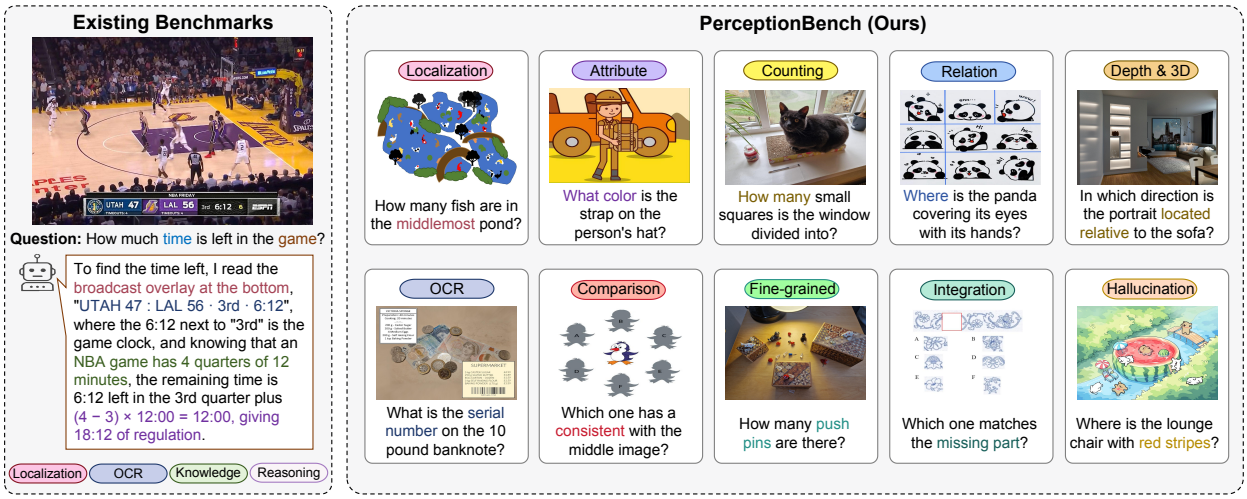



Figure 2: **PerceptionBench vs. existing benchmarks.** PerceptionBench decomposes perception into fine-grained atomic abilities, enabling more detailed and comprehensive evaluation compared to existing benchmarks.

Despite this foundational role, current evaluations rarely both isolate visual perception from the capabilities built on top of it and systematically cover the range of perceptual capabilities that models lack. Benchmarks such as MMMU (Yue, Ni, et al. 2024) and MathVista (P. Lu et al. 2024) grade final answers on multi-step problems, where performance depends not only on perception but also on reasoning and external knowledge. Neither a wrong nor a correct answer can therefore be cleanly attributed to perception alone (Figure 2, left). Meanwhile, benchmarks targeting a single perceptual application, such as OCRBench (Yuliang Liu et al. 2024) and ScreenSpot (K. Li et al. 2025), provide more direct measures of perception but are built top-down: their categories are predefined and populated with questions from a specific application domain. Each benchmark therefore measures only one slice of perception, with its coverage largely shaped by designer priors rather than by the perceptual capabilities that models lack.

We introduce PerceptionBench, a benchmark of 3,000 verified questions organized around ten atomic perceptual capabilities (Section 2). Rather than defining these capabilities from designer priors, we derive them from the failures of frontier MLLMs on 42 existing benchmarks: tracing each failure to its earliest erroneous step and clustering the resulting error labels yields a taxonomy whose perception branch defines the ten capabilities (Section 2.1.1). Some questions are decomposed from attributed failures into atomic sub-questions, inheriting their capability labels from the attribution stage; others are authored anew against the taxonomy. Every question is verified so that its difficulty stems from perception rather than reasoning or external knowledge (Figure 2, right). Because every capability is grounded in an empirically observed weakness, PerceptionBench directly measures those that models lack, providing a fine-grained diagnosis that task-oriented benchmarks are not designed to offer.

Evaluations of sixteen frontier MLLMs on PerceptionBench show that these models remain far from mastering atomic perception (Figure 1): no model reaches 60% overall accuracy, and performance is lowest, on average, on perception-related hallucination. Models with nearly identical overall scores exhibit sharply divergent capability profiles, so aggregate accuracy conceals which perceptual capabilities a model has actually acquired. A substantial fraction of correct answers are not consistently reproduced across repeated runs (Section 3.5), indicating that per-sample perception remains unstable even when aggregate scores are stable.

Our contributions are threefold. (1) We conduct a failure-attribution study of frontier MLLMs on 42 open-source benchmarks, inducing an error taxonomy whose perception branch defines ten atomic perceptual capabilities, and show that existing benchmarks probe narrow, weakly overlapping slices of the failure space. (2) We construct PerceptionBench, 3,000 verified questions organized around these capabilities and difficulty-balanced so that performance reflects perception rather than reasoning or knowledge. (3) We evaluate sixteen frontier MLLMs and provide a fine-grained analysis of their perceptual capabilities, showing that atomic perception remains a key bottleneck and that aggregate scores conceal substantial differences in capability profiles.

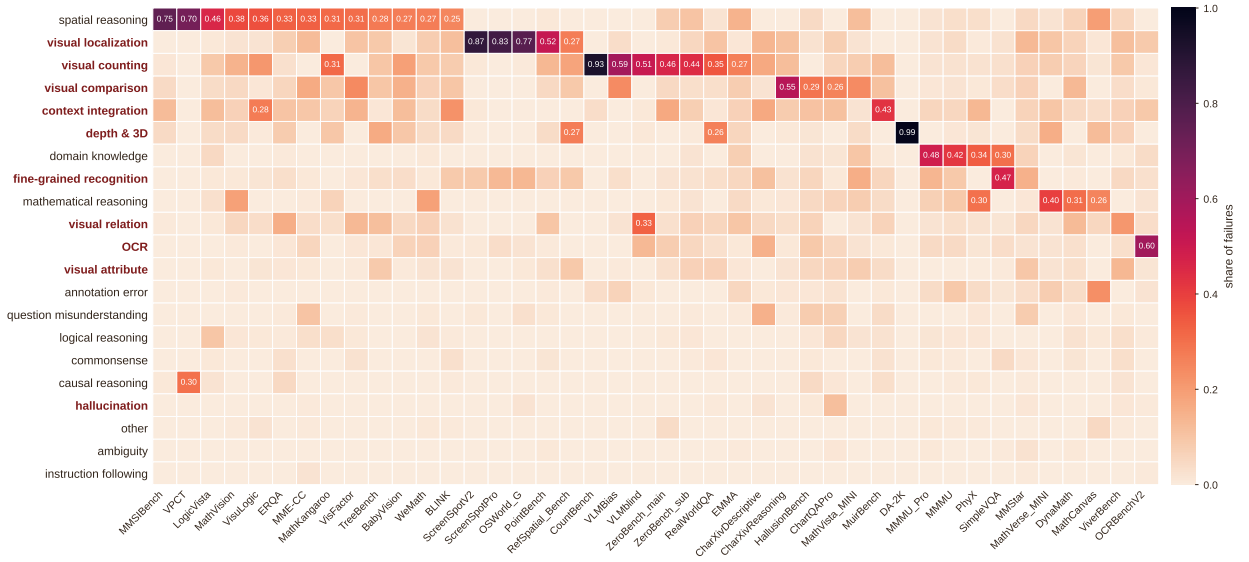



Figure 3: **Failure-type coverage of existing benchmarks.** Distribution of attributed failures across error types for each of the 42 aggregated open-source benchmarks (rows: error types, sorted by global frequency; columns: benchmarks, ordered by their dominant error type). Each benchmark's failures concentrate on one or a few error types, so existing suites probe narrow, complementary slices of the failure space, while the ten perception-branch types (bold labels) recur across nearly all benchmarks. The never-observed over-refusal type is omitted.

## 2 PerceptionBench

PerceptionBench is designed to identify and evaluate the atomic perceptual capabilities required by MLLMs. Because each downstream task exercises a combination of these capabilities, each existing benchmark observes only a sparse, application-biased slice of the capability space. Our failure-attribution study confirms this empirically: failures in each of the 42 aggregated benchmarks concentrate on one or a few error types (Figure 3), and their error-type distributions overlap only weakly across benchmarks (Figure 8). As a result, within this framework, no individual benchmark or small group of benchmarks covers the space of perceptual failures. PerceptionBench therefore follows a three-stage bottom-up pipeline that grounds every evaluation category in empirically observed model failures rather than designer priors. First, we attribute the failures of frontier MLLMs on 42 existing benchmarks to the earliest erroneous step of their reasoning trajectories and cluster the resulting error labels into a unified taxonomy, whose perception branch defines ten atomic perceptual capabilities (Section 2.1.1). Second, guided by this taxonomy, we select informative failures, decompose them into atomic sub-questions, and author additional questions, balancing the pool across categories and difficulty (Section 2.1.2). Third, every sample undergoes multi-stage verification combining automated capability-alignment and visual-grounding checks with independent human validation (Section 2.2). The released benchmark comprises 3,000 verified questions, subsampled from the constructed portion—decomposed or newly authored—of an in-house pool of more than 17,000 samples (Section 2.3).

### 2.1 Benchmark Construction

The construction of PerceptionBench proceeds in two stages, followed by the multi-stage verification in Section 2.2. We first discover a taxonomy of atomic perceptual capabilities from the failures of current MLLMs (Section 2.1.1), and then select and construct evaluation samples that faithfully and comprehensively cover this taxonomy (Section 2.1.2).

#### 2.1.1 Failure-driven Capability Discovery

Instead of manually defining perceptual capabilities from prior knowledge, we recover this structure from the empirical failures of current MLLMs. Failures reveal the perceptual phenomena that existing models have not yet reliably acquired, making them substantially more informative than correctly solved examples for understanding perceptual limitations. This construction is therefore grounded in observed failures rather than solely in designer priors; it remains conditioned on the current generation of models and the coverage of the source benchmarks, and can be re-induced as models evolve.

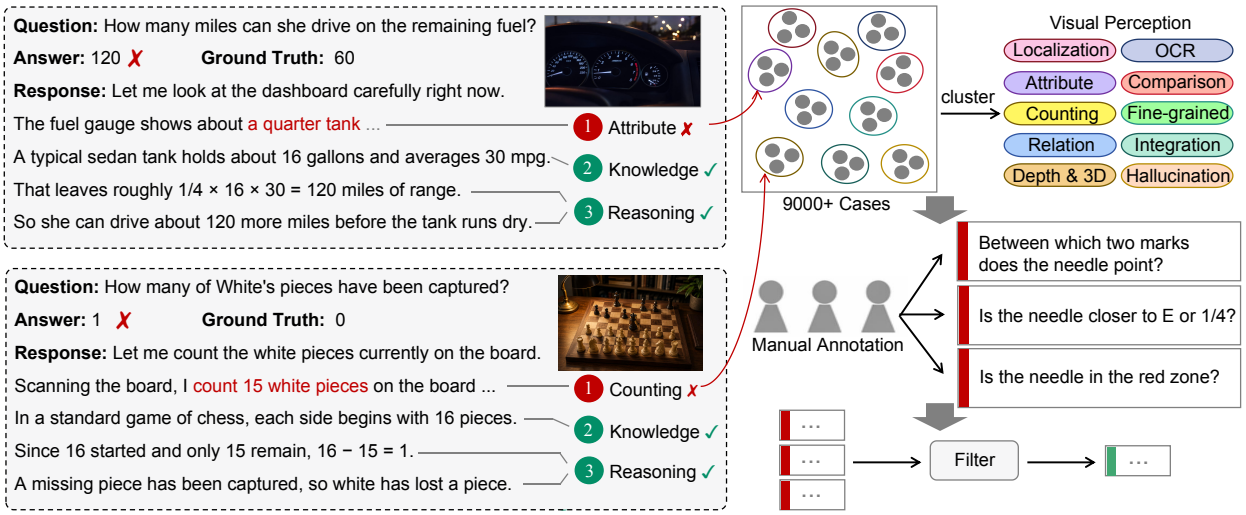



Figure 4: **Construction pipeline of PerceptionBench.** (1) Each model failure on the 42 source benchmarks is attributed to the earliest erroneous step of its reasoning trajectory (**left**); (2) cluster the attributed error labels into a unified taxonomy whose perception branch defines the ten atomic perceptual capabilities (**top right**); (3) manual annotation produces new perception questions targeting the ten atomic capabilities (**middle right**); (4) a final screening filter then removes unsuitable items, yielding the suitable questions of PerceptionBench (**bottom right**).

To obtain these informative failures, we first collect 42 benchmarks spanning diverse visual domains and application scenarios, including document understanding, optical character recognition, chart understanding, GUI interaction, visual grounding, scientific diagrams, remote sensing, and natural images. Although these benchmarks are designed for different downstream tasks, together they expose a broad range of perceptual phenomena encountered by MLLMs. We run a set of frontier MLLMs on these benchmarks and collect the samples they answer incorrectly, forming a pool of candidate failures. However, an incorrect prediction only indicates that the end-to-end inference process has failed, without revealing where the failure occurs: it may originate from visual perception, but also from reasoning, external knowledge, instruction following, or annotation ambiguity.

To identify failures that genuinely reflect perceptual limitations, we introduce a perceptual failure attribution procedure. For each candidate failure, we collect the complete prediction context, including the question, image, reference answer, and the model's reasoning trajectory. A stronger frontier model then analyzes the entire inference process, produces a free-form error analysis that pinpoints the earliest stage responsible for the incorrect answer, and assigns the failure an open-vocabulary error label. We then cluster and merge semantically equivalent labels with an LLM, consolidating them into a unified taxonomy of recurring error types. The resulting taxonomy consists of five major error classes, namely perception, reasoning, knowledge, premise, and other errors, which are further divided into 22 fine-grained error types (the complete taxonomy is provided in Appendix B). Attribution follows a first-error-priority rule: whenever the error trajectory contains a misread visual fact, the failure is attributed to the earliest erroneous step along the perceptual chain, regardless of whether downstream reasoning also fails; only when all visual facts are correctly extracted is a failure attributed to the reasoning or knowledge classes.

Since PerceptionBench targets visual perception, we retain only the perception error class, whose ten error types we adopt as the ten atomic perceptual capabilities evaluated by PerceptionBench: visual localization, visual attribute recognition, visual counting, visual relation understanding, depth and 3D perception, OCR, visual comparison, fine-grained recognition, context integration, and perception-related hallucination. The remaining classes characterize non-perceptual limitations and are used only for attribution, not for benchmark construction.

### 2.1.2 Benchmark Selection and Construction

Guided by the induced taxonomy, we select the most informative failures and construct the evaluation samples of PerceptionBench.

**Difficulty-aware selection.** Directly using the original benchmark distributions is insufficient, as a considerable proportion of existing samples are now consistently solved by frontier MLLMs and contribute little diagnostic value. We therefore evaluate all candidate samples with an ensemble of four frontier MLLMs. A sample is discarded as saturated

only when it is solved correctly by every ensemble model across all evaluation runs; that is, we remove items judged trivial by the whole ensemble. The remaining samples are stratified into three difficulty tiers according to the number of models that answer them correctly, and we sub-sample the retained items across source benchmarks and difficulty (pass rate) to counteract the skewed source distribution. The four ensemble members are disjoint from the sixteen models evaluated in Section 3.1. Difficulty is thus calibrated against current frontier models as a group rather than against any individual model. We then re-balance the retained perception samples according to the induced error-type distribution, forming the initial in-house pool of PerceptionBench.

**Perception decomposition.** To further smooth the difficulty distribution, we decompose many perception questions from the source benchmarks into finer atomic sub-questions. For every sample attributed to a perception error type, we provide annotators with its original question, the error analysis generated during attribution, and the assigned perception error type, and ask them to decompose the item into one or more atomic sub-questions, each isolating the specific perceptual capability responsible for the failure. Every constructed sample is thus grounded in an empirically observed perceptual weakness and inherits its fine-grained capability label from the attribution stage.

**Visual data and deduplication.** To increase visual diversity and improve coverage of underrepresented capabilities, we supplement the benchmark with additional images collected from public web sources and internally curated data, from which annotators author perception questions targeting the same ten capabilities. To mitigate potential contamination, all images, including both those reused from source benchmarks and those newly collected, are checked against large-scale pre-training corpora (e.g., LAION and Common Crawl). For each candidate image, we compute the cosine similarity of its Image Similarity Challenge (ISC) descriptor (Yokoo 2021) against these corpora. Any image whose similarity exceeds a strict threshold of $0.95$, i.e., near-duplicate or leaked images, is discarded. All source benchmarks are used in accordance with their original licenses (Appendix C). The decomposition and annotation are performed by more than ten professional annotators with extensive experience in multimodal evaluation. They ensure that each question admits a unique answer directly grounded in the visual content and that its difficulty arises from perception rather than additional reasoning.

**Final screening.** Finally, mirroring the discovery stage, every newly constructed sample is screened by the same ensemble of frontier MLLMs; samples solved correctly by all of them are removed, and survivors are stratified by difficulty. This keeps the benchmark challenging for frontier models without collapsing onto either trivial examples or purely adversarial cases, and because the screening is again defined by ensemble consensus, it avoids tailoring the benchmark to any individual model's failure profile.

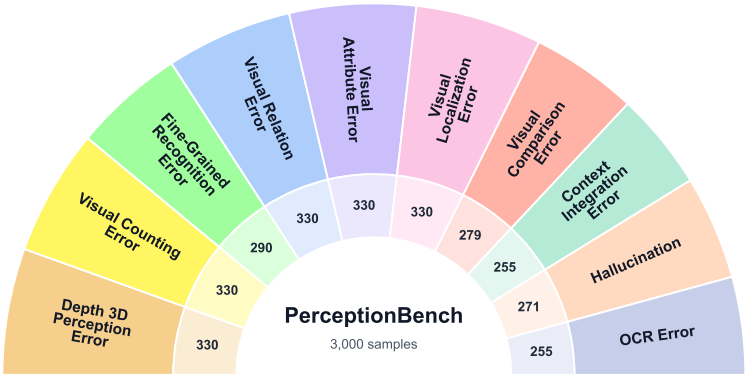
**From the in-house pool to the released benchmark.** Together, the selected samples and the constructed samples form the complete in-house benchmark of more than 17,000 verified samples, which we treat as the full evaluation pool from which the public release is drawn. The released PerceptionBench consists of 3,000 samples subsampled from the *constructed* portion of this pool, with capability-level balancing and difficulty stratification (Section 2.3). Section 3.4 further verifies that the released subset preserves the per-capability difficulty structure and relative model performance of the full pool, so evaluating on the released benchmark remains representative at less than one fifth of the full-pool evaluation cost.

### 2.2 Data Verification

We perform a multi-stage verification process to examine capability alignment, visual grounding, and annotation reliability. Samples that introduce unrelated reasoning requirements, ambiguous visual evidence, or incorrect annotations are removed or revised.

**Capability Alignment Verification.** Each sample is first evaluated by a strong multimodal verifier to determine whether the target perceptual capability is the primary factor required for solving the task. The verifier analyzes the image, question, answer, and capability annotation, and checks whether the failure of a model on this sample can be attributed to insufficient visual information acquisition rather than reasoning complexity, external knowledge, or instruction misunderstanding. Samples where the target capability is not sufficiently isolated are excluded or reassigned.

**Visual Grounding Verification.** For each retained sample, we further verify that the answer can be derived from the visual content itself. The required information must be either explicitly presented in the image or obtained through integration of multiple visual elements within the image. Samples requiring external knowledge, unsupported assumptions, or information beyond the visual input are removed to prevent evaluation from being dominated by non-perceptual factors.




| Category / Metric | Count / Value | Share (%) |
|---|---|---|
| Perception Errors | | |
| Depth 3D Perception Error | 330 | 11.00% |
| Visual Counting Error | 330 | 11.00% |
| Fine-Grained Recognition Error | 290 | 9.67% |
| Visual Relation Error | 330 | 11.00% |
| Visual Attribute Error | 330 | 11.00% |
| Visual Localization Error | 330 | 11.00% |
| Visual Comparison Error | 279 | 9.30% |
| Context Integration Error | 255 | 8.50% |
| Hallucination | 271 | 9.03% |
| OCR Error | 255 | 8.50% |

Figure 5: Distribution of PerceptionBench across the ten atomic perceptual capabilities.

**Independent Human Validation.** After automated verification, human annotators independently review the image-question-answer pairs and capability assignments. Annotators examine whether the visual evidence is sufficient, whether the question matches the intended perceptual challenge, and whether the reference answer is unambiguous. Samples with inconsistent judgments, unclear visual evidence, or annotation errors are manually corrected or discarded.

### 2.3 Benchmark Statistics

As shown in Figure 5, PerceptionBench contains 3,000 verified samples spanning the ten atomic perceptual capabilities identified from the failure-driven capability discovery process. In terms of construction source, 1,800 of the 3,000 questions (60%) are atomic sub-questions decomposed from attributed failures on the source benchmarks, while the remaining 1,200 (40%) are newly authored on supplemented images (Section 2.1.2). By design, the remaining error classes of the induced taxonomy are used exclusively for failure attribution during construction and are excluded from the released benchmark, keeping the evaluation focused on visual perception itself. Overall, the benchmark distribution follows the empirical frequency of perceptual failures observed in current MLLMs while applying capability-level balancing to prevent dominant failure modes from overwhelming the evaluation, and samples are further stratified by difficulty. This design enables PerceptionBench to provide both comprehensive coverage of visual perception capabilities and fine-grained analysis of model limitations.

## 3 Evaluation Results

### 3.1 Experiment Settings

To ensure a rigorous and fair evaluation across the diverse landscape of MLLMs, we maintain strict consistency in our experimental protocols: all models are evaluated with unified prompts and official inference settings where available. We evaluate a suite of representative frontier MLLMs spanning both proprietary and open-source families, including ten proprietary models (Claude-Opus-4.8, Claude-Fable-5, GPT-5.6-Sol, GPT-5.5, Gemini-3.5-Flash, Gemini-3.1-Pro, Qwen3.7-Plus, Seed-2.1-Pro, GLM-5V-Turbo, and Grok-4.5) and six open-source models (Kimi K3, Kimi K2.6, Qwen3.5-397B-A17B, Gemma-4-31B, Minimax-M3, and GLM-4.6V). For each model, we adopt the highest available reasoning budget. Specifically, GPT-5.5 uses “xhigh” mode, GPT-5.6-Sol, Claude-Fable-5, Claude-Opus-4.8 and Kimi K3 use “max” mode, the Gemini series and Seed-2.1-Pro use “high” effort, and Gemma-4-31B runs with thinking mode enabled. Claude-Fable-5 is evaluated with Claude-Opus-4.8 as a fallback on queries refused by its safety filter. In total, 1.1% of its reported answers are produced by Claude-Opus-4.8. All questions in PerceptionBench are open-ended free-form questions whose reference answers are short and uniquely determined. We therefore employ GPT-oss-120B as a unified judge model that compares each model response against the reference answer for automatic evaluation (see Appendix E for the evaluation prompt). On a random sample of 300 predictions, the automatic evaluation agrees with human judgments on 299 cases (99.7%), which is consistent with the design that reference answers are short and uniquely determined.

### 3.2 Main Results

**Atomic perception remains unsolved.** As detailed in Table 1, GPT-5.6-Sol leads at 59.7%, followed closely by Kimi K3 (58.5%), the top-performing open-source model, which surpasses all remaining proprietary systems including

| Model | Overall | VRel | Count | Attr | Depth | Loc | Comp | FGR | Context | OCR | Hallu |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Proprietary MLLMs** | | | | | | | | | | | |
| GLM-5V-Turbo | 39.6 | 41.2 | 40.0 | 41.2 | 41.2 | 43.9 | 45.2 | 36.6 | 32.9 | 43.5 | 28.0 |
| Grok-4.5 | 41.0 | 47.0 | 35.2 | 39.4 | 41.2 | 39.7 | 43.7 | 36.2 | 39.6 | 43.9 | 44.7 |
| Claude-Opus-4.8 | 47.2 | 51.4 | 44.2 | 49.4 | 40.6 | 58.8 | 48.4 | 40.7 | 44.7 | 54.1 | 38.7 |
| Qwen3.7-Plus | 51.1 | 59.1 | 53.3 | 55.8 | 48.5 | 52.7 | 55.9 | 46.8 | 52.2 | 54.5 | 29.5 |
| Gemini-3.5-Flash | 52.0 | 53.6 | 43.6 | 54.5 | 49.7 | 50.6 | 54.8 | 51.7 | 53.3 | 59.6 | 50.6 |
| Seed-2.1-Pro | 55.0 | 57.6 | 51.2 | 58.2 | 43.6 | 50.0 | 59.5 | 56.6 | 60.4 | 66.7 | 49.8 |
| GPT-5.5 | 55.8 | 61.9 | 55.8 | 60.9 | 48.8 | 65.8 | 65.6 | 47.2 | 58.0 | 56.5 | 34.7 |
| Gemini-3.1-Pro | 56.2 | 58.8 | 56.9 | 61.8 | 50.0 | 52.7 | 61.7 | 54.8 | 61.2 | 64.3 | 40.6 |
| Claude-Fable-5 | 57.2 | 58.5 | 52.9 | 60.9 | 51.5 | 70.4 | 56.1 | 51.6 | 59.8 | 64.3 | 45.0 |
| GPT-5.6-Sol | **59.7** | 69.7 | 62.4 | 62.1 | 55.5 | 76.7 | 67.0 | 55.9 | 60.0 | 54.9 | 26.9 |
| **Open-source MLLMs** | | | | | | | | | | | |
| GLM-4.6V | 32.5 | 35.2 | 31.8 | 35.2 | 29.1 | 30.6 | 34.8 | 29.3 | 33.7 | 39.2 | 26.9 |
| Minimax-M3 | 33.1 | 40.0 | 30.3 | 34.6 | 36.7 | 33.3 | 31.2 | 26.6 | 31.0 | 35.7 | 29.9 |
| Gemma-4-31B | 40.7 | 42.7 | 33.9 | 40.3 | 39.1 | 44.9 | 43.7 | 39.0 | 45.9 | 46.7 | 32.1 |
| Kimi K2.6 | 42.6 | 50.9 | 45.2 | 43.6 | 42.4 | 45.2 | 39.1 | 34.5 | 40.4 | 40.8 | 41.0 |
| Qwen3.5-397B-A17B | 47.5 | 55.2 | 49.1 | 53.0 | 44.6 | 46.7 | 49.8 | 44.8 | 50.2 | 52.9 | 26.9 |
| Kimi K3 | **58.5** | 68.2 | 59.7 | 59.4 | 52.4 | 70.3 | 59.1 | 55.9 | 53.3 | 61.2 | 41.7 |

Table 1: **Leaderboard evaluation on PerceptionBench.** The benchmark evaluates MLLMs across ten atomic perceptual capabilities identified from failure-driven capability induction. VRel, Count, Attr, Depth, Loc, Comp, FGR, Context, OCR, and Hallu denote visual relation, counting, attribute, depth and 3D perception, localization, comparison, fine-grained recognition, context integration, optical character recognition, and perception-related hallucination, respectively.

Gemini-3.1-Pro (56.2%) and GPT-5.5 (55.8%). On questions that require only the faithful acquisition of visual information, this shortfall underscores that atomic perception remains a key bottleneck of current MLLMs. Overall accuracy spans 32.5% (GLM-4.6V) to 59.7%, so the benchmark retains strong discriminative power far from any ceiling effect.

**Perceptual competence is unevenly developed.** Across models, perception-related hallucination has the lowest average accuracy (36.7%), well below visual relation (53.2%), OCR (52.4%), and localization (52.0%). This weakness is not limited to weaker models: GPT-5.6-Sol, the overall leader, reaches 76.7% on localization but only 26.9% on hallucination, one of the lowest hallucination scores in the leaderboard. Models with nearly identical overall scores also diverge sharply: GPT-5.5 (55.8%) and Gemini-3.1-Pro (56.2%) differ by less than one point overall, yet GPT-5.5 leads in localization by 13 points (65.8 vs. 52.7), while Gemini-3.1-Pro leads in OCR by 8 points (64.3 vs. 56.5) and in fine-grained recognition by 8 points (54.8 vs. 47.2).

**Perceptual capability and reliability decouple.** The hallucination capability exposes an instructive inversion: the overall leader, GPT-5.6-Sol, records one of the lowest hallucination scores (26.9%), whereas Gemini-3.5-Flash, which sits mid-pack overall (52.0%), achieves the best score (50.6%), closely followed by Seed-2.1-Pro (49.8%). We hypothesize that models tuned for aggressive answer commitment tend to assert plausible but non-existent visual content, inflating their scores on answerable questions while failing precisely on the samples designed to probe false-positive perception. This observation is echoed by the pass@4 versus pass$^4$ analysis in Section 3.5.

**The open-source frontier is closing in.** Kimi K3 trails the best proprietary model by only 1.2 points while outperforming every other proprietary system. Within the open-source group, perception scales with model capacity (Qwen3.5-397B-A17B at 47.5% vs. GLM-4.6V at 32.5%), yet even the strongest models leave every capability below 80%.

### 3.3 Qualitative Analysis

We further inspect individual inference samples of four frontier models (Kimi K3, GPT-5.6-Sol, Claude-Fable-5, and Gemini-3.1-Pro) on PerceptionBench. Figure 6 shows four representative cases, each decomposed from a single source-benchmark item. Although the original questions admit multi-step solutions, the models already err on many of the decomposed atomic questions. These errors are directly attributable to specific perceptual capabilities, such as counting instances or localizing objects, rather than to reasoning or knowledge. The cases echo the aggregate results: on PerceptionBench, failures are often perceptual before they are cognitive.

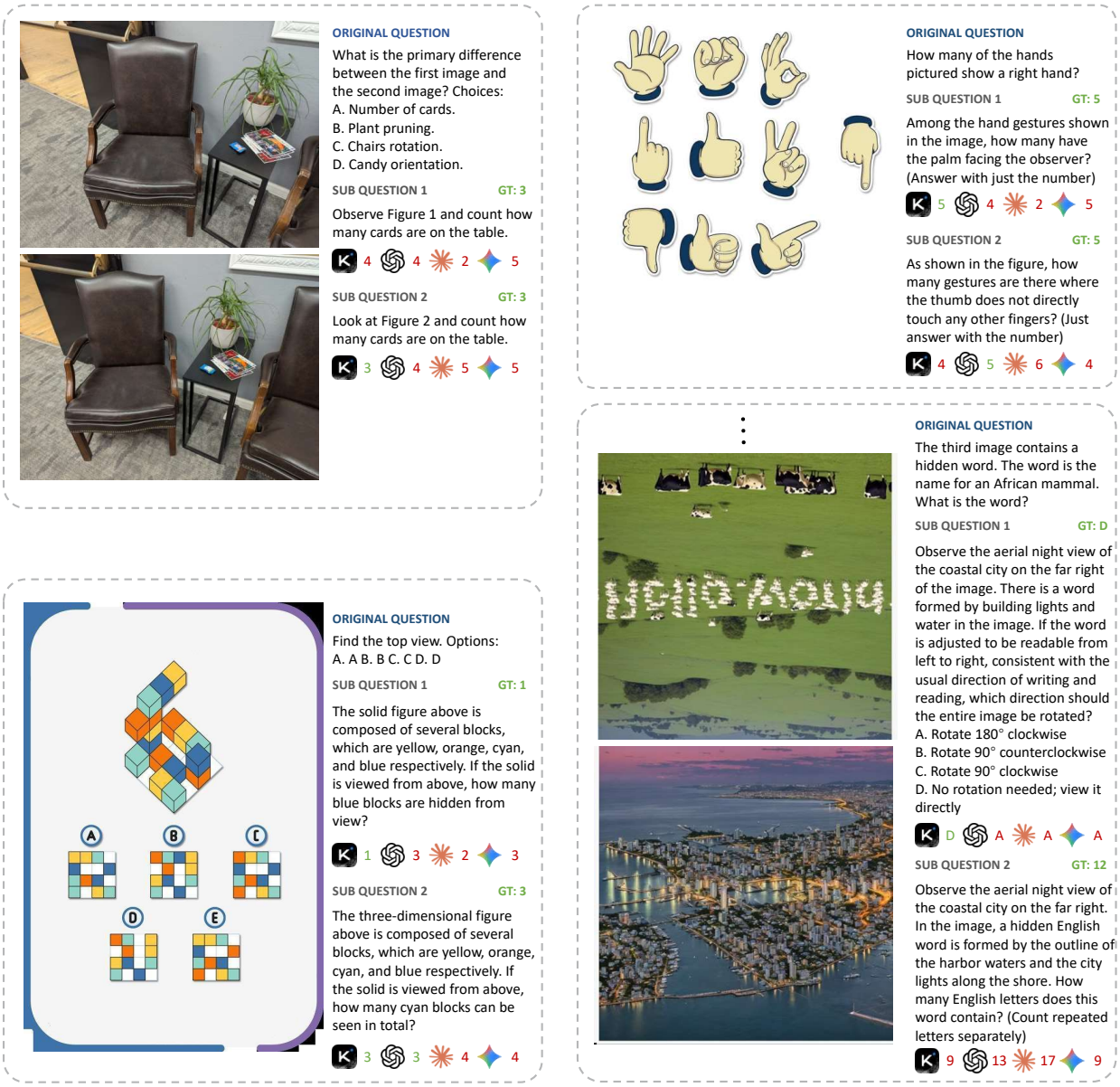



Figure 6: **Qualitative results on PerceptionBench.** Four source-benchmark items whose original questions require multi-step solutions, decomposed into atomic, perception-only questions (GT: ground truth; colored icons: answers of Kimi K3, GPT-5.6-Sol, Claude-Fable-5, and Gemini-3.1-Pro). Sources: ERQA (Gemini Robotics Team et al. 2025) (top left), MathVision (K. Wang et al. 2024) (top right), MMStar (Lin Chen et al. 2024) (bottom left), and ZeroBench (J. Roberts et al. 2025) (bottom right).

### 3.4 Benchmark Representativeness

The released benchmark is subsampled from the constructed samples of the complete in-house benchmark, which contains more than 17,000 verified samples. To evaluate whether the released version faithfully preserves the evaluation characteristics of the full in-house benchmark, we compare per-capability accuracies on the released subset against those on the full benchmark.

As shown in Figure 7, the two are strongly correlated across the ten atomic perceptual capabilities (Pearson $r = 0.84$), and the full benchmark is slightly harder than the released subset for most capabilities, as the released version smooths the difficulty distribution. This suggests that the released subset preserves the per-capability difficulty structure and relative model performance of the full benchmark while substantially reducing evaluation cost.

| Model | Run1 | Run2 | Run3 | Run4 | Mean | Std | pass@4 | pass$^4$ |
|---|---|---|---|---|---|---|---|---|
| GPT-5.6-Sol | 59.9 | 59.5 | 59.8 | 59.8 | 59.7 | 0.17 | 72.9 | 46.9 |
| Gemini-3.5-Flash | 51.4 | 51.5 | 52.5 | 52.6 | 52.0 | 0.64 | 70.2 | 32.6 |
| Kimi K3 | 58.2 | 58.2 | 58.8 | 58.7 | 58.5 | 0.32 | 73.9 | 42.7 |

Table 2: **Evaluation reliability on representative frontier MLLMs.** Each model is evaluated with four independent runs. We report the accuracy of each run together with the mean and standard deviation, as well as pass@4 (solved in at least one run) and pass$^4$ (solved in all runs). Lower standard deviation indicates higher evaluation stability, while the gap between pass@4 and pass$^4$ reflects the behavioral inconsistency of model perception.

### 3.5 Evaluation Reliability

Since MLLMs may exhibit stochastic behaviors during inference, we evaluate several frontier models using four independent runs. For each model, we report the accuracy of each run together with the mean and standard deviation. In addition, we report pass@4 (the proportion of samples solved in at least one of the four runs) and pass$^4$ (the proportion of samples solved in all four runs), which characterize the capability upper bound and the behavioral consistency of each model, respectively.

The observed standard deviations remain consistently small (Table 2): for example, the four runs of Kimi K3 span only 0.6 points (58.2–58.8, std 0.32), indicating that benchmark scores are stable and insensitive to sampling randomness, and that PerceptionBench provides reliable evaluation despite the diversity of visual tasks. Meanwhile, the gap between pass@4 and pass$^4$ reveals that a considerable fraction of samples are solved only intermittently across runs: although the aggregate score is stable, per-sample perceptual behavior is not. This suggests that current models have not robustly acquired the underlying perceptual capabilities, and part of their measured accuracy stems from unstable predictions rather than reliable visual perception.

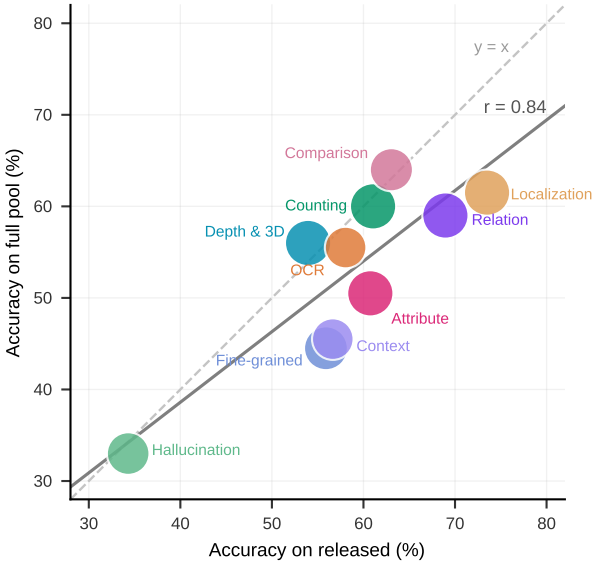



Figure 7: **Per-capability representativeness of the released PerceptionBench.** Each point represents one perceptual capability. The $x$- and $y$-axes show accuracy on the released and full benchmarks, respectively, with point size proportional to the number of samples.

## 4 Related Work

**General-purpose MLLM benchmarks.** Evaluation of MLLMs has largely been organized around tasks rather than perceptual capabilities. General-purpose benchmarks grade end-to-end answers on broad task collections. Early VQA suites (Antol et al. 2015; Hudson et al. 2019; Y. Zhu et al. 2016) test scene understanding through question answering, while recent holistic benchmarks (Yue, Ni, et al. 2024; Yuan Liu et al. 2024; Lin Chen et al. 2024; Yue, Zheng, et al. 2025) extend this paradigm to college-level reasoning and knowledge-intensive problems. Because performance on these benchmarks jointly depends on perception, reasoning, and external knowledge, their aggregate scores do not reveal which of these components accounts for a model's success or failure.

**Application-oriented visual benchmarks.** A second line of work moves closer to perception by targeting individual visual applications, including text recognition (Yuliang Liu et al. 2024; L. Fu et al. 2025; Yubo Ma et al. 2024), structured document and chart understanding (Mathew, Karatzas, et al. 2021; Masry, Long, et al. 2022; Mathew, Bagal, et al. 2022; Zirui Wang et al. 2024), GUI interaction (K. Li et al. 2025; T. Xie, D. Zhang, et al. 2024), spatial interaction (Gao, Qu, et al. 2026), and visual grounding (L. Yu et al. 2016; Krishna et al. 2017; L. Cheng et al. 2025). These suites probe perception more directly, yet each samples only the perceptual skills exercised by its target application, and many items still require domain-specific reasoning or specialized output formats. As a result, task-oriented benchmarks measure perception only indirectly and through narrow, imbalanced slices of the perceptual space. We quantify this fragmentation using the per-benchmark error-type distributions in Figure 3.

**Perception-centric benchmarks.** A more recent line of work evaluates visual perception directly using short questions designed to minimize reasoning and knowledge demands. BLINK (Xingyu Fu et al. 2024), TET (Gao, Zihao Huang, et al. 2025), and VisFactor (J.-T. Huang et al. 2025) assemble core visual cognition tasks that are relatively easy for humans but challenging for MLLMs. Similarly, Rahmanzadehgervi et al. (2024) evaluate MLLMs using deliberately simple visual questions. Other benchmarks isolate specific perceptual phenomena, including visual illusion and hallucination (T. Guan et al. 2024), multi-image spatial reasoning (S. Yang et al. 2025), infant-level visual understanding (Liang Chen et al. 2026), and extremely difficult visual queries (J. Roberts et al. 2025). Together, these studies show that frontier MLLMs continue to struggle with basic visual perception. However, their categories are defined top-down from designer priors, and each benchmark focuses on a particular slice of perception. Our failure-attribution study further shows that the failures exposed by these benchmarks overlap only weakly (Appendix D), suggesting that each captures a fragment of the perceptual space rather than its broader structure.

PerceptionBench differs from these lines of work primarily in how its evaluation categories are obtained. Rather than adopting task boundaries or designer-defined categories, it derives ten atomic perceptual capabilities bottom-up from the attributed failures of frontier MLLMs on 42 existing benchmarks. The question pool is then balanced across these capabilities and difficulty tiers. In this way, PerceptionBench retains the directness of perception-centric evaluation while grounding its capability coverage in empirically observed model weaknesses. This design provides a level of diagnostic granularity that task-oriented benchmarks are not designed to offer.

## 5 Conclusion

In this work, we introduce PerceptionBench, a benchmark for evaluating atomic visual perception in MLLMs. Its ten capability categories are derived bottom-up from attributed failures on 42 existing benchmarks, and each of its 3,000 verified questions isolates a single capability. Evaluations of sixteen frontier MLLMs suggest that atomic perception remains largely unsolved. No model reaches 60% overall accuracy, and even the strongest models leave every capability below 80%. Perception-related hallucination is the weakest capability on average, including for the top-ranked systems. Models with nearly identical overall scores exhibit sharply divergent capability profiles, so aggregate accuracy conceals which capabilities a model has actually acquired. Aggregate scores are also stable across repeated runs while per-sample behavior is not, suggesting that part of the measured accuracy stems from unstable predictions rather than reliable perception. The open-source frontier is closing in: Kimi K3 trails the overall leader by only 1.2 points.

## Limitations

Our taxonomy is induced from the failures of the current model generation, so it reflects today's weaknesses rather than a fixed structure of perception. As models improve, the taxonomy should be re-induced and the ensemble-based difficulty calibration recalibrated. Failure attribution also relies on a stronger analyzer model, and residual attribution errors may propagate into capability labels. A related caveat is that each question's capability label is inherited from first-error attribution on the current model generation. When the first erroneous step of frontier models shifts, the same item may be attributed to a different category. Finally, PerceptionBench isolates perception by design, so capability scores do not directly predict end-to-end task performance; the benchmark is best used alongside task-oriented suites. We hope that this capability-level diagnosis guides more targeted improvements of visual perception, and that the failure-driven pipeline allows the benchmark to evolve together with the models it measures.

## References


Antol, Stanislaw et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

Brower, Chase. *Visual Physics Comprehension Test*. `https://epoch.ai/benchmarks/vpct`. 2025.

Buschoff, Luca M Schulze et al. "Visual cognition in multimodal large language models". In: *Nature Machine Intelligence* 7.1 (2025), pp. 96–106.

Chen, Liang et al. "BabyVision: Visual Reasoning Beyond Language". In: *arXiv preprint arXiv:2601.06521* (2026).

Chen, Lin et al. "Are we on the right way for evaluating large vision-language models?" In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 27056–27087.

Cheng, Long et al. "PointArena: Probing Multimodal Grounding Through Language-Guided Pointing". In: *arXiv preprint arXiv:2505.09990* (2025).

Cheng, Xianfu et al. "SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 4637–4646.

Cherian, Anoop et al. "Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads". In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2024.

Fu, Ling et al. "Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning". In: *Advances in Neural Information Processing Systems* 38 (2025).

Fu, Xingyu et al. "BLINK: Multimodal Large Language Models Can See but Not Perceive". In: *European Conference on Computer Vision*. Springer. 2024, pp. 148–166.

Gao, Hongcheng, Zihao Huang, et al. "Pixels, patterns, but no poetry: To see the world like humans". In: *arXiv preprint arXiv:2507.16863* (2025).

Gao, Hongcheng, Hailong Qu, et al. "SpatialWorld: Benchmarking Interactive Spatial Reasoning of Multimodal Agents in Real-World Tasks". In: *arXiv preprint arXiv:2606.09669* (2026).

Gemini Robotics Team et al. "Gemini Robotics: Bringing AI into the Physical World". In: *arXiv preprint arXiv:2503.20020* (2025).

Guan, Tianrui et al. "HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14375–14385.

Hao, Yunzhuo et al. "Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark". In: *arXiv preprint arXiv:2501.05444* (2025).

Huang, Jen-Tse et al. "Human Cognitive Benchmarks Reveal Foundational Visual Gaps in MLLMs". In: *arXiv preprint arXiv:2502.16435* (2025).

Hudson, Drew A and Christopher D Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 6700–6709.

Krishna, Ranjay et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". In: *International journal of computer vision* 123.1 (2017), pp. 32–73.

Li, Kaixin et al. "Screenspot-pro: Gui grounding for professional high-resolution computer use". In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 8778–8786.

Liu, Yuan et al. "Mmbench: Is your multi-modal model an all-around player?" In: *European conference on computer vision*. Springer. 2024, pp. 216–233.

Liu, Yuliang et al. "Ocrbench: on the hidden mystery of ocr in large multimodal models". In: *Science China Information Sciences* 67.12 (2024), p. 220102.

Lu, Pan et al. "MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts". In: *The Twelfth International Conference on Learning Representations*. 2024.

Ma, Yubo et al. "Mmlongbench-doc: Benchmarking long-context document understanding with visualizations". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 95963–96010.

Masry, Ahmed, Mohammed Saidul Islam, et al. "ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering". In: *Findings of the Association for Computational Linguistics: ACL 2025*. 2025, pp. 19123–19151.

Masry, Ahmed, Do Xuan Long, et al. "Chartqa: A benchmark for question answering about charts with visual and logical reasoning". In: *Findings of the association for computational linguistics: ACL 2022*. 2022, pp. 2263–2279.

Mathew, Minesh, Viraj Bagal, et al. "Infographicvqa". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1697–1706.

Mathew, Minesh, Dimosthenis Karatzas, and CV Jawahar. "Docvqa: A dataset for vqa on document images". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 2200–2209.

Paiss, Roni et al. "Teaching CLIP to Count to Ten". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3170–3180.

Qiao, Runqi et al. “We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning?” In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 20023–20070.

Rahmanzadehgervi, Pooyan et al. “Vision Language Models Are Blind”. In: *Proceedings of the Asian Conference on Computer Vision*. 2024, pp. 18–34.

Roberts, Jonathan et al. “ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models”. In: *arXiv preprint arXiv:2502.09696* (2025).

Shen, Hui et al. “PhyX: Does Your Model Have the “Wits” for Physical Reasoning?” In: *arXiv preprint arXiv:2505.15929* (2025).

Shi, Weikang et al. “Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning”. In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2026, pp. 27933–27954.

Vo, An et al. “Vision Language Models are Biased”. In: *arXiv preprint arXiv:2505.23941* (2025).

Wang, Fei et al. “Muirbench: A comprehensive benchmark for robust multi-image understanding”. In: *International Conference on Learning Representations*. Vol. 2025. 2025, pp. 62624–62650.

Wang, Haochen et al. “Traceable Evidence Enhanced Visual Grounded Reasoning: Evaluation and Method”. In: *arXiv preprint arXiv:2507.07999* (2025).

Wang, Ke et al. “Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset”. In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2024.

Wang, Zirui et al. “Charxiv: Charting gaps in realistic chart understanding in multimodal llms”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 113569–113697.

Wu, Zhiyong et al. “OS-Atlas: A Foundation Action Model for Generalist GUI Agents”. In: *arXiv preprint arXiv:2410.23218* (2024).

xAI. *RealWorldQA*. `https://huggingface.co/datasets/xai-org/RealworldQA`. 2024.

Xiao, Yijia et al. “Logicvista: Multimodal llm logical reasoning benchmark in visual contexts”. In: *arXiv preprint arXiv:2407.04973* (2024).

Xie, Tianbao, Jiaqi Deng, et al. “Scaling Computer-Use Grounding via User Interface Decomposition and Synthesis”. In: *arXiv preprint arXiv:2505.13227* (2025).

Xie, Tianbao, Danyang Zhang, et al. “Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 52040–52094.

Xu, Weiye et al. “VisuLogic: A Benchmark for Evaluating Visual Reasoning in Multi-modal Large Language Models”. In: *arXiv preprint arXiv:2504.15279* (2025).

Yang, Lihe et al. “Depth Anything V2”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 21875–21911.

Yang, Sihan et al. “MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence”. In: *arXiv preprint arXiv:2505.23764* (2025).

Yokoo, Shuhei. “Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection”. In: *arXiv preprint arXiv:2112.04323* (2021).

Yu, Licheng et al. “Modeling context in referring expressions”. In: *European conference on computer vision*. Springer. 2016, pp. 69–85.

Yue, Xiang, Yuansheng Ni, 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*. 2024, pp. 9556–9567.

Yue, Xiang, Tianyu Zheng, 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)*. 2025.

Zhang, Kaiyuan et al. “MME-CC: A Challenging Multi-Modal Evaluation Benchmark of Cognitive Capacity”. In: *arXiv preprint arXiv:2511.03146* (2025).

Zhang, Renrui et al. “MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?” In: *European Conference on Computer Vision*. Springer. 2024, pp. 169–186.

Zhang, Xinchen et al. “Generative Universal Verifier as Multimodal Meta-Reasoner”. In: *arXiv preprint arXiv:2510.13804* (2025).

Zhou, Enshen et al. “RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics”. In: *arXiv preprint arXiv:2506.04308* (2025).

Zhu, Yuke et al. “Visual7w: Grounded question answering in images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4995–5004.

Zou, Chengke et al. “DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models”. In: *International Conference on Learning Representations*. Ed. by Y. Yue et al. 2025, pp. 48337–48383.

# A Contributions

Zichao Lin* Yifeng Xie* Bowen Qu* Haiming Wang Jia Li Haoning Wu Yuhao Dong Zuhao Yang Jinguo Zhu Haoyu Lu Zijia Zhao Tongtian Yue Zhangyang Qi Junwei Yang Mengfan Dong Peizhou Cao Chenzhuang Du Zaida Zhou Haotian Yao Hao Yang Hongcheng Gao Lin Sui Weihong Li Xinxing Zu Jia Chen Yao Wang Xiaoxue Wu Yalin Wang Y. Charles Yiping Bao Yangyang Liu Zhiqi Huang† Xinyu Zhou

*Core contributors.

†Project lead.

# B Full Error Attribution Taxonomy

This appendix presents the complete error attribution taxonomy derived from the failure analysis described in Section 2.1.1. The taxonomy comprises five high-level error classes and 22 fine-grained error types, providing a comprehensive categorization of failure sources in multimodal visual understanding. Among them, the perception-related errors constitute the capability taxonomy adopted by PerceptionBench, while the remaining error classes are introduced solely to facilitate error attribution during taxonomy induction.

Table 3 summarizes the ten perception error types that define the atomic perceptual capabilities evaluated by PerceptionBench. Table 4 presents the remaining four error classes, namely reasoning, knowledge, premise, and other errors. These error types are used to distinguish non-perceptual failures during taxonomy induction, but are not included in the released benchmark.

**Attribution rule (first-error priority).** Whenever the error trajectory contains a misread visual fact, the failure is attributed to the *first* erroneous step along the perceptual chain, and mapped to the corresponding perception error type, regardless of whether downstream reasoning is also incorrect. Only when all visual facts in the trajectory are verifiably extracted correctly can the failure be attributed to the reasoning or knowledge classes.

| Perception Error Types | Definition and Disambiguation Criteria |
|---|---|
| Visual localization | Fails to locate or anchor the correct region or object in the image. (Misidentifying an object as something else belongs to *fine-grained recognition*.) |
| Visual attribute | Misjudges low-level attributes of a single object, such as color, shape, size, or texture. Attribute comparison across objects belongs to *visual comparison*. |
| Visual counting | Miscounts the number of objects or instances. |
| Visual relation | Misreads directly visible 2D spatial relations (left/right, above/below, inside/outside, etc.). |
| Depth & 3D perception | Errors in perceiving stereoscopic information such as depth ordering, occlusion, viewpoint or orientation, and 3D shape structure. |
| OCR | Fails to correctly read text, digits, or symbols in the image. |
| Visual comparison | Draws wrong conclusions when comparing two or more visual elements (larger/smaller, same/different, etc.). |
| Fine-grained recognition | Misrecognizes the category of an object or entity, or fails to distinguish visually similar categories, subtypes, and fine-grained differences. |
| Context integration | Each individual region or sub-figure is perceived correctly, but associating and aggregating information across regions, sub-figures, or images fails. |
| Hallucination | Asserts objects, text, or values that do not exist in the image. (If a corresponding real element exists but is misread, the failure belongs to the corresponding specific perception error type.) |

Table 3: Perception error types in the induced error attribution taxonomy. These ten error types define the atomic perceptual capabilities evaluated by PerceptionBench.

# C Source Benchmark List and License

Table 5 lists the 42 open-source benchmarks (including benchmark splits) aggregated for the failure-driven capability discovery pipeline described in Section 2.1.1. These benchmarks span diverse visual tasks and application domains, and after ensemble-based difficulty filtering they yield approximately 9,000 informative failure cases, which serve as

| Error Type | Definition and Disambiguation Criteria |
|---|---|
| **Reasoning Errors** | |
| Spatial reasoning | Spatial information itself is perceived correctly, but mental transformation or simulation of the spatial structure fails (rotation, folding, imagined viewpoint change, path planning, etc.). Directly misreading a visible relation belongs to *visual relation*; misperceiving depth or viewpoint belongs to *depth & 3D perception*. |
| Logical reasoning | All visual facts are extracted correctly (verifiable in the trajectory), but pure logical inference or deduction fails. |
| Mathematical reasoning | All values involved in the computation are read correctly, but pure arithmetic, geometric, or quantitative computation fails. Misreading a value itself belongs to *OCR*. |
| Causal reasoning | Facts are extracted correctly, but causal or temporal inference is wrong. |
| **Knowledge Errors** | |
| Domain knowledge | Lacks specialized domain knowledge (science, medicine, law, art, etc.). |
| Commonsense | Lacks everyday commonsense about objects, behaviors, or situations. |
| **Premise Errors** | |
| Question misunderstanding | Misunderstands the question itself before any visual step: wrong referent, missed constraints, or misinterpreted task requirements. |
| Instruction following | Understands the task correctly, but the output format, length, or structure violates the requirements. |
| Over-refusal | Refuses to answer or excessively hedges although a definite answer exists. |
| **Others** | |
| Ambiguity | The question admits multiple reasonable answers; the model answer is reasonable but inconsistent with the ground truth. |
| Annotation error | The ground truth itself is wrong; the model answer may actually be correct. |
| Other | Does not fall into any of the categories above. |

Table 4: Non-perceptual error types in the induced taxonomy. These error types are used for failure attribution during taxonomy induction and are excluded from PerceptionBench by design.

the basis for inducing the error attribution taxonomy and identifying the atomic perceptual capabilities evaluated in PerceptionBench.

| Benchmark | Benchmark | Benchmark |
|---|---|---|
| MathVision (K. Wang et al. 2024) | ChartQAPro (Masry, Islam, et al. 2025) | ERQA (Gemini Robotics Team et al. 2025) |
| PhyX (Shen et al. 2025) | HallusionBench (T. Guan et al. 2024) | MuirBench (F. Wang et al. 2025) |
| VLMBias (Vo et al. 2025) | ScreenSpot-Pro (K. Li et al. 2025) | We-Math (R. Qiao et al. 2025) |
| OCRBench-v2 (L. Fu et al. 2025) | EMMA (Y. Hao et al. 2025) | LogicVista (Xiao et al. 2024) |
| VisFactor (J.-T. Huang et al. 2025) | MathCanvas (Shi et al. 2026) | VLMs-are-Blind (Rahmanzadehgervi et al. 2024) |
| PointBench (L. Cheng et al. 2025) | BabyVision (Liang Chen et al. 2026) | CharXiv-Descriptive (Zirui Wang et al. 2024) |
| ViVerBench (X. Zhang et al. 2025) | BLINK (Xingyu Fu et al. 2024) | ZeroBench-main (J. Roberts et al. 2025) |
| MMSI-Bench (S. Yang et al. 2025) | MMStar (Lin Chen et al. 2024) | ScreenSpot-v2 (Z. Wu et al. 2024) |
| VisuLogic (W. Xu et al. 2025) | MME-CC (Kaiyuan Zhang et al. 2025) | MathVista (P. Lu et al. 2024) |
| OSWorld-G (T. Xie, Deng, et al. 2025) | CharXiv-Reasoning (Zirui Wang et al. 2024) | MathVerse (R. Zhang et al. 2024) |
| MMMU-Pro (Yue, Zheng, et al. 2025) | MMMU (Yue, Ni, et al. 2024) | VPCT (Brower 2025) |
| SimpleVQA (X. Cheng et al. 2025) | TreeBench (Haochen Wang et al. 2025) | MathKangaroo (Cherian et al. 2024) |
| DA-2K (L. Yang et al. 2024) | ZeroBench-sub (J. Roberts et al. 2025) | CountBench (Paiss et al. 2023) |
| DynaMath (Zou et al. 2025) | RealWorldQA (xAI 2024) | RefSpatial-Bench (E. Zhou et al. 2025) |

Table 5: The 42 source benchmarks aggregated for failure-driven capability discovery.

All 42 source benchmarks aggregated in our failure-attribution study are publicly available research artifacts, and we use them in accordance with their original licenses. Below, the two ZeroBench splits and the two CharXiv splits are counted under their parent benchmarks. The majority are released under permissive licenses: Apache-2.0 (MMMU, MMMU-Pro, BLINK, ScreenSpot-v2, OSWorld-G, SimpleVQA, TreeBench, DA-2K, DynaMath, RefSpatial-Bench, LogicVista, VisFactor, MathCanvas, ViVerBench, VisuLogic), MIT (ChartQAPro, PhyX, VLMBias, ScreenSpot-Pro, VLMs-are-Blind, BabyVision, ZeroBench, MathVerse, VPCT, MathKangaroo, OCRBench-v2), or BSD-3-Clause

(HallusionBench). Nine benchmarks use Creative Commons licenses that require attribution: CC-BY-4.0 (ERQA, MuirBench, MMSI-Bench, MME-CC, CountBench) or CC-BY-SA-4.0 (MathVista, MathVision, MMStar, EMMA, and CharXiv). Two benchmarks carry restrictive terms: We-Math (CC-BY-NC-4.0) and RealWorldQA (CC-BY-ND-4.0). PointBench ships without an explicit license; we use it strictly for non-commercial research. Every sample in PerceptionBench carries provenance metadata (its source benchmark and original index), and source-derived samples remain subject to the licenses of their origin benchmarks. RealWorldQA is used only for failure analysis and does not contribute any samples to PerceptionBench. Our own annotations and newly authored questions are released under Apache-2.0.

## D Pairwise Overlap Between Source Benchmarks

Figure 8 reports the pairwise weighted Jaccard overlap between the error-type distributions of the 42 source benchmarks (benchmarks are ordered as in Figure 3). The mean off-diagonal overlap is only 0.20, and the few visible off-diagonal clusters correspond to near-duplicate splits of the same benchmark family (e.g., ScreenSpot-v2/ScreenSpot-Pro, ZeroBench-main/ZeroBench-sub, CharXiv-Reasoning/CharXiv-Descriptive) or to groups built around the same task (e.g., math word problems), while overlap across distinct families remains low. This indicates that existing benchmarks provide fragmented and weakly-overlapping views of model failures, and that aggregating them recovers complementary slices of the failure space.

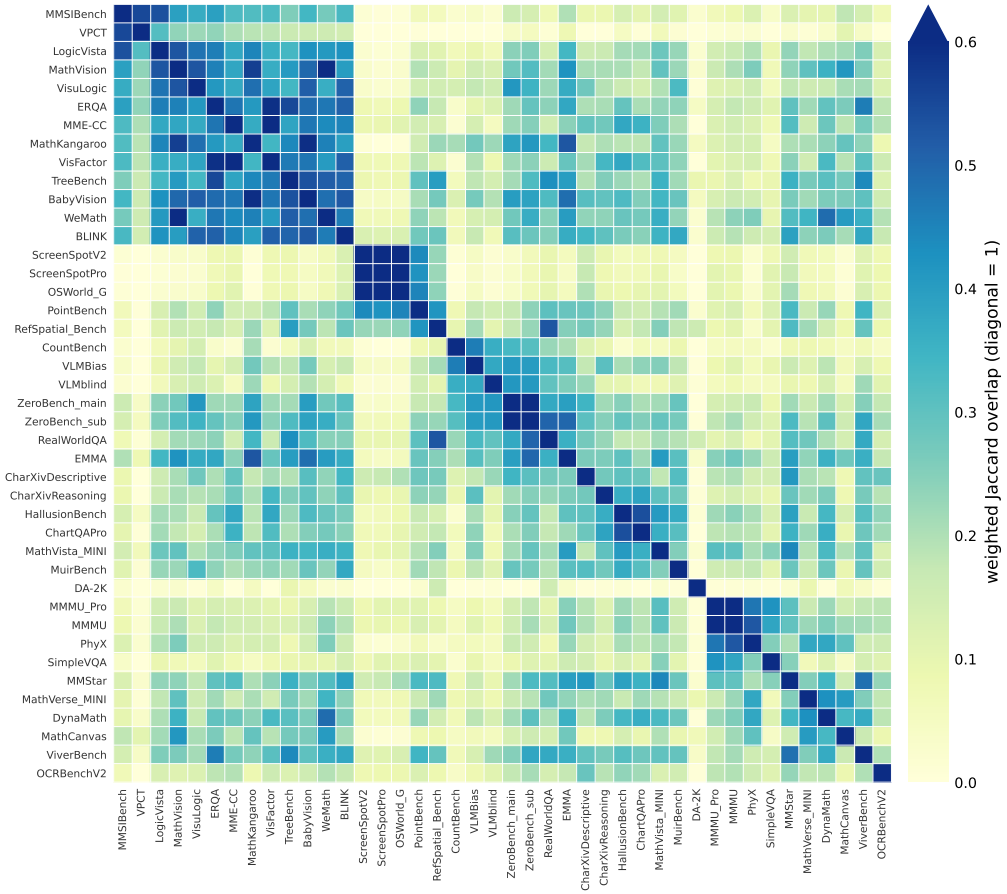



Figure 8: **Pairwise weighted Jaccard overlap between the error-type distributions of the 42 aggregated benchmarks.** The mean off-diagonal overlap is 0.20; off-diagonal clusters are limited to same-family splits and same-task groups.

# E Evaluation Prompt

**Judge Prompt**

Please act as a professional teacher and grade the student's answer. Below are the question, the student's answer, and the reference answer. Based on the question and the reference answer, analyze the student's answer and judge whether it correctly answers the question. I will provide several examples to help you understand how to judge.
========== [Output Format] ==========
Return your judgment in the following format. The first field [reason] gives the reason for your judgment; the second field [judge] gives your verdict as a single boolean, i.e., True or False. Make sure your output ends with True or False.
[reason]
<a brief justification of no more than 100 tokens>
[judge]
False
========== [Notice] ==========
{NOTICE}
========== [Examples] ==========
{EXAMPLES}
========== [Notice] ==========
{NOTICE}
Now you may begin grading.
========== [Question] ==========
{problem}
========== [Student Answer] ==========
{assistant_answer}
========== [Reference Answer] ==========
{reference_answer}
========== [Your Judgment] ==========

**Judge Rules ({NOTICE})**

0. You only need to compare the consistency between the reference answer and the student's answer, focusing especially on the content after summarizing phrases such as "Final answer:". The reference answer is absolutely correct; judge solely by whether the student's final result matches it, even if the solution process is correct.
1. If the question contains multiple sub-questions, output the student's answer, the reference answer, and the consistency for each sub-question in [reason]; judge correct only when all sub-questions are consistent.
2. For multiple-answer questions, the student's answer must contain all correct answers without any extra ones. Pay special attention when the reference answer contains "or": decide whether all answers must be given based on the specific question.
3. If you believe the student's answer equals the reference answer after simplification, you must provide the complete simplification process in [reason] until the two are equal; otherwise it cannot be judged correct. You may not vaguely claim that "they can be made equal".
4. For numerical answers, integers or values with at most 4 significant figures must match exactly; values with more than 4 significant figures must match within 4 significant figures. In [reason], convert both to standard scientific notation, first compare the order of magnitude, then compare the first 4 significant figures. If the units differ, convert to a common unit first.
5. For English writing questions, if the student does not answer in English, judge it incorrect.
6. For physics, chemistry, and biology questions, if the reference answer contains a technical term, the student's answer must contain that exact term; synonyms are not accepted.
7. When the question asks to explain a term, redundant explanation is not penalized, but missing key points must be judged incorrect.
8. For multiple-choice questions, judge incorrect whenever the student's selected option differs from the reference answer, regardless of the content.

# F PerceptionBench Showcases

Figures 9–11 present representative samples covering the ten atomic perceptual capabilities in PerceptionBench, with the complete question text (including all answer options) shown for each sample.

**Visual Localization**

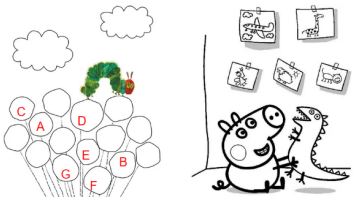
**Question:** At what o'clock position on the dial is the Gemini symbol in the image? Answer with just the number.
**Answer:** 6

**Visual Localization**

**Question:** The red lines in the image divide the picture into nine sections, which are numbered as regions 1–9 in order from left to right and then from top to bottom. Which of the following regions contains no trees at all?
A. Region 1
B. Region 2
C. Region 5
D. Region 8
**Answer:** B

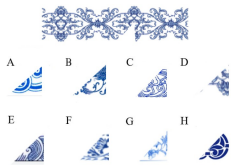
**Visual Attribute**

**Question:** From the perspective shown in the image, compare the two pen holders with pink on the table. Which of the following conclusions is correct?
A. The left pen holder is pure pink with a cartoon character pattern on its surface; the right pen holder is gray-pink patchwork
B. The left pen holder is gray-pink patchwork; the right pen holder is pure pink with a cartoon character pattern on its surface
C. The left pen holder is pure pink; the right pen holder is gray-pink patchwork with a cartoon character pattern on its surface
D. The left pen holder is gray-pink patchwork with a cartoon character pattern on its surface; the right pen holder is pure pink
**Answer:** B




**Visual Attribute**

**Question:** Observe the two cartoon characters in the lower right corner of the picture. Viewing from the perspective presented in the image, which statement correctly describes the composition of these two characters?
A. Both characters are composed entirely of curves
B. Both characters are composed of a combination of straight lines and curves
C. Both characters are composed entirely of straight lines
D. The character on the left is composed entirely of straight lines, while the character on the right is composed entirely of curves
**Answer:** B



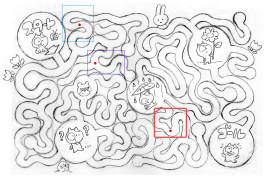

**Fine-grained Recognition**

**Question:** Observe the outline shape of the notch in the main pattern above. From the eight options A, B, C, D, E, F, G, H, which is the only option whose edge contour can perfectly match the notch in the main body and seamlessly fill the missing area?
**Answer:** D

**Fine-grained Recognition**

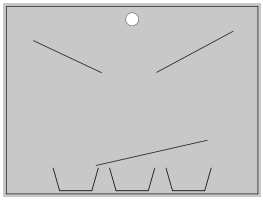
**Question:** Observe the image and compare the styling and outfits of the person in the center position (C position) with the person immediately to their right. Which of the following statements is correct:
A. The two have different hair colors, and their clothing color schemes and decorations are exactly the same
B. The two have different hair colors, and their clothing color schemes and decorations are clearly distinct
C. The two have the same hair color, and their clothing color schemes and decorations are exactly the same
D. The two have the same hair color, and their clothing color schemes and decorations are clearly distinct
**Answer:** A

Figure 9: Representative examples from the ten atomic perceptual capabilities in PerceptionBench (1 of 3).



**Visual Relation**

**Question:** Locate the solid dot inside the red box. As the dot travels along the existing route of the maze, which cat area does it reach first?
A. The puzzled cat at the lower left
B. The happy cat at the lower right
C. The cat holding an umbrella in the center
D. The cat admiring flowers at the upper right
**Answer:** C

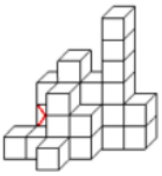


**Visual Relation**

**Question:** In the figure, for the longest black diagonal line segment, in which direction does the line segment extending outward from its right endpoint point?
A. Upper right
B. Directly to the right
C. Lower right
D. Directly downward
**Answer:** A

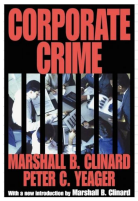
**Depth & 3D**

**Question:** Based on the current observer's perspective, with the direction closer to the camera defined as the front, between the person in red and the black electric fan in the center of the image, which one is further back?
A. The person in red
B. The black electric fan
C. Both are the same
D. Cannot be determined
**Answer:** A



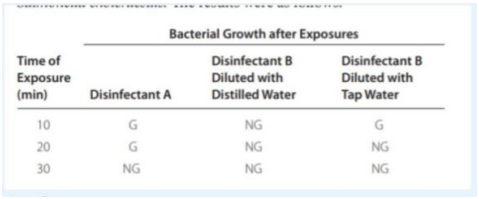
**Depth & 3D**

**Question:** If the small cubes stacked in the figure must be placed on the ground or on another small cube, how many small cubes are there in total in the figure (including the ones that cannot be seen)?
**Answer:** 26 cubes




**OCR**

**Question:** Observe the image. The white text appears in two different font sizes. For the white text in the largest font, read the text content from left to right, paying attention to uppercase and lowercase letters, and preserving all punctuation marks and spaces. Write down the answer.
**Answer:** Marshall B. Clinard



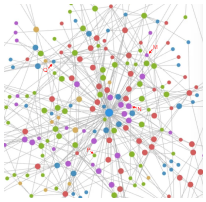
Bacterial Growth after Exposures

| Time of Exposure (min) | Disinfectant A | Disinfectant B Diluted with Distilled Water | Disinfectant B Diluted with Tap Water |
|---|---|---|---|
| 10 | G | NG | G |
| 20 | G | NG | NG |
| 30 | NG | NG | NG |

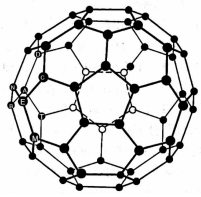
**OCR**

**Question:** (Note: bounding boxes are given in [x1, y1, x2, y2] format.) Observe the image, establish normalized coordinates, read the text content in the coordinate region [0.697, 0.429, 0.875, 0.521] from left to right, distinguish case, and preserve spaces and punctuation, then write the answer.
**Answer:** Tap Water

Figure 10: Representative examples from the ten atomic perceptual capabilities in PerceptionBench (2 of 3).



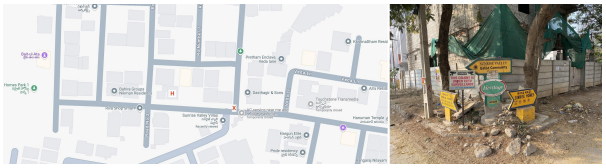
**Visual Comparison**
**Question:** Among the circles corresponding to P, Q, M, and N, which is the largest circle? (Answer with the letter corresponding to the circle)
**Answer:** N



**Visual Comparison**
**Question:** Let the line segment connecting data points A and D be denoted as segment AD, and let the line segment connecting data points E and M be denoted as segment EM. Which of the following is correct regarding the thickness of segments AD and EM?
A. Segment AD is thicker
B. Segment EM is thicker
C. Segments AD and EM are equally thick
D. Cannot be determined
**Answer:** B

**Visual Counting**
**Question:** Observe the red box in the picture. How many flowers have their main body inside the red box?
**Answer:** 2

**Visual Counting**
**Question:** How many potted green plants can be seen in the image in total? (excluding the green plants reflected in the glass)
**Answer:** 5 pots



**Context Integration**
**Question:** Please look at the map in the first image. I took another photo at the location marked with the red cross, which is the second image. Please determine which of the following buildings is on both sides of the road ROAD NO.2 SUNRISE HOMES.
A. Riva shop smart
B. Dasthagir & Sons
C. Bait-ul-Ata
D. Sunrise Valley Villas
**Answer:** B

**Context Integration**
**Question:** Compared with figure 1, what has been added to the scene in figure 2?
A. black office desk
B. semi-transparent storage basket
C. light yellow filing cabinet
D. cardboard box marked with an arrow
**Answer:** D

**Hallucination**
**Question:** How many white tables are there in the office in the picture?
A. 0
B. 1
C. 2
D. 3
**Answer:** A

**Hallucination**
**Question:** Among the 15 patterns in the image, how many patterns have a dog logo?
**Answer:** 0

Figure 11: Representative examples from the ten atomic perceptual capabilities in PerceptionBench (3 of 3).