QUIVER: Benchmarking and Enhancing Physical Reasoning Abilities of Vision-Language Models

Varchita Lalwani, Arbaaz Shafiq, Pratham Arora,
Mukundan Kailash Gurumurthy, Pankaj Pansari

Plaksha Univeristy, Mohali, India

Abstract

Estimating physical relations and attributes is an important ability for AI systems to understand scenes and navigate the world. These skills emerge in humans from a lifetime of physical interaction and underpin our commonsense knowledge. Vision-language models (VLMs) are trained on text-image pairs without physical interaction, and it remains unclear whether genuine grounding can emerge from such supervision. Meanwhile, current reasoning benchmarks predominantly test symbolic reasoning (mathematics, programming) rather than perceptual reasoning in natural environments. We present QUIVER (QUantitative Intuitive-physics Visual Estimation and Reasoning), a diagnostic benchmark that tests whether VLMs can estimate five physical properties from images. These properties are weight or volume of an object in image, a specified angle, degree of stability of some object configuration or whether one object can fit within another object permitting deformations. QUIVER has 459 hand-annotated real-world images with measured numerical ground truth. Our evaluations show that state-of-the-art closed-weight frontier models are competitive with human baseline on these tasks. We improve the weight estimation performance of small and medium open-weight models by fine-tuning them on a dataset automatically derived from COCO, together with reasoning traces generated by large closed-source models.

Introduction

Estimation of physical quantities is fundamental to how we navigate the world. For instance, we routinely judge how heavy an object is before picking it up, whether an object will fit within some given space, or whether a stack of some items will topple. Such judgements may not be always accurate, yet they are fast and usually good enough for us to act on. Such an ability is acquired from a lifetime of interaction with the physical world. An autonomous agent acting in the physical world needs the same ability.

Vision-language models (VLMs) acquire world knowledge by training on image-text pairs, with no physical interaction. It is unclear whether a quantitative sense of weight, capacity or stability can emerge from such supervision. Current reasoning benchmarks do not evaluate such an ability very well. They largely reward symbolic reasoning in mathematics and programming, or perceptual tasks whose answers are already written into the pixels.

Physical reasoning benchmarks come closer, but three limitations remain. First, most items are discrete or ordinal: multiple-choice questions about the physical world, or relative labels such as '*heavier than*' (Chow et al. 2025; Gao et al. 2024). Such formats hide the size of a model's error. Second, benchmarks that do ask for numbers often ask for quantities that are directly measurable in the image, such as an on-screen angle or a gauge reading (Weng et al. 2025; Lin et al. 2025). Third, a growing body of work shows that multimodal items are frequently answerable without the image at all, from memorised magnitudes attached to familiar labels (Chen et al. 2024; Li et al. 2024). A benchmark that does not control for recall cannot separate estimation from lookup.

We present QUIVER (QUantitative Intuitive-physics Visual Estimation and Reasoning), a diagnostic benchmark of 459 hand-annotated real-world images. Each item asks for a physical property of an annotated object or configuration in a single uncalibrated photograph of a cluttered scene: its weight, the liquid volume it holds or can hold, a marked angle, the perturbation needed to make a stable arrangement fall, or whether one object fits inside another when deformation is allowed. Precise ground truth is measured using appropriate measuring instruments. Scenes are built to support estimation rather than recognition. To this end, we avoid standard packaged quantities, keep reference objects visible for scale, and prefer atypical viewpoints. Ease of measurement dictated the choice of physical properties - for instance, we have avoided temperature or friction.

To test whether models estimate or recall, we re-evaluate every item under three levels of Gaussian blur, which removes brand and text cues while leaving shape, identity and layout legible. We report Median Absolute Percentage Error (MdAPE) for the continuous categories and precision and recall for fit.

Our evaluation covers three frontier closed-weight model families, two open-weight models, and a human baseline collected from 95 participants through a web application. We observe that the frontier models are close to each other, with per-category errors typically within five to six percentage points, and they match or exceed the human baseline on four of the five categories. Fine-tuning small and medium open-weight models with QLoRA on data generated automatically over COCO images cuts that error substantially and improves robustness to blur.

Our contributions are:

- QUIVER, a benchmark of 459 images that scores continuous estimates (continuous for four properties and categorial for one property) of material- and constraint-dependent physical properties against in-scene measurements, from a single uncalibrated image. Sample instances from our dataset are shown in top row of figure 1
- A blur-based robustness protocol that separates estimation from label recall, applied to every item.
- A human baseline of 795 responses from 95 participants, collected under the same interface and questions as the model evaluation.
- An automatic question-and-reasoning trace generation pipeline over COCO (Lin et al. 2014) images, and a Quantized Low-Rank Adaptation (QLoRA) fine-tuning recipe that markedly reduces weight estimation error for open-weight models. The pipeline for synthetic data generation and finetuning is shown in bottom row of figure 1

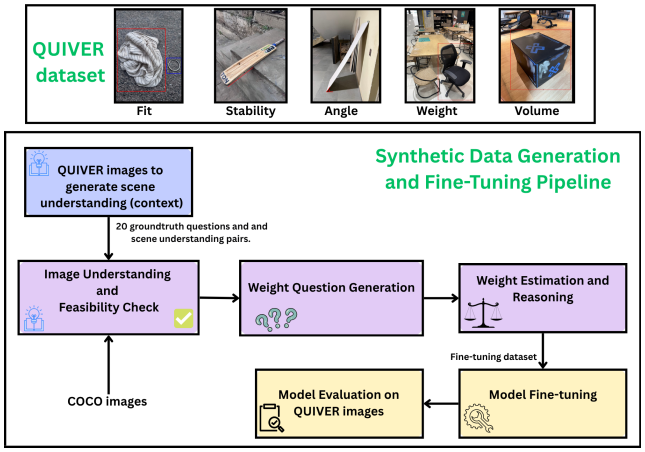


Figure 1: Overview of the QUIVER benchmark construction and fine-tuning pipeline. Top: QUIVER dataset construction. We introduce a benchmark of 459 real-world images spanning five physical reasoning tasks: fit, stability, angle, weight, and liquid volume. Bottom: Automatic data generation and fine-tuning pipeline for open-weights models. QUIVER images are first used to generate scene understanding descriptions. Synthetic training data is then generated using COCO images. Each COCO image is first filtered to retain only suitable weight estimation examples. Using the scene understanding–question pairs from QUIVER as in-context examples (without the corresponding QUIVER images), the model generates weight estimation questions, followed by weight estimates and reasoning traces. The resulting image–question–answer–reasoning tuples make fine tuning dataset and it is used to fine-tune open-source models, which are then evaluated on the original QUIVER benchmark using the benchmark questions without object names.

Related Work

Estimating physical properties from images Recovering geometric and non-geometric properties from images is a long-standing problem in vision (Criminisi, Reid, and Zisserman 2000; Standley et al. 2017; Zhai et al. 2024; Xu et al. 2025; Lan et al. 2025). Two recent systems target mass alone. One fuses a depth-derived volume latent with a VLM-derived density latent under mass-only supervision (Lee et al. 2026). The other overlays detection, scale and cross-section cues on the input image, and releases multi-view RGB-D video with measured mass (Yokomizo et al. 2026). Our setting differs in input and in scope: one uncalibrated image of a cluttered scene, a general model, and five property types.

Benchmarks for physical reasoning PhysBench reports a broad VLM deficit in physical world understanding, but its items are discrete (Chow et al. 2025). A second group asks for numbers that are already in the pixels: angle, length and size (Weng et al. 2025), and instrument reading (Lin et al. 2025). The closest quantitative benchmark scores kinematic estimates from video (Li et al. 2025). Physics problem sets test deduction, where the quantities are given or read off a schematic (Shen et al. 2025; Yue et al. 2024). We test the estimation that precedes it. Recent surveys find the same asymmetry: perception is competent, numerical estimation is weak, and existing benchmarks work with simulated environments (Yu et al. 2025; Melnik et al. 2023). The property annotations closest to ours are ordinal (Gao et al. 2024). QUIVER is, to our knowledge, the first benchmark to score continuous estimates (continuous for four properties and categorial for one property) of material-dependent and constraint-dependent properties against in-scene measurements from a single image.

Answering from memory rather than from the image A persistent concern is that models recall magnitudes instead of reasoning from the given image. MMStar shows that many samples in popular multimodal benchmarks are answerable with no image at all (Chen et al. 2024). NaturalBench pairs items so that a language-only prior cannot beat chance (Li et al. 2024). For quantities, removing the video degrades performance only modestly, and models fall back on real-world magnitudes (Li et al. 2025). A conservation benchmark reports near-chance accuracy (Luo et al. 2026). BLINK makes a related point: recognition is not perception (Fu et al. 2024). Two of our design decisions follow. We avoid standard packaged quantities, so the answer cannot be recalled from a label. And we report performance under three levels of Gaussian blur, which suppresses brand and text cues while leaving shape, identity and layout legible.

Fine-tuning on generated data Visual instruction tuning established the practice of generating multimodal training data with a strong model (Liu et al. 2023). Trace distillation extends it to multi-step reasoning, with a verifier discarding weak traces (Huh et al. 2025). PhysObjects instead fine-tunes on human-annotated physical concepts (Gao et al. 2024). We follow the generated-data route, using QLoRA (Hu et al. 2022) on questions and reasoning traces produced over COCO images (Lin et al. 2014). Our labels are teacher estimates, so the teacher's own error limits what fine-tuning can achieve.

Method

Ground Truth Dataset Collection

Each item consists of an image, a question, and a measured value. We collect five categories of property.

Weight - We ask for the weight of a highlighted object, avoiding standard packaged quantities. For instance, instead of a sealed bag of sugar, which comes in standard weights, we decant an arbitrary amount and ask about that. Answering correctly requires two capabilities: a sense of relative material density, and the ability to infer size or quantity from other objects in the scene. Ground truth comes from a kitchen scale (precision: 0.1g), a hook-type luggage scale (precision: 100g), or a bathroom scale (precision: 1g), depending on the object. Examples are in figure 2

Liquid volume - We ask how much liquid a container currently holds, or how much it holds when full. Containers vary widely in shape. Recognising the object is therefore insufficient; the answer depends on its three-dimensional shape and capacity. We measured volumes using beakers having precision of 1ml. Examples are in figure 3

Angle - We ask for the angle of inclination between two objects, marked in the image with a hand-drawn arc. We measure it with a protractor or a digital inclinometer (precision: 0.01 degrees), depending on the setting. Examples are in figure 2

Stability - We show a stable configuration and ask for the degree of perturbation needed to make it fall. Two considerations motivate this format. The answer is numerical. And an image almost always captures a configuration at rest, which makes the binary version of the question trivial. Ground truth is obtained by perturbing the configuration and recording the point of collapse. The perturbation is expressed in whichever unit the setup affords: displacement, angle, added mass, added volume, or object count. Examples are in figure 5

Geometric fit - We ask whether one object fits entirely inside another. For rigid pairs, this reduces to estimating dimensions and reasoning about spatial arrangement. For most of our datapoints, the inserted object may be bent, folded, or twisted, provided it is not damaged. These demand a further judgement: how far a given material deforms. The benchmark mixes both types. Ground truth is obtained by physically attempting the insertion. Examples are in figure 4

Questions refer to a target object visually annotated (boxed in red) rather than naming directly. For fit, the container is additionally boxed in blue. For angle, the angle is marked with a red arc. Questions then take the form "Estimate the weight of the object highlighted in red" or "Estimate the angle marked in red." Stability questions instead describe the event to be predicted, such as "How far must the book extend beyond the stack before it falls?"

The number of samples in each category is shown in Figure 3. The value ranges and distributions for each category are shown in Figure 6, which illustrate the mean, median, and minimum–maximum ranges.

Our collection criteria were:

- **Range:** Values span small, medium and large instances. This was most feasible for weight and angle; the difficulty of measuring ground truth restricted volume, stability and fit to small and medium objects.
- **Priors over pixels:** Correct answers require priors the image cannot fully convey, such as relative density and flexibility, which humans acquire through physical interaction. Spatial reasoning also matters, strongly for angle and to a lesser degree for volume.
- **Visible references** Every scene contains clearly visible reference objects, and we exclude close-ups of a single object. References anchor scale, as they do in human estimation.
- **Atypical viewpoints** We prefer lateral to frontal views wherever the question stays answerable, forcing greater reliance on priors.
- **Outdoor capture** Where feasible we shot outdoors, for heterogeneity in lighting and texture.
- **Answerability** An item is included only if a human could make a meaningful attempt from the image. There are no trick questions or visual illusions.

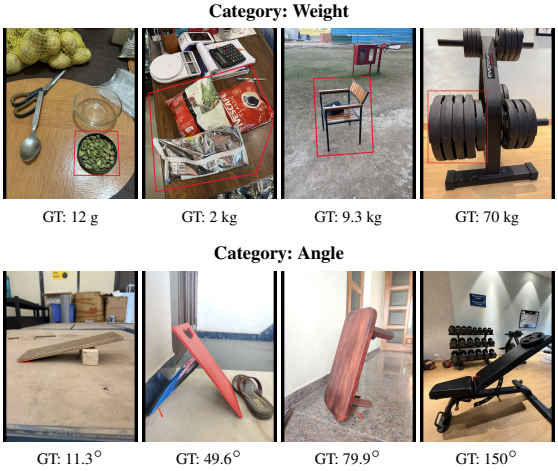


Figure 2: Example images from the Weight and Angle categories with their corresponding ground-truth (GT) answers.

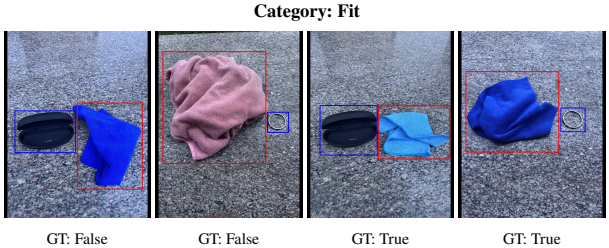
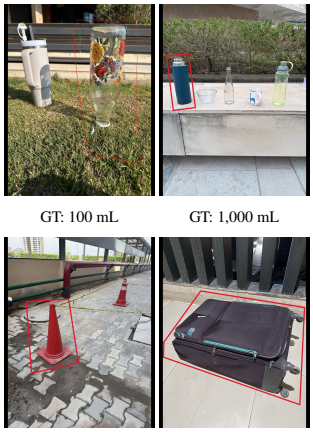


Figure 4: Example images from the Fit category with their ground-truth (GT) answers.

Liquid Volume Examples



GT: 100 mL GT: 1,000 mL

GT: 15,858 mL GT: 105,840 mL

Dataset Distribution

Category	N	%
Fit	116	25.27
Stability	54	11.76
Angle	100	21.79
Liquid Vol.	76	16.56
Weight	113	24.62
Total	**459**	**100**

Figure 3: Liquid Volume examples with ground-truth (GT) answers (left) and the category-wise distribution of the dataset (right). Here, N denotes the number of images.

Stability

GT: The vertical block falls when displaced by 9.5 cm.

GT: The marked angle must be 28° for the object in the blue box to slide.

GT: A weight of 163 g is required to tip the balance.

GT: 100 mL of water balances the highlighted structure.

GT: Two more books are needed to destabilize the structure.

Figure 5: Example images from the Stability category. GT denotes the ground-truth answer.

Human Baseline

We collect a human baseline through a web application. On each trial, a participant answers one question drawn from each of the five categories. Questions come from a fixed subset of 17 items per category, 85 items in total. A participant never sees the same image twice, in any variant. To sustain participation, each trial is presented as a game in which the participant compares their answers against frontier model responses, collected as described in the following section. Model responses are revealed only after the participant commits an answer.

The study drew 95 unique participants, yielding 159 responses per category and 795 in total. Figure 9 shows the distribution of responses per item; most items reached the target of ten. We aggregate at the question level. For the continuous categories, we average participant responses for each item, then compute MdAPE across the 17 items. For fit, we take the majority participant answer for each item, then compute accuracy across the 17 items.

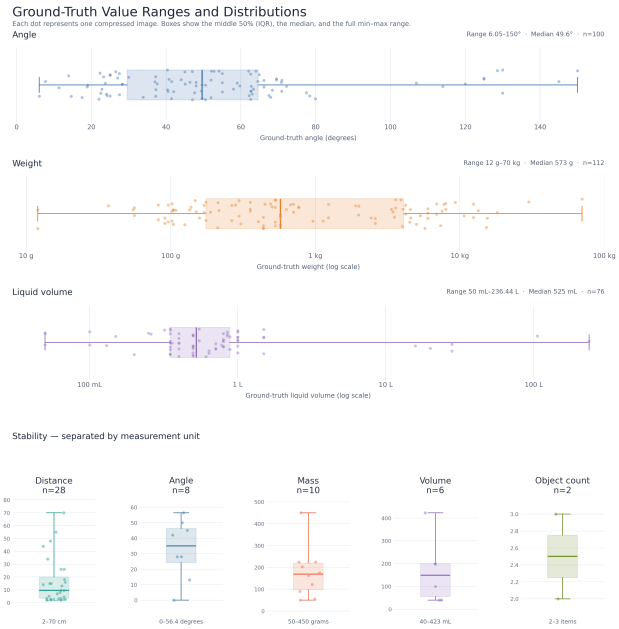


Figure 6: Ground truth value ranges and distributions for angle, weight, liquid volume and stability. Each dot is image, boxes show the middles 50% (IQR), the median and the full min-max range.

Simple Evaluation of Frontier LLMs

We evaluate three closed-weight families, taking the largest model in each with reasoning effort set to High. Each model receives the image and the question, and returns a reasoning trace and a final answer. We request structured output, so the predictions need no parsing. We sample three responses per item and report their median.

For the four numerical categories we report Median Absolute Percentage Error (MdAPE) over all items in the category. Fit is binary, so we report precision and recall.

Robustness Evaluation of Frontier LLMs

Our tasks need only the coarse shape, identity and orientation of the objects in a scene, and these survive mild blurring — a human can still attempt the question. Finer cues do not survive, among them brand names and printed text, which are what let a model recall a magnitude from a label instead of estimating it. We therefore re-evaluate every item under Gaussian blur at three levels: low, medium and high, with σ as 8.2, 12, 15.8 respectively. Figure 7 shows an original image beside its three blurred versions. The blurred subset contains 201 images, comprising 45 weight estimation images, 42 angle estimation images, 39 fit images, 39 stability images, and 36 liquid volume estimation images.

Fine-tuning open weight VLM To build our fine-tuning dataset, we started with approximately 45,000 images from the COCO (Lin et al. 2014) validation and randomly sampled 11,000 images for processing. We used a three-stage pipeline to construct the dataset as shown in lower section of Figure 1.

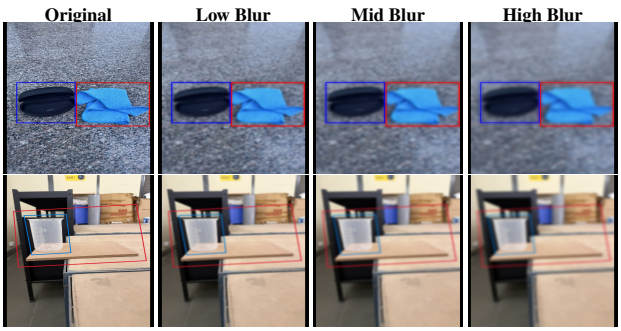


Figure 7: Original images and their low-, mid-, and high-blur versions.

In the 0th stage (blue), QUIVER images are used to generate scene understanding descriptions using Gemini 3.1 Flash-Lite. The model also identifies the objects enclosed by the red and blue bounding boxes and substitutes these object names into the generated questions.

In the first stage (purple), COCO images are processed to generate synthetic training data. The Gemini 3.1 Flash-Lite first determines whether an image is suitable for weight estimation by verifying that some multiple objects are present for reference, objects are clearly visible, have sufficient resolution, are not truncated by the image boundaries. This gaves us 6,426 images. For images that satisfy these criteria, the model (Gemini 3.5 Flash) is provided with in-context examples consisting of scene understanding and question pairs from QUIVER (without the corresponding QUIVER images) and is prompted to generate weight estimation questions for the COCO image which gave us 9,069 questions. The generated image-question pairs are then passed to Gemini 3.5 Flash to estimate the object's weight and produce reasoning traces. Representative examples of coco images for feasibility check are shows in from Stage 1 are shown in Figure 8, and the corresponding questions and answers generated in Stages 2 and 3 are presented in Table 1. We used parameter-efficient fine-tuning using low-rank adaptation (LoRA).

Experiments

We evaluate model performance using the Median Absolute Percentage Error (MdAPE). The closed-source models evaluated are GPT-5.4, gemini-3.1-pro-preview, and Gemini Robotics (gemini-robotics-er-1.6-preview). For the open-source setting, we use Gemma 4 31B and Qwen 3.5 9B, which are subsequently fine-tuned on our generated dataset.

Results in Table 2 compare the performance of three closed-source models (GPT 5.4, Gemini 3.1, and Gemini Robotics) and two open-source models (Gemma-4-31B and Qwen 3.5) across the five physical estimation tasks. Results are reported for both non-blurred and blurred images, while the open-source models are evaluated in their raw and fine-tuned settings.

Overall, the closed-weights models consistently outperform the open-weights models across all categories, demonstrating stronger physical reasoning capabilities. However, within each group, no single model consistently achieves the best performance across all tasks. Instead, the best-performing model is category-dependent, with different models excelling at different physical estimation tasks. Among the closed-source models, performance is relatively similar, with MdAPE values typically differing by only 5–6 percentage points across categories, indicating comparable overall capabilities despite category-specific strengths. A similar trend is observed for the open-source models, where both Gemma-4-31B and Qwen 3.5-9B exhibit task-dependent performance, and neither model consistently outperforms the other across all categories.

(a) Bright living and dining room with a television, dining table, chairs, and fireplace. A person near the kitchen provides scale. Suitable: Yes.

(b) Close-up of a brown bear's head resting on grass. Suitable: No.

(c) Two cats are looking through a window screen from inside a house. The scene includes a stone wall, a window frame, and the two cats as the primary subjects. Suitable: No.

(d) Bedroom with a metal-framed bed, table, chair, windows, purple wall, and stone column. Suitable: Yes.

Figure 8: Representative images evaluated during Stage 1, together with generated scene descriptions and suitability decisions.

Human Baseline

We sampled 17 questions in each category. Our study involved 95 unique users and yielded 159 responses for each catgeory , hence making 759 responses in total.

For continuous categories, we first average the user responses for each question and then compute the MdAPE across the 17 questions. For the boolean Fit category, we first determine the majority user response for each question and then compute the accuracy across the 17 questions.

Category	Metric	Users	Gemini 3.1	GPT 5.4
Angle	MdAPE	**8.32%**	10.43%	13.64%
Fit	Accuracy	**82.35%**	**82.35%**	70.59%
Liquid volume	MdAPE	14.68%	**5.71%**	20.00%
Stability	MdAPE	49.21%	**43.18%**	60.00%
Weight	MdAPE	37.48%	**12.74%**	15.23%

Table 3: Comparison of human and LLM performance. Lower MdAPE and higher accuracy indicate better performance.

Image	Generated Question	Answer	Reasoning
	**Weight** Estimate the weight of the flat-screen television.	10,000 g	The television appears to be approximately 42 inches in size. A modern flat-screen LCD/LED TV of this size typically weighs around 10 kg, giving an estimated weight of 10,000 g.
	Weight Estimate the weight of the wooden dining chair.	5,000 g	some reasoning trace
	Liquid Volume Estimate the full capacity of the lime-green vase on the coffee table in the foreground.	300 mL	The lime-green vase is approximately 22 cm tall with an average diameter of about 4 cm. Approximating it as a cylinder gives a volume of about 276 cm^3. Accounting for the tapered neck and wall thickness, its liquid capacity is estimated to be approximately 250–300 mL.
	Liquid Volume Estimate the full capacity of the purple glass vase next to the television.	1.5 L	some reasoning trace
	**Weight** Estimate the weight of the brown leather shoulder bag carried by the woman.	800 g	some reasoning trace
	Weight Estimate the weight of the swan on the riverbank.	10,000 g	some reasoning trace
	Fit Can the smartphone fit inside the shoulder bag?	True	some reasoning trace

Table 1: Examples of images, questions generated in Stage 2, their corresponding answers generated in Stage 3, and reasoning traces.

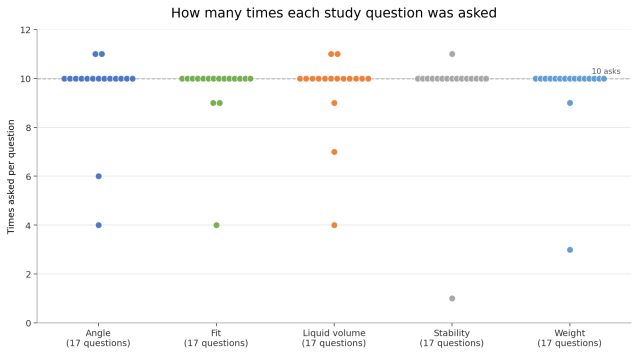


Figure 9: Distribution of how many times each study was asked. Each dot represents one question; the dashed line marks 10 asks.

The distribution of the number of responses collected per question is shown in Figure 9. Table 3 compares human performance with that of the evaluated closed-source models. Gemini matches human performance on the fit task. For angle estimation, humans outperform all evaluated models, although Gemini exhibits only a small performance gap relative to the human baseline. In contrast, Gemini achieves lower MdAPE than humans on liquid volume, stability, and weight estimation, indicating more accurate estimates on these tasks.

Finetune results

We fine-tuned two open-weight vision-language models, Qwen3.5-9B and Gemma4-31B, on a dataset of 9,096 weight estimation samples generated using our question generation pipeline. Each training sample consisted of an image paired with a reasoning trace, with the model trained to predict the corresponding weight value through supervised fine-tuning (SFT). For both models, input images were resized to 512 × 512 pixels, and 4-bit quantization was employed to reduce memory requirements during training.

Qwen3.5-9B was fine-tuned on an NVIDIA DGX Spark (128 GB) system using QLoRA (Dettmers et al. 2023) adapters and Llama-Factory (Zheng et al. 2024). Only the query and value projection layers within the Transformer attention blocks were adapted. The QLoRA configuration used a rank of 4, an alpha value of 8, and a dropout of 0.05. Training was performed with a learning rate of 1×10^{-4}, 5 training epochs, 4 gradient accumulation steps, a cosine learning rate scheduler, and a warm-up ratio of 0.05.

Gemma4-31B was fine-tuned on an NVIDIA H100 GPU hosted on RunPod using the same training configuration. We used QLoRA and the base model weights were loaded in 4-bit quantized form, while LoRA adapter training used BF16 mixed precision. QLoRA adapters were with rank 4, alpha 8, and dropout 0.05, a learning rate of 1×10^{-4}, 5 training epochs, and 4 gradient accumulation steps.

For the weight estimation task, fine-tuning substantially improves the performance of both models, leading to a drastic reduction in MdAPE compared to their corresponding

Image Type	Category	GPT 5.4	Gemini 3.1	Gemini Robotics	Gemma-4-31B	Qwen 3.5-9B
Original	Weight	**23.77%**	28.67%	24.72%	122.50%	**99.91%**
	Liquid Volume	28.02%	25.00%	**17.74%**	**50.20%**	88.80%
	Angle	11.65%	**9.13%**	10.08%	79.09%	**76.13%**
	Stability	**41.27%**	41.86%	53.37%	**82.10%**	90.61%
	Fit	P: 80.78% R: 85.37%	P: 88.68% R: 79.27%	**P: 84.65%** **R: 89.43%**	**P: 77.23%** **R: 93.98%**	P: 75.72% R: 95.12%
Blur	Weight	38.77%	**37.12%**	44.06%	130.00%	**98.87%**
	Liquid Volume	29.48%	30.56%	**20.17%**	**60.40%**	70.00%
	Angle	12.13%	**10.97%**	15.43%	81.79%	**55.70%**
	Stability	53.63%	64.85%	**37.20%**	85.20%	**81.81%**
	Fit	P: 81.63% R: 88.89%	P: 100.00% R: 97.78%	**P: 92.80%** **R: 100.00%**	**P: 95.15%** **R: 83.56%**	P: 75.67% R: 93.33%

Table 2: Category-wise performance of the evaluated closed and open models on original and blurred images. Original denotes evaluation on unmodified images, whereas Blur denotes evaluation after applying a blurring transformation. For the Fit category, P and R represent precision and recall, respectively. Only precision is available for Gemma-4-31B and Qwen 3.5-9B. Bold values indicate the best result within each model group.

Model	Raw	Blur	Fine-tuned	Fine-tuned (Blur)
Qwen 3.5	99.91%	98.87%	**74.11%**	**48.42%**
Gemma-4-31B	122.5%	130%	**95.32%**	**85.29%**

Table 4: Performance comparison of the raw and fine-tuned versions of Qwen3.5-9B and Gemma-4-31B for weight category.

original (non-finetuned) version. An interesting observation is that, after fine-tuning, both models achieve lower MdAPE on blurred images than on the original raw images. Results are shown in Table 4.

Conclusion

Our work introduces QUIVER, a diagnostic benchmark for physical-property estimation by VLMs. It comprises 459 hand-annotated real-world images across five categories — weight, liquid volume, angle, stability, and fit — each with measured numerical ground truth (apart from fit property that is categorical). To pose realistic and interesting estimation questions, scenes use atypical viewpoints and visible reference objects. A blurred version of every item tests robustness to the loss of object identifying text cues - a way to force estimation rather than recall.

We evaluated three frontier closed-weight models and two open-weight models against a human baseline. No single model dominates across tasks, but the frontier models are competitive overall and match or exceed humans in several categories. To improve performance of open-weight models Gemma 4 and Qwen 3.5, we fine-tuned them on an automatcially generated dataset for weight estimation task. The dataset was constructed using a three-stage pipeline that generates weight-estimation questions, estimate answers, and reasoning traces over COCO images. Fine-tuning on this data sharply reduced MdAPE, and the fine-tuned models stayed robust under blur.

The generated supervision uses no human-annotated traces or ground truth, so it inherits the noise and bias of the closed models that produce it. Under compute and budget constraints, we fine-tuned for weight only. Future work will extend the pipeline to the other four properties, collect higher-quality supervision where feasible, and demonstrate how better physical estimation improves downstream robotics tasks such as task planning.

References

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; and Zhao, F. 2024. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, volume 37, 27056–27087.

Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V. C.; and Wang, Y. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *International Conference on Learning Representations*.

Criminisi, A.; Reid, I.; and Zisserman, A. 2000. Single View Metrology. *International Journal of Computer Vision*, 40(2): 123–148.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 10088–10115.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. BLINK:

Multimodal Large Language Models Can See but Not Perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Gao, J.; Sarkar, B.; Xia, F.; Xiao, T.; Wu, J.; Ichter, B.; Majumdar, A.; and Sadigh, D. 2024. Physically Grounded Vision-Language Models for Robotic Manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 12462–12469.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huh, G.; Sheth, D.; Zirvi, R.; and Xiao, F. 2025. SpatialTraceGen: High-Fidelity Traces for Efficient VLM Spatial Reasoning Distillation. *arXiv preprint arXiv:2511.00054*.

Lan, H.; An, Z.; Li, H.; Singh, V.; and Shangguan, L. 2025. Accelerating Physical Property Reasoning for Augmented Visual Cognition. *arXiv preprint arXiv:2511.03126*.

Lee, S.; Jeong, J.; Hong, Y.; and Kim, K. I. 2026. Physically Guided Visual Mass Estimation from a Single RGB Image. *arXiv preprint arXiv:2601.20303*.

Li, B.; Lin, Z.; Peng, W.; Nyandwi, J. d. D.; Jiang, D.; Ma, Z.; Khanuja, S.; Krishna, R.; Neubig, G.; and Ramanan, D. 2024. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. In *Advances in Neural Information Processing Systems*, volume 37.

Li, P.; Xiang, T.; Mao, E.; Wei, S.; Chen, X.; Masood, A.; Fei-Fei, L.; and Adeli, E. 2025. QuantiPhy: A Quantitative Benchmark Evaluating Physical Reasoning Abilities of Vision-Language Models. *arXiv preprint arXiv:2512.19526*.

Lin, F.; Liu, Y.; Xu, H.; Yue, C.; He, Z.; Zhao, M.; Chen, M. H.; Liu, J.; Yao, J. G.; and Yang, X. 2025. Do Vision-Language Models Measure Up? Benchmarking Visual Measurement Reading with MeasureBench. *arXiv preprint arXiv:2510.26865*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 34892–34916.

Luo, D.; Li, Y.; Wang, M.; Zhao, T.; Wang, B.; Wang, S.; Feng, P.; Rahmanzadehgervi, P.; Ma, Z.; and Deng, H. 2026. Vision Language Models Cannot Reason About Physical Transformation. In *Proceedings of the 43rd International Conference on Machine Learning*.

Melnik, A.; Schiewer, R.; Lange, M.; Muresanu, A.; Saeidi, M.; Garg, A.; and Ritter, H. 2023. Benchmarks for Physical Reasoning AI. *Transactions on Machine Learning Research*.

Shen, H.; Wu, T.; Han, Q.; Hsieh, Y.; Wang, J.; Zhang, Y.; Cheng, Y.; Hao, Z.; Ni, Y.; Wang, X.; et al. 2025. PhyX: Does Your Model Have the "Wits" for Physical Reasoning? *arXiv preprint arXiv:2505.15929*.

Standley, T.; Sener, O.; Chen, D.; and Savarese, S. 2017. image2mass: Estimating the Mass of an Object from Its Image. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, 324–333.

Weng, T.; Wang, J.; Jiang, W.; and Ming, Z. 2025. VisNumBench: Evaluating Number Sense of Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3830–3840.

Xu, X.; Ge, W.; Qiu, D.; Chen, Z.; Yan, D.; Liu, Z.; Zhao, H.; Zhao, H.; Zhang, S.; Liang, J.; et al. 2025. GaussianProperty: Integrating Physical Properties to 3D Gaussians with LMMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7231–7240.

Yokomizo, H.; Miyanishi, T.; Gang, Y.; Kurita, S.; Inoue, N.; and Iwasawa, Y. 2026. PhysQuantAgent: An Inference Pipeline of Mass Estimation for Vision-Language Models. *arXiv preprint arXiv:2603.16958*.

Yu, S.; Chen, Y.; Ju, H.; Jia, L.; Zhang, F.; Huang, S.; Wu, Y.; Cui, R.; Ran, B.; Zhang, Z.; Zheng, Z.; Zhang, Z.; Wang, Y.; Song, L.; Wang, L.; Li, Y.; Shan, Y.; and Lu, H. 2025. How Far are VLMs from Visual Spatial Intelligence? A Benchmark-Driven Perspective. *arXiv preprint arXiv:2509.18905*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. 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*, 9556–9567.

Zhai, A. J.; Shen, Y.; Chen, E. Y.; Wang, G. X.; Wang, X.; Wang, S.; Guan, K.; and Wang, S. 2024. Physical Property Understanding from Language-Embedded Feature Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28296–28305.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 400–410. Association for Computational Linguistics.