HUMAN-LEVEL ON WHICH QUESTIONS? A RANKING REVERSAL IN VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

As frontier VLMs increasingly achieve near- or super-human performance on specific benchmarks, there has been increasing speculation that VLMs have super-human capabilities, based on the assumption that performance should generalize. We think that it is only healthy to question this assumption. We find that 8 frontier VLMs match or exceed human performance on AI-generated visual-reasoning questions (AIGEN) but fall substantially behind humans on human-generated questions (HUMANGEN). This happens because frontier VLMs are sensitive to the difference between AIGEN and HUMANGEN, while humans are robust to it. We carry out an in-depth analysis of both audits to see what causes this trend. We find that surface-level phrasing does not explain it. Instead, the result is due to differences in the reasoning paths required by the two audits. HUMANGEN tends to have longer paths that involve finding visual cues that frontier VLMs are less likely to look for. Providing these cues in oracle form restores performance for frontier VLMs, but to varying degrees, suggesting that some VLMs still lack world knowledge necessary to deduce an answer from these cues.

1 INTRODUCTION

Whenever a frontier vision-language model (VLM) matches or exceeds human performance on a specific benchmark, the result is often interpreted as evidence of some general level of capability. We think it is only healthy to question this interpretation, especially as AI is increasingly used as a data-*generation* tool. An AI generator and a human generator may not necessarily cover the same regimes of visual cues, world knowledge, and reasoning paths; hence, it is worth asking whether near- or super-human performance on AI-generated data implies the same on human-generated data.

To investigate this question, we construct a visual question answering (VQA) audit dataset comprising an AI-generated subset (AIGEN) and a human-authored counterpart (HUMANGEN). Both follow the same annotation principle: create questions that require multiple visual cues to answer. As illustrated in Figure 1, even for the same visual scene, AI-generated and human-authored questions can target different objects and relationships. These complementary question sets allow us to examine whether a model's performance advantage over humans holds when the questions, and the reasoning demands they impose, change.

Our audits reveal a reversal of rankings (see Figure 1). Eight frontier VLMs, from multiple different families, match or exceed human performance on AIGEN but fall substantially behind humans on HUMANGEN. To better understand this reversal, we systematically examine what distinguishes the two audits (3.2) We have humans rewrite all AIGEN questions and observe little change in aggregate VLM accuracy, suggesting that surface wording alone does not explain the gap. We then analyze the visual observations and knowledge steps required to answer each question. HUMANGEN questions typically require more steps, and VLM accuracy tends to decline as the number of steps increases, while human accuracy remains comparatively stable. Nevertheless, even among questions requiring the same number of steps, VLMs perform worse on HUMANGEN than on AIGEN, indicating that step count alone is also insufficient to explain the discrepancy. Counting the steps *not attempted by the VLMs* yields a more powerful explanatory signal, but a gap still exists.

We conclude our investigation by conducting component-level probes and oracles (3.6) for HUMANGEN questions, inspired by Selvaraju et al. (2020). Component-level probes decompose the original questions into constituent questions about relevant objects and relationships, while the component

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.

(a) AI-generated question



What must happen for the docked vessel to pass through the channel in front of it?

Reasoning demand: infer bridge operation from vessel clearance.

(b) Human-authored question



I want to explore the castle on the left. Which boat should I take?

Reasoning demand: combine destination, route, bridge clearance, and boat choice.

(c) Participant-balanced accuracy

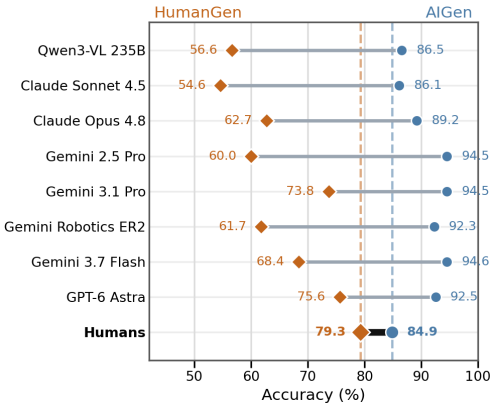


Figure 1: **(a–b)** Two example questions, one AI-generated and one human-generated, illustrating how human and AI generators can select different reasoning demands for the same image. **(c)** VLMs that outperform humans at answering AI-generated questions fall behind when given human-generated questions, as they appear to be more sensitive to the difference between generators.

oracle provides the ground-truth answers of these probes to the VLMs to see how much performance can be recovered on the original questions. VLMs do well (upto saturation) on the component probes, while the component oracles allow VLMs to recover some performance, but to a varying extent.

All in all, our study highlights the importance of being careful about AI-generated benchmarks, which are becoming increasingly common. These can mislead not only about absolute performance, but also about performance relative to humans.

Our contributions are as follows:

- We show that a favorable human-relative assessment does not transfer across two visual-audit construction procedures. Frontier VLMs that meet or exceed the observed humans on caption-constrained AIGen fall below them on complementary HumanGen questions, under matched within-audit support and participant weighting.
- We characterize the shift using a wording control, independently interpretable reasoning-path measurements, cue-attempt probes, and component/fact-support interventions, which separate several candidate explanations without attributing the reversal to a single causal mechanism.

2 Evaluation Setup

In this section we discuss our protocols for constructing our two audits and running VLM and Human evaluations on them.

2.1 Audit Construction

Our two audits use images from PixMo-Cap Deitke et al. (2024), which is a set of web-scraped images from a variety of sources that come with extremely dense captions which are summarized from one or more human annotator voice transcripts. From these, we construct a 153-question AIGen audit using Gemini-2.5-Pro with human review, and a 61-question HumanGen audit using only a human. See Appendix I for example questions from both audits.

2.1.1 AIGen Construction

To construct AIGen, we sample images from PixMo-Cap and feed each one to a Gemini-2.5-Pro generator, along with its provided caption. The generator is instructed to come up with binary-choice questions that involve higher-level reasoning, require multiple visual cues to answer, do not leak visual cues in the text of the question, and do not reduce to caption-matching tasks. The generator is given examples of types of questions it could ask, as well as in-context examples of specific image and questions. The generator has the option of flagging the image as unsuitable for generating a

question. For interpretability purposes, the generator also asked to write down the reasoning for solving the question. The generation prompt is available in Appendix C.

The purpose of the caption is twofold. First, it is meant to mitigate the risk that the generator might hallucinate details that it perceives incorrectly. And second, it is meant to ensure that the generator has access to perceptual cues that humans would perceive, as opposed to only the cues that Gemini-2.5-Pro would perceive; if it had access to only the latter, then the tasks it generates would be definitionally solvable by Gemini-2.5-Pro, and likely other VLMs as well.

Two human researchers review the generated questions, discarding any that do not comply with the stated requirements and lightly editing a few to make them compliant. This review step has roughly 10% yeild rate. We end up with 153 images, each with one question.

2.1.2 HUMANGEN CONSTRUCTION

To construct HUMANGEN, we randomly sample images from PixMo-Cap and have a human researcher follow the same instructions to make questions. Captions are available to the human, but in practice the human typically uses their own judgment to create questions. Like the AI generator, the human generator is also asked to write down their reasoning for interpretability purposes. Yield rate is around 1-2%, as humans tend to be less creative and persistent than AI in trying to find a question that works for each image. We end up with 40 images and 61 questions, as some image get multiple questions. No single image gets more than 4 questions.

2.2 EVALUATION PROTOCOLS

We evaluate both VLMs and human subjects on both AIGEN and HUMANGEN. We describe the evaluation protocols below.

2.2.1 MODEL EVALUATION

We evaluate 15 VLMs from 5 different model families. These are: Gemma-3-4B, Gemma-3-12B, and Gemma-3-27B (Gemma Team, 2025); Llama3.2-11B (Meta AI, 2024) and Llama4-Scout and Llama4-Maverick (Meta AI, 2025); Qwen3-VL-235B (Bai et al., 2025); Claude-Haiku-4.5, Claude-Sonnet-4.5, and Claude-Opus-4.8 (Anthropic, 2025a;b; 2026); Gemini-2.5-Pro, Gemini-3.1-Pro, Gemini-Robotics-ER2, and Gemini-3.7-Flash (Google DeepMind, 2025; 2026a;c;b); and GPT-6-Astra (OpenAI, 2026). In the interest of space, the main paper focuses on the 8 VLMs that match or exceed human performance on AIGEN. Results for the other 7 VLMs are restricted to Appendix H.

Each VLM is given the image, question, and options and is asked to output a JSON containing an answer (A or B), and also reasoning for interpretability purposes. Each VLM is run for 4 repeat trials on each question. If a VLM does not give a valid answer, it is rerun until it does. Options are always presented in random order to avoid positional bias.

2.2.2 HUMAN EVALUATION

We recruit 12 human subjects to evaluate on both HUMANGEN and AIGEN. Subjects consist of undergraduate and graduate students, as well as faculty members and industry professionals. Each subject is given a random sample of AIGEN and HUMANGEN questions, in random order, with a ratio of 2 AIGEN for every HUMANGEN question. Option order is always randomized, and humans are not told if the question they are looking at is from AIGEN or HUMANGEN. All 153 AIGEN questions, and 58 out of the 61 HUMANGEN questions, are answered by at least one human, with the total number of human answers being 210 for HUMANGEN and 420 for AIGEN. See Appendix D for details and extended analysis. For the primary human–model comparison, model accuracy is computed over the same participant-observed question support and with the same participant weighting as human accuracy within each audit.

3 RESULTS

3.1 DOES AIGEN PERFORMANCE IMPLY HUMANGEN PERFORMANCE?

We see in Fig. 1(c) that 8 of the VLMs outperform humans on the AIGEN audit, with GPT-6-Astra and all the Gemini's outperforming by 7% or more. However, all 8 of these models underperform humans on the HUMANGEN audit, with only 2 of them (Gemini-3.1-Pro and GPT-6-Astra) coming within 7% of human performance.

This reversal of fortunes, so to speak, happens because **VLMs are more sensitive than humans** to the shift between AIGEN and HUMANGEN. Humans only do 5.6% worse on HUMANGEN than on AIGEN, while the VLMs degrade by around 17% or more (>3x the human amount), with Gemini-2.5-Pro suffering the greatest collapse at 34.5% (>6x the human amount).

We show some expanded results in Appendices H and D, including results for all 15 VLMs; results for the same but weighted to balance image contribution rather than human subject contribution; and leave-one-subject-out analysis of the human study to demonstrate stability.

Given that superhuman performance on AIGEN does **not** imply the same on HUMANGEN, it is worth looking more closely at the two audits to see why.

3.2 WHAT DISTINGUISHES THE TWO AUDITS?

Question phrasing: To answer this question, we start by looking at the phrasing of the questions, as AI models are known to use idiosyncratic phrasing that is statistically distinct from human writing. To measure the influence of phrasing, we have a human researcher, independent of the HUMANGEN question-writing process, rewrite the questions from AIGEN while preserving their meaning. We see in Fig. 2c that this only has a marginal effect on the Gemini VLMs, changing their accuracy by less than a percent or two. Thus, phrasing does not explain our main result.

Reasoning paths: We next consider the **reasoning path** that a human or VLM would have to follow in order to answer each question. To quantify this, we manually annotate all of HUMANGEN and roughly half of AIGEN (chosen at random) with reasoning paths, representing each path as a list of atomic reasoning steps. Each reasoning step has a free-text description, a type (VISUAL, KNOWLEDGE, or SYNTHESIS), and an explicitness level (EVERYTHING, SOME_THINGS, or NOTHING) indicating the fraction of concepts in the step that are explicitly mentioned in the question or answer options. The minimum necesary set of steps is considered the MAIN path, and any other supporting steps, if present, are considered an AUXILIARY path. We find in our analysis that auxiliary paths have very little predictive power, so we only consider main paths in our analysis. We describe the reasoning-path annotation process, including the WebUI used, in the Appendix.

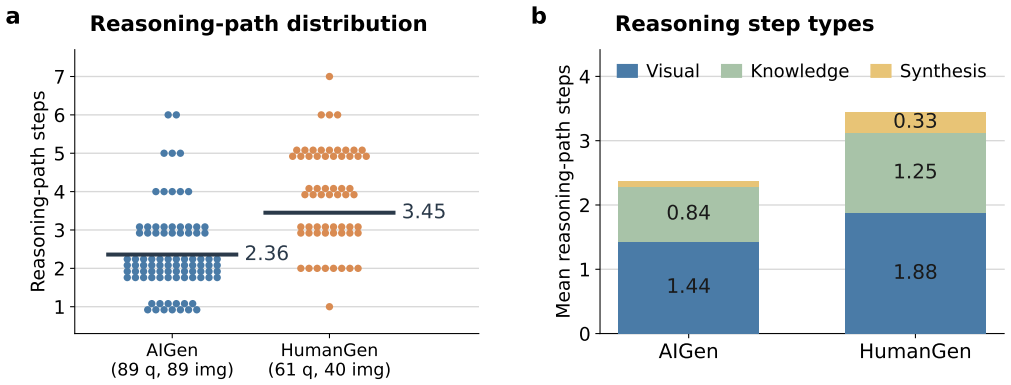
Fig. 2(a-b) shows coarse statistics about the reasoning paths for AIGEN and HUMANGEN. We see that HUMANGEN has an average path length of 3.45 steps, compared to AIGEN's 2.36, and that the disparity applies to visual, knowledge, and synthesis steps alike.

3.3 ARE HUMANS MORE ROBUST TO LONG PATHS?

VLM	L	L_w	L_V	L_K	L^U	L^U_w
Qwen3-VL 235B	-0.282	-0.274	-0.291	-0.058	-0.405	-0.417
Claude Opus 4.8	-0.341	-0.359	-0.270	-0.206	-0.361	-0.376
Gemini 2.5 Pro	-0.325	-0.344	-0.285	-0.188	-0.458	-0.484
Gemini 3.1 Pro	-0.165	-0.175	-0.139	-0.123	-0.317	-0.308
Gemini Robotics ER2	-0.177	-0.198	-0.144	-0.152	-0.541	-0.569
Gemini 3.7 Flash	-0.165	-0.193	-0.129	-0.124	-0.451	-0.419
GPT-6 Astra	-0.110	-0.142	-0.210	-0.020	-0.517	-0.501

Table 1: Pearson correlation of different reasoning-path length metrics (see 3.3 for definitions) against VLM correctness on AIGEN $\bigcup$ HUMANGEN.

The findings from the previous section warrant a deeper dive. Is reasoning path length a good predictor of VLM performance? What is the best way to quantify reasoning path length? And



c **Human rewrite of AIGen**

Gemini model	2.5 Pro	3.1 Pro	Robotics ER2	3.7 Flash
Original (%)	94.3	94.3	92.7	94.6
Rewrite (%)	93.5	96.1	94.6	93.5
Change (pp)	−0.8	+1.8	+1.9	−1.1

Figure 2: **HUMANGEN has longer reasoning paths than AIGEN** (a) Distribution of reasoning-path step count for AIGEN and HUMANGEN, showing that the latter has longer paths on average. (b) Breakdown by step type, showing that the difference applies to all types. (c) Question-phrasing ablation in which a human rewrites the AIGEN questions. Phrasing has negligible impact.

does the rank reversal between humans and VLMs happen because humans are more robust to long reasoning paths? To begin answering these questions, we start by probing the VLMs to see how much of the reasoning paths they actually cover.

Step-attempt probe: We ask each VLM to list all visual observations that it would need in order to determine the answer to each question, given each image. We use an LLM-judge (Opus 4.8) compare their outputs against the reasoning-paths and mark whether or not each visual step was mentioned (i.e. "attempted") by the VLM. Probe and judge prompts are available in the Appendix.

Reasoning-path length metrics: Given a reasoning-path p with steps $s_1, ..., s_N$ and step-attempt probe results, we can quantify path length in several different ways, which we define below:

- $L(p) := N$, i.e. plain path length.
- $L_w(p) := \sum_{i=1}^{N} w(s_i)$, where w is an explicitness-based weighting function. We choose:

$$w(s) := \begin{cases} \frac{1}{3} & \text{for EVERYTHING-explicit steps,} \\ \frac{2}{3} & \text{for SOME_THINGS-explicit steps,} \\ 1 & \text{for NOTHING-explicit steps.} \end{cases}$$

- $L_K(p) := \sum_{i=1}^{N} \mathbf{1}\{s_i \text{ is a knowledge step}\}$
- $L_V(p) := \sum_{i=1}^{N} \mathbf{1}\{s_i \text{ is a visual step}\}$
- $L^U(p) := \sum_{i=1}^{N} \mathbf{1}\{s_i \text{ is visual and was not attempted in the step-attempt probe}\}$
- $L^U_w(p) := \sum_{i=1}^{N} w(s_i)\, \mathbf{1}\{s_i \text{ is visual and was not attempted in the step-attempt probe}\}$

Predictive power of different length metrics: Tab. 1 shows the Pearson correlation of each of the above metrics with the correctness rate (across the 4 trials) of each VLM. We see that all path-length metrics consistently have a negative correlation with correctness rate for all VLMs. L_w almost always has a stronger correlation than L, suggesting that explicitness weighting has predictive power

because models struggle with cues not explicitly mentioned in the question. We also see a sizeable improvement from L and L_w to L^U and L_w^U, suggesting that **VLMs struggle most when they fail to attempt reasoning steps** needed for answering the question. In particular, L_w^U alone explains roughly a third and a quarter of the variance in performance of Gemini-Robotics-ER2 and GPT-6-Astra, respectively.

Path length vs correctness: Fig. 3 takes a deeper dive by plotting a Gaussian-kernel average of human and VLM accuracy against each path length metric. We see that several of the VLMs outperform humans at low values of L_w, but as reasoning steps are added, their performance immediately drops in a near-linear fashion and never recovers to its initial level. Human performance, by contrast, is almost flat. The relationship between L_K and performance is interesting as well; the strongest models can tolerate a single knowledge step, but a second one is associated with lower performance.

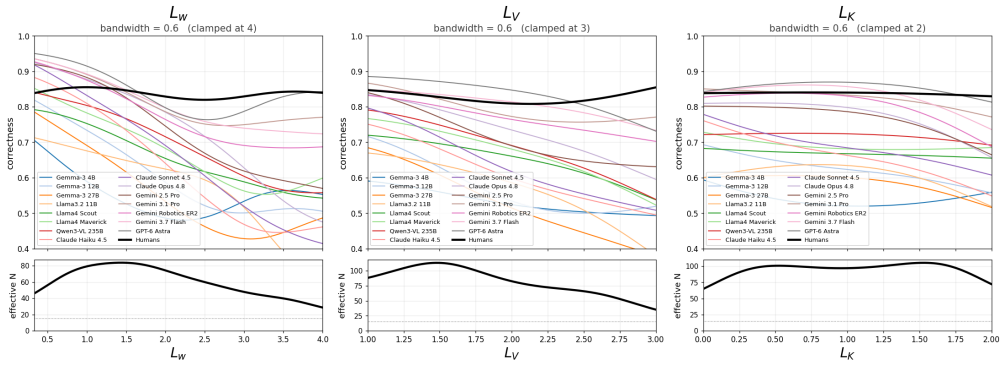


Figure 3: Relationship reasoning-path metrics (L_w, L_V, and L_K) and VLM performance on AIGEN $\bigcup$ HUMANGEN.

3.4 DOES REASONING PATH LENGTH TELL THE WHOLE STORY?

We see that VLM performance correlates with reasoning-path length. But does path length fully account for the disparity between AIGEN and HUMANGEN. If that were the case, then controlling for path length would cause the disparity to collapse.

Fig. 4 shows the accuracy of VLMs on AIGEN and HUMANGEN if we include all path-annotated questions, on questions with length $L_w \leq 2$, and on questions where the VLM attempted all visual reasoning steps (i.e. $L^U = 0$). We see that restricting to $L_w \leq 2$ does *not* cause disparity to collapse. Setting $L^U = 0$, on the other hand, does cause most of the models to have a significantly narrower gap. This suggests that VLMs struggle on HUMANGEN in part because **it contains more steps that they are unlikely to attempt**.

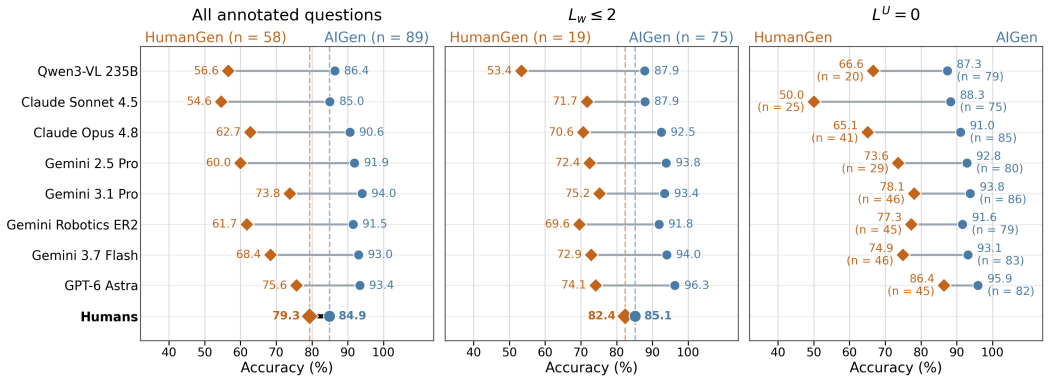


Figure 4: Controlling for path length L_w does not collapse the disparity in VLM performance between AIGEN and HUMANGEN, but controlling for unattempted steps L^U does reduce the disparity.



Native L3 question
What kind of object is preferable for this situation?

A. A cable lock. **B.** A U-lock.
Reference: A (cable lock).

Annotated reasoning-path
1. The rack is too low for a U-lock to reach more than the wheel (visual; only some concepts mentioned in question).
2. The wheel has a quick release (visual; no concepts mentioned in the question).
3. Locking only a removable wheel leaves the bicycle unsecured (knowledge).

Figure 5: An illustrative example in which the reasoning path is short but relies heavily on implicit cues.

3.5 A WORKED EXAMPLE: WHEN THE SCENE OVERRIDES A GENERIC RULE

Consider the HUMANGEN bicycle task in Fig. 5. The image shows a bicycle at a low wheel rack; the question asks which lock is preferable, a cable lock or a U-lock. Although a U-lock is harder to cut, this particular rack would only allow it to secure the front wheel, which is only attached via a quick-release, and so a bicycle thief without any tools could walk off with the bicycle after just a few seconds of effort.

This question is interesting because it is one that any bike commuter would immediately understand, and one that was solved correctly by all five of our human participants who saw it. However, it appears to be challenging for some frontier VLMs, as seen in Tab. 2, where Gemini-3.7-Flash succeeds but Qwen3-VL-235B fails to choose the correct answer. We see in Fig. 5 that the reasoning path has only a few steps, but most of the cues are not explicitly mentioned in the question or options. Hence, a successful VLM must look for these cues without being told to look for them. We see from the step coverage in Tab. 2 that Gemini-3.7-Flash does cover most of the steps, while Qwen3-VL-235B misses most of them.

Configuration	Native question correctness rate	Step coverage	VLM's rationale (condensed)
Qwen3-VL-235B	0/4	0.50	Favors U-lock's resistance to cutting.
Gemini-3.7-Flash	4/4	0.125	Notes that U-lock would not reach frame from this rack.

Table 2: Correctness and step coverage rates for two frontier VLMs that match or exceed human performance on AIGEN. Gemini covers most steps and reliably answers the question. Qwen misses most steps and consistently gets the wrong answer.

3.6 WHAT DO COMPONENT PROBES AND ORACLES REVEAL?

We have found so far that VLMs do worse on AIGEN than on HUMANGEN, in part because HUMANGEN has longer reasoning-paths and in part because the steps in those paths tend to be outside the realm of what VLMs usually attempt. But a gap still remains.

As a final analysis, we consider **component probes**, which are binary-choice perceptual questions about objects and relationships in the image. We use Gemini-2.5-Pro to generate component probes for each question and then manually review and filter them for relevance and quality. We give full details for this process in Appendix G.

We also construct a **component oracle**, in which we ask the VLM the original question but supply it with the components probes and their ground-truth answers. This oracle eliminates the need to do perception *and also* the need to figure out which perceptual tasks to attempt. All that remains is

the task of combining perceptual facts with world knowledge in order to answer the question. See Appendix G for implementation details for the component oracle.

We see on the left side of Fig. 6 that VLMs do much better on component probes than on the corresponding native questions, in fact they saturate the component probes in HUMANGEN. This suggests that perception is not the main hurdle for HUMANGEN, at least not for the strongest VLMs.

The right side of Fig. 6 shows VLM performance on the component oracle compared to the native question. Supplying component ground-truths to the VLMs improves their performance, although the effect size varies between VLMs. For example, Gemini-2.5-Pro is only 3.4% better than Qwen3-VL-235B on native questions, but their oracle performance differs by 16%, suggesting that Gemini-2.5-Pro is able to make better use of component ground-truths. Meanwhile, Gemini-3.1-Pro, Gemini-3.7-Flash, and GPT-6-Astro are able to surpass human performance when they (but not the humans) are given component ground-truths. It remains to be seen whether our human subjects fall short here due to a lack of world knowledge, a failure to look for the right cues, or a combination of both.

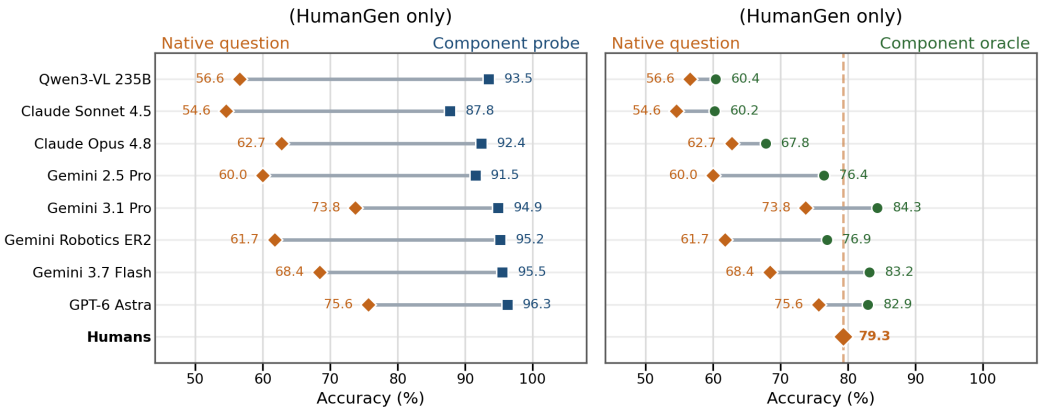


Figure 6: **(left)** VLMs achieve saturated accuracy on component probes, indicating that perception is not the main hurdle for HUMANGEN. **(right)** Supplying VLMs with ground-truth component probe answers improves their performance, allowing some of them to surpass humans on HUMANGEN.

4 DISCUSSION

To summarize our analysis, we find that frontier VLMs that match or outperform humans on our AIGEN audit fall substantially behind humans on our HUMANGEN audit, because the models degrade much more across the two audits than the human participants do. We then ask what measured differences between the audits are associated with this reversal. Surface-level phrasing is unlikely to account for the effect: human rewriting of AIGEN questions produces little aggregate change in VLM accuracy. HUMANGEN questions have longer recorded reasoning paths on average, and VLM accuracy declines as those paths become longer, while human accuracy is comparatively stable. Measures that count visual steps not mentioned under a separate cue-elicitation probe are more predictive of VLM error than raw path length, although they do not fully account for the AIGEN–HUMANGEN gap. Finally, VLMs perform strongly when directly queried about many component-level visual facts, and supplying selected ground-truth component facts improves native-question performance to varying degrees across models. These diagnostics constrain possible explanations for the reversal, but they do not identify a single causal mechanism.

An important direction for future work is to characterize precisely which visual evidence and inferential relationships are most likely to be omitted, misused, or difficult to compose. This could help distinguish failures of evidence selection from failures of perception, knowledge, or inference. It would also be useful to compare a broader range of benchmark-construction procedures, including alternative model-generated and human-generated pipelines, to determine which task demands they systematically emphasize or omit.

Broadly, our results suggest that evaluation construction selects not only how questions are phrased, but also which visual evidence and inferential relationships models are asked to use. A favorable human-relative assessment on one audit therefore need not persist under a complementary construction procedure, even when humans and models are compared on matched support within each audit. As model-generated evaluation data become increasingly common, the central issue is not whether such benchmarks should be used, but how broadly their task demands cover the capabilities about which we hope to make claims. Human-relative capability claims are strongest when the underlying evaluation procedures probe complementary forms of visual evidence and inference, rather than relying on a single procedure. The task demands in such cases may systematically favor one system over another.

5 Related work

Visual reasoning and human comparison. VCR tests image-grounded commonsense answers and rationales (Zellers et al., 2019); GQA generates compositional questions from scene graphs and functional programs (Hudson & Manning, 2019). BLINK compares humans and models on core visual tasks (Fu et al., 2024); MVP-Bench includes human evaluation across low/high-level questions on natural/manipulated images (Li et al., 2024b). AdVQA demonstrates differential human/model difficulty through model-in-the-loop adversarial collection (Sheng et al., 2021), distinct from our audit comparison, although HumanGen curation was model-informed. SpinBench (Zhang et al., 2026) provides a contemporary general-purpose-VLM precedent for human-relative sign reversal across task subtypes including viewer-perspective face-rotation classification and own-perspective classification. **Our question is different:** we ask whether a favorable human-relative assessment obtained under one audit-construction procedure transfers to a complementary construction procedure. Unlike fixed-benchmark human–model comparisons such as VCR and BLINK, or SpinBench's reversal across task subtypes, our comparison varies how the evaluation questions are constructed while matching model and human support within each audit.

Component consistency and reasoning diagnostics. VQA-Introspect evaluates reasoning/perception-subquestion inconsistency in both directions and human answering from supplied subquestions/answers without images (Selvaraju et al., 2020). CURE conditions high-level correctness on correct subanswers (Chen et al., 2024); LENS evaluates perception, understanding and reasoning on shared images (Yao et al., 2026). SPLICE and PercepTax supply event descriptions and ground-truth object properties, respectively (Ballout et al., 2025; Lee et al., 2026). Component/native inconsistency, multilevel diagnostics and supplied-information interventions thus have direct precedent. BLINK-Twice combines challenging questions, adversarial image pairs and annotated reasoning steps, scoring answer-associated reasoning for visual details and scene interpretation (Ye et al., 2025); our S probe scores separately elicited cue lists. Our component/native and image-preserving selected-fact comparisons support characterization of the audit contrast.

What evaluation construction tests. Hand-curated Winoground tests relational meaning with captions rearranging the same words across paired images (Thrush et al., 2022). SugarCrepe reduces linguistic distractor artifacts through language-model generation and adversarial refinement (Hsieh et al., 2023); like GQA, it demonstrates deliberate automatic construction of compositional tests. CLEVR-Humans establishes historical synthetic-to-human-written transfer difficulty for CLEVR-trained, task-specific systems, also evaluated after target-data finetuning (Johnson et al., 2017). This training-transfer setting differs from human-relative evaluation of contemporary general-purpose VLMs. NaturalBench constructs challenging questions through caption-based generation, confounding-pair selection and human verification (Li et al., 2024a): generated questions need not be easy or model-favoring. We measure differential degradation and human-relative sign change across two audit procedures. AIGen restricts solving cues to caption content; rewriting and reference annotations examine wording and recorded solution structure, respectively.

Xu et al. (2025) identify testset-generation and evaluator self-bias in translation, with related self-bias in open-ended generation. Our human-referenced visual comparison does not isolate that mechanism: it measures differential degradation and reversed human-relative assessment across audit procedures, supported by task-structure and elicitation/intervention diagnostics.

REFERENCES

Anthropic. Introducing claude haiku 4.5, October 2025a. URL https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. Claude sonnet 4.5 system card, September 2025b. URL https://www.anthropic.com/system-cards.

Anthropic. Introducing claude opus 4.8, May 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Mohamad Ballout, Okajevo Wilfred, Seyedalireza Yaghoubi, Nohayr Muhammad Abdelmoneim, Julius Mayer, and Elia Bruni. Can you SPLICE it together? a human curated benchmark for probing visual reasoning in VLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11288–11309. Association for Computational Linguistics, November 2025. doi: 10.18653/v1/2025.findings-emnlp.604. URL https://aclanthology.org/2025.findings-emnlp.604/.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 192–210. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.naacl-long.11. URL https://aclanthology.org/2024.naacl-long.11/.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024. URL https://arxiv.org/abs/2409.17146.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2404.12390.

Gemma Team. Gemma 3. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Google DeepMind. Gemini 2.5 pro model card, June 2025. URL https://deepmind.google/models/model-cards/.

Google DeepMind. Gemini 3.1 pro model card, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.7 flash model card, August 2026b. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Google DeepMind. Gemini robotics er 2 model card, July 2026c. URL https://deepmind.google/models/model-cards/gemini-robotics-er-2/.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.14610.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1902.09506.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2989–2998, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Johnson_Inferring_and_Executing_ICCV_2017_paper.html.

Jonathan Lee, Xingrui Wang, Jiawei Peng, Luoxin Ye, Zehan Zheng, Tiezheng Zhang, Tao Wang, Wufei Ma, Siyi Chen, Yu-Cheng Chou, Prakhar Kaushik, and Alan Yuille. PercepTax: Benchmarking physical intelligence through cross-property reasoning in vision-language models. In *Computer Vision – ECCV 2026*, volume 17008 of *Lecture Notes in Computer Science*, pp. 73–91, Cham, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-032-37026-6_5. URL https://link.springer.com/chapter/10.1007/978-3-032-37026-6_5.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://arxiv.org/abs/2410.14669. Datasets and Benchmarks Track.

Guanzhen Li, Yuxi Xie, and Min-Yen Kan. MVP-Bench: Can large vision-language models conduct multi-level visual perception like humans? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13505–13527. Association for Computational Linguistics, November 2024b. doi: 10.18653/v1/2024.findings-emnlp.789. URL https://aclanthology.org/2024.findings-emnlp.789/.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. Gpt-6 astra: A new generation of intelligence, September 2026. URL https://openai.com/index/gpt-6-astra/.

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. SQuINTing at VQA models: Introspecting VQA models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2001.06927.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/aa97d584861474f4097cf13ccb5325da-Abstract.html.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2204.03162.

Wenda Xu, Sweta Agrawal, Vilém Zouhar, Markus Freitag, and Daniel Deutsch. When LLMs benchmark themselves: Deconstructing self-bias in automated evaluation. *arXiv preprint arXiv:2509.26600*, 2025. URL https://arxiv.org/abs/2509.26600v3. Version 3, revised August 2026.

Ruilin Yao, Bo Zhang, Jirui Huang, Xinwei Long, Yifang Zhang, Tianyu Zou, Shili Xiong, Yi Rong, Yufei Wu, Shichao Su, Yifan Xu, Wenxi Zeng, Zhaoyu Yang, Guoyou Li, Shilan Zhang, Zichan Li, Yaxiong Chen, Shengwu Xiong, Peng Xu, Jiajun Zhang, Bowen Zhou, David A. Clifton, and Luc Van Gool. LENS: Multi-level evaluation of multimodal reasoning with large language models. In *International Conference on Learning Representations*, pp. 2627–2654,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/054e9f9a286671ababa3213d6e59c1c2-Paper-Conference.pdf.

Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, and Weijia Li. BLINK-Twice: You see, but do you observe? a reasoning benchmark on visual perception. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.09361.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1811.10830.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations*, 2026. URL https://www.merl.com/publications/docs/TR2026-045.pdf.

A AI USE STATEMENT

Large language models were used during manuscript preparation to assist with literature search and synthesis, manuscript organization and editing, and auditing and verification of experimental results and supporting materials. The underlying experiments and primary analyses were conducted independently of this assistance. The authors reviewed and verified all LLM-assisted computations, citations, interpretations, and final manuscript content.

B REPRODUCIBILITY AND SECONDARY MATERIAL

The accompanying tables retain cohorts, support counts and aggregation rules. Human-relative and leave-one-out margins are recomputed from raw human answers, stored question keys and four native model trial credits per valid question; model credits are checked after answer-order normalization. The participant/margin CSV contains all twelve participant summaries, the matched human/model accuracies and margins for every configuration, the equally weighted full-sample results, and all twelve leave-one-participant-out analyses. The accompanying session map records the author-verified grouping of all twenty files. Bundled code reproduces these records without inferring values from rounded rates. Annotation and S summaries retain frozen CSV calculations; baseline and older diagnostic tables retain source reporting precision.

The supplementary analysis separates source-level descriptions from trial-level evidence. Four independently elicited responses to a task do not provide four independent images; similarly, native answering, selected-fact support, and cue listing do not form a sequence of observed internal processing stages. Reported comparisons use their stated observational units and retain available missingness rather than converting missing outputs to incorrect answers.

We evaluate 15 VLMs from five providers. These are: Gemma-3-4B, Gemma-3-12B, and Gemma-3-27B (Gemma Team, 2025); Llama3.2-11B (Meta AI, 2024) and Llama4-Scout and Llama4-Maverick (Meta AI, 2025); Qwen3-VL-235B (Bai et al., 2025); Claude-Haiku-4.5, Claude-Sonnet-4.5, and Claude-Opus-4.8 (Anthropic, 2025a;b; 2026); Gemini-2.5-Pro, Gemini-3.1-Pro, Gemini-Robotics-ER2, and Gemini-3.7-Flash (Google DeepMind, 2025; 2026a;c;b); and GPT-6-Astra (OpenAI, 2026).

C AUDIT CONSTRUCTION

The AIGEN audit is constructed using the prompt in Listings 1 and 2. Gemini-2.5-Pro is used as the generator. We manually filter the generations to make sure they comply with the generation guidelines. We construct the HUMANGEN entirely by hand, following the same guidelines described in the generation prompt.

```
You are an expert in generating challenging visual reasoning questions for VLMs.
Given an image and a very detailed caption describing it, you must come up with an
interesting visual reasoning question about the image.
Solving this question should require complex, high-level reasoning about the image, such as:
 * figuring out why something is occuring in the image
 * predicting what a specific action in the scene might lead to
 * figuring out what the POV agent should do in this scene
 * speculating about possible objects out-of-frame
 * speculating about some likely attribute of the POV agent
 * speculating about the intent of a person/animal/agent in the scene

 The question should have two possible choices, A and B, one of them correct and one
incorrect.

 Your question MUST fulfill the following requirements:
 * The question must be answerable: When given (question, image, caption), it should be
obvious which of A vs B is correct.
 * The question must require VISUAL reasoning: When given only the question (and no image or
caption), both A and B should sound equally plausible.
 * All visual cues required for solving the question MUST be explicitly mentioned in the
caption. You should only use the image to double-check and ground them.
 * The question should ideally rely on multiple visual cues. For example, a relationship or
comparison between multiple objects, mulitple attributes of an object, multiple scene
attributes, a combination of object attributes and scene attributes, etc.
 * The visual cues involved must NOT be explicitly mentioned in the question or A or B.
```

```
 The image given to you may be unsuitable for generating such questions (it comes from a
source dataset that has MANY unsuitable images!). You should look at both the image AND the
caption, and if EITHER one suggests the image is unsuitable, flag it as unsuitable. It is
better to flag an image as unsuitable than to generate an easy or low-quality question.

If you can generate a good question, please give your output as a JSON dict with the
following keys:
 * "question": the text of the question
 * "options": a dict with keys "A" and "B", mapping to the text of the possible choices
 * "reasoning": an explanation of how to solve the question, mentioning the visual cues that
are used
 * "correct_answer": the correct answer, either "A" or "B"

If the image is unsuitable, or you cannot generate a high-quality question, you should
instead output a JSON dict with the following keys:
 * "unsuitable": an explanation of why the image is unsuitable.

Here are some examples:

<<< IN-CONTEXT EXAMPLE BLOCK -- see Listing 2 >>>


Now please generate a question (or flag as unsuitable) for the following image and caption:

Image:
<IMAGE query>
Caption:
"<CAPTION query>"
```

Listing 1: AIGen question-generation prompt, outer structure. The image and the PixMo-Cap caption for the query are supplied alongside the text shown here; the in-context block is given in Listing 2.

```
Image:
<IMAGE 01>            % PixMo-Cap index 000000000 (suitable)
Caption:
"<CAPTION 01>"
Output:
{
  "question": "Which out-of-frame object is more likely?",
  "options": {
    "A": "a prey animal to the left",
    "B": "a potential mate to the right"
  },
  "reasoning": "The image shows a bird looking right. Both prey and mate would be attractive
to it, but only the mate on the right explains why it is looking right, so the answer is
CORRECT_ANSWER.",
  "correct_answer": "A"
}

Image:
<IMAGE 02>            % PixMo-Cap index 000000001 (suitable)
Caption:
"<CAPTION 02>"
Output:
{
  "question": "Does the person taking the picture have their foot on the ground?",
  "options": {
    "A": "yes",
    "B": "no"
  },
  "reasoning": "The image is from the POV of a motorcyclist, and the background is blurry,
suggesting the motorcyclist is moving, so they would not have their foot on the ground.
Therefore the answer is CORRECT_ANSWER.",
  "correct_answer": "B"
}

Image:
<IMAGE 03>            % PixMo-Cap index 000000041 (suitable)
Caption:
"<CAPTION 03>"
Output:
{
  "question": "Which house is the \"before\" in this \"before-and-after\" picture?",
  "options": {
    "A": "top",
    "B": "bottom"
  },
```

```
  "reasoning": "The house in the bottom is in better condition, the lawn looks healthier, the
 bushes are less overgrown. This suggests that the bottom is \"after\" and the top is \"
before\", therefore the answer is CORRECT_ANSWER.",
  "correct_answer": "A"
}

Image:
<IMAGE 04>            % PixMo-Cap index 000000041 (suitable)
Caption:
"<CAPTION 04>"
Output:
{
  "question": "Why does the image on top have more bushes?",
  "options": {
    "A": "it is a more vegetated part of the country",
    "B": "it is from before they were removed"
  },
  "reasoning": "The two images are views of the same house, therefore they are the same part
of the country, before and after the makeover in which the overgrown bushes were removed. The
 correct answer is CORRECT_ANSWER.",
  "correct_answer": "B"
}

Image:
<IMAGE 05>            % PixMo-Cap index 000000048 (suitable)
Caption:
"<CAPTION 05>"
Output:
{
  "question": "What is the person in this image likely more concerned about?",
  "options": {
    "A": "getting shot",
    "B": "getting mauled"
  },
  "reasoning": "The person is in a forest, where they could get shot by a hunter or mauled by
 an animal. However, they are wearing an orange fluorescent vest, likely to be visible to any
 hunters. They do not have anything to defend themselves from wild animals, suggesting that
mauling is less of a concern. So the correct answer is CORRECT_ANSWER.",
  "correct_answer": "A"
}

Image:
<IMAGE 06>            % PixMo-Cap index 000000147 (suitable)
Caption:
"<CAPTION 06>"
Output:
{
  "question": "Who is more likely to win in a fight?",
  "options": {
    "A": "the person with the arm tattoo",
    "B": "the person with the \"good for health\" jacket"
  },
  "reasoning": "The person with the \"good for health\" jacket is standing and has a gun. The
 person with the arm tattoo is seated and has a sword. Guns are more powerful than swords,
and standing gives more mobility than sitting. Therefore the correct answer is CORRECT_ANSWER
.",
  "correct_answer": "B"
}

Image:
<IMAGE 07>            % PixMo-Cap index 000000148 (suitable)
Caption:
"<CAPTION 07>"
Output:
{
  "question": "As the POV driver, what should I do?",
  "options": {
    "A": "keep driving forward to keep the road clear",
    "B": "stop and wait for instructions"
  },
  "reasoning": "It is important to keep the road clear during an emergency. However, there is
 a firehose across the road, and driving over it might damage it or impede the firefighting
effort. So the correct answer is CORRECT_ANSWER",
  "correct_answer": "B"
}

Image:
<IMAGE 08>            % PixMo-Cap index 000000026 (unsuitable)
Caption:
"<CAPTION 08>"
Output:
```

```
{
  "unsuitable": "There is only one object, the Jojoba oil, and neither it nor the scene have
any semantically meaningful attributes that could lead to an interesting question, making
this an unsuitable image."
}

Image:
<IMAGE 09>            % PixMo-Cap index 000000027 (unsuitable)
Caption:
"<CAPTION 09>"
Output:
{
  "unsuitable": "Almost all of the information in this image is in the form of text, making
it more like a table than an image, and unsuitable for creating visual reasoning tasks."
}

Image:
<IMAGE 10>            % PixMo-Cap index 000000030 (unsuitable)
Caption:
"<CAPTION 10>"
Output:
{
  "unsuitable": "The image is a single object on a white background. That format is
unsuitable for visual reasoning."
}

Image:
<IMAGE 11>            % PixMo-Cap index 000000042 (unsuitable)
Caption:
"<CAPTION 11>"
Output:
{
  "unsuitable": "The image is a like a bag of cliparts pasted onto a white background. They
all follow a theme but do not have any meaningful relationships with each other or attributes
 that lead to interesting questions, so the image is unsuitable."
}

Image:
<IMAGE 12>            % PixMo-Cap index 000000039 (unsuitable)
Caption:
"<CAPTION 12>"
Output:
{
  "unsuitable": "Almost all of the information in this image is in the form of text, making
it more like a table than an image, and unsuitable for creating visual reasoning tasks."
}

Image:
<IMAGE 13>            % PixMo-Cap index 000000008 (unsuitable)
Caption:
"<CAPTION 13>"
Output:
{
  "unsuitable": "The image is a series of charts and text, making it unsuitable for creating
visual reasoning tasks."
}

Image:
<IMAGE 14>            % PixMo-Cap index 000000050 (unsuitable)
Caption:
"<CAPTION 14>"
Output:
{
  "unsuitable": "The image depicts a turtle in some grass and leaves. Although the scene does
 have some attributes (outdoors; daytime; likely autumn), these are not enough to lead to
interesting questions, so the image is unsuitable."
}

Image:
<IMAGE 15>            % PixMo-Cap index 000000059 (unsuitable)
Caption:
"<CAPTION 15>"
Output:
{
  "unsuitable": "The image only has a logo on a burlap texture. It is not possible to
generate interesting questions from this, so the image is unsuitable."
}

Image:
<IMAGE 16>            % PixMo-Cap index 000000064 (unsuitable)
Caption:
```

```
"<CAPTION 16>"
Output:
{
  "unsuitable": "The image is in meme format (picture with text above/below/next to it),
making it unsuitable for our task."
}

Image:
<IMAGE 17>            % PixMo-Cap index 000000073 (unsuitable)
Caption:
"<CAPTION 17>"
Output:
{
  "unsuitable": "The image has some baking molds of different colors, some lemon slices, and
herb sprigs. Although there is color and spatial information, none of it leads to interesting
 visual reasoning questions, making the image unsuitable."
}

Image:
<IMAGE 18>            % PixMo-Cap index 000175757 (unsuitable)
Caption:
"<CAPTION 18>"
Output:
{
  "unsuitable": "The image is a collage with many tiles. Collages are not a suitable format
for visual reasoning."
}

Image:
<IMAGE 19>            % PixMo-Cap index 000241247 (unsuitable)
Caption:
"<CAPTION 19>"
Output:
{
  "unsuitable": "The image is a single object on a white background. That format is
unsuitable for visual reasoning."
}

Image:
<IMAGE 20>            % PixMo-Cap index 000453418 (unsuitable)
Caption:
"<CAPTION 20>"
Output:
{
  "unsuitable": "The image is primarily text, making it unsuitable for visual reasoning."
}

Image:
<IMAGE 21>            % PixMo-Cap index 000000063 (unsuitable)
Caption:
"<CAPTION 21>"
Output:
{
  "unsuitable": "The image is primarily text, making it unsuitable for visual reasoning."
}
```

Listing 2: The in-context block, 21 examples (7 suitable, 14 unsuitable). Each example's image and caption are supplied in place of the placeholders. At query time the examples are randomly shuffled and correct/incorrect option ordering is randomly swapped to avoid positional bias.

D HUMAN EVALUATION PROTOCOL

Human subjects consist of undergraduate and PhD students from multiple institutions, as well as a faculty member and an industry professional. Human subjects are given random subsets from HUMANGEN and AIGEN and are asked to answer questions through the WebUI in Fig. 7.

Tab. 3 and Fig. 8 show statistics regarding coverage. There were a small number of instances in which the same participant answered the same question twice during evaluation. We count these as two individual answers in the main paper, but we show in Tab. 4 that these have negligible impact, and results would be essentially the same no matter how duplicates were treated. Tab. 4 also does a leave-one-participant-out analysis to demonstrate the stability of the human study results.

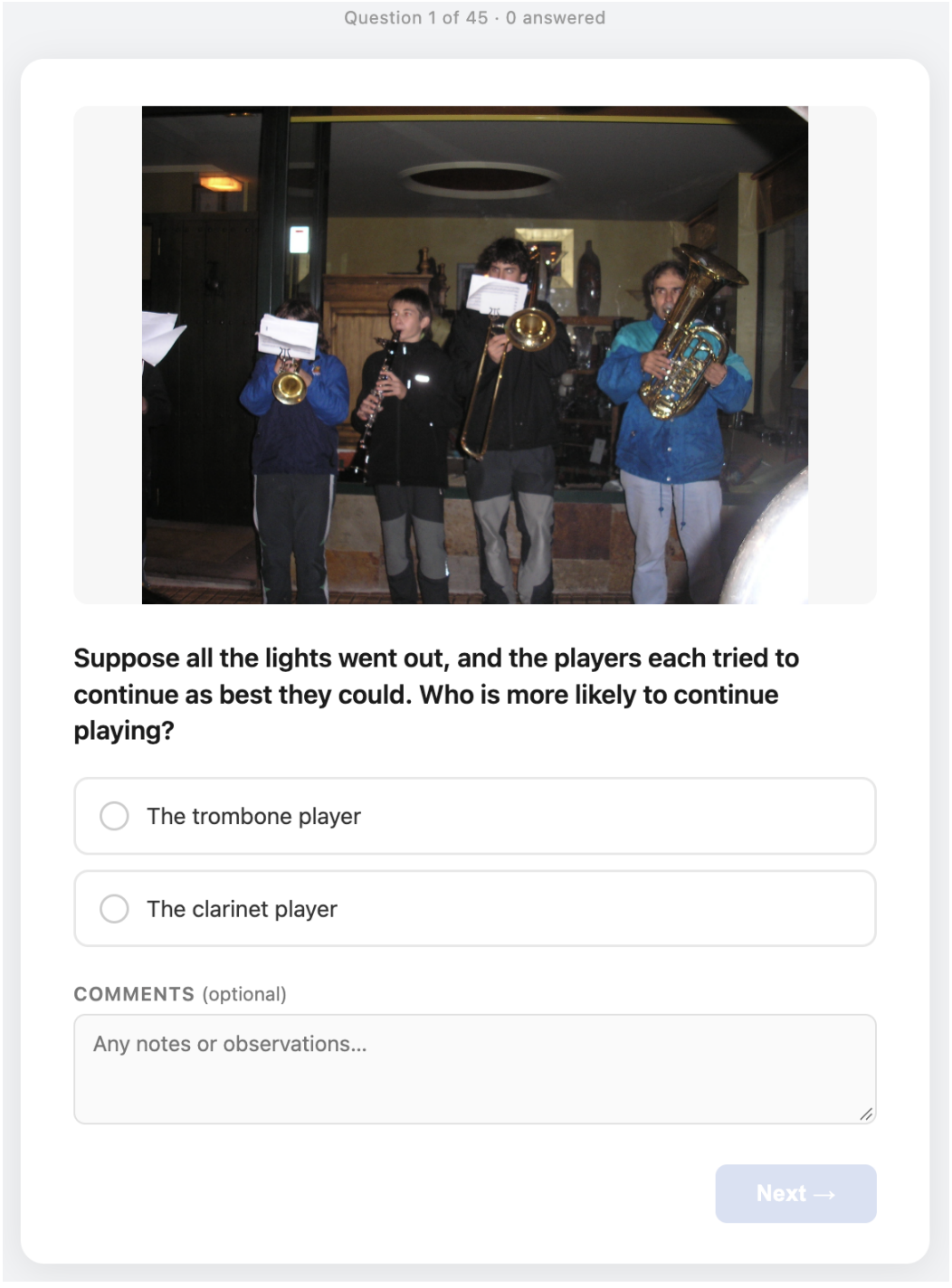


Figure 7: WebUI for human evaluation study.

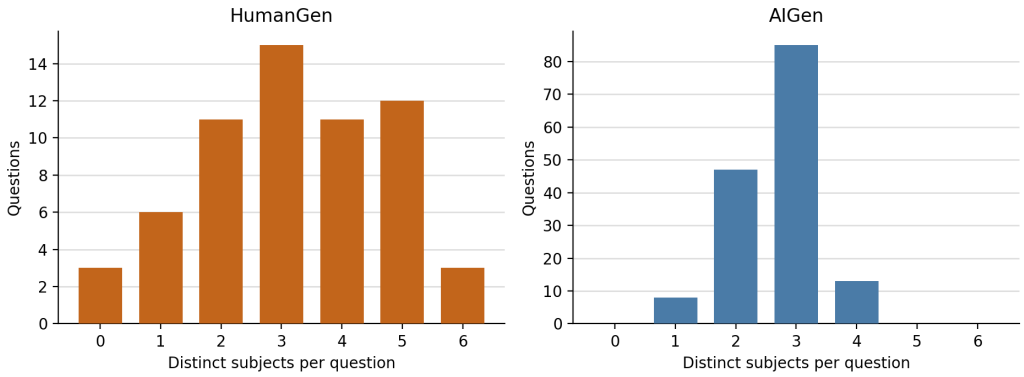


Figure 8: Histogram of questions by participant coverage.

Subject	HumanGen	HumanGen (unique)	AIGen	AIGen (unique)
A	30	25	60	60
B	30	27	60	54
C	20	17	40	40
D	30	27	60	55
E	20	19	40	40
F	10	10	20	20
G	10	10	20	20
H	10	10	20	20
I	10	10	20	20
J	10	10	20	20
K	15	15	30	30
L	15	15	30	30

Table 3: Questions answered by each subject. The plain columns count every answer, so a question answered twice counts twice; the unique columns count distinct questions, collapsing those repeats.

	AIGen	HumanGen
LOO analysis	84.9 [83.9, 85.8]	79.3 [78.0, 81.1]
LOO analysis (collapse duplicates)	84.9 [84.0, 85.8]	79.1 [77.8, 80.8]

Table 4: Human accuracy (%) under senate weighting, with the range over the 12 leave-one-participant-out analyses in brackets. The first row treats a repeated answer to the same question by the same participant as a separate observation; the second averages those repeats so the pair contributes a single correctness value. The two rows differ by at most a few tenths of a point, indicating that repeated answers have negligible impact.

E REASONING-PATH ANNOTATION

A head annotator (an author of this paper) and 3 assistant annotators (PhD students and faculty) each annotated random subsets of AIGEN and HUMANGEN. These were ordered randomly, with the two audits randomly mixed together. The WebUI illustrated in Fig. 9 was used. The head annotator reviewed all assistant annotations and revised for consistency as necessary.

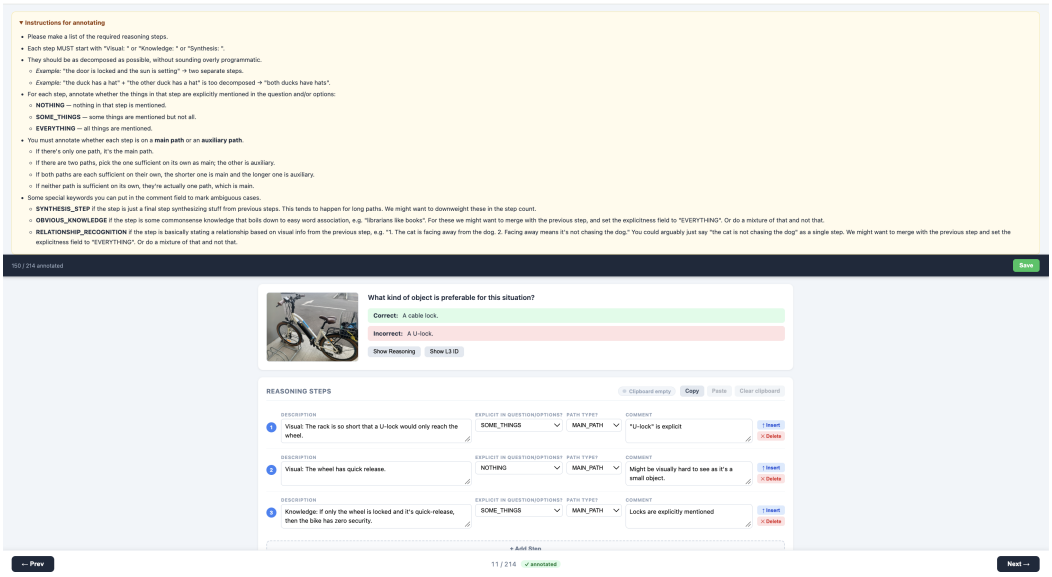


Figure 9: Reasoning-path annotation WebUI.

F STEP-ATTEMPT PROBE

Lst. 3 shows the prompt that we give to each VLM in order to run the step-attempt probe. Lst. 4 shows the prompt that we give to Opus 4.8 as an LLM-judge to determine which of the visual steps in each annotation were attempted in the step-attempt probe.

```
<QUESTION>

A: <OPTION A>
B: <OPTION B>

List all the visual observations that you would need to determine the answer to this question,
 based on this image.
```

Listing 3: Step-attempt probe. The image is supplied alongside this text.

```
You are evaluating whether a set of visual observations attempts each of several required
visual reasoning steps.

You are given two things:
1. A list of required VISUAL REASONING STEPS. Each step describes something that must be
visually determined from an image in order to answer a question.
2. A block of VISUAL OBSERVATIONS: free-text output that another model produced when it was
asked to list the visual observations needed to answer that question (it was told not to
actually answer the question).

For EACH visual reasoning step, decide whether ANY part of the observations mentions or
attempts that step.

The observations do NOT need to get the step correct in order to count as an attempt. They
only need to mention or engage with the same thing the step is about.
- Example: if the step is "the shadows go from left to right", then it COUNTS as an attempt
if the observations say "the shadows go from right to left" (wrong direction), or even "there
 are no shadows" (denies them) -- because in both cases they engage with the shadows.
- However, it does NOT count as an attempt if the observations make no mention of shadows at
all.

REQUIRED VISUAL REASONING STEPS:
1. <VISUAL STEP 1>
2. <VISUAL STEP 2>
...
N. <VISUAL STEP N>

VISUAL OBSERVATIONS (between the markers):
<<<BEGIN OBSERVATIONS>>>
<OBSERVATIONS>
<<<END OBSERVATIONS>>>

Respond with a single JSON object and nothing else. It must have exactly two fields:
- "reasoning": a string briefly explaining, step by step, why each step was or was not
attempted.
- "attempted": a JSON list with exactly N items, one per step in the same order, where each
item is "yes" (the observations mention or attempt that step) or "no" (they do not).

Output only the JSON object.
```

Listing 4: Step-attempt probe LLM-judge prompt.

G COMPONENT PROBES AND ORACLE

Lst. 5 shows the prompt used by Gemini-2.5-Pro to generate component probes. Fig. 10 shows the WebUI that we used to manually review and filter the generated component probe questions for relevance and quality. Lst. 6 shows the prompt given to each VLM to run the component oracle.

```
You are an expert at generating visual perception and reasoning questions for evaluating
vision-language models (VLMs).

Given an image and an existing "Level 3" reasoning question (which requires complex multi-
step inference), your job is to generate simpler
"Level 1" and "Level 2" questions that test the VISUAL UNDERSTANDING that underlies the Level
 3 answer.

DEFINITIONS:
- **Level 1 -- Nodes + Attributes**: Perceptual questions about individual objects and their
attributes.
  Examples: "What is the main object?", "What color is X?", "What shape is Y?", "What is Z
made of?"

LEVEL 1 TAXONOMY -- every Level 1 question must have a "category" field set to exactly one of:


  fine_grained        -- Detailed identification beyond basic recognition (e.g. species,
breed, model, variety).
  ocr                 -- Reading or identifying text that is visible in the image.
  counting            -- How many instances of something are present in the image.
  action              -- What action or activity is being performed by a subject.
  color               -- The color of an object, region, or surface.
  scene_state         -- The overall scene class OR the state/condition of an object (e.g.
cooked vs. raw, open vs. closed).
  object_recognition  -- Identifying what a particular object is.
  localization        -- Providing or referencing the location of an object.
  texture_material    -- The texture pattern or material composition of a surface or object.
  orientation         -- The orientation, facing direction, or pose of an object.
  emotion             -- The emotion expressed by a person or subject in the image.
  other               -- Does not fit into any category.

Choose the single most appropriate category for each Level 1 question. Do NOT invent new
category names.

- **Level 2 -- Relationships**: Questions about relationships between objects in the scene.
  These should cover one or more of:
  - spatial: "Where is X relative to Y?"
  - interactional: "What is the interaction between X and Y?"
  - social/emotional: "How is X being affected by Y?"
  - functional: "What is X being used for in relation to Y?"

REQUIREMENTS for ALL questions:
- Each question must have exactly TWO choices: A and B, one correct and one plausible
distractor.
- The question must be answerable by looking at the image (visual cues must be present).
- The question should NOT be answerable from the question text alone (both options should
sound plausible without seeing the image).
- Do NOT reference the Level 3 question or its answer explicitly.
- Questions should be about concrete, observable visual facts -- NOT speculation or inference.


OUTPUT FORMAT:
Return ONLY a JSON object with two keys: "level1" (list) and "level2" (list).
Each item in the Level 1 list is a question dict with keys:
  - "level": 1
  - "category": one of the Level 1 taxonomy strings listed above (e.g. "object_recognition",
"color", "action", ...)
  - "question": string
  - "options": {"A": "...", "B": "..."}
  - "correct_answer": "A" or "B"

Each item in the Level 2 list is a question dict with keys:
  - "level": 2
  - "relation_type": one of "spatial", "interactional", "social/emotional", "functional"
  - "question": string
  - "options": {"A": "...", "B": "..."}
  - "correct_answer": "A" or "B"

You should generate as many Level 1 and Level 2 questions as you see fit to determine true
understanding of the Level 3 question.
You also don't need to cover all 4 relationship types in the Level 2 questions, only the
relationships that are suitable.
```

```
Similarly, you do not need to cover all Level 1 categories -- only generate questions for
categories that are relevant and observable in the image.

Here are in-context examples showing the Level 3 QA and the corresponding Level 1 and Level 2
 questions:
=== GOOD EXAMPLE 1 (Knife shaped like a whale) ===
IMAGE:
<IMAGE EXAMPLE 1>
Level 3 QA (for context):
{
  "question": "What part of the animal is used to perform the object's main function?",
  "options": {
    "A": "Its head.",
    "B": "Its tail."
  },
  "correct_answer": "A",
  "reasoning": "The object is a knife that is shaped like a whale. The main function of a
knife is to cut, which is done with the sharp edge of the blade. The handle (the non-
functional part) has a wavy shape, suggesting it is the tail and the blade is the head. The
only other relevant visual cue is the hole near the top of the blade resembling an eye, which
 does not conflict its role as the head."
}
Generated LEVEL 1 questions (note the 'category' field for each):
[
  {
    "level": 1,
    "category": "object_recognition",
    "question": "What is the main object in the image?",
    "options": {
      "A": "A knife.",
      "B": "A wooden floor."
    },
    "correct_answer": "A"
  },
  {
    "level": 1,
    "category": "scene_state",
    "question": "What is the shape of the main object?",
    "options": {
      "A": "It is shaped like a ship.",
      "B": "It is shaped like a whale."
    },
    "correct_answer": "B"
  },
  {
    "level": 1,
    "category": "fine_grained",
    "question": "What does the hole in the blade resemble?",
    "options": {
      "A": "A whale's eye.",
      "B": "A spot on a whale's tail."
    },
    "correct_answer": "A"
  }
]
Generated LEVEL 2 questions:
[
  {
    "level": 2,
    "relation_type": "spatial",
    "question": "Where is the handle relative to the blade?",
    "options": {
      "A": "To the right of the blade.",
      "B": "To the left of the blade."
    },
    "correct_answer": "A"
  },
  {
    "level": 2,
    "relation_type": "spatial",
    "question": "Where is the small hole located on the object?",
    "options": {
      "A": "Near the blade.",
      "B": "Near the base of the handle"
    },
    "correct_answer": "A"
  },
  {
    "level": 2,
    "relation_type": "spatial",
    "question": "Where is the hole relative to the handle?",
```

```
    "options": {
      "A": "To the right",
      "B": "To the left."
    },
    "correct_answer": "B"
  }
]
=== GOOD EXAMPLE 2 (Monkey interacting with humans) ===
IMAGE:
<IMAGE EXAMPLE 2>
Level 3 QA (for context):
{
  "question": "Based on the repeated behavior shown across these four scenes, what is a
likely future outcome?",
  "options": {
    "A": "The monkey will learn to avoid humans as a potential threat.",
    "B": "The monkey might will learn to be friendly and calm around humans."
  },
  "correct_answer": "B",
  "reasoning": "The four images (A, B, C, and D) consistently show a monkey interacting
closely with humans, reaching out its hand as if begging for or taking something, likely food.
 This repeated positive reinforcement from humans leads to habituation, where the monkey
loses its natural fear and becomes comfortable around people. "
}
Generated LEVEL 1 questions (note the 'category' field for each):
[
  {
    "level": 1,
    "category": "object_recognition",
    "question": "Who are the main subjects in the image?",
    "options": {
      "A": "A monkey and a human.",
      "B": "Two monkeys."
    },
    "correct_answer": "A"
  },
  {
    "level": 1,
    "category": "counting",
    "question": "How many distinct scenes are shown in the image?",
    "options": {
      "A": "Two.",
      "B": "Four."
    },
    "correct_answer": "B"
  },
  {
    "level": 1,
    "category": "action",
    "question": "What is the monkey doing in the scenes?",
    "options": {
      "A": "Sitting still and observing the human from a distance.",
      "B": "Reaching its hand out toward the human."
    },
    "correct_answer": "B"
  }
]
Generated LEVEL 2 questions:
[
  {
    "level": 2,
    "relation_type": "interactional",
    "question": "What is the interaction between the human and the monkey?",
    "options": {
      "A": "The human is chasing the monkey away.",
      "B": "The monkey is reaching out toward the human, as if begging or taking something."
    },
    "correct_answer": "B"
  },
  {
    "level": 2,
    "relation_type": "social/emotional",
    "question": "How is the monkey being affected by the human's presence?",
    "options": {
      "A": "The monkey appears fearful and is trying to flee.",
      "B": "The monkey appears comfortable and is approaching the human."
    },
    "correct_answer": "B"
  }
]
```

```
---
Now generate Level 1 and Level 2 questions for this new image.

The existing Level 3 QA for this image (use it to understand what visual elements matter, but
 do NOT reproduce it):
<NATIVE QUESTION, OPTIONS, CORRECT ANSWER AND REASONING, AS JSON>

Return ONLY a JSON object formatted exactly like:
{
  "level1": [
    {
      "level": 1,
      "category": "object_recognition",
      "question": "...",
      "options": {"A": "...", "B": "..."},
      "correct_answer": "A"
    },
    ...
  ],
  "level2": [
    {
      "level": 2,
      "relation_type": "spatial",
      "question": "...",
      "options": {"A": "...", "B": "..."},
      "correct_answer": "B"
    },
    ...
  ]
}
<TARGET IMAGE>
```

Listing 5: Component probe generation. One prompt per native question: the model is given two worked examples and then the target image together with the question and its generator reasoning trace, and returns the perceptual probes as JSON.

```
The following questions about the image have been answered:

Question: <PROBE QUESTION 1>
Correct Answer: <PROBE 1 CORRECT ANSWER>
Incorrect Answer: <PROBE 1 INCORRECT ANSWER>

...

Question: <PROBE QUESTION N>
Correct Answer: <PROBE N CORRECT ANSWER>
Incorrect Answer: <PROBE N INCORRECT ANSWER>

Based on the above, answer this question:
<QUESTION>

A: <OPTION A>
B: <OPTION B>

Reply with a JSON object and nothing else -- no markdown, no backticks, no extra text.
The JSON must have exactly two keys: "answer" (the letter A or B) and "reasoning" (a brief
explanation of why you chose that answer).
Example: {"answer": "A", "reasoning": "The image shows ..."}
```

Listing 6: Component oracle, image plus probe answers. The ground-truth answer to every perceptual probe attached to the question is supplied, and the image is attached.

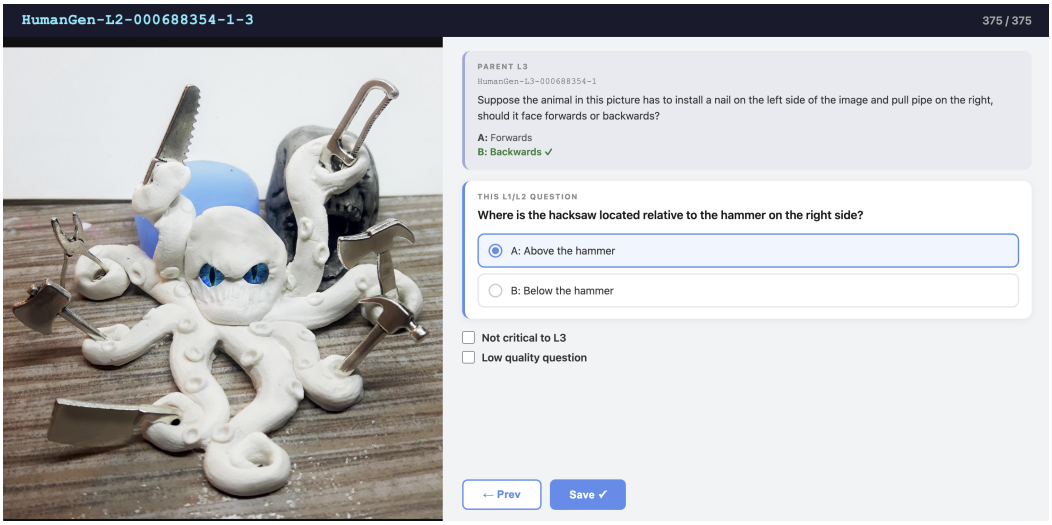


Figure 10: Component probe review WebUI.

H Extended VLM Results on AIGen vs HumanGen

Fig. 11 shows extended results with all 15 VLMs, and the same with questions weighted so that each image has equal contribution to the accuracy.

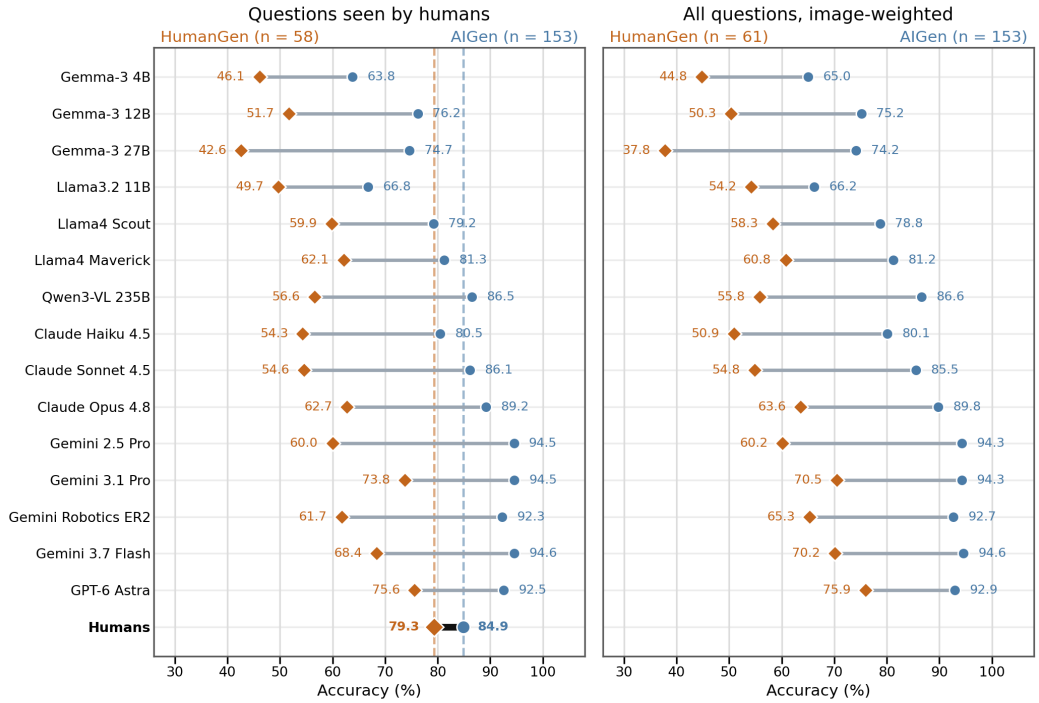


Figure 11: **(left)** Extended version of Fig. 1(c) that includes all 15 VLMs. **(right)** The same, but questions are weighted so that each image contributes equally to accuracy. Results show the same trend.

I EXAMPLES FROM HUMANGEN AND AIGEN

HumanGen example 1



Question: What does the dog most likely want to do?

CORRECT: Go to sleep.
INCORRECT: Keep the cat safe.

Generator reasoning: The cat is lying on a bed that is much bigger than it, suggesting that it is the dog's bed. The dog is standing by the bed and likely wants to use it but cannot because the cat is occupying it.

Reasoning path

MAIN PATH

Visual: The cat is lying on a bed much bigger than it is. Explicitness: SOME THINGS

Knowledge: Beds that are much bigger than cats are usually meant for dogs. Explicitness: SOME THINGS

Synthesis: The dog is most likely standing by the bed because it wants to use it and is being blocked by the cat. Explicitness: SOME THINGS

Perceptual probes

What two types of animals are seen in the image?
CORRECT: A dog and a cat.
INCORRECT: Two different breeds of dog.

Where is the cat located?
CORRECT: Inside a soft pet bed.
INCORRECT: On a patterned rug.

What is the cat doing in the pet bed?
CORRECT: It is lying down with its eyes open.
INCORRECT: It is curled up and sleeping.

Where is the large, light-colored dog looking?
CORRECT: Towards the camera.
INCORRECT: Directly at the cat.

Where is the dog positioned relative to the pet bed?
CORRECT: Sitting on the rug next to the bed.
INCORRECT: Lying down in front of the bed.

How does the size of the cat compare to the pet bed it is occupying?
CORRECT: The cat is small and takes up only a portion of the bed.
INCORRECT: The cat is large and fills the entire bed.

What is the physical interaction between the dog and the cat?
CORRECT: They are near each other but not touching.
INCORRECT: They are playing with each other.

HumanGen example 2



Question: What kind of object is preferable for this situation?

CORRECT: A cable lock.
INCORRECT: A U-lock.

Generator reasoning: A U-lock is stronger than a cable lock. However, in this case, the rack is too short for a U-lock to protect anything other than the wheel, which is held on by a quick-release that could easily be opened by hand. A cable lock would at least require the bike thief to have tools, so it is preferable.

Reasoning path

MAIN PATH

Visual: The rack is so short that a U-lock would only reach the wheel. Explicitness: SOME THINGS

Visual: The wheel has quick release. Explicitness: NOTHING

Knowledge: If only the wheel is locked and it's quick-release, then the bike has zero security. Explicitness: SOME THINGS

Perceptual probes

What is the metal object on the ground in the bottom left corner?
CORRECT: A bike rack.
INCORRECT: A shopping cart corral.

What is a key physical characteristic of the bike rack?
CORRECT: It is short, only reaching the lower part of the wheel.
INCORRECT: It is tall, extending above the bike's frame.

What type of mechanism is visible at the center of the front wheel, attaching it to the fork?
CORRECT: A quick-release lever.
INCORRECT: A set of large hex bolts.

Which part of the bicycle is currently positioned inside the bike rack?
CORRECT: The front wheel.
INCORRECT: The rear wheel.

Given the low height of the rack, which part of the bike could be most directly secured to it with a small, rigid lock?
CORRECT: The front wheel.
INCORRECT: The main frame of the bike.

HumanGen example 3



Question: Suppose I got on the boat and rowed, what would most likely happen?

CORRECT: It wouldn't move at all.
INCORRECT: It would move forward a couple feet.

Generator reasoning: The front of the boat is at least a few feet from the water's edge, but there is a bird wading in the water, suggesting that the water is too shallow for the boat to float.

Reasoning path

MAIN PATH

Visual: The bird is wading in the water. Explicitness: NOTHING

Knowledge: This suggets the water is very shallow. Explicitness: NOTHING

Knowledge: Boats cannot be rowed through very shallow water. Explicitness: SOME THINGS

Perceptual probes

What is the main red object in the center of the image?
CORRECT: A boat
INCORRECT: A bathtub

What kind of animal is standing in the water?
CORRECT: A bird
INCORRECT: A penguin

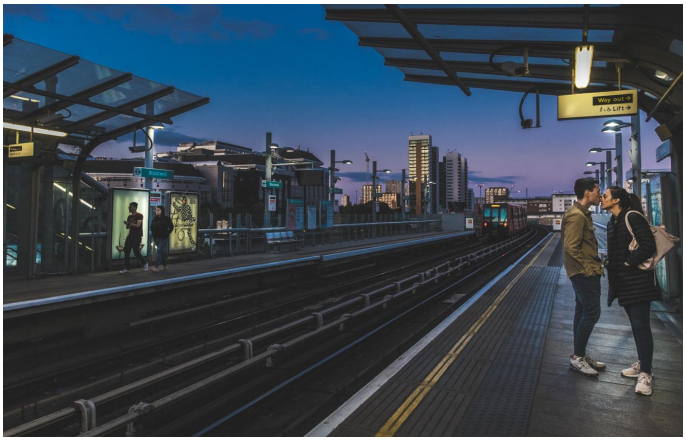
What is the nature of the body of water shown?
CORRECT: It is a small, round puddle.
INCORRECT: It is the edge of a large lake.

What is the bird doing in the water?
CORRECT: It is standing or wading.
INCORRECT: It is swimming or floating.

What is the position of the boat relative to the puddle and the grass?
CORRECT: The boat is entirely floating in the puddle.
INCORRECT: The boat is partially resting on the grass and partially in the puddle.

How is the bird positioned within the water?
CORRECT: It is standing on the bottom of the shallow puddle.
INCORRECT: It is floating on the surface of deep water.

HumanGen example 4



Question: Suppose one of the kissing people playfully shoved the other one away. Who would likely create a less dangerous situation by doing that?

CORRECT: The man
INCORRECT: The woman

Generator reasoning: The man has his back to the train tracks, so if the woman shoved him he might fall in. The woman likely has her back to a railing, so it is less dangerous for the man to shove her.

Reasoning path

MAIN PATH

Visual: kissing couple is man-left woman-right
Explicitness: SOME THINGS

Visual: There is drop-off on left of couple only.
Explicitness: SOME THINGS

Knowledge: Drop-offs are more dangerous than non-dropoffs.
Explicitness: SOME THINGS

Synthesis: man-shove goes from left to right, pushing away from the greater danger. Explicitness: SOME THINGS

Perceptual probes

What is the primary setting of this image?
CORRECT: A train station platform
INCORRECT: An airport runway

What are the man and woman on the right side of the platform doing?
CORRECT: Kissing
INCORRECT: Arguing

Which direction is the man's back facing?
CORRECT: Towards the train tracks
INCORRECT: Towards the city buildings

Which direction is the woman's back facing?
CORRECT: Away from the train tracks
INCORRECT: Towards the train tracks

What is located directly behind the man who is kissing the woman?
CORRECT: The train tracks
INCORRECT: The rest of the platform and a wall

How is the woman positioned relative to the man and the train tracks?
CORRECT: The man is standing between her and the train tracks.
INCORRECT: She is standing between the man and the train tracks.

HumanGen example 5



Question: Suppose it is late afternoon, which is likely more dangerous?

CORRECT: A very strong wind from the East.
INCORRECT: A very strong wind from the West.

Generator reasoning: The shadows are pointing right, suggesting left is West and right is East. A wind blowing from right to left presents more danger as it could blow the people off of the cliff, so a wind from the East is more dangerous.

Reasoning path

MAIN PATH

Knowledge: Sun is West in afternoon, casting shadows East. Explicitness: SOME THINGS

Visual: Shadows are left to right. Explicitness: NOTHING

Visual: People have cliff on left only. Explicitness: NOTHING

Knowledge: Cliffs are more dangerous than non-cliffs. Explicitness: SOME THINGS

Synthesis: Wind from East goes right to left, blowing towards danger. Explicitness: SOME THINGS

Perceptual probes

What is the primary setting shown in the image?
CORRECT: A winding road in a mountainous area.
INCORRECT: A straight highway through a flat plain.

In which general direction are the shadows from the people and cars pointing?
CORRECT: Towards the right side of the image.
INCORRECT: Towards the left side of the image.

Where is the group of people in the foreground standing in relation to the edge of the overlook?
CORRECT: They are standing close to the unguarded edge.
INCORRECT: They are standing in the middle of the paved area, far from the edge.

What is located on the other side of the guardrail along the main road?
CORRECT: A steep drop-off down the mountainside.
INCORRECT: A flat, grassy shoulder.

The shadows cast by the people fall onto what surface?
CORRECT: The asphalt of the overlook they are standing on.
INCORRECT: The side of the mountain across the valley.

HumanGen example 6



Question: Who is more likely to shoot at the goal in this image?

CORRECT: The player farthest left in the image.
INCORRECT: The player immediately left of the one with number 34 on their jersey.

Generator reasoning: The teams appear to be differentiated by jersey color (red vs black). The player with 34 is wearing red and appears to be a goalie who is near their goal. The player immediately left of 34 is also wearing red and therefore on the team defending the goal. The player farthest left is wearing black and therefore on the opposing team, which is trying to score on the goal.

Reasoning path

MAIN PATH

Visual: The player farthest left is wearing black.
Explicitness: SOME THINGS

Visual: The player with number 34 and the player immediately left of them are both wearing red. Explicitness: SOME THINGS

Visual: The player with number 34 appears to be the goalie for the goal that is shown. Explicitness: SOME THINGS

Knowledge: Players can generally be grouped into teams by jersey color.
Explicitness: SOME THINGS

Synthesis: Farthest left player is opposite team of goalie, while player next to 34 is same team, so farthest left player is more likely to shoot.
Explicitness: SOME THINGS

Perceptual probes

What sport is being played in this image?
CORRECT: Ice hockey
INCORRECT: Field hockey

What are the two main colors of the jerseys worn by the players?
CORRECT: Red and black
INCORRECT: Blue and white

What number is on the back of the goalie's red jersey?
CORRECT: 34
INCORRECT: 31

Where is the player wearing number 34 positioned?
CORRECT: Directly in front of the goal net.
INCORRECT: Near the center of the ice rink.

Based on the large pads and specific helmet, what is the role of the player wearing number 34?
CORRECT: Goalie
INCORRECT: Forward

Based on their jersey colors, what is the relationship between the player with number 34 and the player farthest to the left?
CORRECT: They are on opposing teams.
INCORRECT: They are on the same team.

What is the relationship between the player wearing number 34 and the other player in a red jersey to their left?
CORRECT: They are on the same team.
INCORRECT: They are on opposing teams.

Where is the goal net in relation to the goalie wearing number 34?
CORRECT: To the player's right.
INCORRECT: To the player's left.

How are the players in red jerseys positioned relative to the goal?
CORRECT: They are positioned between the player in black and the goal.
INCORRECT: They are positioned behind the goal.

HumanGen example 7



Question: Suppose I am standing in the arena and the wind picks up, which way should I face to avoid getting sand in my eyes?

CORRECT: Away from the sun
INCORRECT: Towards the sun

Generator reasoning: The flags and the shadows are both pointing to the left, indicating that the sand would come from the same direction as the sun. Therefore I should face away from the sun.

Reasoning path

MAIN PATH

Visual: Flags are to the left of their poles. Explicitness: NOTHING

Knowledge: That means wind is blowing from the right. Explicitness: SOME THINGS

Visual: Shadows are right to left. Explicitness: NOTHING

Knowledge: That means the sun is on the right. Explicitness: SOME THINGS

Knowledge: I want to be away from the wind to protect my eyes. Explicitness: SOME THINGS

Synthesis: Wind blowing from the left, which is opposite the sun, therefore I want to be away from the sun to be away from the wind. Explicitness: SOME THINGS

Perceptual probes

What are the red objects attached to the tops of the poles in the background?
CORRECT: Flags
INCORRECT: Umbrellas

In which general direction are the red flags blowing?
CORRECT: Towards the left of the image.
INCORRECT: Towards the right of the image.

Where are the shadows of the people in the arena cast?
CORRECT: To the left of the people.
INCORRECT: Directly underneath the people.

What material covers the ground in the central arena?
CORRECT: Sand
INCORRECT: Grass

How does the direction the flags are blowing compare to the direction the shadows are cast?
CORRECT: They are pointing in the same general direction.
INCORRECT: They are pointing in opposite directions.

Based on the shadows, from which direction is the sunlight originating?
CORRECT: From the right side of the image.
INCORRECT: From the left side of the image.

What is the relationship between the wind's source (inferred from flags) and the sun's position (inferred from shadows)?
CORRECT: The wind is blowing from the same direction that the sun is in.
INCORRECT: The wind is blowing from the opposite direction of the sun.

HumanGen example 8



Question: Suppose all the lights went out, and the players each tried to continue as best they could. What would most likely happen to the ratio of woodwind to brass sound?

CORRECT: It would increase
INCORRECT: It would decrease

Generator reasoning: The trumpet and trombone player have sheet music, while the clarinet and euphonium player do not. This implies that the clarinet and euphonium players are the only ones who memorized their parts and are likely the only ones who could keep consistently playing in the dark. This would result in one woodwind against one brass instead of one woodwind against three brasses, so the ratio of woodwind to brass sound would increase.

Reasoning path

MAIN PATH

Visual: The trumpet and trombone players have sheet music in front of them. Explicitness: SOME THINGS

Visual: The clarinet and euphonium players do not have sheet music in front of them. Explicitness: SOME THINGS

Knowledge: People with a music sheet likely have not memorized the music and need to see their sheet. Explicitness: NOTHING

Knowledge: People without a music sheet likely have memorized the music and do not need to see anything. Explicitness: NOTHING

Knowledge: It is hard to see things in the dark. Explicitness: SOME THINGS

Knowledge: Trumpet, trombone, and euphonium are brass, while clarinet are woodwind. Explicitness: NOTHING

Synthesis: Before the lights go out, we have 1 woodwind and 3 brass. After the lights go out, only the clarinet and euphonium can play from memory, which is 1 woodwind and 1 brass, which is a greater ratio of woodwind to brass. Explicitness: SOME THINGS

Perceptual probes

How many people are visibly playing musical instruments in the foreground?
CORRECT: Four
INCORRECT: Three

What is the black instrument being played by the person second from the left?
CORRECT: A clarinet
INCORRECT: A flute

What is the large, shiny brass instrument being held by the person on the far right?
CORRECT: A tuba/euphonium
INCORRECT: A French horn

What is the instrument played by the person in the center whose face is mostly obscured by a piece of paper?
CORRECT: A trombone
INCORRECT: A saxophone

What appears to be attached to the instruments of the leftmost and center players?
CORRECT: Sheet music on a holder
INCORRECT: Decorative white flags

Which player is holding their instrument without any visible sheet music attached?
CORRECT: The clarinet player second from the left.
INCORRECT: The trumpet player on the far left.

HumanGen example 9



Question: What pose is the taker of this photo most likely in?

CORRECT: Lying on their side.
INCORRECT: Sitting upright.

Generator reasoning: The pose of the cat relative to the legs indicates that it's A. The pose of the couch relative to the legs is another possible reasoning path.

Reasoning path

MAIN PATH

Visual: The legs are stacked parallel to the up-down direction of the couch in the background. Explicitness: NOTHING

Knowledge: This means the POV person must be on their side. Explicitness: SOME THINGS

AUXILIARY PATH

Visual: The legs are stacked parallel to the up-down direction of the cat. Explicitness: NOTHING

Knowledge: This means the POV person must be on their side. Explicitness: SOME THINGS

Perceptual probes

What is the cat doing?
CORRECT: Sleeping.
INCORRECT: Stretching.

Where is the cat located?
CORRECT: On a person's lap.
INCORRECT: On the arm of a couch.

What is the cat resting on top of, which is itself on the person's lap?
CORRECT: A soft, gray blanket.
INCORRECT: A fluffy pillow.

How is the cat's body positioned relative to the person's legs?
CORRECT: It is lying across their legs.
INCORRECT: It is curled up between their legs.

AIGen example 1



Question: Where is the photographer most likely located?

CORRECT: In a large indoor space, like a workshop or studio.
INCORRECT: On a nearby mountain peak, observing the scene from afar.

Generator reasoning: The image and caption describe a man who is "disproportionately larger than the landscape" and is interacting with what is explicitly called a "detailed and intricately crafted model of a snow-covered countryside." His actions, such as pouring "fake snow" from a container and holding a paintbrush, along with the presence of a miniature building and satellite, confirm that the entire landscape is a small-scale diorama. Since the scene is a model, it must be located indoors. Therefore, the photographer must also be in that same indoor space, such as a studio or workshop, to take the picture.

Reasoning path

MAIN PATH

Visual: The man is larger than the landscape. Explicitness: SOME THINGS

Knowledge: Thus, the scene is likely a small-scale diorama
Explicitness: SOME THINGS

Knowledge: Its likely that a diorama would be located indoors in a studio.
Explicitness: SOME THINGS

Perceptual probes

Based on the relative scale of the man and the environment, what is being depicted?
CORRECT: A miniature model of a mountain landscape.
INCORRECT: A real, full-sized mountain landscape.

What is the man holding in his hands?
CORRECT: A bucket and a large brush.
INCORRECT: A backpack and a walking stick.

What is the man doing in the scene?
CORRECT: He is sprinkling a substance onto the landscape.
INCORRECT: He is digging in the snow.

How does the size of the man compare to the mountains and trees around him?
CORRECT: He is disproportionately large compared to the mountains and trees.
INCORRECT: He is very small compared to the mountains and trees.

What is the interaction between the objects the man is holding and the environment?
CORRECT: He is using the brush to apply a substance from the bucket onto the model mountain.
INCORRECT: He is using the bucket to collect rock samples from the mountain with a brush.

Where are the small buildings located relative to the man?
CORRECT: To the left and far behind him.
INCORRECT: Directly in front of him at his feet.

AIGen example 2



Question: What is the most likely reason for the foreground fox's expression and posture?

CORRECT: It is investigating an unfamiliar object with curiosity.
INCORRECT: It is preparing to confront the other fox.

Generator reasoning: The foreground fox's nose is extremely close to the camera lens, and its large, wide eyes are looking directly into it. This indicates its full attention is on the camera. The other fox is in the background, out of focus, and the foreground fox shows no signs of directing its attention or aggression towards it (e.g., turning its head, snarling). The ears are pointed backward, which, combined with the intense focus on the camera, suggests a mixture of curiosity and caution towards an unknown object, rather than aggression towards the other animal.

Reasoning path

MAIN PATH

Visual: The foreground fox is facing away from the foreground fox.
Explicitness: EVERYTHING

Knowledge: Facing away is not consistent with preparing to investigate.
Explicitness: EVERYTHING

AUXILIARY PATH

Visual: The fox has an ears pointed backwards and intensely focused eyes.
Explicitness: EVERYTHING

Knowledge: This expression is consistent with curiosity about unfamiliar object Explicitness: EVERYTHING

Perceptual probes

How many foxes are visible in the image?
CORRECT: Two
INCORRECT: One

In which direction is the fox in the foreground looking?
CORRECT: Directly at the camera.
INCORRECT: Towards the other fox in the background.

How do the eyes of the foreground fox appear?
CORRECT: Wide open and round.
INCORRECT: Narrowed and squinting.

How are the ears of the fox in the foreground positioned?
CORRECT: They are angled slightly backward.
INCORRECT: They are pointed straight forward.

Where is the main fox located in relation to the camera?
CORRECT: It is extremely close to the camera lens.
INCORRECT: It is several yards away from the camera.

Where is the second fox located relative to the fox in the foreground?
CORRECT: In the background, behind the foreground fox.
INCORRECT: In the foreground, next to the main fox.

Based on their body language, what is the interaction between the two foxes?
CORRECT: The foreground fox appears to be ignoring the background fox.
INCORRECT: They are preparing to fight each other.

AIGen example 3



Question: What is the most probable next action for the person whose meal is in the foreground?

CORRECT: They will use the contents of the small dark container.
INCORRECT: They will use the contents of the shakers next to the water bottle.

Generator reasoning: The meal in the foreground is a large stack of sweet pancakes, topped with banana slices, powdered sugar, and chocolate syrup. On the table, there is a small dark container with additional syrup, and there are also salt and pepper shakers. Given that pancakes are a sweet dish, it is highly probable that the person will add more syrup. It is very improbable that they would add salt or pepper. Therefore, the most probable next action is to use the contents of the small dark container.

Reasoning path

MAIN PATH

Visual: The foreground meal is pancake with a syrup on a plate.
Explicitness: SOME THINGS

Visual: The shakers next to the water bottle have salt and pepper.
Explicitness: SOME THINGS

Visual: The small dark container contains syrup.
Explicitness: SOME THINGS

Knowledge: Pancakes are sweet and eaten using syrup. Shakers contain salt. So, they will use contents on small dark container.
Explicitness: SOME THINGS

Perceptual probes

What is the main food item on the plate in the foreground?
CORRECT: A stack of pancakes.
INCORRECT: A pile of waffles.

What is in the small white pitcher on the plate?
CORRECT: Syrup.
INCORRECT: Cream.

What are the two shakers on the far left of the table?
CORRECT: Salt and pepper shakers.
INCORRECT: Sugar and cinnamon shakers.

What kind of fruit is sliced on top of the pancakes?
CORRECT: Bananas.
INCORRECT: Apples.

What toppings are drizzled and sprinkled over the pancakes?
CORRECT: Chocolate syrup and powdered sugar.
INCORRECT: Caramel sauce and nuts.

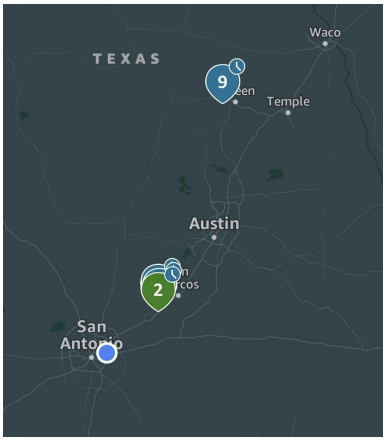
Where are the salt and pepper shakers located in relation to the water bottle?
CORRECT: Next to the water bottle.
INCORRECT: Behind the water bottle.

Where is the small pitcher of syrup located?
CORRECT: On the same plate as the pancakes.
INCORRECT: On the plate in the background.

What is the purpose of the syrup in relation to the pancakes?
CORRECT: It is a condiment to be poured on the pancakes.
INCORRECT: It is an ingredient to be mixed into the water.

Which dish is in the foreground, closer to the camera?
CORRECT: The dish with the stack of pancakes.
INCORRECT: The dish with the fried egg on top.

AIGen example 4



Question: What is the user's likely plan for the immediate future?

CORRECT: To travel north from their current position.
INCORRECT: To travel further south from their current position.

Generator reasoning: The caption identifies the blue and white icon over San Antonio as the user's "current location". The other location pins, such as the green one marked '2' and the blue one marked '9', are all positioned north of San Antonio. This arrangement strongly suggests a travel route heading north from the current location. Therefore, the user's immediate plan is to travel north.

Reasoning path

MAIN PATH

Visual: The picture is of a map with pinned numbers on it and a blue dot.
Explicitness: NOTHING

Knowledge: Numbers in map normally means the user's planned stops in iteration. Explicitness: NOTHING

Knowledge: Blue dot on map means user's current position.
Explicitness: NOTHING

Visual: There is two numbers, 2 and 9, above the blue dot in the image
Explicitness: NOTHING

Knowledge: Top of the map generally means north.
Explicitness: SOME THINGS

Synthesis: As pinned stops are top from user's current position, user is travelling north. Explicitness: SOME THINGS

Perceptual probes

What does this image primarily display?
CORRECT: A digital map with a planned route.
INCORRECT: A weather radar showing storm systems.

What is the name of the city located near the bottom-left of the map?
CORRECT: San Antonio
INCORRECT: Austin

What number is displayed inside the green location pin?
CORRECT: 2
INCORRECT: 5

What number is displayed inside the blue location pin?
CORRECT: 9
INCORRECT: 6

What type of icon is marking the location of San Antonio?
CORRECT: A plain blue and white circle.
INCORRECT: A green pin with a number in it.

How many numbered location pins are shown on the map?
CORRECT: Two
INCORRECT: Three

Where is the city of Austin located relative to San Antonio?
CORRECT: To the north.
INCORRECT: To the south.

Where is the green pin marked '2' positioned relative to the blue and white icon over San Antonio?
CORRECT: North of it.
INCORRECT: South of it.

How is the blue pin marked '9' positioned relative to the green pin marked '2'?
CORRECT: Further north.
INCORRECT: Further south.

Which city is located between San Antonio and Temple on the map?
CORRECT: Austin
INCORRECT: Waco

AIGen example 5



Question: What would a supervising adult likely be most concerned about in this situation?

CORRECT: The child getting his clothes dirty.
INCORRECT: The child running out of paint.

Generator reasoning: The caption notes that the boy is wearing a "pristine white polo shirt". However, it also describes him as being deeply absorbed in a messy painting process, with his "left arm is almost entirely covered in the green, textured paint". This strong contrast between his clean clothes and his messy arms makes it highly probable that his shirt will get paint on it soon, which would be a primary concern for a supervisor. While running out of paint is a possibility in any art project, the caption mentions two jars of purple paint, and they appear to have a good amount of paint left, making this a less immediate concern.

Reasoning path

MAIN PATH

Visual: Paint is not in running-out state. Explicitness: EVERYTHING

AUXILIARY PATH

Visual: Paint is near clothes. Explicitness: EVERYTHING

Visual: Clothes are currently clean, therefore dirtyable. Explicitness: EVERYTHING

Perceptual probes

What is the color of the polo shirt the boy is wearing?
CORRECT: White
INCORRECT: Green

What is the condition of the boy's left arm?
CORRECT: It is covered in green paint.
INCORRECT: It is mostly clean.

What is the condition of the boy's white shirt?
CORRECT: It appears to be clean.
INCORRECT: It has large paint splatters on it.

What activity is the boy engaged in?
CORRECT: Painting on a large canvas.
INCORRECT: Playing with modeling clay.

How many open jars of purple paint are visible in the foreground?
CORRECT: Two
INCORRECT: One

What is the boy holding in his right hand?
CORRECT: A paintbrush.
INCORRECT: A small toy.

Where is the boy's paint-covered arm in relation to his clean shirt?
CORRECT: It is positioned right next to his torso and sleeve.
INCORRECT: It is held far away from his body.

How is the boy interacting with the green paint on the canvas?
CORRECT: He is applying it with his bare hands and arm.
INCORRECT: He is carefully applying it with a fine-tipped brush.

Where is the large painting surface located relative to the boy?
CORRECT: It is on the ground in front of him.
INCORRECT: It is propped up on an easel.

AIGen example 6



Question: How many total clocks are most likely on the building?

CORRECT: Three or four.
INCORRECT: Two or three.

Generator reasoning: The image shows two full clock faces on two adjacent sides of the tower. The caption also mentions that "A third clock face can be partially seen around the corner". This strongly suggests the tower is designed to have a clock on each of its four sides to show the time in all directions. Therefore, it is highly probable that the unseen fourth side also has a clock face.

Reasoning path

MAIN PATH

Visual: There are two clocks clearly showing, and a third partially seen on the right Explicitness: SOME THINGS

Perceptual probes

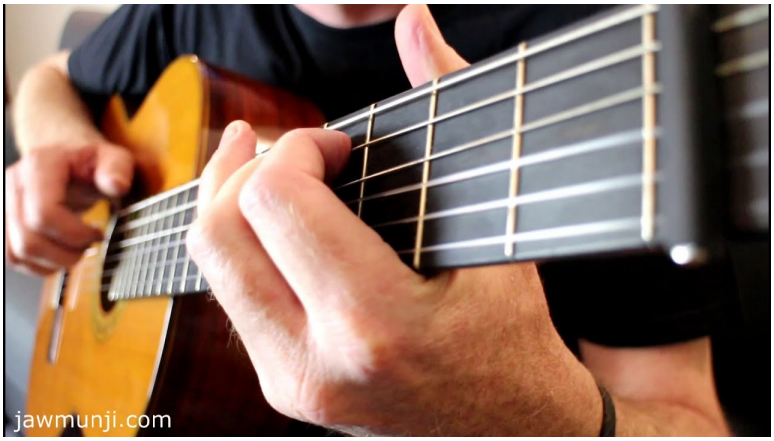
What is the primary structure shown in the image?
CORRECT: A clock tower.
INCORRECT: A lighthouse.

How many clock faces are entirely visible in the image?
CORRECT: Two.
INCORRECT: Three.

Looking to the far left of the image, what can be partially seen on the tower?
CORRECT: The edge of a third clock face.
INCORRECT: A decorative window.

How are the two fully visible clocks positioned relative to each other on the tower?
CORRECT: On two different, adjacent sides.
INCORRECT: On the same side, one above the other.

AIGen example 7



Question: What is a likely characteristic of the person in the image?

CORRECT: He is right-handed.
INCORRECT: He is left-handed.

Generator reasoning: The image and caption describe a person playing a guitar in the standard way for a right-handed individual. The caption states that his 'left hand forming a chord on the guitar's neck and his right hand strumming the strings near the base.' This division of labor, where the non-dominant hand handles the complex task of fretting and the dominant hand handles the rhythm (strumming), is the conventional method for right-handed players. Therefore, it is most likely that the man is right-handed.

Reasoning path

MAIN PATH

Visual: The person is strumming with their right hand and/or fretting with their left hand (attempting to locate either hand counts as an attempt at this step). Explicitness: SOME THINGS

Knowledge: Guitarists usually strum with their dominant hand and fret with their non-dominant hand. Explicitness: SOME THINGS

Perceptual probes

What musical instrument is being played in the image?
CORRECT: A guitar
INCORRECT: A cello

What is the hand in the foreground doing on the guitar's neck?
CORRECT: Pressing down on the strings against the frets.
INCORRECT: Strumming the strings to produce sound.

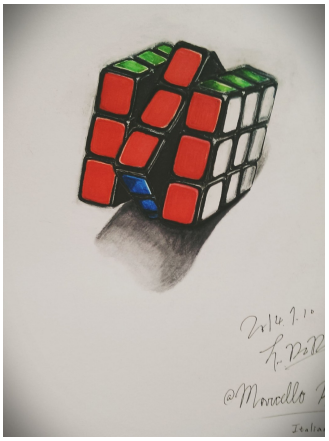
What is the hand in the background doing near the guitar's body?
CORRECT: Strumming the strings.
INCORRECT: Adjusting the tuning pegs.

Where is the person's hand that is closer to the camera located on the guitar?
CORRECT: On the neck.
INCORRECT: Over the sound hole.

How do the person's two hands work together to play the instrument?
CORRECT: One hand presses strings on the neck to form chords, while the other hand strikes the strings.
INCORRECT: Both hands are used to hold the neck of the guitar firmly in place.

What is the function of the hand on the guitar's neck in relation to the strings?
CORRECT: To change the pitch of the strings by pressing them against frets.
INCORRECT: To create the rhythm by striking the strings.

AIGen example 8



Question: If the last action performed on the object was reversed, what would be the most likely visual change?

CORRECT: The blue pieces would no longer be visible.
INCORRECT: The green pieces would no longer be visible.

Generator reasoning: The caption explains that the cube is in its current state because of a 'twisted orientation of the middle row'. This twist has a specific consequence mentioned in the caption: it 'causes part of the blue bottom face to become visible', revealing 'two blue squares'. The question asks what would happen if this action were reversed. Reversing the action that made the blue squares visible would, logically, hide them again. The caption also mentions that the twist is 'obscuring the top green row'. Reversing this twist would therefore make the obscured green pieces visible again, not make all green pieces disappear. Therefore, the most direct outcome of reversing the move is that the blue pieces would be hidden from view.

Reasoning path

MAIN PATH

Visual: The center column has been rolled back. Explicitness: NOTHING

Visual: The blue pieces at the bottom are showing. Explicitness: SOME THINGS

Knowledge: Rolling forward would hide the bottom pieces. Explicitness: SOME THINGS

Perceptual probes

What object is depicted in the drawing?
CORRECT: A Rubik's cube
INCORRECT: A set of building blocks

What is the color of the squares on the top face of the cube?
CORRECT: Green
INCORRECT: Blue

What is the color of the two squares revealed by the twist, located on the bottom part of the cube?
CORRECT: Blue
INCORRECT: Yellow

What is the state of the cube's middle horizontal layer?
CORRECT: It is twisted out of alignment with the other layers.
INCORRECT: It is perfectly aligned with the other layers.

On which face are the green squares primarily located?
CORRECT: The top face.
INCORRECT: The front face.

What is the spatial relationship between the twisted middle row and the green squares on the top face?
CORRECT: The twist of the middle row partially obscures some of the green squares.
INCORRECT: The twist of the middle row makes more green squares visible.

What is the visual effect of the middle row being twisted in this specific orientation?
CORRECT: It reveals squares from the bottom face.
INCORRECT: It completes the pattern on the red face.

Where are the visible blue squares located relative to the twisted middle row?
CORRECT: They are directly below it.
INCORRECT: They are directly above it.

AIGen example 9



Question: Which of these two dogs has more likely just finished a long run?

CORRECT: The one on the left.
INCORRECT: The one on the right.

Generator reasoning: The dog on the left, the Foxhound, has its tongue out, which is a sign of panting to cool down after physical exertion like running. The dog on the right, the Beagle, has its mouth closed. The caption also notes the Foxhound is 'taller and leaner, with longer legs', which are physical traits suited for running, further supporting this conclusion.

Reasoning path

MAIN PATH

Visual: The dog on the left is panting. Explicitness: SOME THINGS

Visual: The dog on the right is not panting. Explicitness: SOME THINGS

Knowledge: Dogs pant when they've exercised. Explicitness: SOME THINGS

Perceptual probes

What is the state of the mouth of the dog in the left panel?
CORRECT: It is open with its tongue hanging out.
INCORRECT: It is closed.

What is the state of the mouth of the dog in the right panel?
CORRECT: It is closed.
INCORRECT: It is open and appears to be barking.

How does the physical build of the dog on the left compare to the dog on the right?
CORRECT: It appears taller and leaner.
INCORRECT: It appears shorter and stockier.

Where is the dog with its tongue out located relative to the dog with the closed mouth?
CORRECT: To the left.
INCORRECT: To the right.