OPTICALBENCH: A PHYSICS-GROUNDED BENCHMARK FOR VISION-LANGUAGE REASONING ON OPTICAL EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence is moving into the experimental loop of the physical sciences, proposing experiments, driving instruments and reading what they return. However, an optical system is still set up and aligned by hand, and whether a measurement is sound is judged by a human experimenter rather than by anything automatic. Vision-language models (VLMs) are a promising tool for such systems, given their increasingly strong reasoning. Yet no benchmark measures their capabilities on optical experiments across different levels of reasoning and various optical tasks. To bridge this gap, we introduce OpticalBench, a benchmark of setup–result pairs spanning six representative optical tasks, with each instance comprising the system components, their poses and parameters, the setup diagram, and the corresponding optical results. OpticalBench contains 1,961 scenes and 31,887 image inputs in total. Based on these setup–result pairs, we synthesise 7,664 question–answer pairs at four progressive levels of reasoning: Perception, Derivation, Prediction, and Decision, all grounded in physically meaningful quantities and relationships. We evaluate ten proprietary and open-weight deployments under a unified protocol, with per-item chance floors and controlled visual ablations. Experimental results reveal substantial differences across models: overall accuracy ranges from 13.3% to 65.0%, averaging 38.0% across the ten deployments. Performance also varies markedly across reasoning levels, averaging 47.9% for Perception, 39.3% for Derivation, 28.6% for Prediction, and 36.3% for Decision, revealing substantial room for improvement in reasoning over optical experiments. OpticalBench serves as a physics-grounded benchmark for measuring and advancing the capabilities of VLMs toward automated optical experimentation.

1 INTRODUCTION

Artificial intelligence is increasingly entering the experimental loop of the physical sciences, where models can propose experiments, interact with instruments, and interpret their outputs (Wang et al., 2023; Boiko et al., 2023; Szymanski et al., 2023). Vision-language models (VLMs), with their growing ability to reason jointly over visual and textual information, offer a promising route toward such systems. Yet optical experiments still rely heavily on human experimenters to set up and align components, interpret measurements, and decide whether an experimental result is physically sound. Participating in this process requires more than recognising what appears in an image: VLMs need to connect observations to the underlying physical system, reason about their relationships, predict the consequences of interventions, and determine what measurement should come next.

Existing multimodal benchmarks for science and physics primarily collect their items from examinations, textbooks, and published figures (Yue et al., 2024; Lu et al., 2022; He et al., 2024; Chow et al., 2025; Xiang et al., 2025; Wang et al., 2025; Dai et al., 2025; Ding et al., 2026), where optics appears only as isolated problems, collected images, or generated illustrations (Zhu et al., 2025). None of them presents a complete optical experiment, in which a model is given an optical setup alongside the results it produces. Consequently, whether current VLMs can reason over such experiments remains entirely unmeasured.

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.

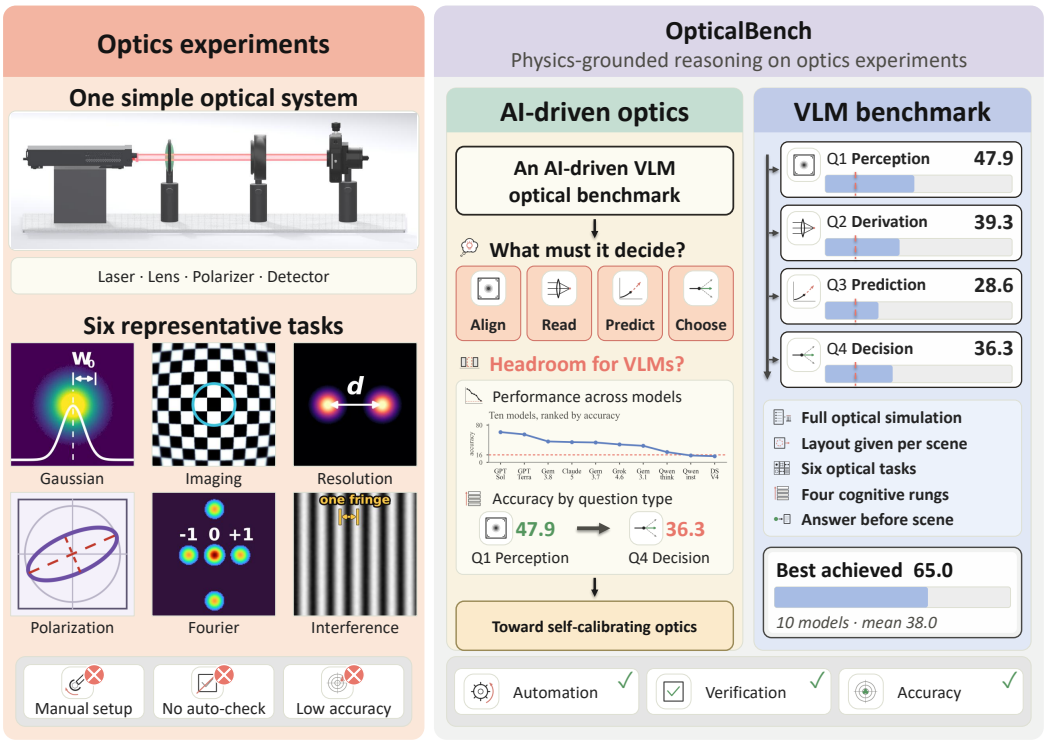


Figure 1: OpticalBench in one view. **Left:** an optical experiment and the six representative task families considered in this work. **Right:** the benchmark constructed from these experiments. Each scene pairs a declared optical setup with its corresponding results, from which physics-grounded question–answer pairs are generated at four progressive levels of reasoning: Perception, Derivation, Prediction, and Decision.

Building such a benchmark raises an annotation problem that collected figures never face: the quantities that matter in an optical experiment, such as a beam's $1/e^2$ radius, a fringe visibility, or a Stokes parameter (Figure 1, left), cannot be labelled reliably by a person looking at a camera frame. We therefore let physics annotate the data. Each instance is generated from a declared optical setup, with its components, poses, and parameters, whose results are then simulated, and every answer is computed from the simulated physical state. This physics-grounded annotation yields exact labels at scale and allows the same experiment to be regenerated at new parameter values.

We introduce **OpticalBench** (Figure 1), a benchmark of optical experiments for VLMs spanning six representative optical tasks: Gaussian, Imaging, Resolution, Polarization, Fourier, and Interference. OpticalBench contains 1,961 setup–result pairs and 31,887 image inputs, each pairing the system components, their poses and parameters, and the setup diagram with the corresponding optical results. From these pairs we synthesise 7,664 question–answer pairs at four progressive levels of reasoning: Perception, Derivation, Prediction, and Decision, forming a 6×4 grid of optical tasks and reasoning levels, together with a held-out split whose physical parameters lie outside the ranges of the main splits.

Evaluating ten proprietary and open-weight VLMs under a unified protocol yields three main findings. *First, current VLMs are far from reliable optical experimenters.* Overall accuracy ranges from 13.3% to 65.0% and averages 38.0%; the two strongest models lead all others by more than 15 points, while the open-weight models stay within six points of the 16.2% chance floor. *Second, VLMs see but do not yet measure.* The seven proprietary models tell in which direction a beam spot is displaced on the sensor in 83% of cases, yet infer the mirror tilt that displaced it in only 29%, and recover a beam's $1/e^2$ radius from its image in only 10%; averaged over all ten models, accuracy falls from 47.9% at Perception to 28.6% at Prediction. *Third, longer reasoning improves calculation but not measurement.* Letting Gemini 3.8 Flash reason at its default length instead of the minimal setting raises its Derivation accuracy by 12.6 points but leaves its Perception accuracy unchanged. These failures point to measurement tools and training on physics-grounded data as next steps.

Our contributions are threefold:

- **Physics-grounded optical benchmark:** We introduce OpticalBench, to our knowledge the first benchmark of optical experiments for VLMs, with 1,961 setup–result pairs, 31,887 image inputs, and 7,664 question–answer pairs across six optical tasks and four progressive levels of reasoning.
- **Annotation by simulation:** We compute every answer from the simulated physical state rather than labelling it by hand (Figure 2). Generating rather than collecting items yields exact labels, a chance floor computable per item, and experiments regenerated at new parameter values, as in our held-out parameter split.
- **Evaluation, analysis and insights:** We evaluate ten VLM deployments under a unified protocol. The analysis separates visual measurement from physical reasoning and locates where test-time reasoning helps: current VLMs see optical phenomena but do not yet measure them. This points to measurement tools and physics-grounded training as next steps toward automated optical experimentation.

2 Related Work

AI for physical experimentation. AI is increasingly being integrated into scientific experimentation, from proposing experimental conditions and controlling instruments to interpreting measurements and selecting subsequent actions (Wang et al., 2023; Boiko et al., 2023; Mandal et al., 2025). Autonomous and self-driving laboratories further close this loop by coupling machine learning with physical instrumentation for iterative experiment design and execution (Burger et al., 2020; Szymanski et al., 2023; Abolhasani & Kumacheva, 2023). These developments shift the role of AI from analysing scientific data after an experiment to reasoning within the experimental process itself. In optical experimentation, this requires models to connect a physical setup with its measured consequences and to reason from observation toward prediction and experimental decision-making. Purpose-built algorithms already design quantum-optical setups (Krenn et al., 2016; Klimesch et al., 2026) and align interferometers from fringe images (Sorokin et al., 2020). OpticalBench focuses on evaluating these capabilities for VLMs rather than on automating the instrumentation itself.

Multimodal benchmarks for science and physics reasoning. General multimodal science benchmarks assemble their items from examinations, textbooks and published papers (Yue et al., 2024; Lu et al., 2022; He et al., 2024; Lu et al., 2024; Hao et al., 2025), and their physics-specific successors narrow the subject while keeping the sourcing (Chow et al., 2025; Dai et al., 2025; Xiang et al., 2025; Wang et al., 2025; Ding et al., 2026). Two consequences follow, and both are where OpticalBench begins. First, the configuration behind a collected figure is not part of the item, so the item cannot be regenerated at a new parameter value, resampled to flatten a skewed answer distribution, or rerendered with one input removed, and its chance level is not knowable from the item alone. Second, optics appears here only as a subtopic of a physics chapter: GOBench, the one optics-centred multimodal benchmark we find, draws its images with generative image models rather than an optics solver, so its ground truth is a judgement of plausibility rather than a solved field (Zhu et al., 2025). We take our four cognitive rungs from this line rather than inventing them (Dai et al., 2025; Ding et al., 2026); what we add is a declared scene that stays fixed while a model climbs them.

Procedural generation for controlled physical reasoning. Generating an item rather than collecting it buys ground truth known by construction (Johnson et al., 2017; Bear et al., 2021; Rahmanzadehgervi et al., 2024), freedom from publication bias and label noise (Bentham & Srikumar, 2026), and a chance level computable per item, because the option pool is something we emit rather than something we inherit. The contamination literature supplies the motive: models answering with no image beat random choice by over 24 points on average across six benchmarks (Chen et al., 2024), text-answerable items survive until filtered out (Yue et al., 2025), removing language shortcuts costs 35 to 42 points (Burgess et al., 2025), and models lean on prior world knowledge rather than on the inputs they are given (Li et al., 2026). Every one of these controls is subtractive: neither a filter nor a rewrite can produce an item the collected pool never held, such as one drawn from a withheld parameter interval. Table E.1 in Appendix E compares OpticalBench with fourteen prior benchmarks along these lines.

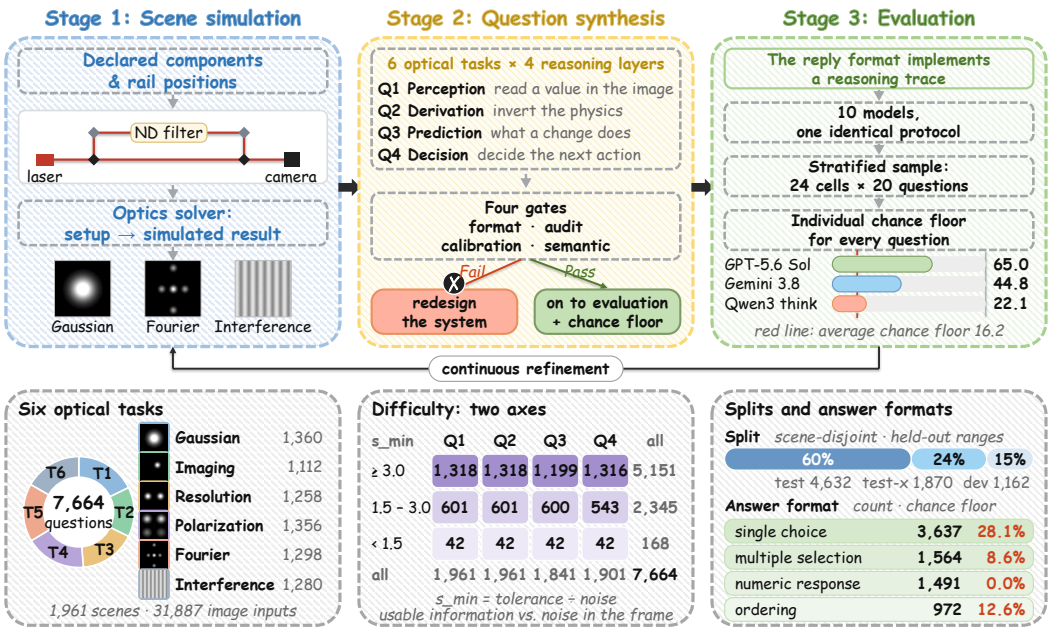


Figure 2: **Overview of the OpticalBench pipeline and benchmark composition. Top:** the three-stage pipeline for constructing and evaluating OpticalBench: scene simulation generates the optical setup and corresponding results, question synthesis constructs physics-grounded question–answer pairs through four quality-control gates, and evaluation scores VLM deployments under a unified protocol with per-item chance floors. **Bottom:** composition of the resulting benchmark, showing the distribution across six optical tasks, the two difficulty axes, and the dataset splits and answer formats.

3 METHOD

OpticalBench is constructed through three stages, illustrated in Figure 2: *scene simulation*, which defines an optical setup and computes its corresponding results; *question synthesis*, which generates physics-grounded question–answer pairs at progressive levels of reasoning; and *evaluation*, which scores VLM deployments under a unified protocol with per-item chance floors. This section describes the construction and validation of the benchmark.

3.1 REPRESENTATIVE OPTICAL TASKS

First, we introduce six representative optical tasks covering Gaussian beam propagation, imaging, resolution, polarization, Fourier optics, and interference (Born & Wolf, 1999; Saleh & Teich, 2019). Across these tasks, an optical system is assembled from common components such as light sources, lenses, and cameras, with task-specific elements introduced to realise different optical experiments. Mirrors steer the beam for imaging, wave plates and polarizers transform and analyse its polarization, gratings and a $4f$ system map spatial frequencies into diffraction orders, and beamsplitters form two coherent paths for interference (Hecht, 2017).

These configurations produce distinct optical results governed by the corresponding physical principles. Gaussian propagation produces beam profiles whose widths evolve along the optical axis (Kogelnik & Li, 1966); imaging converts beam steering into displacement of the observed spot (Saleh & Teich, 2019); resolution determines whether two nearby sources remain distinguishable (Rayleigh, 1879); polarization optics converts the state set by the wave plates into intensity measured behind an analyzer (Goldstein, 2011; Lin et al., 2025); Fourier optics produces spatially separated diffraction orders determined by the grating (Goodman, 2017); and interference produces fringes whose geometry and visibility depend on the relative angle and amplitudes of the two beams (Hariharan, 2007).

Table 1: Composition of OpticalBench across six representative optical tasks and four progressive levels of reasoning, forming 24 evaluation cells. Each cell reports the number of question–answer pairs. The final two columns report the total number of question–answer pairs and optical scenes for each task.

Family	Q1 Perception	Q2 Derivation	Q3 Prediction	Q4 Decision	Total	Scenes
T1 Gaussian	340	340	340	340	1,360	340
T2 Imaging	308	308	188	308	1,112	308
T3 Resolution	315	315	315	313	1,258	315
T4 Polarization	339	339	339	339	1,356	339
T5 Fourier	339	339	339	281	1,298	339
T6 Interference	320	320	320	320	1,280	320
Total	**1,961**	**1,961**	**1,841**	**1,901**	**7,664**	**1,961**

3.2 Optical Experiment Generation

For each optical task, we sample a set of physical parameters to define a specific optical setup, including its optical components, their parameters, positions, and poses. From this configuration, we render the setup diagram and simulate the corresponding optical results, forming a setup–result pair in which the configuration and observations remain physically consistent. The optical results are computed using complementary paraxial and wave-optics models. A paraxial ABCD solver propagates the complex q parameter through the optical system to compute Gaussian beam evolution analytically, while a diffraction solver propagates the optical field numerically using angular-spectrum and scaled-Fresnel methods (Schmidt, 2010). Ground truth is then computed directly from the simulated optical system rather than manually annotated. We validate the simulation against closed-form solutions, the FFT propagation core, and an independent quadrature reference, and verify agreement between the paraxial and diffraction solvers at camera planes where both apply.

We then convert the simulated optical intensity into camera observations using a sensor model. The intensity is integrated over the camera pixel grid at the specified pixel pitch, followed by read noise and signal quantisation to produce monochrome images.

3.3 Question–Answer Pair Generation and Dataset Composition

From each setup–result pair, we generate question–answer pairs at four progressive levels of reasoning, with task-specific physical quantities withheld from the declared setup. Across all six tasks, *Perception* identifies information directly presented in the setup and optical results; *Derivation* infers latent physical quantities from the observations and underlying physics; *Prediction* predicts the effect of an intervention; and *Decision* determines the next experimental action. Following PhysicsArena (Dai et al., 2025) and SPUR (Ding et al., 2026), these levels represent progressively deeper reasoning rather than predefined empirical difficulty.

Before entering the benchmark, each question–answer pair passes four quality-control gates for format, record consistency, physical calibration, and semantic dependence on the intended visual evidence. The resulting benchmark contains 7,664 question–answer pairs over 1,961 optical scenes and 31,887 image inputs, forming 24 cells across six optical tasks and four reasoning levels (Table 1). As summarised in Figure 2, the six tasks are broadly balanced, while difficulty is characterised along two independent axes: reasoning depth and physical resolvability. The latter is quantified by $s_{\min}$, which measures usable information relative to sensor noise and grading tolerance; configurations with $s_{\min} < 1$ are treated as unanswerable and excluded.

The benchmark is further divided into three splits: `dev` (15%) for development, `test` (60%) for the main model evaluation, and `test-x` (24%) as a held-out split with disjoint physical parameter ranges for evaluating generalisation. Questions are presented in four answer formats: single choice, multiple selection, numeric response, and ordering. The model receives the layout, the camera frames and the question, and returns one answer in a fixed JSON format (Appendix B); accuracy is the fraction of answers that match the simulated key exactly, or, for numeric responses, fall within the tolerance stated in the question. These formats accommodate different physical quantities and reasoning tasks. Each item carries a chance floor derived from its possible answer space. Numeric responses have a chance floor of zero because their answers are continuous values rather than choices from a finite set of candidates.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

With OpticalBench constructed, we next evaluate the capabilities of current VLMs across its optical tasks and reasoning levels. We evaluate ten vision-language deployments in total, including three open-weight or open-lineage deployments and seven proprietary ones. The three open deployments are DeepSeek V4 Flash Vision served by DeepInfra and the instruct and thinking variants of Qwen3-VL-235B served by Alibaba.

To ensure a fair and consistent evaluation, each question is evaluated once at temperature 0 with a completion budget of 16,000 tokens. Responses follow a strict JSON schema enforced by the API (Appendix B). Any response that cannot be parsed is scored as incorrect and remains in the denominator, preventing malformed outputs from receiving an advantage. To ensure reproducibility across open-weight deployments, we pin a single serving provider for each model and disable fallback. We request minimal reasoning of every deployment, `effort low`; DeepSeek V4 Flash Vision is the only exception: it ignores this setting and exhausts the completion budget with hidden reasoning on all three evaluator retries without returning any visible response. We therefore run this deployment with reasoning disabled.

For statistical evaluation, the average chance floor of the evaluation sample is 16.2%. We test each cell against its own chance floor using an exact one-sided binomial test and compare deployments pairwise on the same items using McNemar's test (McNemar, 1947). Both tests are Holm-corrected (Holm, 1979) for multiple comparisons. Given the evaluation budget, we construct the leaderboard from a stratified sample of the `test` split, with 20 questions from each of the 24 cells and 480 questions in total. This sample resolves paired differences of 5.7 to 8.9 percentage points. As a validation, a full-split run of one deployment achieves 44.9%, compared with 44.8% on the sampled evaluation. However, the 95% interval for an individual 20-question cell spans up to about ± 20 percentage points, so we treat per-cell results as descriptive. The leaderboard phase costs $117.63, with the breakdown reported in Appendix D; Appendix C describes the four quality-control gates and their verdicts on the frozen bank.

4.2 RESULTS ACROSS TASK FAMILIES AND COGNITIVE RUNGS

Performance comparison across VLM deployments. As shown in Table 2, GPT-5.6 Sol achieves the highest overall accuracy among all evaluated deployments at 65.0%, followed by GPT-5.6 Terra at 60.0%. Both substantially outperform the 16.2% chance floor. The next-best deployment scores 44.8%, leaving a 15.2-point gap below Terra. Performance across optical tasks, however, shows complementary strengths between the two leaders. Sol leads across all four reasoning levels and three of the six optical tasks, while Terra performs better on Gaussian (27.5% vs. 26.2%), Imaging (66.2% vs. 56.2%), and Fourier (100.0% vs. 93.8%).

Pairwise comparisons further distinguish the performance of different model groups. After Holm correction, GPT-5.6 Sol and GPT-5.6 Terra perform comparably, while both significantly outperform the other eight deployments. Gemini 3.8 Flash, Claude Sonnet 5, and Gemini 3.7 Flash also show comparable performance. Reasoning improves open-weight performance but does not close the gap to proprietary deployments. Qwen3-VL-235B improves from 14.8% in its instruct form to 22.1% in its thinking form, yet remains 13.5 points below the weakest proprietary deployment.

Five of the 24 cells use numeric responses with a chance floor of zero. Excluding these cells, GPT-5.6 Sol and GPT-5.6 Terra significantly exceed their chance floors in 15 and 14 cells, respectively, while neither instruct-level open-weight deployment does so in any non-trivial cell. These counts should be interpreted cautiously because each cell contains only 20 questions, and failure to exceed the floor significantly does not imply chance-level performance. Moreover, the chance floor does not account for imbalance in the observed answer distribution. For single-choice questions, always selecting the most frequent answer position would achieve 34%, compared with the 28.1% chance floor.

Performance across reasoning levels and optical tasks. As shown in Table 2, Perception is the highest rung for six of the seven proprietary rows, and every one of the seven reads lower at Prediction than at Perception, by between 14.1 (Grok 4.6, 45.8 to 31.7) and 37.5 (Gemini 3.7 Flash, 63.3 to

Table 2: Accuracy across four reasoning levels, in percent; higher is better. The *Above floor* column reports how many of the 24 evaluation cells each deployment performs significantly above its chance floor under a Holm-corrected one-sided binomial test. The two numbers are reported as *all/non-trivial*: the first counts all 24 cells, while the second excludes the five numeric cells whose chance floor is zero. The first row reports the average chance floor. We highlight the best and second-best result in each column.

Method	Perception	Derivation	Prediction	Decision	Above floor	**Average**
Chance floor	*11.8*	*14.8*	*16.3*	*22.0*	–	*16.2*
Open-weight and open-lineage						
DeepSeek V4 Flash Vision	10.8	15.0	7.5	20.0	2 / 0	13.3
Qwen3-VL-235B instruct	10.8	20.0	10.0	18.3	1 / 0	14.8
Qwen3-VL-235B thinking	20.8	20.0	16.7	30.8	3 / 1	22.1
Proprietary						
Gemini 3.1 Pro	56.7	26.7	24.2	35.0	10 / 7	35.6
Grok 4.6	45.8	40.8	31.7	35.8	10 / 7	38.5
Gemini 3.7 Flash	63.3	44.2	25.8	37.5	15 / 10	42.7
Claude Sonnet 5	58.3	45.8	28.3	41.7	14 / 9	43.5
Gemini 3.8 Flash	65.8	39.2	36.7	37.5	15 / 10	44.8
GPT-5.6 Terra	66.7	69.2	51.7	52.5	19 / 14	60.0
GPT-5.6 Sol	80.0	72.5	53.3	54.2	19 / 15	65.0

Table 3: Accuracy across six optical tasks, in percent; higher is better. The final column reports the overall average for each deployment. The first row reports the average chance floor for each optical task. We highlight the best and second-best result in each column.

Method	Gauss.	Imag.	Resol.	Polar.	Four.	Interf.	**Avg**
Chance floor	*10.1*	*7.2*	*12.1*	*25.0*	*23.6*	*19.2*	*16.2*
Open-weight and open-lineage							
DeepSeek V4 Flash Vision	16.2	8.8	16.2	10.0	15.0	13.8	13.3
Qwen3-VL-235B instruct	12.5	10.0	12.5	22.5	13.8	17.5	14.8
Qwen3-VL-235B thinking	8.8	17.5	25.0	33.8	25.0	22.5	22.1
Proprietary							
Gemini 3.1 Pro	17.5	26.2	38.8	60.0	45.0	26.2	35.6
Grok 4.6	17.5	25.0	32.5	63.7	41.2	51.2	38.5
Gemini 3.7 Flash	21.2	33.8	43.8	57.5	61.3	38.8	42.7
Claude Sonnet 5	21.2	38.8	47.5	52.5	72.5	28.7	43.5
Gemini 3.8 Flash	20.0	32.5	45.0	70.0	56.2	45.0	44.8
GPT-5.6 Terra	27.5	66.2	58.8	67.5	100.0	40.0	60.0
GPT-5.6 Sol	26.2	56.2	67.5	81.2	93.8	65.0	65.0

25.8); Sol falls 80.0 to 72.5 to 53.3. The ladder does not fall monotonically, though: all seven read higher at Decision than at Prediction, and Decision carries the highest floor of the four (22.0 against 16.3). The finding is a step between Perception and the three rungs above it rather than a monotone decline, the expected shape when every rung above Perception carries a just-read quantity through a relation the model must supply.

As shown in Table 3, Gaussian is the lowest-scoring family for all seven, the leader reaching 26.2 there against 93.8 on Fourier, and part of that is grading rather than physics: its Derivation cell asks for a verdict on four interval claims at once and is graded all-or-nothing: the seven average 27.1 there, although each single-claim verdict matches the key 66–81% of the time and 65 of their 140 replies assert all four claims, true together in only 4 of 20 items. We therefore name the family the lowest-scoring rather than the hardest; Imaging Perception, the sign of the displacement along each axis, is the highest-scoring cell at the other end (82.9). Figure E.1 in Appendix E resolves the same result per family and rung.

4.3 Input Ablation and Reasoning Budget

Next, we evaluate how different visual inputs affect model performance using Gemini 3.7 Flash on a separate sample of 461 questions drawn from all three splits (Table E.4). All conditions use the provider's default reasoning setting and are graded identically.

Input ablation. With the full input, accuracy reaches 51.2%. Removing all images reduces it to 17.8%, close to the 15.9% chance floor, while the layout alone achieves 17.6%. Replacing the corresponding frames with those from another optical bench yields only 21.0%. In contrast, retaining the correct frames without the layout reaches 33.0%. These results show that performance depends primarily on the visual evidence of the corresponding optical experiment rather than on textual or visual shortcuts.

Reasoning budget. We further evaluate the effect of reasoning budget using Gemini 3.8 Flash on the leaderboard questions (Table E.3, Appendix E). The provider's default reasoning setting improves overall accuracy by 4.8 points over `effort low`. The improvement is largest for Derivation (+12.6) and Decision (+8.4), while Perception remains unchanged. The overall pattern across reasoning levels is preserved under both settings, indicating that the observed gap beyond perception is not caused by the minimal reasoning setting.

4.4 Generalisation to held-out parameter ranges

We further examine whether model performance depends on the physical parameter ranges used in the standard evaluation split. For example, a model may learn the typical range of beam waists rather than generalise the underlying physical relationship. To test this, we evaluate Gemini 3.8 Flash and GPT-5.6 Sol on the `test-x` split, which contains disjoint physical parameter ranges. Following the same evaluation protocol, we sample 20 questions from each cell for a total of 480 questions.

As shown in Table E.6, neither model shows a significant performance degradation on the held-out parameter ranges. Gemini 3.8 Flash achieves 44.0% on `test-x`, compared with 44.8% on the standard evaluation split. GPT-5.6 Sol increases from 65.0% to 69.8%. This increase does not imply that held-out parameters are easier or harder in general. For example, the held-out Gaussian beam waists lie toward the larger end of the parameter range, producing larger spots that can be easier to measure. Accordingly, `test-x` evaluates generalisation across different regions of the physical parameter space rather than increased task difficulty.

4.5 Discussion

What VLMs can and cannot do in an optical experiment. Across the 24 cells, the seven proprietary models score highest where the answer is a direction, a peak position or an intensity order, such as the diffraction orders of Fourier Perception (77.9%), and lowest where a continuous quantity must be measured and then carried through a relation, such as a Gaussian beam's width predicted at another plane (11.4%) or the spot displacement predicted after a mirror tilt (19.3%). Fourier and Polarization, the only two families without a numeric-response cell, are also the only two in which these models average above 60%, and they stay on top after correcting for chance. The contrast is sharpest between two Perception cells that both ask for a length read off a camera frame in micrometres: the separation of two peaks is recovered in 80.0% of cases and a beam's $1/e^2$ radius in 10.0%, and GPT-5.6 Sol answers none of its 20 radius questions within tolerance. The wrong radii are not random: 100 of the 126 are too small, their geometric mean is 0.71 of the true radius, and 16 coincide with the half width or the full width at half maximum, consistent with placing the edge where the spot fades visibly rather than at the $1/e^2$ intensity level that an experimenter obtains from a fitted profile. The same gap separates reading from inferring within one scene: the proprietary models tell in which direction an Imaging spot is displaced in 82.9% of cases but infer the mirror tilt that displaced it in only 28.6%. Corrected for chance, they reach 45.0 on single-choice items, 38.8 on ordering, 29.8 on multiple selection and 29.9 on numeric responses.

Where the errors come from. Prior work shows that models can answer visual questions above chance without images (Chen et al., 2024), while subsequent benchmarks reduce such shortcuts through filtering (Yue et al., 2025; Burgess et al., 2025). Our controlled ablation provides a complementary test. Resolving it by task family shows what each experiment requires (Table E.5): for Gaussian, Imaging, Resolution and Fourier, adding the layout diagram to the frames raises accuracy by 13 to 37 points, consistent with the pixel pitch, distances and focal lengths that only the declared setup provides being needed to turn pixels into physical quantities, whereas Polarization is solved from the frames alone (70.3% with or without the layout), and the layout alone stays within 5.5 points of the chance floor in every family. Above Perception, the failures lie in the physics rather than in reading the frame. On the full-split run of Gemini 3.8 Flash (4,629 questions), answering a

scene's Perception question correctly barely changes the chance of answering its higher-rung questions correctly within a task family (Mantel–Haenszel odds ratios 1.05, 1.09 and 1.14 for Derivation, Prediction and Decision, each with a 95% interval containing 1), and accuracy moves by at most 6.7 points across within-cell tertiles of the usable signal $s_{\min}$. Corrected for chance, the seven proprietary models fall from 57.3 at Perception to 39.4 at Derivation and 23.5 at Prediction, with Decision at 25.7, so the rebound at Decision in raw accuracy is largely its higher chance floor.

How model responses affect accuracy. Accuracy alone may not fully reflect physical reasoning when model responses exhibit systematic biases or generation failures. Qwen3-VL-235B Instruct places 93% of its single-choice answers in the first two positions, compared with 51% of the ground-truth answers, and preserves the input order in 37% of ordering questions compared with 3% in the ground truth. These biases are substantially reduced in its thinking variant, with corresponding rates of 63% and 10%. DeepSeek V4 Flash Vision exhibits a different failure mode, reaching the completion limit without producing a valid answer on 30 of the 480 questions. In contrast, GPT-5.6 Sol shows answer distributions close to those of the ground truth. These observations suggest that benchmark scores should be interpreted together with model response behaviour, particularly when errors may arise from systematic output biases rather than physical reasoning alone. Test-time reasoning, in turn, helps with derivation but not with measurement: beyond the gains reported in Section 4.3, default reasoning raises Imaging, whose questions turn a displacement into a tilt, by 16.2 points ($p = 0.011$), but improves neither Gaussian nor Interference. The remaining headroom is large and the failures are model-specific. Only 9.6% of the leaderboard questions defeat all ten models, half of them numeric measurements and half from the Gaussian family, whereas 87.5% are answered correctly by at least one proprietary model, against 65.0% for the best single model.

Insights and future directions. Our results highlight several directions for future benchmark and model development. Controlled generation enables stronger tests of visual dependence and physical generalisation through observation substitution and held-out parameter ranges. However, our held-out evaluation currently covers only two deployments, motivating broader evaluation across different model families. Reasoning budget also remains an important evaluation dimension because its benefit varies across reasoning levels. More broadly, the largest performance gap appears between perception and higher levels of physical reasoning, highlighting derivation, prediction, and experimental decision-making as key directions for model improvement. Our analysis suggests two concrete steps. Because current VLMs cannot measure precisely by eye and more reasoning does not supply that ability, they could be paired with measurement tools such as profile fitting and centroiding on the raw frame, and trained on physically annotated measurements, which a generator such as ours can produce at scale; and because their errors are complementary, verification against the declared setup, for instance by re-predicting a measured quantity, may turn partial competence into reliable answers. For evaluation, chance floors differ across formats and rungs, so we recommend reporting chance-corrected accuracy per rung alongside raw accuracy. Looking forward, extending these capabilities to real optical experiments presents an important direction for future research. Such systems would require VLMs to move beyond visual reasoning toward embodied interaction with physical instruments, including experimental manipulation, measurement, and adaptive decision-making. All scenes in OpticalBench are simulated: real benches add alignment drift, stray light and detector artefacts that our sensor model does not reproduce, so the benchmark measures reasoning over idealised experiments.

5 Conclusion

We introduced OpticalBench, to our knowledge the first benchmark of optical experiments for VLMs, built from 1,961 simulated scenes and 31,887 image inputs, in which every answer is computed from a simulated optical setup rather than annotated by hand. Across 7,664 questions on six optical tasks and four levels of reasoning, ten VLMs reach 13.3% to 65.0% and average 38.0%, and the average across deployments falls from 47.9% at Perception to 39.3% at Derivation and 28.6% at Prediction before recovering to 36.3% at Decision. They read directions and peak positions but not photometric measurements, their failures above perception lie in the physics rather than in reading the frame, and test-time reasoning helps derivation but not measurement, so the gap the benchmark exposes is a measurement gap rather than a reading one, and it persists whether or not a deployment is allowed to think for longer. OpticalBench offers a physics-grounded testbed for measuring and advancing VLMs toward automated optical experimentation.

AI USE STATEMENT

We used a large language model to polish the writing of this paper and to help develop the generator that builds the benchmark: its scene rules, parameter ranges and validity constraints were drafted with AI assistance and then checked by the authors. The benchmark is a synthetic dataset, but it is not model-generated: every scene, image, question and answer is computed by our deterministic pure-Python optics simulator. The authors have reviewed all AI-assisted work and take responsibility for the final content.

ETHICS STATEMENT

All scenes, images and questions in OpticalBench are produced by simulation, so the benchmark involves no human subjects, no personally identifiable information and no third-party imagery, and no institutional review was required. The generator, the rendered images and the question bank will be released for research use.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the simulation that computes every answer, its validation and the sensor model. Section 3.3 describes question synthesis and grading, and Appendix C the quality-control gates and their verdicts on the frozen bank. Section 4.1 states the evaluation protocol and how the leaderboard sample is drawn, Appendix B prints the answer-format instruction verbatim and describes the per-row protocol record, and Appendix D checks that sample against its full split. We will release the generator, the rendered images and the question bank for research use.

REFERENCES

Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2:483–492, 2023. URL https://doi.org/10.1038/s44160-022-00231-0.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating Physical Prediction from Vision in Humans and Machines. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2106.08261.

Oliver Bentham and Vivek Srikumar. InfiniteScienceGym: An Unbounded, Procedurally-Generated Benchmark for Scientific Analysis. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2604.13201.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. URL https://doi.org/10.1038/s41586-023-06792-0.

Max Born and Emil Wolf. *Principles of Optics*. Cambridge University Press, Cambridge, 7th edition, 1999.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583(7815): 237–241, 2020. URL https://doi.org/10.1038/s41586-020-2442-2.

James Burgess, Jeffrey J. Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G. Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D. Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.13399.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2403.20330.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2501.16411.

Song Dai, Yibo Yan, Jiamin Su, Zihao Dongfang, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, and Xuming Hu. PhysicsArena: The First Multimodal Physics Reasoning Benchmark Exploring Variable, Process, and Solution Dimensions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17290–17316, 2025. URL https://aclanthology.org/2025.findings-emnlp.937/.

Junpeng Ding, Zichen Tang, Haihong E, Mengyuan Ji, Yang Liu, Haolin Tian, Haiyang Sun, Pengqi Sun, Yang Xu, Yichen Liu, Haocheng Gao, Zijie Xi, Ruomeng Jiang, Peizhi Zhao, Rongjin Li, Yuanze Li, Jiacheng Liu, Zhongjun Yang, Jintong Chen, and Siying Lin. Decoding Scientific Experimental Images: The SPUR Benchmark for Perception, Understanding, and Reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://arxiv.org/abs/2604.27604.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. IsoBench: Benchmarking Multimodal Foundation Models on Isomorphic Representations. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.01266.

Dennis H. Goldstein. *Polarized Light*. CRC Press, Boca Raton, FL, 3rd edition, 2011.

Joseph W. Goodman. *Introduction to Fourier Optics*. W. H. Freeman, New York, 4th edition, 2017.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2501.05444.

P. Hariharan. *Basics of Interferometry*. Academic Press, 2nd edition, 2007.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://aclanthology.org/2024.acl-long.211/.

Eugene Hecht. *Optics*. Pearson, 5th edition, 2017.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1612.06890.

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. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2412.00947.

Jonathan Klimesch, Sören Arlt, Carlos Ruiz-Gonzalez, Carla Rodríguez, Xuemei Gu, Philipp Haslinger, Pietro Vischia, Christian Haack, Yehonathan Drori, Rana Adhikari, Markus Arndt, Michael Kagan, Lukas Heinrich, and Mario Krenn. Designing physics experiments with artificial intelligence. *Nature*, 657(8130):47–58, 2026. URL https://doi.org/10.1038/s41586-026-10898-6.

Herwig Kogelnik and Tingye Li. Laser Beams and Resonators. *Applied Optics*, 5(10):1550–1567, 1966. URL https://doi.org/10.1364/AO.5.001550.

Mario Krenn, Mehul Malik, Robert Fickler, Radek Lapkiewicz, and Anton Zeilinger. Automated Search for new Quantum Experiments. *Physical Review Letters*, 116(9):090405, 2016. URL https://doi.org/10.1103/PhysRevLett.116.090405.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Fei-Fei Li, and Ehsan Adeli. QuantiPhy: A Quantitative Benchmark Evaluating Physical Reasoning Abilities of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2512.19526.

Beibei Lin, Zifeng Yuan, and Tingting Chen. RGB-to-Polarization Estimation: A New Task and Benchmark Study. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/30697d9ef8ce55de6ccc38e043a94142-Abstract-Datasets_and_Benchmarks_Track.html.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2209.09513.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.02255.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M. Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and N. M. Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16:9104, 2025. URL https://doi.org/10.1038/s41467-025-64105-7.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. URL https://doi.org/10.1007/BF02295996.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision Language Models Are Blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, 2024. URL https://arxiv.org/abs/2407.06581.

Lord Rayleigh. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49):261–274, 1879. URL https://doi.org/10.1080/14786447908639684.

Bahaa E. A. Saleh and Malvin Carl Teich. *Fundamentals of Photonics*. Wiley, Hoboken, NJ, 3rd edition, 2019.

Jason D. Schmidt. *Numerical Simulation of Optical Wave Propagation with Examples in MATLAB*. SPIE Press, Bellingham, WA, 2010. URL https://doi.org/10.1117/3.866274.

Dmitry Sorokin, Alexander Ulanov, Ekaterina Sazhina, and Alexander Lvovsky. Interferobot: aligning an optical interferometer by a reinforcement learning agent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/99ba5c4097c6b8fef5ed774a1a6714b8-Abstract.html.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990): 86–91, 2023. URL https://doi.org/10.1038/s41586-023-06734-w.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023. URL https://doi.org/10.1038/s41586-023-06221-2.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. PhysUniBench: A Multi-Modal Physics Reasoning Benchmark at Undergraduate Level. *arXiv preprint arXiv:2506.17667*, 2025. URL https://arxiv.org/abs/2506.17667.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. SeePhys: Does Seeing Help Thinking? – Benchmarking Vision-Based Physics Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2505.19099.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2409.02813.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2403.14624.

Xiaorong Zhu, Ziheng Jia, Jiarui Wang, Xiangyu Zhao, Haodong Duan, Xiongkuo Min, Jia Wang, Zicheng Zhang, and Guangtao Zhai. GOBench: Benchmarking Geometric Optics Generation and Understanding of MLLMs. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, 2025. URL https://doi.org/10.1145/3746027.3758207.

A APPENDIX

This appendix collects the material referenced in the main text. Appendix B prints the answer-format instruction verbatim and describes the per-row protocol record; Appendix C describes the four quality-control gates and their verdicts on the frozen bank; Appendix D compares the leaderboard sample against the full split and itemises what the evaluation costs; Appendix E reports the composition of the bank, the per-family results, the held-out ranges and the reasoning-budget contrast; Appendix F shows one example item per task family, with its images, question, key and the answer of every deployment.

B ANSWER FORMAT AND THE PER-ROW PROTOCOL RECORD

The answer-format instruction, verbatim. Every question stem ends with the response shape, shown here for a single-choice item, printed exactly as the generator writes it into the bank:

```
{"reasoning": "<your working>", "answer": <n>}
```

with `<n>` replaced by `<number>` or `[<n>, ...]` in the other formats. The API enforces that shape as a strict JSON schema, and the schema orders the fields so that the reasoning field is emitted before the answer field. We send the block unchanged at evaluation time: the prompt is frozen into the bank at generation time, so editing it would change every item's key hash and void every response already purchased against it. A reply we cannot read is scored wrong and stays in the denominator, because every way of wrapping a correct answer in the wrong shape — a unit suffix, a nested `{value: v}` object, a scalar inside a list, option text instead of an option number, a bare value with no object — lands in unparsed rather than in graded-wrong, so dropping unparsed replies would reward the deployments that follow the instruction least.

The per-row protocol record. We log one record per evaluated deployment, complete enough that a reader can tell what computation produced each number and can detect after the fact if they did not reach the same one. Each record carries the model identifier; the evaluation date; the provider requested and the provider that answered, since an identifier on a router names a set of deployments rather than a model and we pin one provider for every open row and disable fallback; the reasoning setting we send beside the setting the deployment actually honours; the median visible and hidden character counts per response, which are what distinguish a long answer from a long silence and turn "this model scores at the floor" into "this model returned nothing to score"; the prompt and completion tokens per request; and the count of unparsed replies. A dash marks a field an endpoint does not expose. The same record separates the two ways a row reaches a 48,000-token completion, which look identical in a billing record and opposite in a usage record: three retries of 16,000 hidden tokens with zero visible characters is a deployment ignoring the effort request, while three retries of 16,000 *visible* tokens is a deployment failing to terminate.

Five checks of the format gate (C1, C2.3, C3.3, C5.3 and M1) now fire nowhere at all, because the matching answer format they test left the bank; the 123 tests that remain cover the four formats that ship — single choice, multiple selection, numeric response and ordering. We record this because every guard that lets a check skip is a place a test suite goes quiet.

C QUALITY CONTROL

The four quality-control gates of Section 3.3 stand between the generator and the frozen bank. The format gate runs 123 checks on the answer formats, option pools and answer distributions of every cell; the record-consistency audit runs 68 checks, including re-deriving every key from the stored scene and matching each declared chance floor to the answers the generator actually emits; the calibration gate checks that every value printed in a question agrees with the value recomputed from the scene; and the semantic gate re-scores items without their camera frames to find questions answerable without the intended evidence. On the frozen bank the first three gates pass with no surviving flag, and the semantic gate flags one of the 24 cells, Imaging Decision, as partly answerable from the printed text. We also inspected 72 questions by hand, three per cell, and found no unreasonable item.

Frames that a question compares share a common exposure scale, so that independent auto-exposure cannot introduce artificial cues about relative intensity or fringe visibility.

D SAMPLE AGAINST POPULATION, AND WHAT THE EVALUATION COSTS

Table D.1 checks the stratified sample the leaderboard runs on against the split it is drawn from, and Table D.2 itemises what the leaderboard phase costs.

Table D.1: The 480-question leaderboard sample against the full `test` split. We report the stratified sample's estimate with its 95% interval beside the population value from a full-split run of one deployment, Gemini 3.8 Flash, together with the difference, for the overall score and for each cognitive rung and task family. The run covers 4,629 of the 4,632 questions in the split; the three missing ones are provider-side HTTP 400 responses, 0.06% of requests. The population value falls inside the sample's interval in every row here and in 24 of 24 cells. The intervals are wide, so this table bounds the sample's aggregate faithfulness rather than certifying any individual cell.

	Sample (480)	95% interval	Population (4,629)	Δ
Overall	**44.8**	40.4–49.3	**44.9**	-0.1
Perception	65.8	57.0–73.7	63.3	$+2.6$
Derivation	39.2	30.9–48.1	43.5	-4.4
Prediction	36.7	28.6–45.6	36.1	$+0.6$
Decision	37.5	29.4–46.4	35.8	$+1.7$
T1 Gaussian	20.0	12.7–30.0	17.0	$+3.0$
T2 Imaging	32.5	23.2–43.4	36.3	-3.8
T3 Resolution	45.0	34.6–55.9	48.4	-3.4
T4 Polarization	70.0	59.2–78.9	66.8	$+3.2$
T5 Fourier	56.2	45.3–66.6	61.6	-5.4
T6 Interference	45.0	34.6–55.9	38.3	$+6.7$

Table D.2: Cost of the leaderboard phase, one row per result file. We report the rows billed and the spend of each purchased run, each figure the sum of that file's per-row usage cost including retries, together with the trial calls that precede a purchase decision and the two paid files we void. Voided files are listed rather than dropped, because they are part of what the phase cost. Per-question unit prices span \$0.0025 for Qwen3-VL-235B instruct to \$0.035 for Qwen3-VL-235B thinking. The per-file figures are rounded to the cent and therefore sum to within a cent of the total. The five-arm input ablation runs on a different sample and is listed separately from the total.

Result file	Rows	Cost
Claude Sonnet 5, `test` 480	480	\$10.38
Claude Sonnet 5, two early trial calls (voided)	4	\$1.37
Gemini 3.1 Pro, `test` 480	480	\$11.44
Gemini 3.7 Flash, `test` 480	480	\$2.34
Gemini 3.8 Flash, full `test` split (contains its 480 sample)	4,632	\$22.52
Gemini 3.8 Flash, `test` 480, provider-default reasoning	478	\$8.16
Gemini 3.8 Flash, `test-x` 480	480	\$2.39
Grok 4.6, `test` 480	480	\$7.85
GPT-5.6 Sol, `test` 480	480	\$9.43
GPT-5.6 Sol, `test-x` 480	480	\$8.86
GPT-5.6 Terra, `test` 480	480	\$12.62
Qwen3-VL-235B instruct, `test` 480	480	\$1.21
Qwen3-VL-235B instruct, mixed-provider rows (voided)	37	\$0.11
Qwen3-VL-235B thinking, `test` 480	480	\$16.72
DeepSeek V4 Flash Vision, `test` 480	481	\$1.75
Kimi K2.6 trial calls (not purchased)	2	\$0.35
Claude Haiku 4.5 and Qwen3.8 Flash trial calls (not purchased)	8	\$0.14
Leaderboard phase total		**\$117.63**
Five-arm input ablation (different sample)		\$13.40

The two heaviest lines of that bill are the full-split run behind the sampling check and Qwen3-VL-235B thinking; the cheapest complete row is Qwen3-VL-235B instruct. Four deployments are tried before a purchase decision, and the rule that produced the split is worth stating because a purchase

taken after seeing a trial call is itself a selection step: a trial call buys a row unless its price or its output shape makes the run unmeasurable. GPT-5.6 Terra is bought after its trial and becomes the tenth leaderboard row; Kimi K2.6, Claude Haiku 4.5 and Qwen3.8 Flash are not bought. Kimi K2.6 emits roughly 52,000 visible characters per question, which prices a 480-question run at $40 to $85 against a leaderboard phase of $117.63 in total, and Kimi K3 at $150 to $300 on the same output length; one of the two Kimi trial calls never answers within 24,000 tokens, so a paid run would record our budget ceiling as a physics error.

E Dataset composition, per-family results, held-out ranges and the reasoning budget

Counted per question, the 7,664 items send 31,887 image inputs, drawn from 10,822 distinct rendered images. Eight artefacts belong to Sections 2, 3 and 4 and are printed here for space. Table E.1 places OpticalBench against fourteen prior benchmarks on the five properties Section 2 argues for; Table E.2 names the bench each task family declares and the latent quantities it withholds; Figure E.1 resolves the leaderboard of Table 2 per family and rung; Tables E.4 and E.5 carry the input ablation of Section 4.3, overall and by task family; Table E.3 carries the paired reasoning-budget contrast; Table E.6 the in-distribution against held-out comparison of Section 4.4; and Figure E.2 the same questions sliced by rung, family, answer format and declared depth.

Table E.1: Comparison of recent benchmarks for multimodal scientific and physical reasoning. We evaluate each by whether its images come from an optics simulator its authors run, whether every scene ships a machine-readable component list with positions that is also rendered for the model, whether items carry three or more declared strata of cognitive depth reported per stratum, whether a chance level is computed per item from the pool the generator can emit, and whether the same items are also scored under a degraded or substituted input (the five tick columns, in that order). A $\surd$ marks a property provided and a $\times$ one absent, ∞ a procedural benchmark with no fixed item count; a $\times$ under per-cell chance floor may still publish one pooled random baseline, which this column does not count. Ours is the last row, in **bold** on a gray background. The compared benchmarks are, in the order of the rows, Johnson et al. (2017), Yue et al. (2024), He et al. (2024), Zhang et al. (2024), Fu et al. (2024), Yue et al. (2025), Chow et al. (2025), Kamoi et al. (2025), Dai et al. (2025), Xiang et al. (2025), Zhu et al. (2025), Li et al. (2026), Ding et al. (2026) and Bentham & Srikumar (2026).

Benchmark (venue, year)	Sim.	Layout	Rungs	Floor	Ablation	Data type	Amount
CLEVR (CVPR, 2017)	$\times$	$\times$	$\times$	$\times$	$\surd$	Rendered 3-D shapes	1.0M
MMMU (CVPR, 2024)	$\times$	$\times$	$\surd$	$\times$	$\surd$	Mixed exam imagery	11.5k
OlympiadBench (ACL, 2024)	$\times$	$\times$	$\times$	$\times$	$\times$	Competition diagrams	8.5k
MathVerse (ECCV, 2024)	$\times$	$\times$	$\times$	$\times$	$\surd$	Math diagrams	15.0k
IsoBench (COLM, 2024)	$\times$	$\times$	$\times$	$\times$	$\surd$	Rendered plots and boards	1.9k
MMMU-Pro (ACL, 2025)	$\times$	$\times$	$\times$	$\times$	$\surd$	Mixed exam imagery	3.5k
PhysBench (ICLR, 2025)	$\times$	$\times$	$\times$	$\times$	$\times$	Video, image, simulation	10.0k
VisOnlyQA (COLM, 2025)	$\times$	$\times$	$\times$	$\times$	$\times$	Synthetic figures	1.2k
PhysicsArena (EMNLP, 2025)	$\times$	$\times$	$\surd$	$\times$	$\times$	Physics problem diagrams	5.1k
SeePhys (NeurIPS, 2025)	$\times$	$\times$	$\surd$	$\times$	$\times$	Textbook and exam diagrams	2.0k
GOBench (ACM MM, 2025)	$\times$	$\times$	$\times$	$\times$	$\times$	Generated optics imagery	1.0k
QuantiPhy (CVPR, 2026)	$\times$	$\times$	$\times$	$\times$	$\surd$	Rendered kinematic video	3.3k
SPUR (ACL, 2026)	$\times$	$\times$	$\surd$	$\times$	$\times$	Experimental imagery	4.3k
InfiniteScienceGym (COLM, 2026)	$\times$	$\times$	$\times$	$\times$	$\times$	Procedural text, no imagery	∞
OpticalBench (Ours, 2026)	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	**Simulated bench: layout + frames**	**7.7k**

Table E.2: The six task families, their declared benches and their withheld latents. We list, for each family, the bench we declare, the named laboratory procedure its questions instantiate, and the physical quantities we withhold from the declaration so that the layout diagram alone cannot settle the answer. HWP and QWP denote a half- and a quarter-wave plate, BS a beamsplitter, and L1 and L2 the two lenses of a $4f$ relay.

Family	Declared bench	Named procedure	Withheld latent
T1 Gaussian	laser → lens → cameras	caustic scan, hyperbolic fit	output waist size and position
T2 Imaging	laser → turning mirror → lens → cameras	near-/far-field centroid walk	two-axis mirror tilt
T3 Resolution	incoherent point pair → lens → cameras	through-focus conjugate search	separation, mode-field radius
T4 Polarization	laser → polarizer → HWP → QWP → analyzer → camera	Stokes polarimetry	the two wave-plate angles
T5 Fourier	grating → L1 → iris → L2 → camera	Fourier-plane order positions	two grating periods
T6 Interference	laser → BS → two arms → BS → camera	fringe visibility measurement	arm angle and amplitude ratio

Table E.3: Effect of the reasoning budget on two model generations across the four cognitive rungs. We report per-rung accuracy in percent and, rightmost, accuracy over the items each model answers under both settings; higher is better. Rows are reasoning settings rather than models, gray rows indicate model grouping, `effort low` is printed above the provider's default, and the italic third row of each group is the McNemar p-value of the paired comparison above it, which is not an accuracy and does not compete for the marks. This within-model contrast carries no chance-floor row. We mark only the better result in each column within each group, with tied values sharing a mark.

Setting	Perc	Deriv	Pred	Dec	**Avg**
Gemini 3.7 Flash					
Effort low	63.3	44.2	25.8	37.5	35.8
Provider default	61.3	49.0	45.2	49.3	43.2
McNemar p	–	–	–	–	*0.31*
Gemini 3.8 Flash					
Effort low	65.8	39.5	36.7	37.0	44.8
Provider default	65.8	52.1	35.0	45.4	49.6
McNemar p	*1.0*	*0.014*	*0.86*	*0.11*	*0.045*

Table E.4: Input ablation across five visual conditions, in percent; higher is better. Δ reports the accuracy difference from the full-input condition, and the McNemar p-value measures the significance of this difference on the same questions. The first row reports the average chance floor. We highlight the best and second-best result in each column.

Input	Δ vs full	McNemar p	Perc	Deriv	Pred	Dec	**Avg**
Chance floor	–	–	–	–	–	–	*15.9*
Layout only (no frames)	-33.6	5.6×10^{-33}	11.7	15.3	17.2	26.3	17.6
Text only (no image)	-33.4	2.0×10^{-31}	13.3	16.2	16.4	25.4	17.8
Layout + frames of another scene	-30.2	6.1×10^{-27}	16.7	19.8	19.8	28.1	21.0
Frames only (no layout)	-18.2	1.5×10^{-12}	45.0	35.1	24.1	27.2	33.0
Full input (layout + frames)	–	–	66.7	48.6	42.2	46.5	51.2

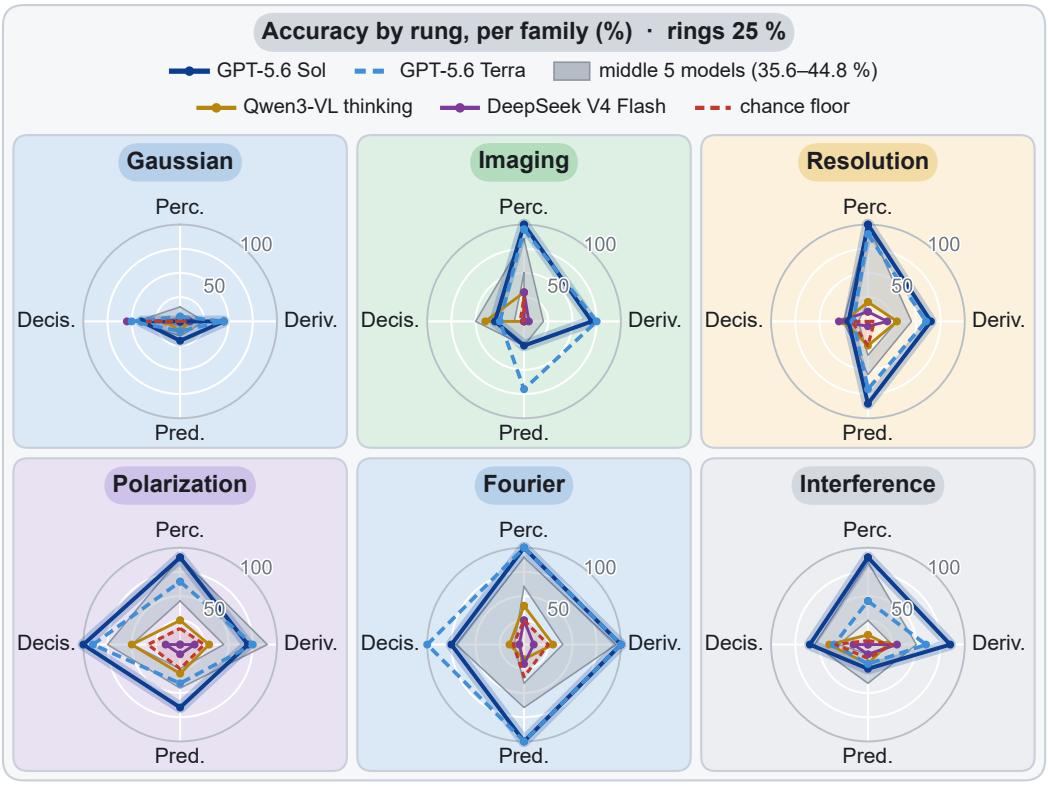


Figure E.1: Per-family results across the cognitive ladder. One radar panel per task family, the four cognitive rungs on the axes (Perception, Derivation, Prediction, Decision, clockwise from the top; rings every 25 points), and a red dashed contour at each cell's own chance floor, so the area between a polyline and that contour is the accuracy earned beyond guessing. GPT-5.6 Sol (solid) and GPT-5.6 Terra (dashed) are drawn as lines, as are Qwen3-VL-235B thinking and DeepSeek V4 Flash Vision; the five rows between them, from Gemini 3.1 Pro at 35.6 to Gemini 3.8 Flash at 44.8, are drawn as one grey band spanning their minimum and maximum per cell, since several of their pairs do not separate after Holm correction. Qwen3-VL-235B instruct is omitted, for the reason given in Section 4.5. Accuracy drops between Perception and the rungs above it and does not continue falling into Decision: the step out of perception is this benchmark's dominant difficulty axis, and it is not the axis a reader would predict from the families.

Table E.5: The input ablation of Table E.4 resolved by task family, in percent. Adding the layout to the frames matters most for Gaussian, Resolution and Fourier; Polarization is solved from the frames alone; the layout alone stays near the chance floor in every family. The rightmost column is each family's average chance floor.

Family	n	Full	Frames only	Layout only	Other scene	Text only	*Floor*
T1 Gaussian	75	25.3	12.0	14.7	16.0	17.3	*10.3*
T2 Imaging	77	48.1	35.1	11.7	20.8	13.0	*7.2*
T3 Resolution	83	56.6	19.3	15.7	16.9	13.3	*11.6*
T4 Polarization	74	70.3	70.3	20.3	24.3	16.2	*24.9*
T5 Fourier	80	70.0	37.5	20.0	25.0	25.0	*23.6*
T6 Interference	72	34.7	25.0	23.6	23.6	22.2	*18.2*
All	461	51.2	33.0	17.6	21.0	17.8	*15.9*

Table E.6: Accuracy on the in-distribution and held-out splits for two deployments, by rung and by family, with the average over each 480-question sample rightmost; higher is better. Rows are evaluation splits rather than models, gray rows group them by model, and the in-distribution split is printed first in each group. The first row, on a darker band, is the average chance floor, 16.2 on both splits; it does not compete for the marks. The two overall differences are the only comparisons we claim: Gemini 3.8 Flash -0.8 (two-proportion $p = 0.80$) and GPT-5.6 Sol $+4.8$ ($p = 0.11$). We mark only the better of the two splits in each column within each group.

Split	Perc	Deriv	Pred	Dec	Gauss.	Imag.	Resol.	Polar.	Four.	Interf.	**Avg**
Chance floor	–	–	–	–	–	–	–	–	–	–	*16.2*
Gemini 3.8 Flash											
`test` (in distribution)	65.8	39.2	36.7	37.5	20.0	32.5	45.0	70.0	56.2	45.0	44.8
`test-x` (held out)	65.8	39.2	39.2	31.7	12.5	37.5	48.8	66.2	62.5	36.2	44.0
GPT-5.6 Sol											
`test` (in distribution)	80.0	72.5	53.3	54.2	26.2	56.2	67.5	81.2	93.8	65.0	65.0
`test-x` (held out)	81.7	70.8	66.7	60.0	42.5	63.8	70.0	83.8	97.5	61.2	69.8

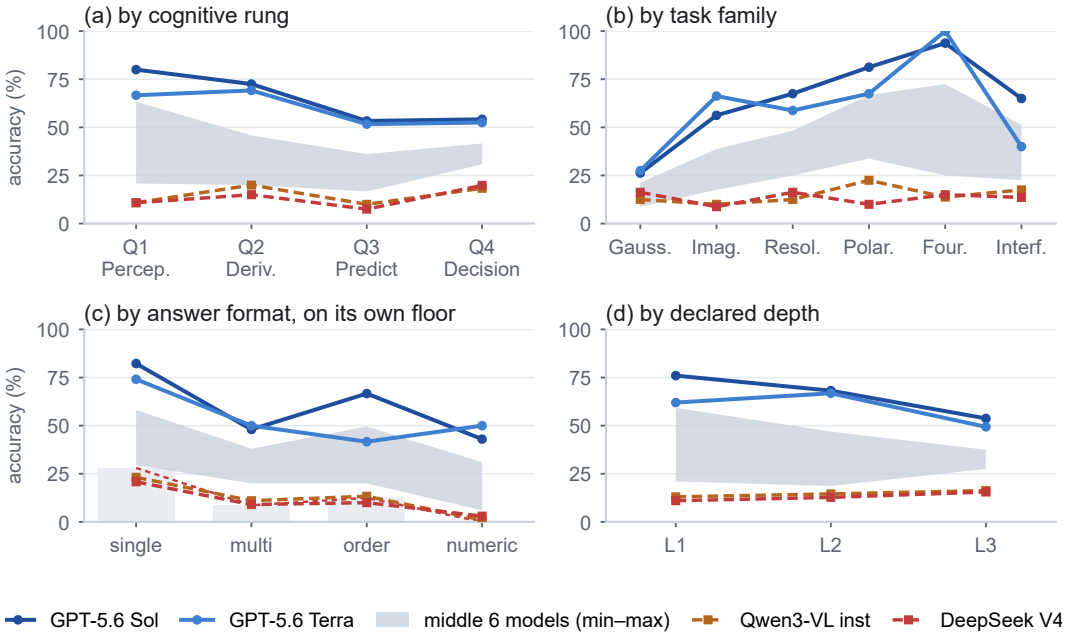
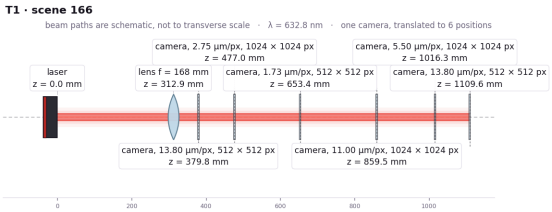


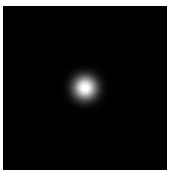
Figure E.2: The same 480 questions sliced four ways, all ten deployments drawn. Panels report accuracy by **(a)** cognitive rung, **(b)** task family, **(c)** answer format, over the grey bars giving that format's own mean chance floor, and **(d)** declared depth. The two strongest and the two weakest deployments carry their own lines; the six between them are one band from minimum to maximum, because several of their pairs do not separate after Holm correction. The bank's other difficulty axis, measured resolvability, is not plotted: its hardest bin holds eight questions.

F EXAMPLE ITEMS

Figures F.1–F.3 show one item from each task family, with its images, question, key and the answer of each of the ten deployments. All six come from the leaderboard sample and together cover the four levels of reasoning and the four answer formats; within each family we chose an item whose frames are legible at print size and whose answers separate the deployments.

(a) Gaussian, Perception, numeric response (`T1-166-Q1`, `test` split; 2 of 10 deployments correct)

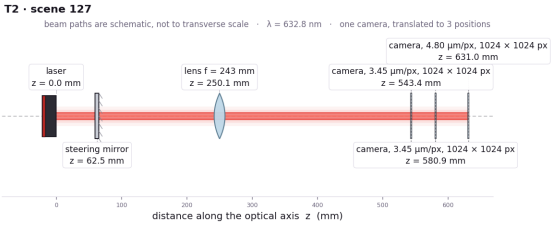


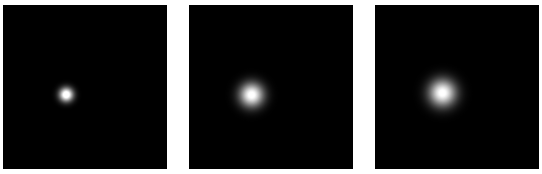


`cam3.png`

Question	The layout shows the set-up; cam3.png is what the third camera recorded. Give the $1/e^2$ intensity radius of the spot.
Key	87.072 micrometres; an answer within 5% of it is graded correct
Answers	GPT-5.6 Sol: 74.4 ×; GPT-5.6 Terra: 70.1 ×; Gemini 3.8 Flash: 102 ×; Claude Sonnet 5: 552 ×; Gemini 3.7 Flash: 90 √; Grok 4.6: 88 √; Gemini 3.1 Pro: 95.1 ×; Qwen3-VL-235B thinking: 6.92 ×; Qwen3-VL-235B instruct: 0 ×; DeepSeek V4 Flash Vision: 68.75 ×.

(b) Imaging, Prediction, numeric response (`T2-127-Q3`, `test` split; 2 of 10 deployments correct)



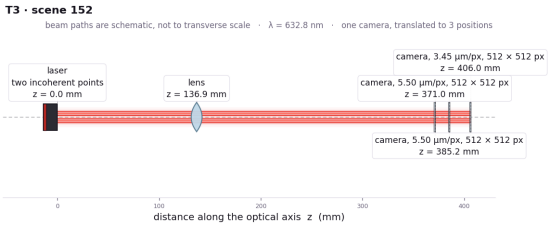


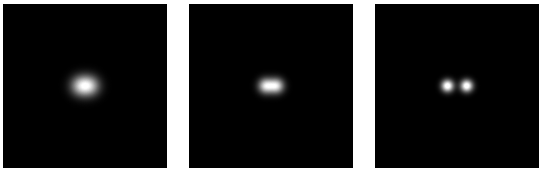
`cam1.png` `cam2.png` `cam3.png`

Question	The steering mirror is adjusted so that the beam leaving it is deflected by a further 400 µrad toward +x (the right of the frames). Where would the spot centre then sit horizontally on the LAST camera, measured from the centre of that frame, taking +x as the right of the frame?
Key	−326.6 micrometres, signed; an answer within 19.6 micrometres of it is graded correct
Answers	GPT-5.6 Sol: −341.4 √; GPT-5.6 Terra: −317.4 √; Gemini 3.8 Flash: −377.4 ×; Claude Sonnet 5: 109.8 ×; Gemini 3.7 Flash: −379.8 ×; Grok 4.6: −428 ×; Gemini 3.1 Pro: −235.1 ×; Qwen3-VL-235B thinking: 227.4 ×; Qwen3-VL-235B instruct: 0 ×; DeepSeek V4 Flash Vision: 227.4 ×.

Figure F.1: Example items: (a) Gaussian and (b) Imaging. Layout and camera frames as sent to the model (blank margins of the layout trimmed); question without its closing reply instruction; each deployment's answer, with √ for correct.

(a) Resolution, Derivation, multiple selection (`T3-152-Q2`, `test` split; 4 of 10 deployments correct)

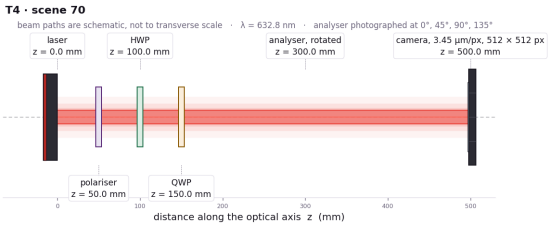


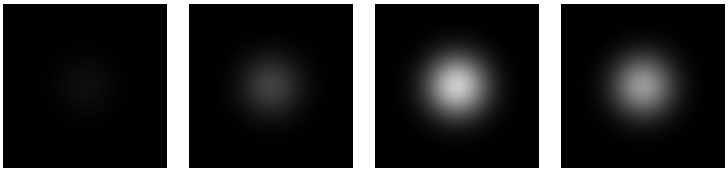


`cam1.png` `cam2.png` `cam3.png`

Question	The three frames are the three camera planes, in the order the planes appear along the axis. Count a plane as resolving the pair when the dip between the two peaks reaches a contrast of at least 0.30, that is (peak − valley)/(peak + valley) $\geq$ 0.30 along the line joining them. Which of the following are true of this set-up? Select every one that applies, and no others: (1) in whichever frame shows the deepest dip, the two spots land more than 69 pixels apart; (2) the three camera planes span more than 4.8 Rayleigh ranges of the beam at the image; (3) at least 2 of the three frames resolve the pair; (4) the lens magnifies the pair by more than 1.7×.
Key	(4)
Answers	GPT-5.6 Sol: (4) $\surd$; GPT-5.6 Terra: (4) $\surd$; Gemini 3.8 Flash: (1, 2, 4) ×; Claude Sonnet 5: (4) $\surd$; Gemini 3.7 Flash: (1, 4) ×; Grok 4.6: (2, 4) ×; Gemini 3.1 Pro: (1, 4) ×; Qwen3-VL-235B thinking: (2, 4) ×; Qwen3-VL-235B instruct: (1, 2, 3, 4) ×; DeepSeek V4 Flash Vision: (4) $\surd$.

(b) Polarization, Perception, ordering (`T4-70-Q1`, `test` split; 7 of 10 deployments correct)



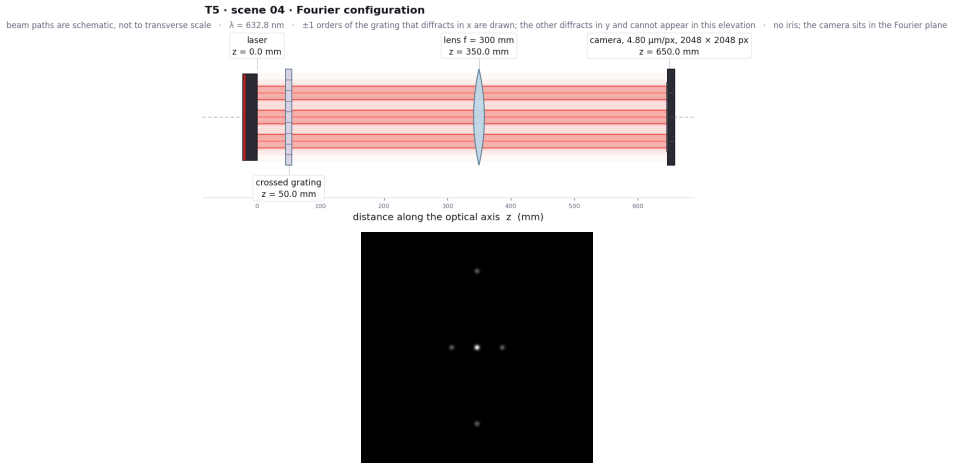


`cam_a0.png` `cam_a1.png` `cam_a2.png` `cam_a3.png`

Question	The layout shows the set-up; the four frames are the same spot photographed through the analyser at 0°, 45°, 90° and 135°, on one shared intensity scale. Rank them by brightness, brightest first. Put these in the order asked: (1) the 0° frame; (2) the 45° frame; (3) the 90° frame; (4) the 135° frame.
Key	(3, 4, 2, 1), first to last
Answers	GPT-5.6 Sol: (3, 4, 2, 1) $\surd$; GPT-5.6 Terra: (3, 4, 2, 1) $\surd$; Gemini 3.8 Flash: (3, 4, 2, 1) $\surd$; Claude Sonnet 5: (4, 3, 1, 2) ×; Gemini 3.7 Flash: (3, 4, 2, 1) $\surd$; Grok 4.6: (3, 4, 2, 1) $\surd$; Gemini 3.1 Pro: (3, 4, 2, 1) $\surd$; Qwen3-VL-235B thinking: (3, 2, 4, 1) ×; Qwen3-VL-235B instruct: (3, 4, 2, 1) $\surd$; DeepSeek V4 Flash Vision: (1, 2, 3, 4) ×.

Figure F.2: Example items: (a) Resolution and (b) Polarization. Layout and camera frames as sent to the model (blank margins of the layout trimmed); question without its closing reply instruction; each deployment's answer, with $\surd$ for correct.

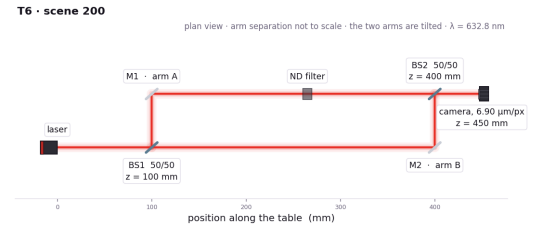
(a) Fourier, Derivation, single choice (`T5-04-Q2`, `test` split; 4 of 10 deployments correct)



`cam_fourier.png`

Question	Using the layout and the Fourier photograph, which of these gives the grating period AT THE OBJECT along x and along y? Choose exactly one: (1) 246.3 µm along x, 81.3 µm along y; (2) 58.4 µm along x, 177.0 µm along y; (3) 81.3 µm along x, 246.3 µm along y; (4) 177.0 µm along x, 58.4 µm along y.
Key	(4) 177.0 µm along x, 58.4 µm along y
Answers	GPT-5.6 Sol: (4) $\surd$; GPT-5.6 Terra: (4) $\surd$; Gemini 3.8 Flash: (1) $\times$; Claude Sonnet 5: (4) $\surd$; Gemini 3.7 Flash: (4) $\surd$; Grok 4.6: (1) $\times$; Gemini 3.1 Pro: (1) $\times$; Qwen3-VL-235B thinking: (3) $\times$; Qwen3-VL-235B instruct: (3) $\times$; DeepSeek V4 Flash Vision: (2) $\times$.

(b) Interference, Decision, single choice (`T6-200-Q4`, `test` split; 2 of 10 deployments correct)



`cam_p1.png` `cam_p2.png` `cam_p3.png`

Question	You may put ONE neutral-density filter in the brighter arm to make the fringes as deep as possible. Which of these filters would do it? Choose exactly one: (1) OD 0.70; (2) OD 0.84; (3) OD 1.55.
Key	(1) OD 0.70
Answers	GPT-5.6 Sol: (1) $\surd$; GPT-5.6 Terra: (2) $\times$; Gemini 3.8 Flash: (2) $\times$; Claude Sonnet 5: (2) $\times$; Gemini 3.7 Flash: (2) $\times$; Grok 4.6: (1) $\surd$; Gemini 3.1 Pro: (2) $\times$; Qwen3-VL-235B thinking: (2) $\times$; Qwen3-VL-235B instruct: (2) $\times$; DeepSeek V4 Flash Vision: (3) $\times$.

Figure F.3: Example items: (a) Fourier and (b) Interference. Layout and camera frames as sent to the model (blank margins of the layout trimmed; in (a) the drawing is re-centred under its line of notes); question without its closing reply instruction; each deployment's answer, with $\surd$ for correct.