VRIQ: Benchmarking and Analyzing Visual-Reasoning IQ of VLMs

Anonymous ACL submission

Abstract

Do vision–language models (VLMs) fail at visual reasoning because they cannot reason, or because they fail to perceive the relevant visual structure? We introduce **VRIQ**, a benchmark of 1,500 IQ-style visual reasoning puzzles spanning five psychometric categories across abstract and natural-image domains. To disentangle perception from reasoning, we additionally introduce **DiagVRIQ**, a 209-item diagnostic subset paired with human-authored descriptions and perception-oriented annotations. Across open-weight and proprietary VLMs, abstract-domain performance remains near random while natural-domain accuracy improves substantially but stays below human level. Our describe-then-solve analysis on DiagVRIQ decouples the two stages and shows that, in several categories, reasoning failures persist even when perception is fully correct, indicating that aggregate accuracy conceals genuinely distinct failure modes. We further test two interventions targeting each stage: (i) augmenting the prompt with ground-truth textual descriptions, and (ii) tool-augmented reasoning that lets the model crop, zoom, and re-examine while thinking over the image. Both yield clear improvements, with tool-augmented thinking producing the largest gains on spatially demanding categories. Together, VRIQ and DiagVRIQ provide a framework for studying visual reasoning as the interaction of separable perception and reasoning stages rather than a single monolithic capability.

1 Introduction

Recent years have witnessed remarkable advances in both Large Language Models (LLMs) and Vision–Language Models (VLMs), driven by new architectures (Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023), improved training methodologies (Chung et al., 2024; Ouyang et al., 2022), and large-scale datasets (Schuhmann et al., 2022; Gadre et al., 2023). Reasoning has become a cornerstone of LLM capability: chain-of-thought prompting (Wei et al., 2022), self-consistency (Wang et al., 2023b), and tree-of-thoughts (Yao et al., 2023) all show that systematic step-by-step reasoning substantially improves performance on complex tasks.

Extending these benefits to the visual domain has proven harder. Visual reasoning, the process of analyzing, interpreting, and drawing logical conclusions from visual content, is increasingly important as VLMs are deployed in high-stakes settings such as medical diagnosis (Li et al., 2023; Phan et al., 2024) and autonomous navigation (Pan et al., 2024; Chen et al., 2024). Recent work has proposed visual chain-of-thought (Shao et al., 2024), tool-augmented reasoning (Gupta and Kembhavi, 2023; Surís et al., 2023), and reasoning-optimized models such as o3 (OpenAI, 2025e). Yet evaluations continue to reveal persistent failures on tasks requiring abstract thinking and multi-step inference (Lu et al., 2024a; Yu et al., 2023).

A central difficulty in evaluating visual reasoning is that existing benchmarks typically treat perception and reasoning as a single monolithic capability (Cai et al., 2025). However, visual reasoning fundamentally depends on two distinct stages: (i) *perception*, the extraction of relevant visual structure from the image, and (ii) *reasoning*, the application of abstract rules or relational operations over the perceived information. Failures in either stage may lead to incorrect answers, yet current evaluations rarely distinguish between them.

Several benchmarks have been proposed to study visual reasoning (Li et al., 2024a; Jiang et al., 2024; Zhang et al., 2025a), but important limitations remain. Some focus primarily on natural-image reasoning with relatively shallow inference requirements (Fu et al., 2024; Ging et al., 2024), while others evaluate abstract IQ-style reasoning without connecting to real-world visual understanding (Zhang et al., 2019; Barrett et al., 2018). More

importantly, existing benchmarks generally evaluate only end-to-end accuracy, making it difficult to determine whether failures arise from incorrect perception or downstream reasoning errors.

To address this gap, we introduce **VRIQ**, a benchmark of 1,500 expert-authored IQ-style visual reasoning puzzles spanning five psychometric categories across both abstract and natural-image domains. Each category is instantiated in paired abstract and natural variants sharing identical reasoning structure, enabling controlled comparison of model behavior across visual representations.

Disentangling perception from reasoning is itself nontrivial, particularly for abstract puzzles whose visual structure resists faithful textual description. We address this through a multi-perspective diagnostic framework. For abstract puzzles, we construct perception probes targeting individual visual primitives (e.g., object identity, count, orientation, spatial position), providing an indirect measure of perceptual access. For natural-domain categories whose rules admit faithful textual description, we construct **DiagVRIQ**, a 209-item diagnostic subset paired with human-authored ground-truth descriptions, on which we apply a describe-then-solve protocol that decomposes each item into perception and reasoning stages.

Building on this attribution framework, we further evaluate two interventions designed to improve performance at each stage: (i) ground-truth perceptual augmentation, which supplies verified textual descriptions alongside the image, and (ii) tool-augmented multimodal reasoning, which enables models to crop, zoom, and re-examine images during inference.

Our experiments show that perception and reasoning act as distinct failure modes whose relative importance depends on the category. In several settings, models continue to fail even when perception is correct, demonstrating persistent reasoning limitations beyond visual-extraction errors. Both interventions yield improvements consistent with this attribution, with tool-augmented reasoning producing the largest gains.

In summary, our contributions are:

1. **VRIQ and DiagVRIQ**: a 1,500-item benchmark spanning paired abstract and natural-image reasoning tasks, together with a 209-item diagnostic subset containing human-authored descriptions enabling item-level perception–reasoning attribution.
2. A multi-perspective diagnostic framework combining end-to-end evaluation, perception probes, and a describe-then-solve protocol for disentangling perceptual and reasoning failures.
3. Two targeted interventions, ground-truth perceptual augmentation and tool-augmented reasoning, together with extensive experiments across open-weight and proprietary VLMs.

We will release VRIQ, DiagVRIQ, all perception probes, ground-truth descriptions, and the methodology for constructing diagnostic probes.

2 Related Work

Multimodal benchmarks. As multimodal foundation models improve on traditional vision tasks, evaluating their reasoning abilities has become increasingly important. To achieve this, many general-purpose Visual Question Answering (VQA) tasks (Mathew et al., 2021; Gurari et al., 2018; Antol et al., 2015; Goyal et al., 2017; Singh et al., 2019; Hudson and Manning, 2019; Marino et al., 2019; Li et al., 2024b; Liu et al., 2023; Liang et al., 2024; Saikh et al., 2022; Yue et al., 2024; Wang et al., 2023a; Chen et al., 2023) have been proposed. To support more targeted evaluation, recent work has introduced specialized multimodal benchmarks: (Lu et al., 2024b; Zhang et al., 2024; Qiao et al., 2024; Wang et al., 2024b; Lu et al., 2021; Li et al., 2025b; Wang et al., 2024a; Doris et al., 2024). Our benchmark differs from prior work by focusing specifically on IQ-style puzzle solving and by providing diagnostic tools for attributing failures to perception or reasoning.

Solving multimodal IQ problems. Efforts have been made to evaluate VLMs on solving puzzle or IQ problems. Zhang et al. (2025c) focus on evaluating simple puzzle solving without symbolic or IQ questions. Li et al. (2025a) introduces more complex puzzle problem. MMIQ (Cai et al., 2025) is the closest to ours, which also evaluates multimodal performance on IQ problems; however, unlike MMIQ, our benchmark introduces both abstract and natural puzzle families in addition to a hierarchical probing framework, enabling fine-grained attribution of failures to different perceptual versus reasoning sources.

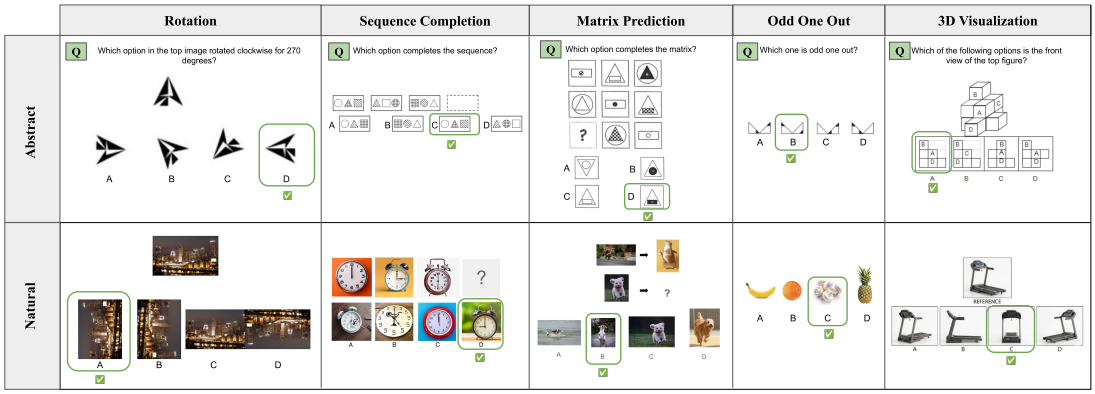


Figure 1: VRIQ overview. Five reasoning categories are instantiated in abstract and natural domains, enabling controlled comparison of visual representation effects.

3 VRIQ: A Diagnostic Visual Reasoning Benchmark

VRIQ (Visual Reasoning IQ) is a benchmark of **1,500 expert-authored visual reasoning puzzles** designed to diagnose perception and reasoning failures in vision–language models. The benchmark spans five psychometric reasoning categories: Sequence Completion (SC), Matrix Prediction (MP), Odd One Out (OO), Figure Rotation (FR), and 3D Visualization (3D). Each category is instantiated in two matched domains: an *abstract* domain built from geometric primitives and a *natural* domain built from real-world objects and scenes. Every category contains 200 abstract and 100 natural items.

3.1 Benchmark Construction

Abstract-domain puzzles. Abstract puzzles consist of geometric shapes, patterns, and spatial transformations with carefully designed distractors. Items were sourced from two channels: (i) adapted public-domain aptitude and civil-service examinations, and (ii) puzzles authored from scratch by the research team. Adapted items were reformulated into a unified VRIQ format and, where appropriate, transformed into novel instances by modifying superficial visual properties (e.g., shape identity, count, orientation, or styling) while preserving the underlying reasoning rule. For the 3D Visualization category, geometric configurations and viewpoint transformations were generated using Tinkercad (Autodesk Inc., 2025). Additional details are provided in Appendix C.

Natural-domain puzzles. Natural puzzles preserve the same category-specific reasoning structure while replacing geometric primitives with real-world entities and scenes. Items were created through a semi-agentic human–AI pipeline: annotators first authored seed examples, GPT-5.2-Thinking (OpenAI, 2025d) proposed candidate task instances, and human reviewers selected, revised, and validated the final examples. Publicly available copyright-compliant images were then assembled into the final puzzle layouts. Full generation details appear in Appendix A.

Quality control and contamination mitigation. All puzzles were independently solved by two annotators to verify answer correctness and uniqueness. Natural items were additionally reviewed for rule consistency, visual clarity, and absence of shortcut cues. Because some abstract puzzles originate from public examination materials, residual training-data contamination cannot be fully ruled out. We mitigate this in three ways: (i) all natural-domain items are author-created and have no public-web counterparts; (ii) adapted abstract items were systematically modified into novel instances (Appendix C); and (iii) approximately 40% of abstract items were manually authored from scratch.

3.2 Diagnostic Annotations

VRIQ provides two complementary diagnostic annotation layers designed for different visual regimes. These annotations are used as evidence in the diagnostic framework introduced in section 4 and in the results analysis in section 6.

Perception probes (P-probes). For abstract puzzles, whose visual structure cannot be faithfully described with text, we annotate each item with atomic probes targeting individual visual primitives independently of the puzzle-solving task. Each

probe is a multiple-choice question about a single perceptual attribute, including object identity, count, orientation, position, color, or 3D structure. For example, a probe may ask about the number of objects in a panel or the orientation of an arrow. Because probes are evaluated independently of the original puzzle question, probe accuracy reflects perceptual access rather than reasoning ability. Probe accuracy is measured over the whole population and results are in section 6 and Appendix H.

DiagVRIQ: a subset for perception–reasoning attribution. Three natural-domain categories, Matrix Prediction (MP), Odd One Out (OO), and Sequence Completion (SC), admit faithful textual descriptions of their visual content. For these categories, we construct DiagVRIQ, a 209-item diagnostic subset (98 MP, 51 OO, and 60 SC) paired with ground-truth descriptions. Descriptions were generated using GPT-5.2 (OpenAI, 2025d) and verified by human annotators to ensure complete coverage of panels, answer options, and relevant visual relationships. To avoid leakage, descriptions are restricted to observable visual content and neither reveal the underlying rule nor indicate the correct answer. Each description was authored and independently reviewed against the source puzzle to ensure completeness and rule neutrality (Appendix B). These annotations support both the describe-then-solve attribution protocol (subsection 4.2) and the ground-truth augmentation experiments (subsection 4.3).

4 Evaluation Framework

VLM failures on visual reasoning may arise from perception, reasoning, or interactions between them, and these stages are tightly coupled in IQ-style puzzles. Rather than relying on a single signal to attribute failures, we use three complementary diagnostic signals and treat converging evidence across them as supporting stronger claims than any single signal in isolation. Additional perceptual evidence on abstract items is drawn from the P-probe annotations introduced in subsection 3.2 and analyzed in section 6.

4.1 End-to-End Accuracy

We measure multiple-choice accuracy across all five reasoning categories in both abstract and natural domains. Because the two domains share logical structure within each category, this enables direct cross-domain comparisons within each category.

4.2 Description-Based Perception–Reasoning Attribution

For the DiagVRIQ subset (subsection 3.2), we adopt a describe-then-solve setup in which the model produces, within a single response, (i) a natural-language description of the puzzle as if explaining it to a non-sighted reader, followed by (ii) the answer to the original multiple-choice question. The description and answer are extracted from the same generation via the prompt template in Appendix E.

We evaluate each model-generated description against the ground-truth description using the rubric in Appendix F, with GPT-5.2 (OpenAI, 2025d) acting as the judge. The judge assigns a binary label indicating whether the description correctly captures the entities, panel structure, and answer options visible in the image. This label evaluates whether the model correctly *perceived* the visual content, not whether the description is sufficient to solve the puzzle, so missing entities, hallucinations, or structural misunderstandings count as perception failures regardless of the final answer.

Combining the perception label with answer correctness enables us to separate two distinct failure modes: (i) cases where the model fails to correctly perceive the visual content, and (ii) cases where the model perceives the content correctly but still fails to solve the puzzle. From this decomposition, we derive two aggregate metrics:

- **Perception accuracy**: the fraction of items whose entities, panel structure, and answer options are correctly identified, independent of the final answer.

- **Reasoning accuracy**: the fraction of correctly perceived items that are also answered correctly. This measures answer accuracy conditioned on successful perception.

To validate the perception labels, one author independently annotated all model descriptions using the same criterion. Author–judge agreement reached **92.7%** overall ($\kappa = 0.84$), comparable to strong LLM-as-judge results in prior work (Kim et al., 2024; Lee et al., 2024); per-model statistics are in Appendix G.

Scope. We emphasize that this protocol is correlational rather than causal. The describe-then-solve setup may omit information that the model uses internally, and textual serialization can alter the effective task representation. We treat the perception/reasoning labels as a diagnostic signal and rely on the intervention-based test in subsection 4.3 for independent corroboration.

4.3 Reasoning Behavior with Ground-Truth Perception

To test whether accurate perceptual content alone is sufficient to unlock reasoning, we re-run DiagVRIQ with the ground-truth description appended to the prompt. We compare two conditions:

- **Baseline**: image + multiple-choice question.
- **GT-augmented**: image + question + ground-truth description, with an instruction to use both modalities.

The descriptions contain only visual information and do not reveal the underlying rule or correct answer (subsection 3.2). The exact prompt is provided in Appendix K. If the subsection 4.2 attribution is meaningful, GT augmentation should preferentially improve categories whose dominant failure mode is perceptual rather than relational.

5 Experimental Setup

We evaluate 17 vision–language models spanning four capability tiers: small open-source (3B–9B parameters), mid-size open-source (32B–34B), proprietary frontier (GPT-4o, GPT-5.1, Gemini-2.5-Pro, Claude Sonnet 4.6), and tool-augmented (OpenAI o3 in "thinking with images" mode (OpenAI, 2025f)). The set covers instruction-tuned and quantized open-source models, reasoning-optimized variants, and frontier API-served models. Full model list and evaluation settings appear in Appendix D.

Tool-augmented evaluation (o3). We evaluate `o3` using the OpenAI Responses API with the `code_interpreter` tool enabled. This tool allows the model to crop, zoom, rotate, and otherwise process images, as well as execute Python code during inference (OpenAI, 2025a). We use this setting to test whether tool-augmented inference improves VRIQ performance, particularly on categories with strong spatial or perceptual demands. Each item is evaluated once.

Human baseline. We evaluate human performance on a stratified subset of 300 questions (30 per category $\times$ domain cell). We recruit 20 participants (ages 18–55, mixed academic backgrounds); each solves 30 questions, and each question is independently answered by two annotators. Participants face no time limit and are instructed to prioritize accuracy. We report mean accuracy with 95% confidence intervals.

6 Results

We evaluate VRIQ across both abstract and natural domains using the models described in section 5. Results are organized around the diagnostic framework of section 4, progressing from end-to-end accuracy to perception–reasoning attribution and intervention-based analyses.

6.1 End-to-End Performance

Abstract visual reasoning remains challenging for current VLMs. Across all evaluated models, performance on abstract IQ-style reasoning remains only marginally above the 25% random baseline (Table 1). The strongest model, `Qwen3-VL-32B-Thinking`, achieves **35.5%** average accuracy, while the mean performance across all VLMs is only **26.1%**. Notably, proprietary frontier models do not consistently outperform strong open-weight alternatives: `GPT-4o` reaches **30.3%**, only modestly exceeding several open-source systems. In contrast, human participants achieve **90.4%** accuracy, indicating that the tasks are solvable and the human–VLM gap remains substantial.

Natural variants improve performance but do not resolve the task. Grounding the same reasoning structures in natural images yields consistent improvements. Sequence Completion improves from **25.9%** to **42.8%**, and Odd One Out from **25.2%** to **43.9%**. Despite these gains, average natural-domain accuracy (**40.6%**) still falls far short of the **98.2%** human level. Natural-image priors reduce difficulty but do not appear sufficient to produce robust visual reasoning.

6.2 Description-Based Perception and Reasoning Attribution

To disentangle perceptual failures from downstream reasoning errors, we evaluate models on DiagVRIQ using the describe-then-solve protocol introduced in subsection 4.2. Table 3 decomposes

Table 1: Accuracy (%) of different vision–language models (VLMs) on the VRIQ benchmark across five categories: SC = Sequence Completion, 3D = 3D Reasoning, FR = Figure Rotation, OO = Odd One Out, and MP = Matrix Prediction.

Model	**SC (%)**	**3D (%)**	**FR (%)**	**OO (%)**	**MP (%)**	**Avg (%)**
Random Guess	*25*	*25*	*25*	*25*	*25*	*25*
Open Source VLMs						
Qwen2.5-VL-3B-Instruct	20.5	28	23	25.5	17	22.8
Qwen2.5-VL-3B-Instruct-AWQ	20.5	30	17	20	19.5	21.4
Qwen2.5-VL-7B-Instruct	40.5	33	25	26.5	31	31.2
Qwen3-VL-32B-Thinking	44.5	34	31.5	33.5	34	35.5
InternVL3-9B	24	32	20	16.5	31	24.7
llava-v1.6-mistral-7b-hf	16	28.5	20	27	20	22.3
llava-v1.6-vicuna-7b-hf	23.5	24	16	15	18.5	19.4
llava-OV-1.5-8B	14.0	32.0	22.0	25.0	41.5	26.9
Bee-8B-RL	31	22	30	25	41.5	29.9
llava-v1.6-34b-hf	25	23	20	20	26	22.8
Proprietary VLMs						
gpt-5.1	28	27	31	33.5	27	29.3
gpt-4o	28.5	31	32	30.5	29.5	30.3
gemini-2.5-pro	27	29.5	31.5	27.5	28	28.7
gpt-5-mini	24.5	22	24	26	11.5	21.6
gpt-4o-mini	23.5	25	25	28.5	23.5	25.1
claude-sonnet-4-6	24	22	30	22.5	32.5	26.2
Avg %	*25.94*	*27.69*	*24.88*	*25.16*	*27.00*	*26.13*
Human Baseline %	*93.43*	*85.82*	*92.54*	*90.92*	*89.09*	*90.36*

performance into perception accuracy and reasoning accuracy conditioned on successful perception.

Perception accuracy. Perception accuracy increases steeply in every category: from **6.1%** (LLaVA) to **70.4%** (Qwen) to **94.9%** (GPT-4o) on Matrix Prediction; from **33.3%** to **84.3%** to **94.1%** on Odd One Out; and from **3.3%** to **30.0%** to **86.7%** on Sequence Completion. End-to-end accuracy follows the same trend. Across all three categories, the smaller model (LLaVA) produces descriptions judged as perceptually correct in fewer than one-third of items, and on Matrix Prediction and Sequence Completion in fewer than 7%.

Conditional answer accuracy. Conditioning on successful perception substantially increases answer accuracy for all models. GPT-4o correctly solves **91.4%** of perception-correct Matrix Prediction items, **89.6%** of Odd One Out items, and **90.4%** of Sequence Completion items. Qwen also exhibits substantially higher conditional accuracies, ranging from **60.5%** to **73.9%**. However, even frontier models fail on a non-trivial subset of perception-correct examples, demonstrating that reasoning limitations remain visible after perceptual errors are removed.

Probe-level accuracy. For abstract puzzles, whose visual structure resists faithful textual description, we measure perception indirectly through atomic P-probes (full details and per-probe results in Appendix H). Failure rates concentrate on counting (73.8%), 3D/depth (67.5%), position (58.2%), and rotation/orientation (52.4%), while color-related probes fail at only 31.4%. These patterns indicate that VLMs are substantially more reliable on appearance-level perception than on structured spatial perception, and align with the category-level difficulty pattern: 3D Visualization and Figure Rotation, the categories with the lowest end-to-end accuracy, also exhibit the highest probe failure rates.

We note that conditional answer accuracy is computed on the perception-correct subset, which varies in size across cells; cells with very small subsets are reported for completeness but are not load-bearing for our interpretation. Notably, expressing the end-to-end accuracy as a sum of conditional probabilities, conditioned on the outcome of the perception test, validates the analysis, see Appendix I. As further validation, note that Author/LLM-judge agreement on the binary perception label is high (**92.7%** with Cohen's $\kappa = 0.84$; Appendix G). Last, representative qualitative examples are provided in Appendix J.

Table 2: Accuracy on natural reasoning categories.

Model	SC (%)	OO (%)	FR (%)	3D (%)	MP (%)	Avg (%)
Random Guess	*25*	*25*	*25*	*25*	*25*	*25*
Open Source VLMs						
Qwen2.5-VL-3B-Instruct	34	41	40	30	43	37.6
Qwen2.5-VL-3B-Instruct-AWQ	37	36	35	25	29	32.4
Qwen2.5-VL-7B-Instruct	25	50	35	30	40	36.0
Qwen3-VL-32B-Thinking	42	54	49	41	50	47.2
InternVL3-9B	40	40	45	40	35	40.0
llava-v1.6-mistral-7b-hf	20	30	18	27	25	24.0
llava-v1.6-vicuna-7b-hf	24	34	15	19	15	21.4
llava-OV-1.5-8B	41	29	45	37	52	40.8
Bee-8B-RL	33	25	30	50	46	36.8
llava-v1.6-34b-hf	40	44	20	27	43	34.8
Proprietary VLMs						
gpt-5.1	59	60	40	41	59	51.8
gpt-4o	64	57	40	48	62	54.2
gemini-2.5-pro	61	50	33	59	65	53.6
gpt-5-mini	49	48	40	28	51	43.2
gpt-4o-mini	44	48	40	39	48	43.8
claude-sonnet-4-6	72	56	26	44	64	52.4
Avg %	*42.81*	*43.88*	*34.44*	*36.56*	*45.44*	*40.62*
Human Baseline %	*98.77*	*98.83*	*98.84*	*97.22*	*97.35*	*98.2*

Table 3: Per-category accuracy decomposition on DiagVRIQ. Perception accuracy denotes successful visual grounding, while reasoning accuracy measures answer correctness conditioned on successful perception.

Model	Category	End-to-end Acc	Perception Acc	Reasoning Acc
llava-v1.6-mistral-7b-hf	MP (N=98)	21/98 (21.4%)	6/98 (6.1%)	2/6 (33.3%)
	OO (N=51)	20/51 (39.2%)	17/51 (33.3%)	10/17 (58.8%)
	SC (N=60)	15/60 (25.0%)	2/60 (3.3%)	1/2 (50.0%)
Qwen2.5-VL-7B-Instruct	MP (N=98)	58/98 (59.2%)	69/98 (70.4%)	51/69 (73.9%)
	OO (N=51)	27/51 (52.9%)	43/51 (84.3%)	26/43 (60.5%)
	SC (N=60)	23/60 (38.3%)	18/60 (30.0%)	13/18 (72.2%)
gpt-4o	MP (N=98)	86/98 (87.8%)	93/98 (94.9%)	85/93 (91.4%)
	OO (N=51)	43/51 (84.3%)	48/51 (94.1%)	43/48 (89.6%)
	SC (N=60)	48/60 (80.0%)	52/60 (86.7%)	47/52 (90.4%)

6.3 Reasoning Improvement Under Ground-Truth Perception

Ground-truth augmentation provides a direct test of the attribution hypothesis. If perceptual failures contribute substantially to model errors, supplying verified perceptual information should improve downstream performance. We therefore compare baseline performance (image + question) with GT-augmented performance (image + question + human-authored description) on the same DiagVRIQ items for LLaVA-1.6, Qwen2.5-VL-7B, and GPT-4o (Figure 2).

All evaluated models improve under GT augmentation, and the magnitude of improvement broadly follows baseline perception accuracy (Table 3). LLaVA exhibits the largest gains, improving by up to **+41.7** points on Sequence Completion and by more than **9** points across all categories. Qwen improves by **+10.2**, **+4.0**, and **+28.4** points on Matrix Prediction, Odd One Out, and Sequence Completion, respectively, while GPT-4o shows smaller but consistent gains (**+2.0**, **+5.9**, and **+3.3**). Models and categories with lower baseline perception accuracy generally benefit more from access to ground-truth descriptions. Nevertheless, residual failures persist even under GT augmentation, indicating that correct perceptual access alone is insufficient to fully solve VRIQ.

We additionally explore whether this intervention generalizes beyond natural-image reasoning. Preliminary experiments on abstract-domain Sequence Completion suggest that augmentation can also improve performance when the visual structure can be faithfully expressed in text without revealing the underlying rule (Appendix L).

Table 4: Performance of `gpt-o3` with and without tool augmentation (horizontal layout).

Category	Natural Domain			Abstract Domain		
	With	**Without**	**Δ**	**With**	**Without**	**Δ**
3D Visualization	100.0	48.5	+51.5	35.0	24.0	+11.0
Figure Rotation	92.0	40.0	+52.0	68.0	42.0	+26.0
Matrix Prediction	75.0	62.0	+13.0	47.0	35.5	+11.5
Odd One Out	70.0	59.0	+11.0	52.5	31.5	+21.0
Sequence Completion	80.0	68.5	+11.5	48.5	37.0	+11.5
Overall	83.4	55.6	+27.8	50.2	34.0	+16.2

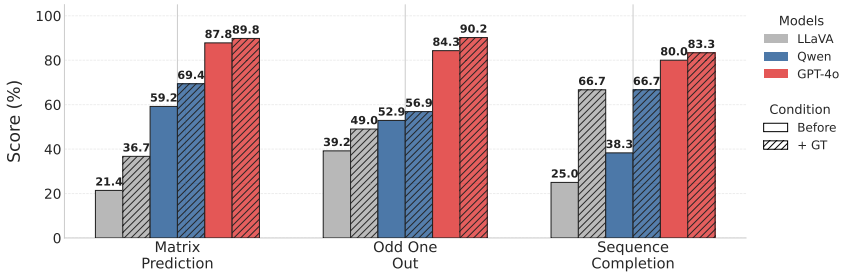


Figure 2: Accuracy before and after augmenting the prompt with ground-truth descriptions, by category and model. GT augmentation produces positive gains across every model–category cell.

6.4 Tool-Augmented Inference

We next investigate whether interactive visual tools can mitigate the perceptual difficulties identified above. Specifically, we compare `gpt-o3` with and without access to image cropping, zooming, rotation, and lightweight Python execution during inference.

Tool augmentation produces large gains in both domains (Table 4): average accuracy rises from **55.6%** to **83.4%** in the natural domain (+27.8) and from **34.0%** to **50.2%** in the abstract domain (+16.2).

The largest improvements occur in categories with strong spatial-perceptual demands. In the natural domain, Figure Rotation improves by **+52.0** points and 3D Visualization by **+51.5** points. Similar gains are observed in the abstract domain, where Figure Rotation improves by **+26.0** points. These are also the categories exhibiting the highest perception-probe failure rates, consistent with the hypothesis that tool use primarily mitigates perceptual difficulties. By contrast, improvements for Odd One Out and Sequence Completion remain comparatively modest (approximately **+11** points across domains), suggesting that additional visual access is less beneficial when perception is already relatively reliable.

7 Discussion and Conclusion

Discussion. Across all analyses, three findings emerge. First, abstract visual reasoning remains challenging for current VLMs despite strong performance on many contemporary multimodal benchmarks. Second, perceptual grounding explains a substantial portion of performance variation, particularly on tasks requiring structured spatial understanding. Third, improving perceptual access through either ground-truth descriptions or interactive visual tools yields substantial gains but does not eliminate errors, indicating that both perception and reasoning contribute to current limitations.

Conclusion. We introduced VRIQ, a benchmark of 1,500 IQ-style visual reasoning puzzles spanning matched abstract and natural domains, together with DiagVRIQ, a diagnostic subset for perception-reasoning attribution. By combining description-based attribution, ground-truth augmentation, and tool-assisted inference, we show that visual reasoning failures arise from both perceptual and reasoning limitations whose importance varies across categories and model capabilities. We hope VRIQ provides a useful framework for moving beyond aggregate accuracy toward diagnosing why vision-language models fail on visual reasoning tasks.

Limitations

The describe-then-solve protocol is a diagnostic approximation rather than a complete decomposition of perception and reasoning, since converting images into text may alter task difficulty or omit information used by the model; accordingly, we restrict it to natural-domain categories where faithful textual descriptions are feasible. Perception labels additionally depend on a judging procedure whose reliability is evaluated through calibration and agreement analyses (Appendix G). Finally, VRIQ focuses on five categories of nonverbal psychometric reasoning and does not cover broader forms of visual reasoning such as temporal, causal, or embodied reasoning.

Ethics Statement

This work introduces VRIQ and DiagVRIQ, a benchmark for evaluating visual reasoning in vision–language models. All puzzle images are either authored from scratch by the research team or sourced from publicly available, copyright-compliant materials; the benchmark and accompanying annotations will be released under CC BY-NC 4.0 for non-commercial research use, and we discourage their use as training data so as to preserve diagnostic value for future evaluations.

The human-baseline study involved adult volunteers (ages 18–55) who participated on a voluntary, unpaid basis after being informed of the task and the use of their responses. No personally identifying information was collected or retained, and the study was determined not to require formal IRB review at our institution. Our use of AI assistants (GPT-5.2) in item generation, ground-truth description authoring, and perception-success judging is fully disclosed in Appendix D; all final items, analyses, and technical claims were verified by human researchers.

We do not foresee meaningful risks of misuse: the benchmark targets nonverbal abstract and natural-image puzzle solving and does not produce, distribute, or facilitate the production of harmful content.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Anthropic. 2026. Introducing claude sonnet 4.6. Accessed: 2026-05-25.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Autodesk Inc. 2025. Tinkercad. Accessed: 2026-01-05.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. 2018. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Huanqia Cai, Yijun Yang, and Winston Hu. 2025. Mm-iq: Benchmarking human-like abstraction and reasoning in multimodal models. *arXiv preprint arXiv:2502.00698*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.

Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. 2024. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*, pages 22–38. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. 2024. DesignQA: A multimodal benchmark for evaluating large language models' understanding of engineering documentation. *Preprint*, arXiv:2404.07917.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, and 15 others. 2023. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Simon Ging, Mar'ia A. Bravo, and Thomas Brox. 2024. Open-ended VQA benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. *Preprint*, arXiv:2402.07270.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Hugging Face. 2025a. llava-hf/llava-v1.6-34b-hf. Model card; accessed: 2026-01-10.

Hugging Face. 2025b. llava-hf/llava-v1.6-mistral-7b-hf. Model card; accessed: 2025-09-15.

Hugging Face. 2025c. llava-hf/llava-v1.6-vicuna-7b-hf. Model card; accessed: 2025-09-15.

Hugging Face. 2025d. OpenGVLab/InternVL3-9B. Model card; accessed: 2025-09-15.

Hugging Face. 2025e. Qwen/Qwen2.5-VL-3B-Instruct. Model card; accessed: 2025-09-15.

Hugging Face. 2025f. Qwen/Qwen2.5-VL-3B-Instruct-AWQ. Model card; accessed: 2025-09-15.

Hugging Face. 2025g. Qwen/Qwen2.5-VL-7B-Instruct. Model card; accessed: 2025-09-15.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 46567–46592. Curran Associates, Inc.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Hengzhi Li, Brendon Jiang, Alexander Naehu, Regan Song, Justin Zhang, Megan Tjandrasuwita,

Chanakya Ekbote, Steven-Shine Chen, Adithya Balachandran, Wei Dai, Rebecca Chang, and Paul Pu Liang. 2025a. PuzzleWorld: A benchmark for multimodal, open-ended reasoning in puzzlehunts. *Preprint*, arXiv:2506.06211.

Ming Li, Jike Zhong, Tianle Chen, Yuxiang Lai, and Konstantinos Psounis. 2025b. EEE-Bench: A comprehensive multimodal electrical and electronics engineering benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. SceMQA: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*. Later published at ECCV 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 6774–6786.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

OpenAI. 2024a. GPT-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-09-15.

OpenAI. 2024b. Hello GPT-4o. Accessed: 2025-09-15.

OpenAI. 2025a. Code interpreter. Accessed: 2026-05-01.

OpenAI. 2025b. GPT-5 mini model. OpenAI API documentation; accessed: 2026-01-10.

OpenAI. 2025c. GPT-5.1: A smarter, more conversational ChatGPT. Accessed: 2026-01-10.

OpenAI. 2025d. Introducing GPT-5.2. Accessed: 2026-01-05.

OpenAI. 2025e. Introducing OpenAI o3 and o4-mini. Accessed: 2025-09-15.

OpenAI. 2025f. Thinking with images. Accessed: 2025-09-15.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. 2024. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W. Verjans. 2024. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pretraining framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11492–11501.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. ScienceQA: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski,

Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642. Curran Associates, Inc.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11854–11864.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *Preprint*, arXiv:2402.14804.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023a. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2406.18521.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5312–5322.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Weichen Zhang, Zile Zhou, Xin Zeng, Xuchen Liu, Jianjie Fang, Chen Gao, Yong Li, Jinqiang Cui, Xinlei Chen, and Xiao-Ping Zhang. 2025a. Open3D-VQA: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *Preprint*, arXiv:2503.11094.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. 2025b. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. *arXiv preprint arXiv:2510.13795*.

Zeyu Zhang, Zijian Chen, Zicheng Zhang, Yuze Sun, Yuan Tian, Ziheng Jia, Chunyi Li, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2025c. PuzzleBench: A fully dynamic evaluation framework for large multimodal models on puzzle solving. *Preprint*, arXiv:2504.10885.

A Procedure for Natural Question Generation

Natural-domain items were generated through a semi-agentic human–AI pipeline. Annotators first authored five seed examples for each reasoning category, which were then provided to GPT-5.2-Thinking (OpenAI, 2025d) as few-shot prompts. The model was asked to (i) propose 15 candidate sub-categories for each reasoning category and (ii) generate 20 candidate puzzle instances per sub-category.

Five students independently ranked the generated sub-categories and instances according to feasibility, reasoning quality, and visual clarity. We retained the top-ranked categories and examples, yielding 100 natural-domain items per reasoning category. Annotators then sourced publicly available copyright-compliant images for each entity in the retained instances and assembled the final puzzle layouts.

Worked example. For Odd One Out, one generated biology-themed instance used the sequence *frog egg, tadpole, tadpole with legs* with candidate options *frog egg, adult frog, fish, lizard*. The intended rule was continuation of the frog life cycle, making *adult frog* the correct answer. Annotators sourced publicly available images for each entity and assembled the final puzzle.

Annotator team and compensation. Throughout this paper, "annotators" refers to the team that contributed to item authoring, image sourcing, layout assembly, and answer-correctness verification across both the natural and abstract domains (Appendix A,Appendix B,Appendix C). The team comprised the paper's authors together with three undergraduate research assistants who provided support across these activities. The undergraduate assistants were compensated as paid student researchers at the standard hourly rate set by our institution and will be acknowledged by name in the non-anonymous version of this paper.

B Ground-Truth Description Authoring Protocol

DiagVRIQ ground-truth descriptions were generated by GPT-5.2 (OpenAI, 2025d) under three constraints embedded in the generation prompt: (i) describe all panels, entities, attributes, and answer options visible in the image; (ii) do not name the underlying transformation, grouping, or sequence rule; and (iii) do not reveal the correct answer or its option label.

Each generated description was then independently verified by an annotator against the puzzle image. The annotator checked **coverage** (all panels and answer options are described), **rule-neutrality** (no statement of the puzzle's rule, transformation, or grouping principle), and **answer-neutrality** (no privileged reference to any specific option). Descriptions that failed any check were revised before inclusion in DiagVRIQ.

For Sequence Completion, descriptions enumerate the contents of each top-row panel and each candidate answer panel using concrete visual vocabulary (shape, position, orientation, count) without indicating how panels relate. For Matrix Prediction, descriptions enumerate each cell and answer option. For Odd One Out, descriptions list each candidate entity together with its visible attributes, without identifying which property the puzzle groups by.

C Abstract Puzzle Sourcing and Modification

C.1 Source Materials and Screening

Abstract-domain puzzles were constructed from two sources: publicly accessible civil-service and standardized aptitude examination materials, and items authored from scratch by the annotator team.

Annotators manually filtered sourced items to remove puzzles that (i) relied on textual content, (ii) admitted ambiguous or non-unique answers, (iii) were visually low quality or poorly legible, or (iv) did not match the target reasoning categories. All retained puzzles were independently reviewed by a second annotator.

C.2 Reformatting and Modification

Sourced puzzles were reconstructed into a unified VRIQ format. Panels and answer options were normalized into consistent configurations, and minor visual artifacts were removed when necessary.

To reduce contamination risk from widely circulated puzzle collections, a substantial subset of sourced items was modified into novel instances by altering superficial visual attributes including shape identities, counts, orientations, and styling while preserving the underlying reasoning rule.

C.3 Manually Authored Items

In addition to sourced materials, annotators authored abstract puzzles from scratch to increase

coverage of underrepresented rule types and perceptual attributes. These items followed the same quality-control pipeline as sourced puzzles.

C.4 Composition and Validation

The final abstract-domain set comprises 600 modified sourced items and 400 manually authored items. All puzzles were independently solved by two annotators and retained only when both agreed on a unique correct answer.

License and intended use. VRIQ, DiagVRIQ, perception probes, and ground-truth descriptions will be released under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. Evaluation code will be released under the MIT license. The benchmark is intended for research use in evaluating and analyzing visual reasoning capabilities of vision–language models. We discourage use as a training set, as this would compromise the benchmark's diagnostic value for future model evaluation.

D Models and Evaluation Details

Table 5 lists all evaluated VLMs with version strings, parameter counts, and source references.

Compute budget. Open-source models were evaluated on a single workstation with 2 NVIDIA RTX A6000 GPUs (48GB VRAM each). Proprietary models (GPT-4o, GPT-4o-mini, GPT-5.1, GPT-5-mini, Gemini-2.5-Pro, Claude Sonnet 4.6, and o3) were accessed through their respective official APIs. All evaluations use deterministic decoding and a single run per item; no hyperparameter search was performed.

Human baseline details. The 300-question subset was selected by stratified sampling (30 items from each of 5 categories $\times$ 2 domains). Participants were recruited from university research groups and screened for normal or corrected-to-normal vision. Each question was independently answered by two annotators.

Participant instructions. Participants were given the following written instructions before beginning the task: "You will see a series of visual reasoning IQ puzzles, each with four answer options. For each puzzle, select the option that best completes the pattern. There is no time limit, and you may take breaks between items. If you are uncertain, select your best guess. Your responses contribute to an aggregate human-accuracy estimate; no personally identifying information is recorded." Participants completed the task in a single session without per-item feedback. No training items or calibration questions were provided.

Recruitment, payment, and consent. Participants were unpaid volunteers and were informed that participation was entirely voluntary. Before beginning the task, they were informed that their responses would be used solely to compute aggregate human-baseline accuracy and that no personally identifying information would be recorded or retained.

Ethics review. The human-baseline data collection involved adult volunteers solving multiple-choice puzzles with no personally identifying information collected or retained, and was determined not to require formal IRB review at our institution.

Use of AI assistants. GPT-5.2-Thinking was used in the semi-agentic natural-domain item generation pipeline (Appendix A); GPT-5.2 was used to generate initial DiagVRIQ ground-truth descriptions, which were subsequently verified and revised by human annotators (Appendix B); and GPT-5.2 was used as the LLM-as-judge for perception-success labeling, with calibration against human annotations reported in Appendix G. AI assistants were also used for minor stylistic editing of the manuscript. All technical claims, analyses, and final benchmark items were authored or verified by the human researchers.

E Describe-Then-Solve Prompt

In the describe-then-solve protocol (subsection 4.2), the model produces a description of the puzzle followed by an answer, in a single generation. We use the prompt in Figure 4 for all evaluated VLMs. The description and answer are extracted from the same generation via tagged delimiters; no second model call is required.

> **System.** You are an expert at solving visual reasoning puzzles. You will be shown an image and a multiple-choice question.
> Before answering, produce a faithful description of the image as if explaining it to a non-sighted reader. Describe all panels, entities, attributes, and answer options. Do not infer the rule or answer during the description.
> Then reason about the puzzle and provide the final answer.

Figure 3: 3D abstract examples built by Tinkercad.

User.
`[image]`
Question: `{question}`
Options: A) `{a}` B) `{b}` C) `{c}` D) `{d}`

Format:
`<description>...</description>`
`<think>...</think>`
`<answer>X</answer>`
where `X` is one of A, B, C, or D.

Figure 4: Describe–then–solve prompt used for perception–reasoning attribution (subsection 4.2). The model produces the description, reasoning, and answer in a single generation.

F Description-Quality Judge Prompt

We use GPT-5.2 as the LLM judge. For each DiagVRIQ item, the judge receives (i) the puzzle image, (ii) the ground-truth description authored under the protocol in Appendix B, and (iii) the model-generated description from the describe-then-solve protocol. The judge returns three scalar scores in $[1, 5]$ and a binary perception-success label. The binary label is the primary attribution signal used throughout the main paper; the scalar dimensions are collected for diagnostic purposes and to enable cross-checking of the binary judgments, but are not used in the perception/reasoning decomposition itself. The full prompt is shown below.

Table 5: Vision–language models evaluated on VRIQ. Open-source models accessed via Hugging Face; proprietary models accessed via official APIs. "Thinking" indicates reasoning-optimized variants; "Tools" indicates models with image-manipulation tool access during inference.

Model	Params	Type	Citation
Small open-source (3B–9B)			
Qwen2.5-VL-3B-Instruct	3B	Instruction-tuned	(Hugging Face, 2025e)
Qwen2.5-VL-3B-Instruct-AWQ	3B	Quantized	(Hugging Face, 2025f)
Qwen2.5-VL-7B-Instruct	7B	Instruction-tuned	(Hugging Face, 2025g)
InternVL3-9B	9B	Instruction-tuned	(Hugging Face, 2025d)
LLaVA-v1.6-Mistral-7B	7B	Instruction-tuned	(Hugging Face, 2025b)
LLaVA-v1.6-Vicuna-7B	7B	Instruction-tuned	(Hugging Face, 2025c)
LLaVA-OneVision-1.5-8B-RL	8B	RL-tuned	(An et al., 2025)
Bee-8B-RL	8B	RL-tuned	(Zhang et al., 2025b)
Mid-size open-source (32B–34B)			
Qwen3-VL-32B-Thinking	32B	Thinking	(Bai et al., 2025)
LLaVA-v1.6-34B	34B	Instruction-tuned	(Hugging Face, 2025a)
Frontier / proprietary			
GPT-4o	–	Instruction-tuned	(OpenAI, 2024b)
GPT-4o-mini	–	Instruction-tuned	(OpenAI, 2024a)
GPT-5.1	–	Frontier	(OpenAI, 2025c)
GPT-5-mini	–	Frontier	(OpenAI, 2025b)
Gemini-2.5-Pro	–	Frontier	(Comanici et al., 2025)
Claude-Sonnet-4.6	–	Frontier	(Anthropic, 2026)
Tool-augmented			
OpenAI o3 (thinking with images)	–	Thinking + Tools	(OpenAI, 2025e)

System. You are an expert evaluator of vision–language model descriptions for visual reasoning tasks. You will be given a puzzle image, a human-authored reference description, and a model-generated description. Your task is to determine whether the model description correctly captures the visual content of the image.

Important. Do *not* evaluate whether the description contains enough information to solve the puzzle or infer the underlying rule. Evaluate only whether it accurately reports what is visually present.

User.
`[image]`

Reference description: `{ground_truth}`

Model description: `{model_description}`

Perception Success (binary).
Does the model description correctly identify the visual entities, panel structure, and answer options shown in the image?

- `"yes"`: entities, panels, and answer options are correctly identified; minor wording differences or small omissions are acceptable.
- `"no"`: entities are misidentified, hallucinated, omitted, or the panel structure is misunderstood.

Provide a brief justification (1–2 sentences), then return the verdict.

Return **only** valid JSON:

```
{
  "reason": "...",
  "perception_success": "yes/no"
}
```

We adapt the Prometheus-Vision (Lee et al., 2024) rubric structure: explicit score-level descriptions for each dimension, chain-of-thought reasoning before scoring, and a separate binary verdict that is not derivable from the scalar scores alone. The binary verdict is the primary signal for attribution (subsection 4.2).

G Calibration of the Perception-Success Judge

To validate the binary perception-success label produced by GPT-5.2, one author independently annotated all 627 model descriptions (3 models $\times$ 209 DiagVRIQ items) using the same criterion stated in the judge prompt: "does the model correctly identify the visual entities, panel structure, and answer options present in the image?". The author labeled each description as either `yes` or `no`, without access to the judge's output. Annotations were performed independently across the three model output sets. Agreement is highest on LLaVA (96.9%), where descriptions are uniformly poor and perception failure is unambiguous. Agreement is lowest on Qwen (89.1%), where descriptions sit at the boundary of acceptable perception and partial misidentification, exactly the situation where a binary judgment is hardest. The judge is mildly conservative on the harder models, classifying borderline cases as per-

Table 6: Author–judge agreement on binary perception-success labels.

Model	N	Agreement (%)	Author $P\checkmark$ (%)	Judge $P\checkmark$ (%)
LLaVA-1.6-Mistral-7B	209	96.9	9.9	12.0
Qwen2.5-VL-7B	209	89.1	46.6	41.5
GPT-4o	209	92.3	90.3	85.6
Combined	627	92.7	—	—

ception failure at a slightly higher rate than the author would. This bias is consistent across categories and does not change the qualitative attribution: under either label source, OO is reasoning-bound and SC is perception-bound.

Cohen's κ across all 627 items is 0.84, indicating substantial agreement well above chance and comparable to published LLM-as-judge calibration figures (e.g., Prometheus reports $r = 0.897$ with humans on text rubrics (Kim et al., 2024); Prometheus-Vision reports $r = 0.786$ on multimodal rubrics (Lee et al., 2024)).

We report the LLM-judge labels as primary numbers throughout the main paper. subsection 6.2 results computed under author labels yield the same three-tier qualitative pattern; per-cell counts differ by at most 4 items per category.

H Abstract Perception Capability Audit

H.1 Motivation

For abstract puzzles, whose visual structure cannot be faithfully described in text without simplifying the task or implicitly revealing the underlying rule, we cannot apply the describe-then-solve protocol used for DiagVRIQ (subsection 6.2). To test whether models reliably see the small visual details that abstract puzzles depend on, specific dot counts, shape orientations, or panel positions, we authored atomic P-probes that target individual visual primitives independently of the puzzle-solving task. P-probes provide a population-level audit of which perceptual primitives current VLMs reliably extract from abstract images.

H.2 Probe Authoring and Coverage

P-probes are multiple-choice questions about a single perceptual attribute of an abstract puzzle image, designed to be answerable from the image alone without any reasoning over the puzzle's rule. Each probe targets exactly one of eight perceptual dimensions: color, shape, count, position, rotation/orientation, 3D/depth, symmetry/pattern, and distractor similarity. The eight dimensions were chosen to span the small visual details that abstract puzzles compositionally rely on.

We authored **20 probes per (probe type, category) cell** where the probe type was applicable to the category, sampled from the abstract subset. Probes were authored by the same team that constructed the abstract puzzles, working from the puzzle images themselves. For each selected puzzle, probe authors identified one perceptual dimension present in the image and authored one probe targeting that dimension. Probes that referenced rule structure, transformation patterns, or the puzzle's answer were excluded. Two authors independently verified each probe to confirm that (i) the question targets a single perceptual attribute, (ii) the correct answer is unambiguously determined by the image, and (iii) the question does not reveal information about the puzzle's underlying rule. Probes that failed verification were revised or discarded.

Some probe types are not applicable to all categories. 3D/depth probes are administered only on Figure Rotation and 3D Visualization items where depth information is present; distractor-similarity probes are administered only on Sequence Completion and Odd One Out items where the puzzle contains explicit distractors; color and shape probes are restricted to categories where these primitives vary across panels. Probe types marked N/A in Table 7 correspond to (probe type, category) cells where the relevant perceptual primitive does not appear, and no probes were authored for those cells. Across all applicable cells, the audit comprises **520 P-probes in total** (20 probes per applicable cell, summed across the table).

H.3 Example P-Probes

Figure 5 shows three representative probes spanning different probe types and categories. Each probe presents the puzzle image together with a single perceptual question and four answer options. The probe is administered independently of the original puzzle question.

Example 1: Count probe (Matrix Prediction)

Image: [Abstract MP puzzle showing a 3×3 matrix of dot patterns.]
Question: How many black dots are in the top-right cell of the matrix?
Options: A) 2 B) 3 C) 4 D) 5

Example 2: Rotation probe (Figure Rotation)

Image: [Abstract FR puzzle showing four geometric shapes in different orientations.]
Question: Which direction is the arrow in the third panel pointing?
Options: A) Up B) Right C) Down D) Left

Example 3: Position probe (Sequence Completion)

Image: [Abstract SC puzzle showing a sequence of squares with shaded regions.]
Question: In the second panel, where is the shaded region located?
Options: A) Top-left B) Top-right C) Bottom-left D) Bottom-right

Figure 5: Three example P-probes spanning different probe types and categories. Each probe targets a small visual detail of the puzzle image and is answerable from the image alone, without inferring the puzzle's rule.

H.4 Evaluation Protocol

Each P-probe is presented to the model as an independent multiple-choice question with the puzzle image attached. The prompt template is shown in Figure 6. Probes are evaluated in a separate context from the puzzle-solving question; the model does not receive the puzzle's original question or answer when answering a probe.

System. You will be shown an image together with a multiple-choice question about a specific visual attribute of the image. Answer based only on what is visible in the image.

User.
`[image]`
Question: `{probe_question}`
Options: A) `{a}` B) `{b}` C) `{c}` D) `{d}`

Answer with a single letter (A, B, C, or D).

Figure 6: Prompt template used for P-probe evaluation. The template is identical across probe types and categories.

H.5 Results

Table 7 reports failure rates averaged across models for each probe type and category. Failure rates concentrate on counting (**73.8%**), 3D/depth (**67.5%**), position (**58.2%**), and rotation/orientation (**52.4%**), all primitives that require precise spatial and quantitative perception. By contrast, color-related probes fail at only **31.4%**. These results suggest that current VLMs are substantially better at appearance-level perception than at extracting structured spatial relationships.

The pattern aligns with the category-level difficulty ordering reported in Table 1: Figure Rotation and 3D Visualization, the categories with the lowest end-to-end accuracy, also exhibit the highest probe failure rates. This consistency supports the interpretation that abstract failures originate at least partly before downstream reasoning occurs.

H.6 Scope and Limitations

We highlight three caveats to interpreting these results:

- **Population-level audit, not item-linked diagnostic.** P-probe accuracy on a population of items does not establish that a model's failure on a specific puzzle was caused by a perceptual error on that puzzle. The describe-then-solve analysis in subsection 6.2 provides item-level perception attribution on the natural-domain DiagVRIQ subset; P-probes provide complementary aggregate evidence on the abstract subset.
- **Probes are derived from puzzle images.** Each P-probe is constructed from the same image used in a VRIQ puzzle. The probe and the puzzle are evaluated in independent contexts (the model does not see the puzzle's original question when answering a probe), but model behavior may correlate across the two questions because they share visual content. We do not claim probe accuracy is fully independent of puzzle accuracy.
- **Probe-type definitions are dataset-specific.** The eight probe types target perceptual primitives that abstract VRIQ puzzles rely on, but other definitions of "perceptual capability" would yield different probe sets. The reported failure rates should be interpreted as evidence about specific perceptual primitives, not as a comprehensive perception evaluation.

I Consistency Between Attribution Metrics and End-to-End Accuracy

We further examine whether the proposed attribution framework is internally consistent with the

Table 7: Detailed perception probe failure rates. Failure rates (%) for each probe type across IQ categories, with averages reported in the final column. N/A indicates probes not applicable to a given task category.

Probe Type	SC (%)	MP (%)	OO (%)	FR (%)	3D (%)	Avg (%)
Color	26.67	47.50	20.00	N/A	N/A	31.39
Shape	35.68	60.00	65.00	80.00	N/A	60.17
Count	64.29	80.00	77.50	75.00	72.21	73.80
Position	56.22	73.33	50.00	51.43	60.00	58.20
Rotation/Orientation	45.39	50.00	50.00	64.00	N/A	52.35
3D/Depth	N/A	N/A	N/A	70.00	65.00	67.5
Symmetry/Pattern	40.91	46.67	75.00	N/A	N/A	54.19
Distractor Similarity	52.00	N/A	46.33	N/A	N/A	49.17

observed end-to-end task accuracy. The framework decomposes successful question answering into two sequential components: (i) successful perception and (ii) successful reasoning conditioned on successful perception. Under this formulation, the probability of producing a correct answer can be expressed as:

$$P(A^{\checkmark}) = P(P^{\checkmark})\,P(A^{\checkmark} \mid P^{\checkmark}) + P(P^{\times})\,P(A^{\checkmark} \mid P^{\times}) \tag{1}$$

where $P(P^{\checkmark})$ denotes perception accuracy, $P(A^{\checkmark} \mid P^{\checkmark})$ denotes reasoning accuracy conditioned on successful perception, and $P(P^{\times}) = 1 - P(P^{\checkmark})$ denotes failed perception. The term $P(A^{\checkmark} \mid P^{\times})$ captures the probability of answering correctly despite incorrect perception.

Since all tasks are four-option multiple-choice questions, a failure in perception should approximately reduce answer selection to random guessing. We therefore assume:

$$P(A^{\checkmark} \mid P^{\times}) \approx 0.25 \tag{2}$$

Substituting this assumption yields:

$$P(A^{\checkmark}) \approx P(P^{\checkmark})P(A^{\checkmark} \mid P^{\checkmark}) + (1 - P(P^{\checkmark}))0.25 \tag{3}$$

Using the measured perception and reasoning accuracies from Table 3, the estimated accuracies closely align with the observed end-to-end results. For example, for GPT-4o on Sequence Completion:

$$0.867 \times 0.904 + (1 - 0.867) \times 0.25 = 0.817$$

which is close to the observed accuracy of 80.0%. Similarly, for Qwen2.5-VL-7B on Odd One Out:

$$0.843 \times 0.605 + (1 - 0.843) \times 0.25 = 0.549$$

which is close to the observed 52.9%.

Overall, the decomposition estimates remain within a small margin of the measured end-to-end accuracies across nearly all evaluated settings (Table 8). This suggests that most cases in which models answer correctly despite failed perception are largely attributable to chance-level guessing rather than successful reasoning under incorrect perceptual grounding.

An exception is observed for LLaVA-1.6 on Matrix Prediction, where the decomposition estimate deviates more substantially from the observed accuracy. This may indicate that the model exhibits non-uniform answer selection behavior under perceptual failure, such as systematic distractor biases, causing the residual accuracy to differ from the random-guess assumption. Nevertheless, the decomposition remains broadly consistent across models and task categories.

Table 8: Consistency check between attribution decomposition and observed end-to-end accuracy. Estimated accuracy is computed as $P(P^{\checkmark})P(A^{\checkmark} \mid P^{\checkmark}) + (1 - P(P^{\checkmark})) \times 0.25$.

Model + Category	Est Acc	Observed Acc	Diff
LLaVA MP	25.5	21.4	4.1
LLaVA OO	36.3	39.2	2.9
LLaVA SC	25.8	25.0	0.8
Qwen MP	59.4	59.2	0.2
Qwen OO	53.7	52.9	0.8
Qwen SC	39.2	38.3	0.9
GPT-4o MP	88.0	87.8	0.2
GPT-4o OO	85.6	84.3	1.3
GPT-4o SC	80.8	80.0	0.8

J Qualitative Examples

Figure 7, Figure 8, and Figure 9 present representative qualitative examples from the description-based attribution and GT-augmentation experiments.

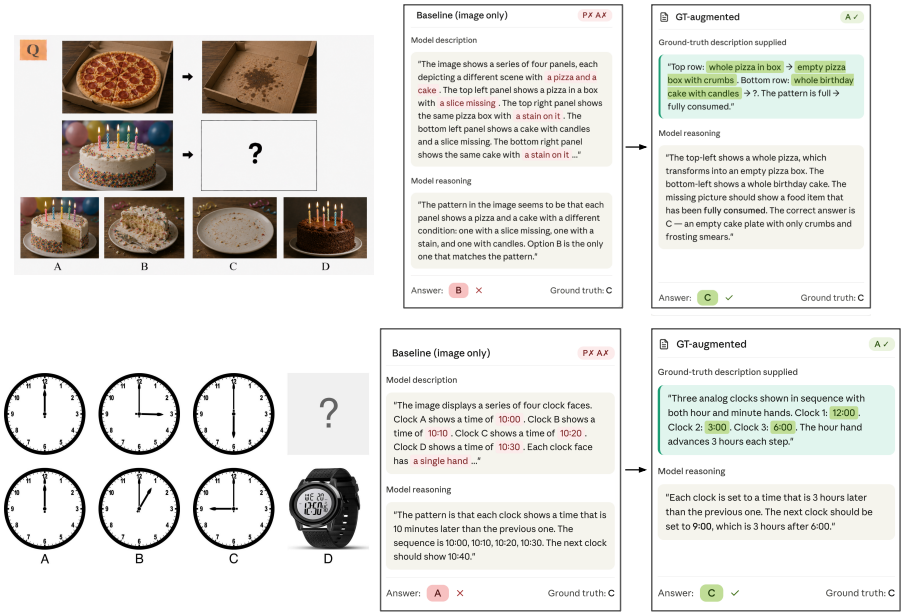


Figure 7: Representative LLaVA-1.6 failures resolved by ground-truth augmentation. In the baseline setting, the model incorrectly parses the visual structure of the puzzle and predicts the wrong answer. Supplying a verified textual description corrects the perceptual error and recovers the correct solution.

K Ground-Truth Augmentation Prompt

The GT-augmented condition (subsection 4.3) supplies the model with the human-authored ground-truth description alongside the image and question. We use the prompt in Figure 10 for all evaluated VLMs.

L Generalizing Ground-Truth Augmentation Beyond Natural Images

The ground-truth augmentation protocol introduced in subsection 4.3 is not inherently limited to natural-image reasoning tasks. In principle, the same methodology can extend to abstract visual reasoning problems whenever the visual structure can be faithfully serialized into text without (i) simplifying the task, (ii) implicitly revealing the underlying rule, or (iii) omitting information necessary for solving the puzzle.

To test this possibility, we conducted a small pilot experiment on abstract-domain Sequence Completion items. For each item, we built the ground-truth textual description enumerating the panel contents, geometric entities, spatial relations, orientations, and answer options while intentionally avoiding explicit statements of the underlying transformation rule. Models were then evaluated using the same GT-augmentation protocol used in DiagVRIQ: the model received both the original image and the verified textual description before answering the multiple-choice question.

Table 9 reports results on 20 abstract Sequence Completion items. All three evaluated models improve under GT augmentation, suggesting that at least part of the failure on these abstract puzzles arises from imperfect visual perception rather than reasoning limitations alone. We caution that the sample is small (n=20). Performance remains substantially below ceiling even with verified perceptual information, indicating that abstract reasoning itself remains challenging after perceptual ambiguity is reduced.

Several VRIQ categories, particularly Figure Rotation and 3D Visualization, rely on dense spatial and geometric structure that cannot be faithfully serialized into text without either severe information loss or implicit rule leakage. For such categories, we instead rely on the P-probe annotations described in subsection 3.2 and Appendix H.

Examples. Figure 11 and Figure 12 present two representative abstract Sequence Completion items. Using only the image, the model incorrectly describes the visual progression and predicts the wrong continuation. After receiving a human-authored description explicitly enumerating the

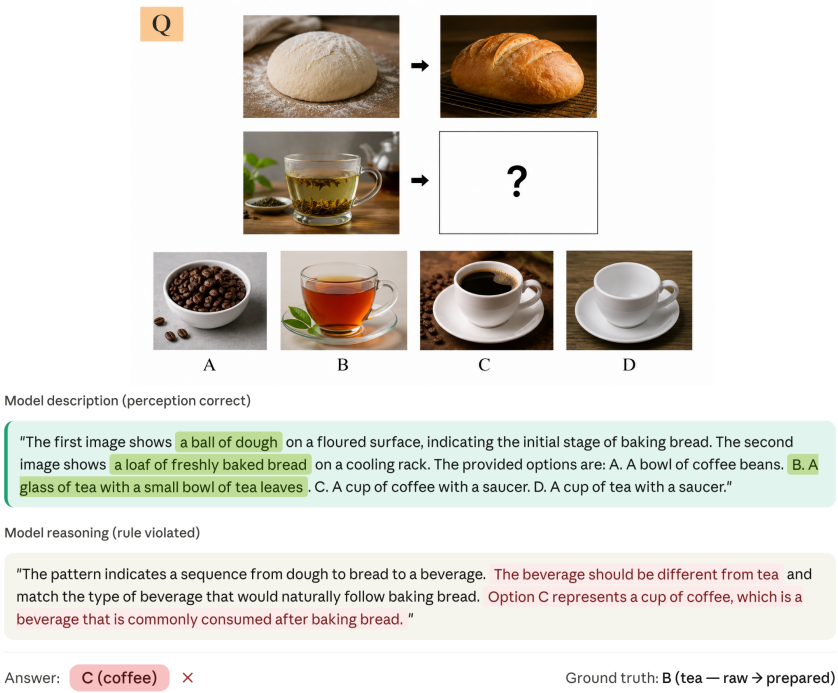


Figure 8: Representative residual reasoning failure from Qwen2.5-VL-7B. The model correctly identifies the visual entities and panel structure but applies an incorrect relational rule, producing the wrong answer even after GT augmentation.

Table 9: Pilot experiment on n=20 abstract-domain Sequence Completion items using human-authored ground-truth descriptions. All three models gain under GT augmentation in the same direction as the natural-domain finding (subsection 6.3). Sample is small; we treat this as illustrative evidence of methodology generalization rather than a quantitative claim.

Model	Baseline	GT-augmented
LLaVA-1.6	25.0%	30.0%
Qwen2.5-VL-7B	25.0%	40.0%
GPT-4o	30.0%	45.0%

panel contents and answer options, the model identifies the correct visual relationships and solves the puzzle successfully. These examples illustrate that, even in abstract domains, some failures originate from incomplete visual parsing rather than deficient symbolic reasoning alone.

M o3 Tool-Use Examples

We provide three representative examples demonstrating how tool-augmented reasoning performs across different task categories and failure modes.

M.1 Example 1: Figure Rotation (Abstract Domain, Success)

Task. Identify the next image in the sequence of compass orientations shown in Figure 13.
Tool call sequence (logged):

```
1. crop(x=40, y=20, w=240, h=240)
2. crop(x=320, y=20, w=240, h=240)
3. crop(x=600, y=20, w=240, h=240)
4. rotate(angle=90)
5. rotate(angle=180)
6. rotate(angle=270)
```

o3 Response

```
The pattern is brutally simple:
Each compass rotates 90° clockwise
each step: 1st has N on top → 2nd
has W on top → 3rd has S on top →
next must have E on top.
Also, the missing wedge on the left
grows each step, so the next one
should still have that left-side cut.
Only Option C matches (E at the top +
left-side cut).
Answer: C
```

Predicted answer: C (correct). **Ground truth:** C.

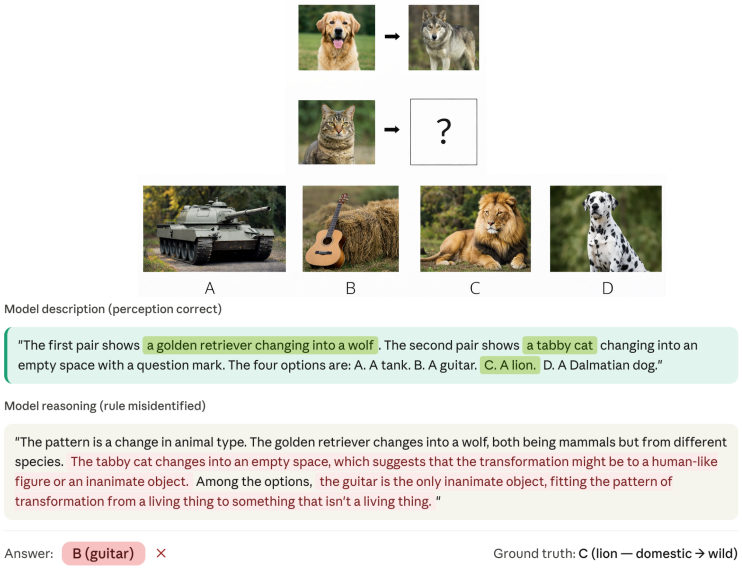


Figure 9: Additional residual reasoning failure under GT augmentation. The model correctly identifies all entities but constructs an invalid grouping strategy that cannot uniquely determine the odd-one-out item.

System. You are an expert at solving visual reasoning puzzles. You will be shown an image, a verified textual description of what the image contains, and a multiple-choice question. Use both the image and the description when reasoning. The description has been prepared by human annotators and is reliable.

User.
`[image]`
Description: `{ground_truth_description}`
Question: `{question}`
Options: A) `{a}` B) `{b}` C) `{c}` D) `{d}`

Think step by step using both the image and the description. Then provide your final answer as a single letter (A, B, C, or D).

Figure 10: GT-augmented prompt. The baseline condition removes the description block and the corresponding sentence in the system message.

Why tools help. Instead of estimating rotation from raw pixels, o3 explicitly generates a rotated reference (Figure 14) and compares it against candidate options. This converts a challenging rotation perception task (52.4% failure rate, Table 7) into direct visual alignment, demonstrating how tools can successfully mitigate certain perceptual bottlenecks.

M.2 Example 2: Sequence Completion with Counting Errors (Natural Domain)

Task. Predict the next image in the log-stacking sequence shown in Figure 15.
Tool call sequence (logged):

```
1. crop(...)  % isolate sequence region
2. crop(...)  % option A
3. crop(...)  % option B
4. crop(...)  % option C
5. crop(...)  % option D
```

o3 Response

```
Count the logs in the top sequence:
1st: 1
2nd: 2
3rd: 2 (same total, just arranged
slightly differently)
4th: 3
5th: 5
That's the Fibonacci pattern starting
as 1, 2, 2, 3, 5, so the next should
be 8 logs.
Among the options, only D shows a
stable stack of 8 logs (arranged as 3
+ 3 + 2).
Answer: D
```

Predicted answer: D (incorrect). **Ground truth:** A.

Failure analysis. The tool trace reveals o3 correctly inferred that the sequence follows a Fibonacci-like pattern and predicted the next configuration should contain eight logs. However, after isolating the answer options (Figure 16), the model miscounted logs in the rendered stacks due

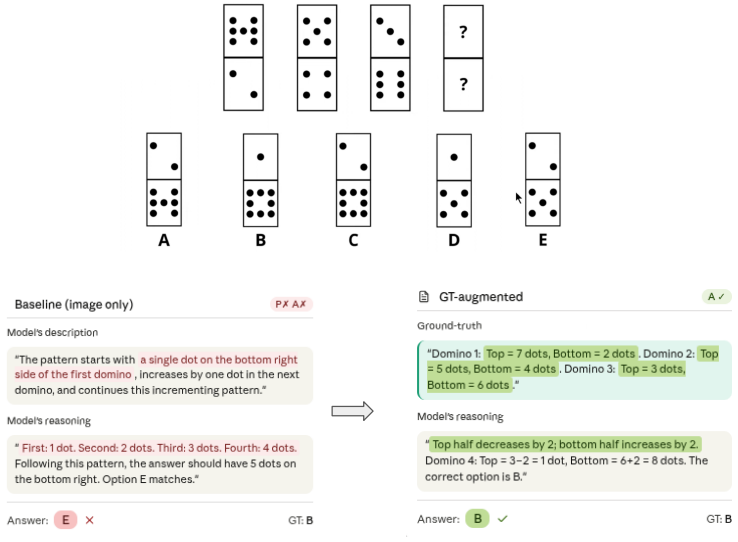


Figure 11: Representative abstract-domain Sequence Completion example illustrating GT-based perceptual augmentation. Red highlights indicate perceptual errors in the baseline description; green highlights indicate corrected visual understanding after GT augmentation.

to occlusion and overlapping cylindrical structures, incorrectly selecting Option D. This demonstrates that tool-assisted isolation improves pattern reasoning but does not resolve fine-grained counting in cluttered natural scenes, consistent with counting having the highest probe failure rate (73.8%, Table 7).

M.3 Example 3: Matrix Reasoning with Persistent Errors (Abstract Domain)

Task. Predict the missing cell in a 3×3 matrix of circle patterns (Figure 17).

Tool execution summary. Over 352 seconds, the model performed:

- 13 cropping operations (matrix cells and answer options),
- 8 circle-detection attempts using HoughCircles with varying parameters,
- Position clustering to classify black vs. white circles.

What went right. The model correctly identified progressive grid size increases (3×3 → 4×3 → 5×4), demonstrating successful structural pattern recognition.

What went wrong. Despite extensive tool use, perception errors persisted:

- Options A and B: 14 circles detected (4 black, 10 white).
- Options C and D: 11 circles detected (1 black, 10 white).
- Cell (Row 1, Col 3): only 5 circles detected (undercount).
- Cell (Row 2, Col 3): 9 circles detected.

These miscounts corrupted the final decision. Additionally, the model failed to detect the actual transformation rule: circles that are black in *both* columns 1 and 2 should be removed in column 3.

o3 Response

```
A. Because in the rightmost column,
the black circles keep the same
"top-right pair" pattern from row 1
→ row 2, and in row 3 that pattern
expands into the larger layout in
the only way that stays consistent
with the build-up seen in the first
two columns. Option A matches that
continuation; the others either drop
required blacks or place them where
```

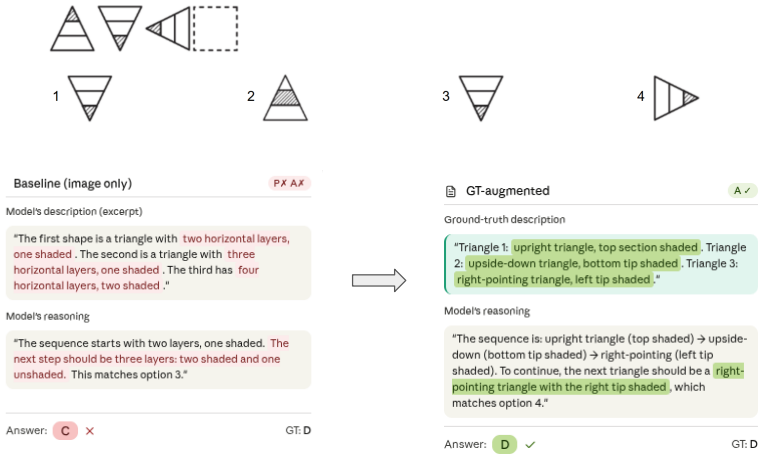


Figure 12: Additional abstract-domain example showing improvement under GT augmentation. In the baseline setting, the model incorrectly interprets the visual structure of the sequence, leading to an incorrect answer. Providing a verified textual description resolves the perceptual ambiguity and enables correct reasoning over the abstract transformation. Highlighted regions show the specific visual attributes that were initially misidentified and later corrected through augmentation.

Figure 13: Input sequence with the missing final image and multiple-choice options (A–D).

Figure 14: Reference image after 270° rotation produced by the tool, used for comparison with the answer options.

```
the progression never puts them.
```

Predicted answer: A (incorrect). **Ground truth:** B.

Key takeaway. Tool augmentation enabled correct structural reasoning but failed to resolve counting (73.8% failure) and position (58.2% failure) bottlenecks (Table 7). This confirms that tools improve reasoning strategies but do not eliminate fundamental perception limitations in abstract visual tasks.

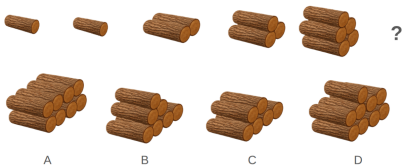


Figure 15: Natural-domain sequence completion task with missing final image and answer options (A–D).

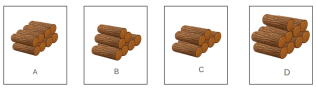


Figure 16: Cropped candidate options used by o3 for visual inspection and counting.

N Dataset Samples

We provide additional sample items spanning the five reasoning categories and both abstract and natural domains. We also include representative DiagVRIQ items paired with their human-verified ground-truth descriptions.

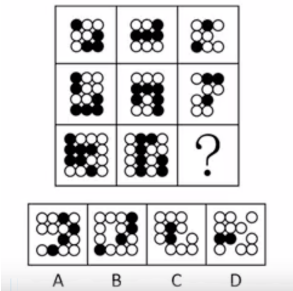


Figure 17: Abstract matrix reasoning task with one missing cell (bottom-right).

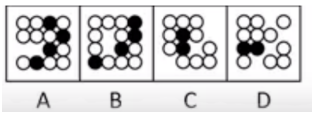


Figure 18: Tool execution trace showing repeated circle-detection attempts. Despite systematic cropping, accurate enumeration and spatial localization remain unreliable under clutter.

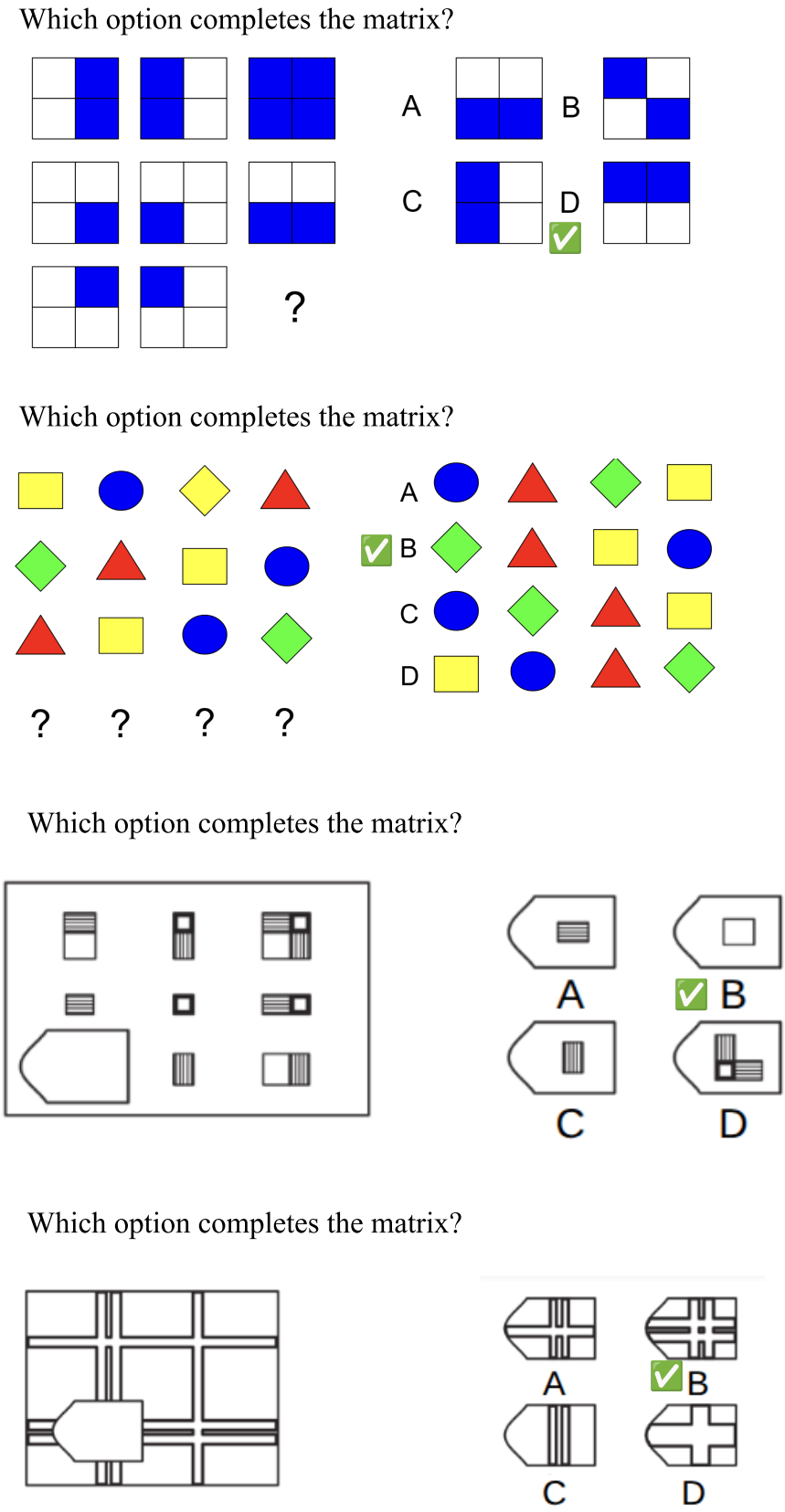
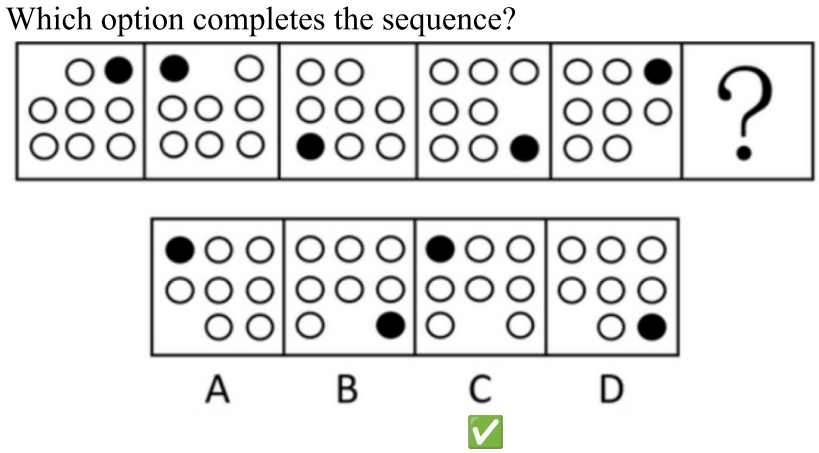
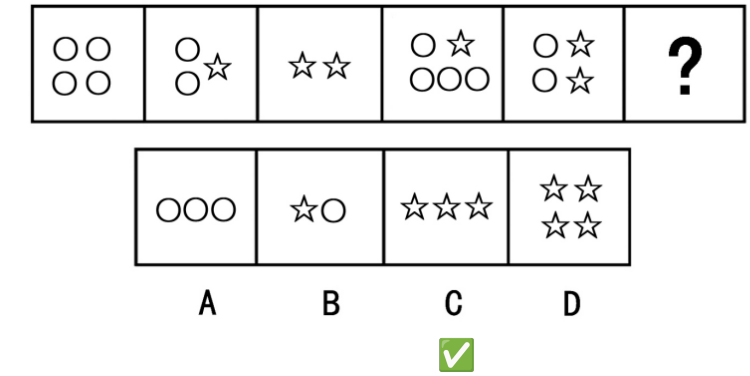
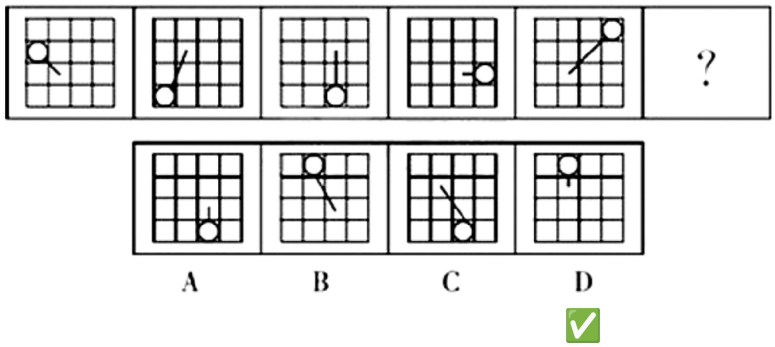


Figure 19: Sample VRIQ abstract Matrix Prediction items. Each puzzle requires identifying the missing panel in a 3×3 matrix based on a transformation rule applied across rows, columns, or both.







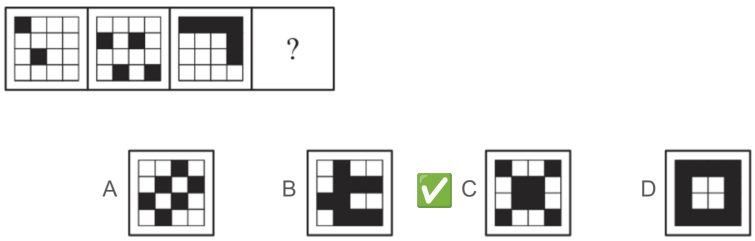


Figure 20: Sample VRIQ abstract Sequence Completion items. Each puzzle requires identifying the next element in a sequence governed by a positional, structural, or arithmetic rule.

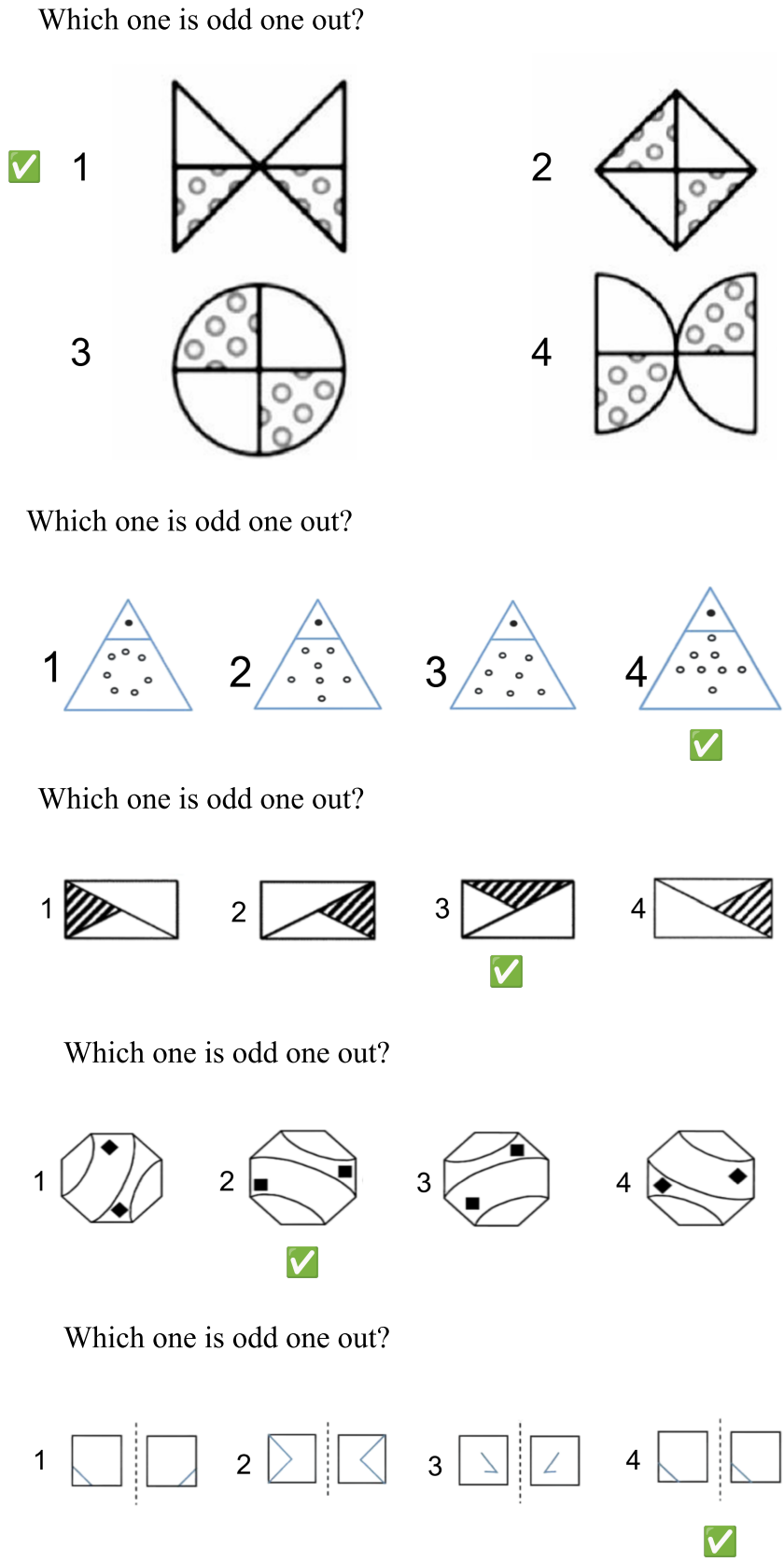
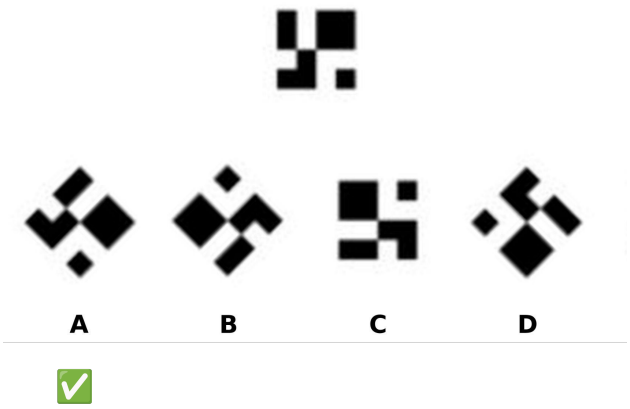
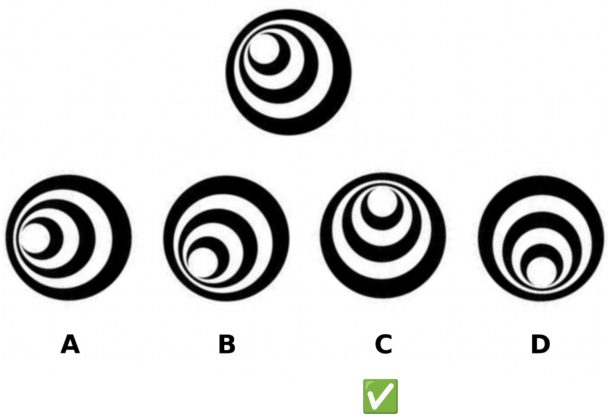


Figure 21: Sample VRIQ abstract Odd One Out items. Each puzzle requires identifying the candidate that does not share a common visual property with the others.

Which of the following 4 options (on the second row) is the original figure (on the first row) rotated counter-clockwise by 315 degrees?



Which of the following 4 options (on the second row) is the original figure (on the first row) rotated clockwise by 45 degrees?



Which of the following 4 options (on the second row) is the original figure (on the first row) rotated counter-clockwise by 315 degrees?

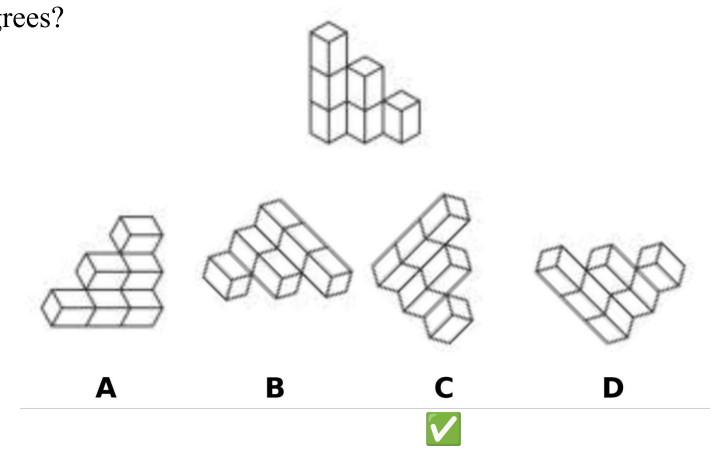


Figure 22: Sample VRIQ abstract Figure Rotation items. Each puzzle requires identifying a rotated version of a geometric figure.

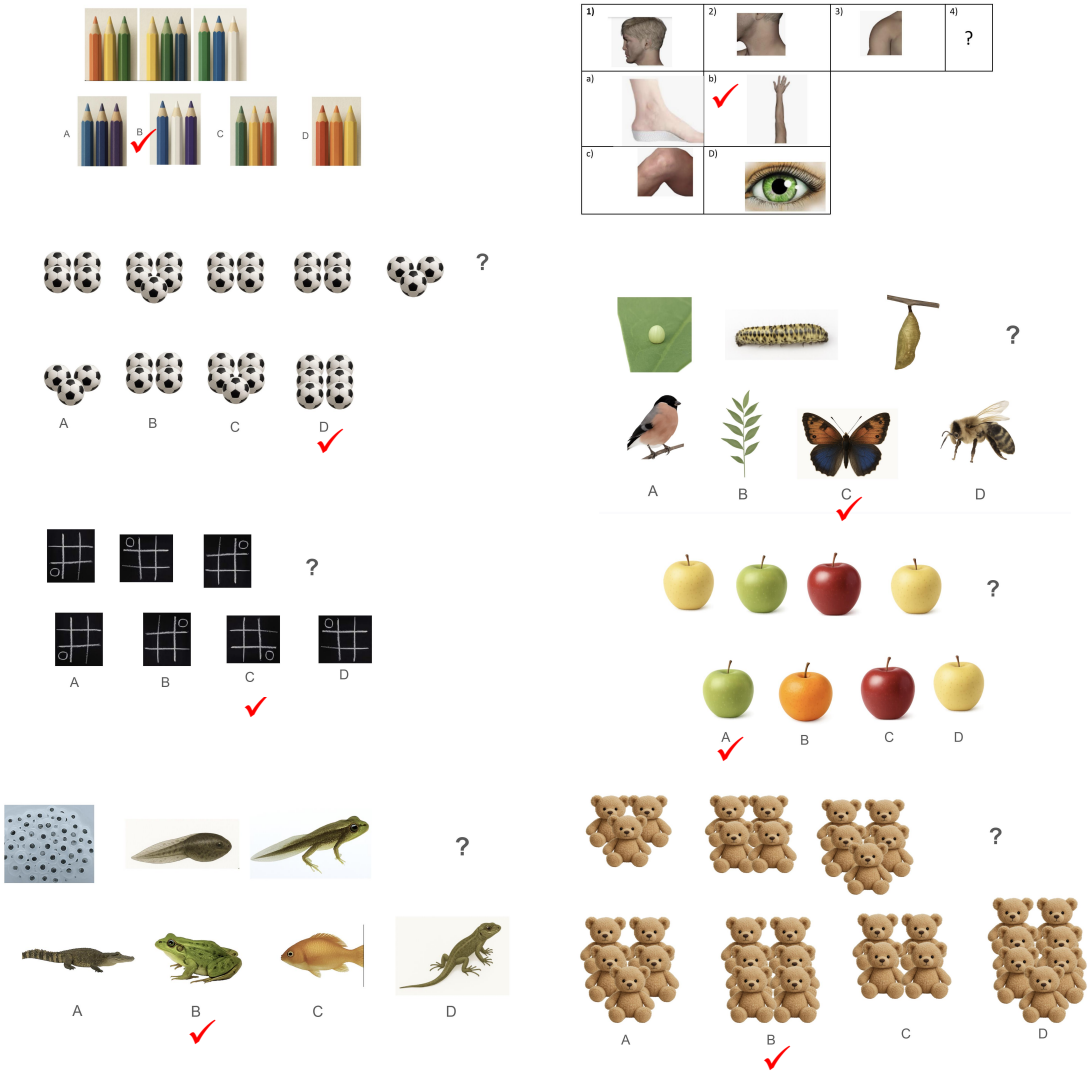


Figure 23: Sample VRIQ natural Sequence Completion questions.

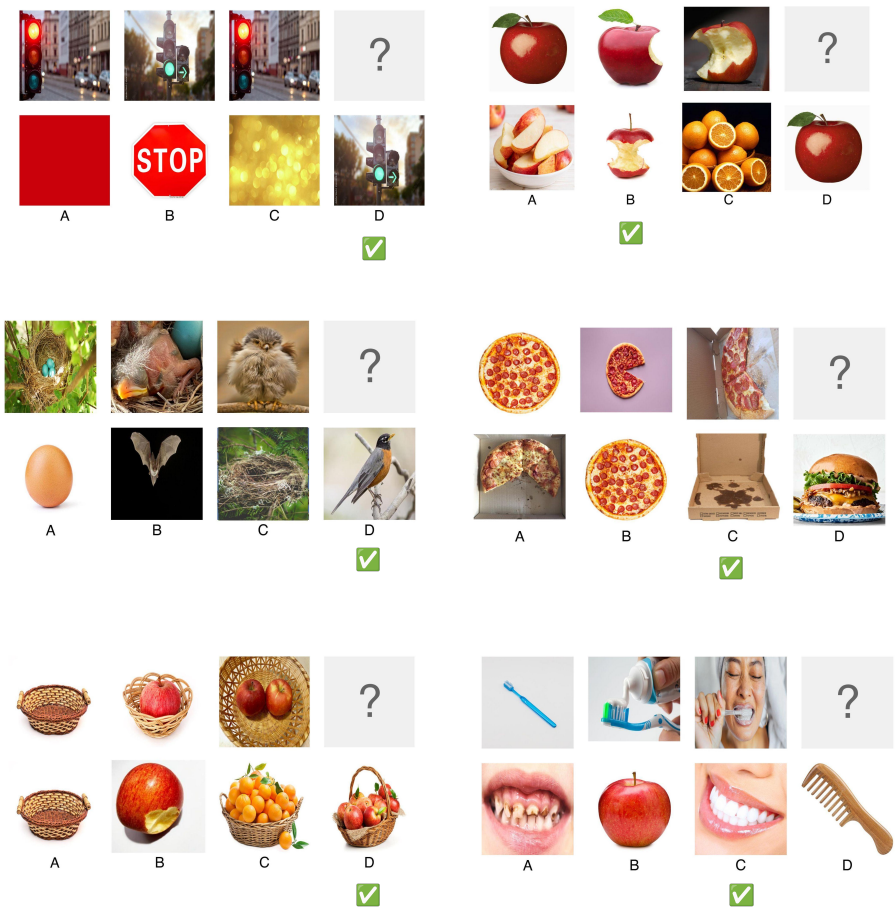


Figure 24: Sample VRIQ natural Sequence Completion items. Each puzzle requires identifying the next element in a sequence of real-world entities or scenes that follow an underlying rule.

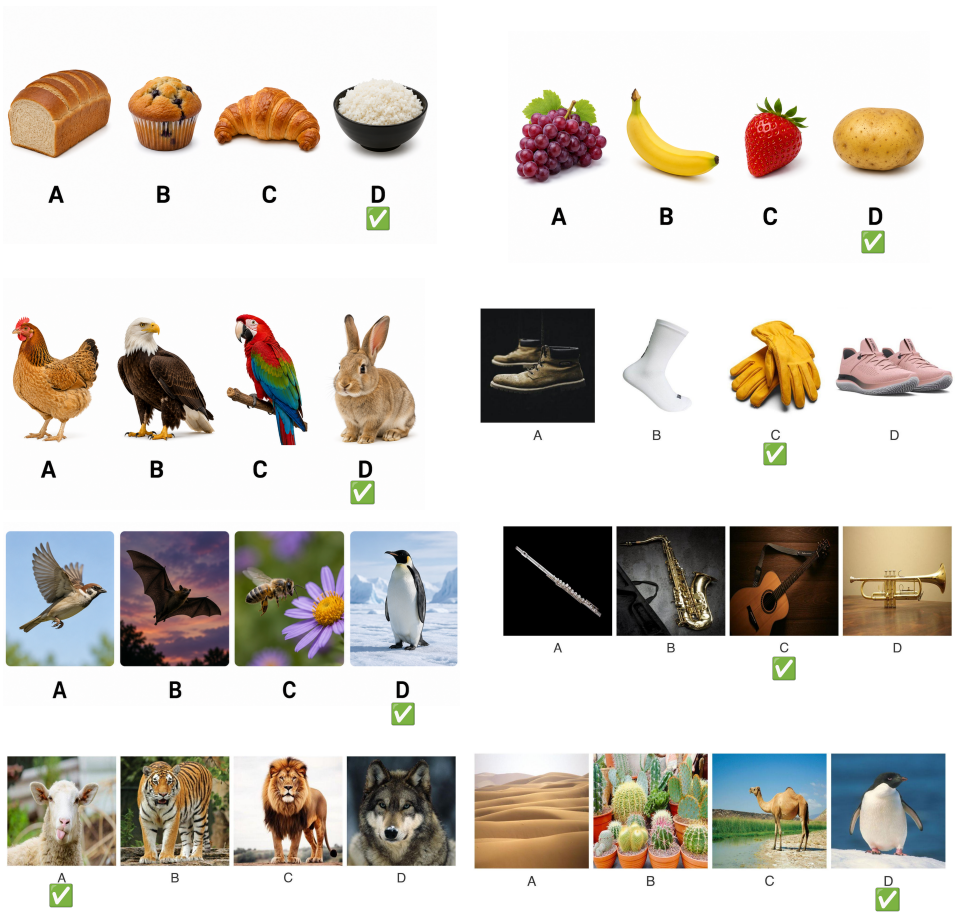


Figure 25: Sample VRIQ natural Odd One Out items. Each puzzle requires identifying the entity that does not belong to a shared category.

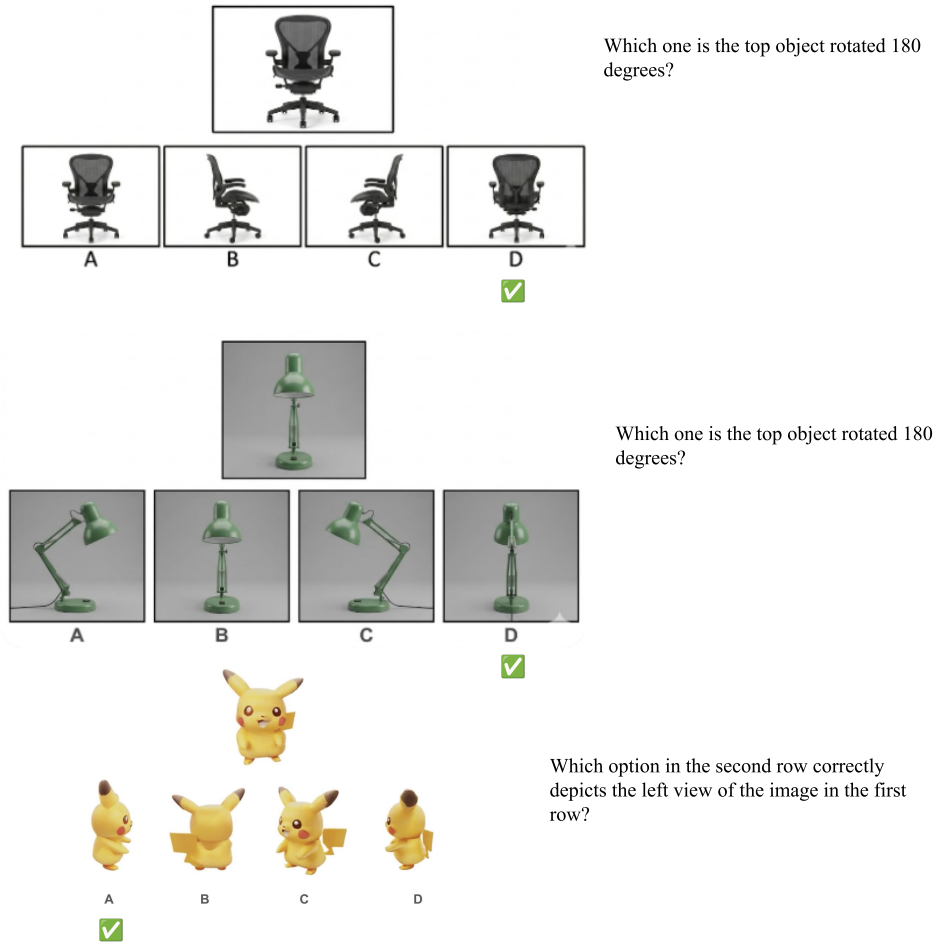


Figure 26: Sample VRIQ natural 3D Visualization items. Each puzzle requires identifying an alternate viewpoint of a real-world object.

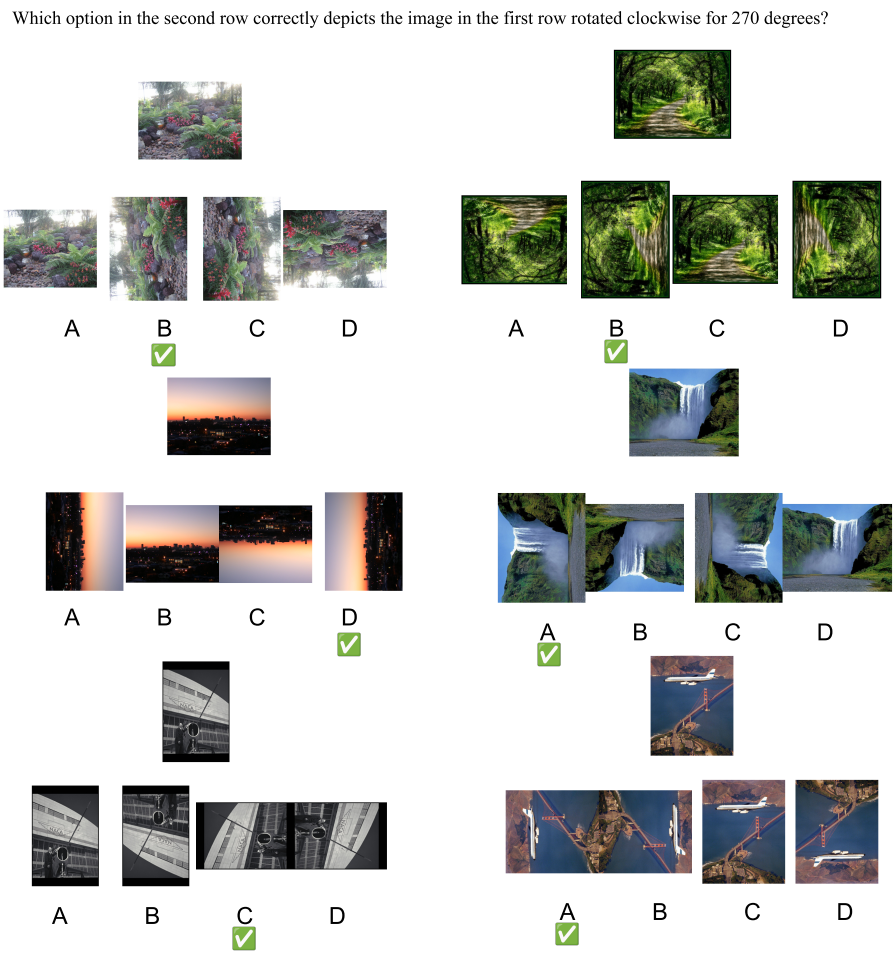
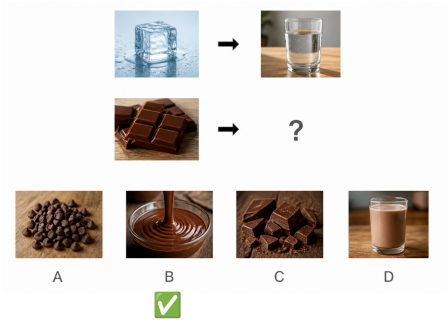
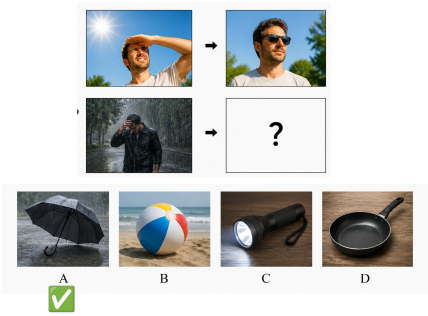


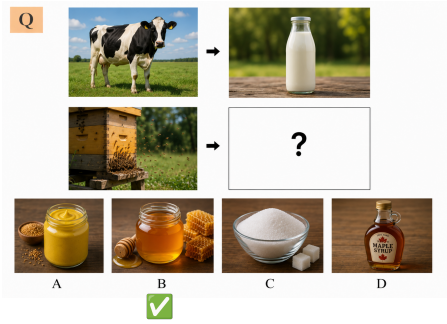
Figure 27: Sample VRIQ natural Figure Rotation items. Each puzzle requires identifying a rotated version of a real-world scene.



In the top row, the top-left picture shows an ice cube. An arrow points to the top-right picture, which shows a glass of water. In the bottom row, the bottom-left picture shows a solid chocolate bar. An arrow points to a question mark. Which option completes the matrix? A. Chocolate chips B. Melted chocolate in a bowl C. Broken chocolate pieces D. Chocolate milk in a glass



In the top row, the top-left picture shows a person under bright sun. An arrow points to the top-right picture, which shows a person wearing sunglasses. In the bottom row, the bottom-left picture shows a person standing in heavy rain. An arrow points to a question mark. Which option completes the matrix? A. An umbrella in the rain B. A beach ball on sand C. A flashlight on a table D. A frying pan on a table



In the top row, the top-left picture shows a dairy cow standing in a field. An arrow points to the top-right picture, which shows a bottle of milk. In the bottom row, the bottom-left picture shows a beehive with bees flying around it. An arrow points to a question mark. Which option completes the matrix? A. A jar of mustard B. A jar of honey C. A bowl of sugar D. A bottle of maple syrup

Figure 28: Three representative DiagVRIQ Matrix Prediction items paired with their human-verified ground-truth descriptions.