PLOTTWIST: DECEPTION-AWARE REASONING IN VISUAL RIDDLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have achieved remarkable progress in visual understanding, yet their ability to reason reliably in the presence of misleading visual information remains largely unexplored. Existing benchmarks mainly evaluate recognition and question answering in settings where relevant visual evidence is explicit and does not intentionally conflict with distracting cues. However, real-world visual interpretation can involve misleading signals and deceptive patterns that provide compelling but unreliable evidence, making it important to understand whether multimodal models can identify trustworthy evidence and resist deceptive interpretations.

Unlike conventional spurious-correlation settings, where models exploit unreliable statistical associations, the misleading clues considered here are semantically relevant and locally plausible. The challenge is therefore to arbitrate among competing pieces of evidence and prioritize the most diagnostic visual cues.

We introduce **PlotTwist**, a diagnostic benchmark for evaluating deception-aware reasoning in visual riddles. PlotTwist focuses on scenarios in which models must distinguish diagnostic evidence from deceptive clues, prioritize informative visual cues, and infer the intended answer beyond superficially plausible interpretations. To obtain sufficiently diverse and creative examples, PlotTwist combines visual riddles curated from online sources with newly developed riddles created through human-designed scenarios with the aid of generative models, which also support image generation. This process enables the construction of diverse visual riddles exhibiting different forms of misleading evidence and deception.

PlotTwist contains 75 visual riddles, accompanied by structured annotations of deceptive visual clues, diagnostic evidence, intended answers, and solution rationales. We evaluate state-of-the-art proprietary and open-weight LVLMs by first measuring their ability to solve the riddles. We then use manually annotated regions to construct visual hints that selectively emphasize either deceptive clues or diagnostic evidence. By comparing model responses before and after these interventions, we examine how strongly models persist in their initial interpretations, whether emphasizing deceptive evidence can induce incorrect answers, and whether highlighting diagnostic evidence can recover initially incorrect predictions.

Our experiments show that current LVLMs are frequently misled by plausible deceptive cues and often fail to identify or sufficiently prioritize the evidence required for the intended solution.

PlotTwist provides a systematic framework for studying how vision-language models reason under misleading visual evidence and offers a challenging benchmark for developing more robust, evidence-grounded, and deception-resistant multimodal systems. The benchmark and code are available at `https://asdafi.github.io/PlotTwist`.

1 INTRODUCTION

Large vision-language models (LVLMs) increasingly support general-purpose interaction with images through natural-language instructions Liu et al. (2023). Their capabilities are commonly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

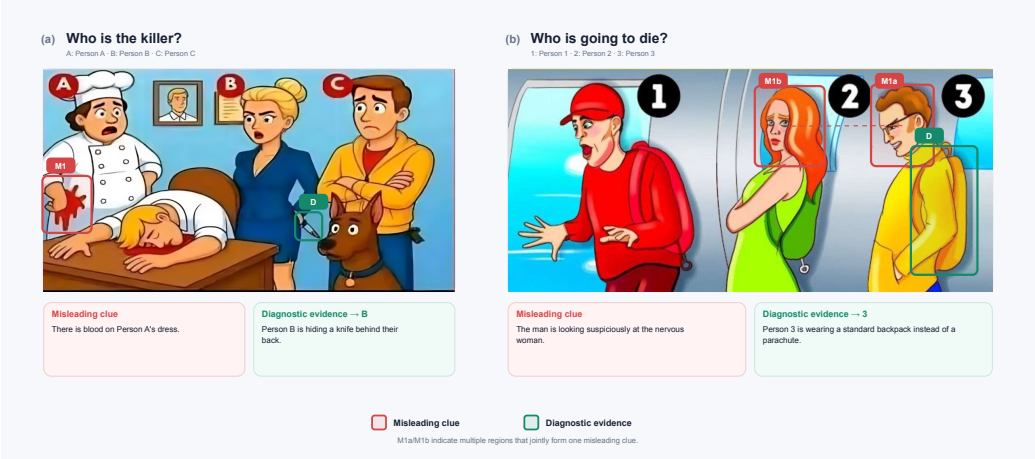


Figure 1: **Competing visual evidence in PlotTwist.** Each riddle contains locally plausible misleading clues (red) that support an incorrect interpretation, together with more diagnostic evidence (green) supporting the reference answer. A misleading clue may span multiple visual regions, as illustrated in (b). PlotTwist explicitly localizes both types of evidence so that their influence can be studied independently.

evaluated on tasks such as reading text in images Singh et al. (2019), mathematical reasoning in visual contexts Lu et al. (2024), and visual riddles that require commonsense or background knowledge Bitton-Guetta et al. (2024). Across these settings, successful reasoning requires more than recognizing objects or attributes: a model must determine which observations are relevant and how strongly they should influence the final answer.

This becomes particularly important when an image contains multiple question-relevant clues that support competing interpretations. Visual riddles provide a natural setting for studying this problem. A conspicuous clue may make an incorrect answer appear plausible, while a different and sometimes less obvious observation provides stronger evidence for another conclusion. We refer to a clue as *misleading* when it supports a plausible incorrect interpretation in isolation but is outweighed by more diagnostic evidence in the complete scene. Importantly, such a clue need not be visually misperceived or irrelevant to the question: a model may recognize it correctly and still assign it excessive evidential weight.

This motivates our central question: *Can LVLMs identify and prioritize diagnostic visual evidence when other, semantically relevant clues support a tempting alternative?*

Figure 1 illustrates this evidential structure in two examples from our benchmark. In each case, the misleading clue is visually meaningful and provides a plausible basis for an incorrect interpretation, while the diagnostic clue provides stronger evidence for the intended answer.

Prior work provides several foundations for studying these failures. Visual Riddles evaluates clue-based multimodal reasoning and the use of textual hints Bitton-Guetta et al. (2024), while Misleading ChartQA and MVI-Bench study failures caused by deceptive chart designs and misleading visual inputs Chen et al. (2025; 2026). VI-Probe further examines model sensitivity to controlled visual modifications and visual guidance Sun et al. (2026). These works establish that multimodal models can be sensitive to misleading visual information and external guidance. Our focus is complementary: rather than treating the misleading signal primarily as corrupted or deceptive visual content, we study cases where the misleading clue itself can be correctly perceived and semantically relevant. The challenge lies in *evidence arbitration*: determining which of several plausible observations should dominate the final interpretation.

We introduce **PlotTwist**, a diagnostic benchmark for studying this setting through visual riddles. Each riddle contains a reference answer and human annotations identifying misleading clues and diagnostic evidence. These annotations include bounding boxes that localize the relevant image

regions and allow individual pieces of evidence to be manipulated without changing the underlying scene or the intended answer.

PlotTwist combines two complementary sources of examples. The first consists of visual riddles collected from online sources and reviewed by humans. The second consists of newly developed riddles built from human-designed scenarios and rendered with the aid of generative models. In the latter case, humans specify the question, candidate answers, intended solution, misleading clues, and diagnostic evidence, while generative models support scene construction and image rendering. Both sources are reviewed under the same inclusion and annotation criteria.

Beyond evaluating models on the original images, we use the annotated regions to construct controlled visual-hint interventions. Image markings have been used previously to direct multimodal models toward particular visual regions Shtedritski et al. (2023); Yang et al. (2023). We use this mechanism diagnostically: rather than merely helping a model localize an object, we selectively emphasize either misleading or diagnostic evidence and observe how the model's answer changes.

This produces two complementary probes. First, highlighting a misleading clue allows us to test whether an initially correct prediction can be displaced by a plausible competing interpretation. Second, highlighting diagnostic evidence allows us to test whether an initially incorrect prediction can be recovered when attention is redirected toward the evidence supporting the reference answer. Since the intervention changes only the visual emphasis of existing content, the underlying question and intended answer remain fixed.

We evaluate state-of-the-art proprietary and open-weight LVLMs from multiple model families and parameter scales. Our analysis first measures baseline riddle-solving performance and then examines answer transitions under misleading and diagnostic visual hints. This formulation allows us to move beyond aggregate accuracy and study how strongly model decisions depend on which evidence is made visually salient.

Our contributions are threefold:

- **A benchmark for reasoning under competing visual evidence.** We introduce PlotTwist, a collection of visual riddles organized around misleading and diagnostic clues. The benchmark combines reviewed online riddles with human-designed, model-rendered scenarios and provides human-annotated clue regions for controlled analysis.
- **A visual-hint protocol for probing evidence arbitration.** We selectively emphasize annotated misleading and diagnostic clues while keeping the underlying scene and answer unchanged, enabling direct measurement of misleading-induced errors, diagnostic recovery, and persistence of model interpretations.
- **A comparative diagnostic evaluation of contemporary LVLMs.** We evaluate proprietary and open-weight models across multiple families and scales, analyzing baseline accuracy together with their sensitivity to competing visual evidence.

2 Related Work

Visual riddles and detective reasoning. Visual riddle benchmarks evaluate whether multimodal models can integrate visual observations with reasoning and background knowledge to resolve nontrivial scenarios. *Visual Riddles* presents image-based puzzles paired with questions, explanations, and textual hints, including controlled modifications of critical visual cues Bitton-Guetta et al. (2024). *DailyClue*, introduced in *Seek-and-Solve*, evaluates clue-driven reasoning in everyday scenes and studies how access to explicit clues affects model performance Li et al. (2026). Related benchmarks investigate abstract visual reasoning, reasoning with reduced dependence on specialized knowledge, and explanations of commonsense violations Yan et al. (2025); Song et al. (2025); Bitton-Guetta et al. (2023).

CaughtCheating frames visual reasoning as a detective task in which models must actively search for subtle, context-dependent suspicious evidence and integrate multiple observations to infer hidden events Li et al. (2025). PlotTwist differs by explicitly constructing cases where one question-relevant clue can plausibly support an incorrect interpretation while other evidence supports the

intended resolution, requiring arbitration among competing clues rather than only discovery of relevant evidence.

Diagnostic evaluation of perception and reasoning. A growing body of work moves beyond aggregate task accuracy to diagnose specific capabilities and failure sources in visual reasoning. *CLEVR* introduced a controlled diagnostic benchmark for compositional visual reasoning, using functional programs and question families to probe capabilities such as attribute recognition, counting, comparison, spatial reasoning, and logical operations Johnson et al. (2017). *PuzzleVQA* evaluates reasoning over abstract visual patterns and diagnoses errors by progressively providing intermediate information for visual perception, inductive reasoning, and deductive reasoning Chia et al. (2024). *VRIQ* separates perceptual and reasoning failures through hierarchical diagnostic probes and further analyzes capabilities such as shape, counting, position, and depth perception Khezresmaeilzadeh et al. (2026). *BLINK* isolates fundamental visual-perception abilities, including relative depth, visual correspondence, object localization, counting, and multi-view reasoning Fu et al. (2024). *Do You See Me* provides controlled diagnostics over 2D and 3D visual perception tasks Kanade & Ganu (2026). At a broader level, *MMMU* evaluates expert-level multimodal problems requiring the joint use of perception, domain knowledge, and reasoning Yue et al. (2024), while *Perception Test* organizes video understanding around multiple perceptual and reasoning skills Patraucean et al. (2023).

Reasoning under misleading visual information. Several recent benchmarks investigate cases in which visual information encourages an incorrect interpretation. *Misleading ChartQA* evaluates reasoning over deliberately misleading visualizations, covering deceptive chart-design practices and providing annotations of misleading regions and explanations Chen et al. (2025). Its region-aware pipeline investigates whether explicitly identifying misleading chart regions improves downstream reasoning. PlotTwist differs because its misleading clue can be a correctly depicted and semantically valid observation; the challenge is determining how strongly it should be weighted relative to other evidence.

MVI-Bench evaluates LVLM robustness to misleading visual inputs across visual concepts, attributes, and relationships, including resemblance, representation confusion, material confusion, reflections, occlusion, and visual illusions Chen et al. (2026). In PlotTwist, a misleading clue may itself be correctly perceived; failure can instead occur during reasoning when a locally plausible clue is weighted more strongly than a more diagnostic one. Thus, PlotTwist tests comparison and prioritization of meaningful evidence, rather than only resistance to misleading visual content.

PlotTwist additionally evaluates *interpretation persistence* through controlled visual-hint interventions. Humans annotate regions corresponding to misleading and diagnostic clues, which are selectively highlighted to measure how strongly models persist in their interpretation when attention is redirected toward competing evidence.

Grounded explanations and abductive inference. Prior work has studied how models connect localized visual evidence to hypotheses and explanations. *Visual Commonsense Reasoning* jointly evaluates answer and rationale selection and uses adversarial matching to construct challenging alternatives Zellers et al. (2019). *Grounding Visual Explanations* investigates whether linguistic explanations are supported by relevant image regions and explores counterfactual explanations for alternative predictions Hendricks et al. (2018).

Sherlock formulates visual abductive reasoning around localized clues and plausible inferences, evaluating evidence localization, inference retrieval, and ranking against human judgments Hessel et al. (2022). PlotTwist differs by explicitly placing clues with different diagnostic value in competition: the model must determine not only which evidence supports an inference, but which evidence should dominate when plausible interpretations conflict.

Shortcut learning and spurious correlations. Shortcut learning concerns models exploiting predictive decision rules that succeed under a particular data distribution without capturing the intended solution strategy Geirhos et al. (2020). A canonical manifestation is a spurious correlation, in which a target becomes associated with a nuisance or non-essential feature. The Waterbirds setting, for example, exposes reliance on backgrounds correlated with bird classes and motivates evaluation on groups where the correlation changes Sagawa et al. (2020). WILDS broadens this perspective by

evaluating models under naturally occurring distribution shifts across several real-world applications Koh et al. (2021).

Visual question answering has also served as a major testbed for shortcut learning. VQA v2 reduces question-conditioned language priors by pairing similar visual situations with different answers Goyal et al. (2017). VQA-CP changes question–answer distributions between training and testing to expose dependence on learned answer priors Agrawal et al. (2018), while VQA-CE constructs counterexamples to predictive rules involving both visual and linguistic features Dancette et al. (2021). VQA-VS extends this setting to multiple shortcut types and out-of-distribution test splits Si et al. (2022). RUBi approaches the problem from the mitigation side by reducing reliance on unimodal question biases during training Cadene et al. (2019).

Recent work examines shortcuts in contemporary vision and vision–language models. UrbanCars and ImageNet-W show that multiple visual shortcuts can coexist and that suppressing one may increase reliance on another Li et al. (2023). RaVL investigates the discovery and mitigation of fine-grained spurious visual–textual correlations in fine-tuned vision–language models Varma et al. (2024). SpuriVerse extends spurious-correlation evaluation to large vision–language models by identifying diverse shortcut patterns from real-world VQA failures and validating them through counterfactual examples Yang et al. (2025).

Spurious correlations typically arise from statistical feature–label associations that are predictive in the observed data but unreliable for the intended task. In contrast, PlotTwist's misleading clues are semantically relevant and can reasonably support an interpretation on their own. The challenge is therefore not to ignore these clues, but to compare competing evidence and prioritize the most diagnostic clues within the same instance.

Visual hints and evidence interventions. Prior work has investigated whether explicit visual guidance can alter how models use image evidence. HINT encourages models to align feature importance with human visual attention during training Selvaraju et al. (2019). At inference time, visual prompting approaches instead modify the image to direct attention toward particular regions. Red-circle prompting studies the effect of simple visual markings Shtedritski et al. (2023), while Set-of-Mark prompting overlays referenceable marks onto visual regions to facilitate grounding and interaction Yang et al. (2023). *VI-Probe* examines model behavior under controlled visual modifications and studies the effect of visual guidance on response stability Sun et al. (2026).

PlotTwist uses visual hints over human-annotated regions corresponding to individual pieces of evidence, including misleading and diagnostic clues. By selectively highlighting these regions, we evaluate how strongly a model persists in its interpretation when attention is explicitly redirected toward competing visual evidence.

3 PLOTTWIST

PlotTwist evaluates how large vision-language models (LVLMs) resolve competing visual evidence. Each riddle contains diagnostic evidence supporting a reference answer and one or more misleading clues that make an incorrect answer plausible. These clues can be correctly perceived and relevant to the question: the challenge is to determine which evidence supports the answer.

3.1 DATASET CONSTRUCTION

The dataset contains 75 riddles: 38 collected online and 37 developed from human-designed scenarios. Both sources follow the same inclusion criteria and annotation scheme.

Inclusion criteria. A riddle must contain a visually identifiable clue supporting a plausible incorrect interpretation and discernible diagnostic evidence supporting a single defensible reference answer. The solution should not require arbitrary assumptions about unobserved events. Human reviewers assess solvability, answer ambiguity, and the plausibility of competing interpretations; candidates are revised or excluded when these criteria are not met.

Online riddles. We collect visual riddles from public puzzle and brain-teaser sources and retain their source URLs. Human reviewers identify the reference answer, diagnostic evidence, and mis-

leading clues. Where needed, revisions preserve the underlying reasoning problem while improving evidence visibility and answer defensibility.

Human-designed scenarios. Human authors specify the question, choices, reference answer, diagnostic evidence, and misleading clues before generation. Generative models assist with scene composition and rendering. Humans inspect and revise the images to check that intended clues are visible and that generation artifacts do not introduce unintended shortcuts or alternative solutions. Thus, humans define and review the reasoning structure while models support its visual realization.

3.2 Clue Annotations

Human annotations include the question, candidate answers, reference answer, and an explanation of its supporting evidence. One or more bounding boxes localize the diagnostic evidence. Each misleading clue has its own associated incorrect answer, explanation, and bounding boxes. Several clues may support the same incorrect answer, and a single clue may span multiple regions. Clue counts, box counts, and reasoning steps are therefore distinct quantities. The formal schema and annotation distributions appear in Appendix B.

4 Evaluation Protocol

4.1 Models and Baseline Evaluation

Models. We evaluate fourteen proprietary and open-weight LVLMs spanning several model families and scales. The proprietary models include GPT-5.6 Sol, GPT-4o, and GPT-4o mini OpenAI (2026; 2024a;b); Gemini 2.5 Pro and Gemini 2.5 Flash Google (2025); and Claude Sonnet 5 Anthropic (2026).

The open-weight models include Llama 4 Maverick and Llama 4 Scout Meta AI (2025); Qwen3-VL 8B, 32B, and 235B Bai et al. (2025a); Qwen2.5-VL 72B Bai et al. (2025b); Mistral Small 3.2 Mistral AI (2025); and Gemma 3 27B Gemma Team et al. (2025). This selection provides coverage across multiple organizations, model families, and model scales rather than focusing exclusively on flagship proprietary systems.

Inference and scoring. All requests use OpenRouter OpenRouter (2026). Each request supplies an image, question, and candidate answers and asks for a brief explanation and a single choice identifier in separate tagged fields. Accuracy uses strict matching of the extracted identifier; missing or invalid identifiers count as incorrect. The exact prompt appears in Appendix A.1.

The baseline run contains 1,050 completed requests across fourteen models and 75 riddles. Only 74 examples have finalized scorable reference choices, yielding 1,036 scored predictions. The table values are consistent with 38 scored online and 36 scored human-designed examples; the remaining human-designed example is excluded from these accuracy calculations. We report source-specific accuracy and *source-macro accuracy*, the unweighted mean of the two source accuracies. Pooled accuracy instead weights each scored model–riddle prediction equally. Operational measurements are reported in Appendix E.

4.2 Targeted Visual Hints

We next vary the salience of annotated evidence while retaining the question. For an initially correct prediction, we highlight the first annotated misleading clue. For an incorrect or invalid prediction, we highlight diagnostic evidence. Small regions are additionally enlarged in an edge panel and linked to their source locations. For model m and riddle i, the second-pass image is

$$I_{mi}^{(1)} = \begin{cases} H_{\mathrm{mis}}(I_i), & \hat{y}_{mi}^{(0)} = y_i, \\ H_{\mathrm{diag}}(I_i), & \text{otherwise.} \end{cases} \tag{1}$$

Here H_{mis} highlights the first misleading clue's boxes and H_{diag} highlights the diagnostic boxes. This asymmetric protocol tests resistance to misleading guidance on baseline successes and recovery under diagnostic guidance on baseline failures. The protocol diagram appears in Appendix D.

4.3 Transition Metrics

Let V_t indicate a valid choice at pass t, and let C_t indicate a correct answer. Invalid answers have $C_t = 0$. For each model and source, we report strict accuracy and

$$R_{\text{corr}} = P(C_1 = 1 \mid C_0 = 0), \tag{2}$$

$$R_{\text{reg}} = P(C_1 = 0 \mid C_0 = 1), \tag{3}$$

$$P_{\text{correct}} = P(\hat{y}^{(1)} = \hat{y}^{(0)} \mid C_0 = 1, V_1 = 1), \tag{4}$$

$$P_{\text{wrong}} = P(\hat{y}^{(1)} = \hat{y}^{(0)} \mid C_0 = 0, V_0 = 1, V_1 = 1). \tag{5}$$

Correction and regression include invalid outputs; persistence excludes invalid revised outputs and, for P_{wrong}, invalid baseline outputs. Consequently, P_{correct} need not equal $1 - R_{\text{reg}}$. Wrong-answer persistence is also not the complement of correction: a model can switch to another incorrect choice. Empty conditioning sets yield undefined rates and must be reported with their support counts.

For a fixed cohort, strict accuracy satisfies

$$A_1 - A_0 = (1 - A_0)R_{\text{corr}} - A_0 R_{\text{reg}}. \tag{6}$$

This identity applies before averaging across models; products of separately averaged rates do not generally reproduce the average accuracy change.

5 Results

5.1 Baseline Performance

Across models, 434 of 1,036 scored predictions are correct (41.9%). GPT-5.6 Sol has the highest observed source-macro accuracy (65.1%), followed by Gemini 2.5 Pro (63.5%) and Gemini 2.5 Flash (55.3%). All remaining models score below 45%. These are descriptive rankings on a small dataset; the reported differences do not establish statistical separation. Figure 2 summarizes the results, and Appendix C gives the full table and source-accuracy heatmap.

Dataset sources. Pooled accuracy is 37.4% on online riddles and 46.6% on human-designed scenarios. Eleven of fourteen models perform better on the latter, whereas the two Gemini models and GPT-4o mini perform better on online riddles. Because content and construction differ between subsets, these differences cannot be attributed to generation method alone. Empirical sample difficulty is reported in Appendix C.2.

5.2 Effects of Visual Hints

Table 1 preserves the supplied intervention summary. Its prior accuracies (36.0% online; 45.4% human-designed) differ from the baseline table (37.4%; 46.6%). The available aggregate results do not establish whether this reflects different cohorts, exclusions, or runs. We therefore treat these as separate reported snapshots pending reconciliation, rather than a verified paired comparison to the baseline table.

Within the reported hint snapshot, mean accuracy increases by 12.8 percentage points on online riddles and 5.6 points on human-designed scenarios. Correction rates are similar (56.2% and 55.1%), but regression is also substantial (62.5% and 53.6%). A net gain therefore does not imply reliable resistance to misleading guidance: recovery and loss of correct answers occur in different baseline-conditioned groups.

5.3 Model-Level Revision Behavior

The reported examples illustrate why final accuracy alone is insufficient. On human-designed scenarios, GPT-5.6 Sol declines from 72.2% to 52.8%, whereas Claude Sonnet 5 improves from 52.8% to 63.9%, with 31.6% regression. On online riddles, Qwen2.5-VL 72B improves from 18.4% to 63.2% despite a 71.4% regression rate: substantial recovery coexists with losses among its initially correct answers. These model-level values are retained from the supplied draft and require verification against the same paired cohort used for the aggregate hint table. Figures 3 and 4 show the model-level accuracy changes and directional transition rates, respectively.

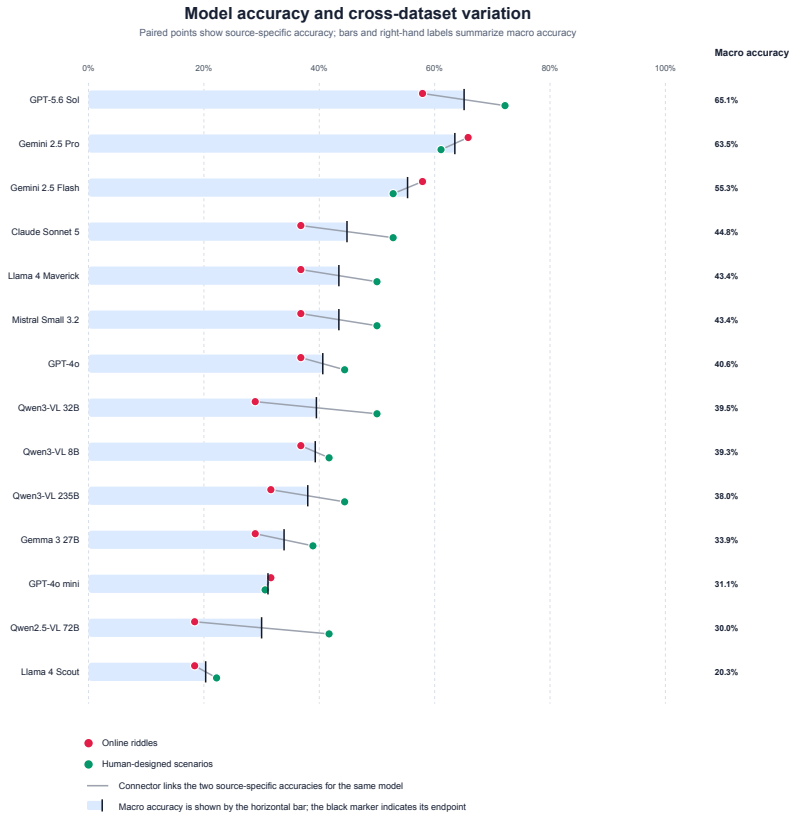


Figure 2: Baseline accuracy across evaluated LVLMs. Source-specific accuracies are shown alongside their unweighted mean (source-macro accuracy).

Table 1: Reported visual-hint snapshot, averaged across fourteen models (percent). These prior accuracies differ from the baseline snapshot in Table 2; cohort reconciliation is pending. Persistence after correct answers measures robustness to misleading evidence, whereas persistence after wrong answers measures undesirable stubbornness despite diagnostic evidence.

Metric (%)	**Human-designed**	**Online**
Prior accuracy	45.4	36.0
Revised accuracy	51.0	48.8
Change (percentage points)	+5.6	+12.8
P_{correct}	48.3	39.0
P_{wrong}	38.3	31.9
R_{corr}	55.1	56.2
R_{reg}	53.6	62.5

6 LIMITATIONS

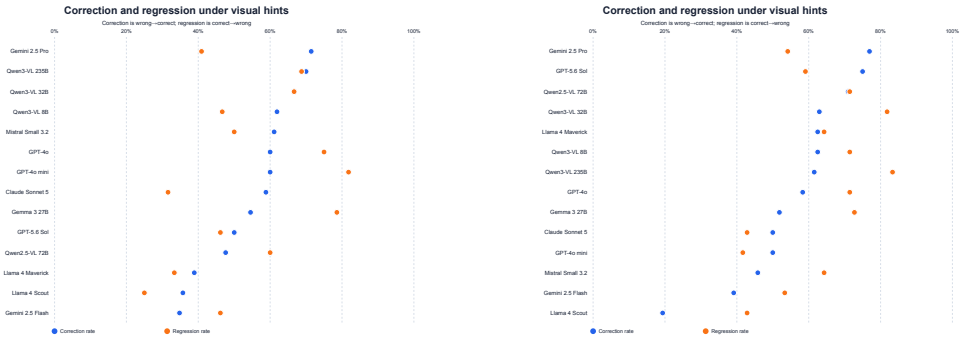
The benchmark is small and includes one example without a finalized scorable reference choice in the baseline snapshot. Source differences combine content and construction effects. Online riddles may also have appeared in model training data; source provenance alone does not establish absence of exposure.

The intervention changes visual presentation, including enlargement for some regions, and is selected using baseline correctness. It therefore characterizes conditional sensitivity to the implemented hints, rather than isolating a causal effect of deception from salience or enlarged visibility. Models can also receive different hint types on the same riddle. Comparisons of transition rates must account for their different eligible examples and sample counts.



(a) Human-designed scenarios (b) Online riddles

Figure 3: **Accuracy before and after targeted visual hints.** Correct baseline predictions receive a misleading visual hint, whereas incorrect or invalid predictions receive diagnostic evidence. Per-model values require verification against the reconciled hint cohort. Invalid answers are counted as incorrect.



(a) Human-designed scenarios (b) Online riddles

Figure 4: **Correction–regression trade-off under targeted visual hints.** Correction measures recovery after diagnostic evidence is highlighted; regression measures loss of an initially correct prediction after misleading evidence is highlighted. Correction and regression must be interpreted jointly.

Answer accuracy does not establish whether a model perceived every clue or used the intended reasoning. Likewise, agreement between two predictions measures answer stability; interpreting it as persistence of a prior belief requires knowing whether the second request included the earlier interaction. Neutral-hint controls, matched enlargement controls, repeated runs, and human comparisons would help distinguish these effects.

7 Conclusion

PlotTwist studies visual riddles in which plausible misleading clues compete with diagnostic evidence. The baseline results show substantial room for improvement across fourteen LVLMs. The reported hint experiments suggest that visual guidance can both recover incorrect answers and overturn correct ones. Evaluating correction together with regression distinguishes responsiveness to guidance from resistance to misleading evidence. Larger datasets and controlled, paired evaluations are needed to establish how these behaviors generalize.

AI Use Disclosure

AI-assisted tools were used in this research for manuscript drafting and language refinement, review and feedback, literature search, development of experimental code, and support for data generation.

In particular, generative models were used to assist with the visual realization of some human-designed riddle scenarios. All such uses were paired with human review, revision, and validation. The authors retained responsibility for the research questions, scenario design, annotations, experimental protocol, analysis, interpretation, and final manuscript.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1712.00377.

Anthropic. Introducing Claude Sonnet 5, 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.07274.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. Visual Riddles: A commonsense and world knowledge challenge for large vision and language models, 2024. URL https://arxiv.org/abs/2407.19474.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases in visual question answering. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.10169.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. MVI-Bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in LVLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2511.14159.

Zixin Chen, Sicheng Song, KaShun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13756–13789. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.695. URL https://aclanthology.org/2025.emnlp-main.695/.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.03149.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Google. Gemini 2.5 Model Family, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-2-5-model-family-expands/.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, 2017. URL https://arxiv.org/abs/1612.00837.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Computer Vision – ECCV 2018*, 2018. URL https://arxiv.org/abs/1807.09685.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of Sherlock Holmes: A dataset for visual abductive reasoning. In *Computer Vision – ECCV 2022*, 2022. URL https://arxiv.org/abs/2202.04800.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Fei-Fei Li, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://cs.stanford.edu/people/jcjohns/clevr/.

Aditya Sanjiv Kanade and Tanuja Ganu. Do you see me: A multidimensional benchmark for evaluating visual perception in multimodal LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

Tina Khezresmaeilzadeh, Jike Zhong, and Konstantinos Psounis. VRIQ: Benchmarking and analyzing visual-reasoning IQ of VLMs, 2026. URL https://arxiv.org/abs/2602.05382.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021.

Ming Li, Chenguang Wang, Yijun Liang, Xiyao Wang, Yuhang Zhou, Xiyang Wu, Yuqing Zhang, Ruiyi Zhang, and Tianyi Zhou. Caughtcheating: Is your mllm a good cheating detective? exploring the boundary of visual perception and reasoning. *arXiv preprint arXiv:2507.00045*, 2025.

Xiaomin Li, Tala Wang, Zichen Zhong, Ying Zhang, Zirui Zheng, Takashi Isobe, Dezhuang Li, Huchuan Lu, You He, and Xu Jia. Seek-and-Solve: Benchmarking MLLMs for visual clue-driven reasoning in daily scenarios, 2026. URL https://arxiv.org/abs/2604.14041.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20071–20082, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.08485.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02255.

Meta AI. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Mistral AI. Mistral Small 3.2. Mistral AI Model Documentation, 2025. URL https://docs.mistral.ai/models/mistral-small-3-2-25-06.

OpenAI. GPT-4o System Card, 2024a. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-4o mini: Advancing Cost-Efficient Intelligence, 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. GPT-5.6 Sol Model. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

OpenRouter. Openrouter pricing and billing. OpenRouter Documentation, 2026. URL https://openrouter.ai/pricing. Accessed September 2026.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1902.03751.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2304.06712.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.08920.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. VisualPuzzles: Decoupling multimodal reasoning evaluation from domain knowledge, 2025. URL https://arxiv.org/abs/2504.10342.

Xiaoxiao Sun, Mingyang Li, Kun Yuan, Min Woo Sun, Mark Endo, Shengguang Wu, Changlin Li, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Do VLMs perceive or recall? probing visual perception vs. memory with classic visual illusions, 2026. URL https://arxiv.org/abs/2601.22150.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. RaVL: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hao Yan, Xingchen Liu, Hao Wang, Zhenbiao Cao, Handong Zheng, Liang Yin, Xinxing Su, Zihao Chen, Jihao Wu, Minghui Liao, Chao Weng, Wei Chen, Yuliang Liu, and Xiang Bai. VisuRiddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning, 2025. URL https://arxiv.org/abs/2506.02537.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V, 2023. URL https://arxiv.org/abs/2310.11441.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Z. Liu, Yulia Tsvetkov, and Bill Howe. Escaping the SpuriVerse: Can large vision-language models generalize beyond seen spurious correlations? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

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

A Evaluation Details

A.1 Evaluation Prompt

All models are evaluated using the same multiple-choice prompt template. For each example, {question} and {choices} are replaced with the corresponding riddle question and answer choices, and the associated image is supplied together with the prompt. Each model–riddle pair is evaluated independently, without conversational history.

```
Look at the image carefully and answer the following
multiple-choice question.

Question:
{question}

Choices:
{choices}

Analyze the visual evidence and determine the correct choice.

Important:
- The correct answer is always one of the provided choices.
- Select exactly one choice ID from the choices above.
- Do not create a new choice ID.
- Do not return the choice text, only the choice ID.

Return your response exactly in the following format:

<reasoning>
Provide a concise explanation of the visual evidence and why
```

```
the selected choice is correct.
</reasoning>

<answer>
Return ONLY the ID of the selected choice.
</answer>
```

The <reasoning> field is retained for qualitative analysis, while benchmark accuracy is determined from the choice identifier returned in <answer>. Choice identifiers are kept in their original form (e.g., A/B/C or 1/2/3) rather than normalized across examples. Missing answers, invalid identifiers, or outputs from which a valid choice ID cannot be extracted are counted as incorrect.

A.2 Inference Configuration

Requests are issued through OpenRouter. The supplied manuscript does not specify exact API model IDs and provider routes, evaluation dates, temperature, reasoning settings, output limits, retry behavior, or second-pass conversation history. These details must be recovered from the experiment configuration before the evaluation is fully reproducible.

B Annotation Schema and Distributions

Notation. Let the dataset be

$$\mathcal{D} = \{r_i\}_{i=1}^{N}.$$

Each riddle is represented as

$$r_i = (I_i, q_i, \mathcal{C}_i, y_i, e_i, \mathcal{B}_i^{\text{diag}}, \mathcal{M}_i), \tag{7}$$

where I_i is the image, q_i is the question, $\mathcal{C}_i$ is the set of candidate answers, $y_i \in \mathcal{C}_i$ is the reference answer, and e_i is its explanation. The set $\mathcal{B}_i^{\text{diag}}$ contains the bounding boxes localizing diagnostic evidence.

The misleading-clue annotations are represented as

$$\mathcal{M}_i = \{m_{ij}\}_{j=1}^{K_i}, \qquad m_{ij} = (\widetilde{y}_{ij}, \rho_{ij}, \mathcal{B}_{ij}^{\text{mis}}), \tag{8}$$

where K_i is the number of misleading clues, $\widetilde{y}_{ij} \in \mathcal{C}_i \setminus \{y_i\}$ is the incorrect answer associated with clue j, ρ_{ij} explains the misleading inference, and $\mathcal{B}_{ij}^{\text{mis}}$ contains the corresponding bounding boxes.

C Additional Baseline Results

C.1 Source-Specific Accuracy

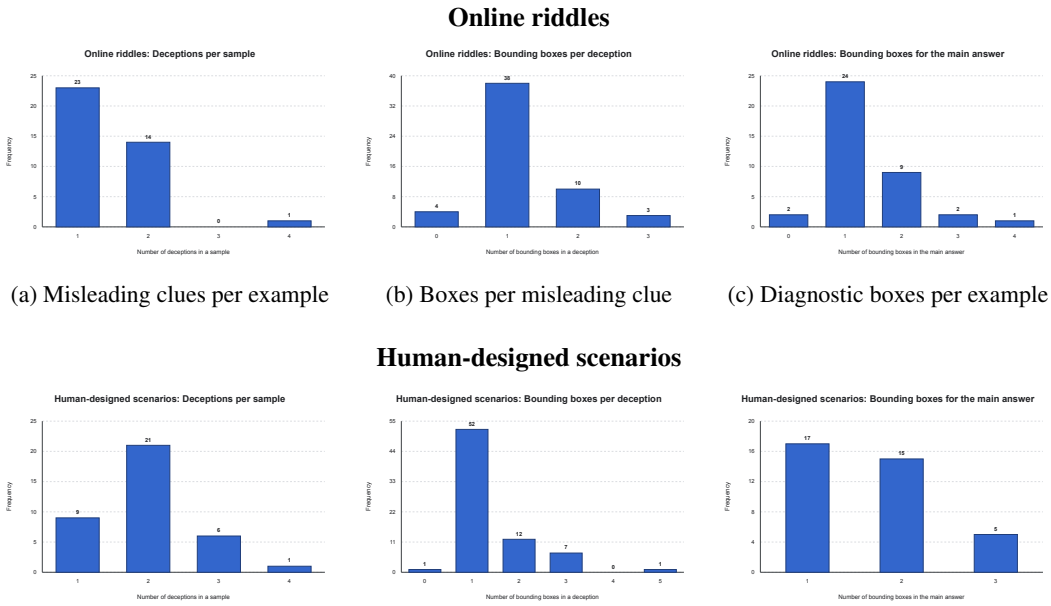
Figure 6 gives the source-specific accuracies in heatmap form.

C.2 Sample Difficulty

Aggregate accuracy does not reveal whether failures arise from only a small number of exceptionally difficult riddles. We therefore characterize the empirical difficulty of each example using the fraction of the fourteen evaluated models that answer it correctly.

As shown in Figure 7, difficult examples are distributed throughout the benchmark. Among online riddles, 60.5% are answered correctly by fewer than 40% of the evaluated models. The corresponding fraction for human-designed scenarios is 41.7%. Conversely, only 5.3% of online riddles and 13.9% of human-designed scenarios are solved by at least 80% of models.

This mixture of baseline successes and failures is particularly useful for our subsequent visual-hint evaluation. Initially correct predictions allow us to measure whether emphasizing misleading evidence can induce an error, whereas initially incorrect predictions provide cases in which emphasizing diagnostic evidence may recover the intended answer.



(a) Misleading clues per example (b) Boxes per misleading clue (c) Diagnostic boxes per example

(d) Misleading clues per example (e) Boxes per misleading clue (f) Diagnostic boxes per example

Figure 5: **Annotation distributions in PlotTwist.** The top row shows online riddles and the bottom row shows human-designed scenarios. From left to right, we report the number of misleading clues per example, the number of bounding boxes required to localize each misleading clue, and the number of regions associated with the diagnostic evidence. A single clue may span multiple image regions, so clue counts and bounding-box counts represent different properties of the annotations.

Table 2: Baseline performance and observed OpenRouter inference cost on the original, unmarked PlotTwist images. Macro accuracy is the unweighted mean of accuracy on the online and human-designed subsets. Cost is the mean observed inference cost per request under our evaluation configuration.

Model	Online	Human-designed	Macro Acc.	Cost/Req.
GPT-5.6 Sol	57.9	72.2	65.1	$0.00577
Gemini 2.5 Pro	65.8	61.1	63.5	$0.01027
Gemini 2.5 Flash	57.9	52.8	55.3	$0.00178
Claude Sonnet 5	36.8	52.8	44.8	$0.00438
Llama 4 Maverick	36.8	50.0	43.4	$0.00064
Mistral Small 3.2	36.8	50.0	43.4	$0.00017
GPT-4o	36.8	44.4	40.6	$0.00288
Qwen3-VL 32B	28.9	50.0	39.5	$0.00019
Qwen3-VL 8B	36.8	41.7	39.3	$0.00021
Qwen3-VL 235B	31.6	44.4	38.0	$0.00046
Gemma 3 27B	28.9	38.9	33.9	$0.00007
GPT-4o mini	31.6	30.6	31.1	$0.00373
Qwen2.5-VL 72B	18.4	41.7	30.0	$0.00128
Llama 4 Scout	18.4	22.2	20.3	$0.00029

D Visual-Hint Protocol Diagram

Figure 8 summarizes the conditional intervention. The accuracy and transition plots are retained in the main results section.

E Additional Operational Information

The baseline evaluation reports $2.41 across 1,050 requests. The rounded per-model costs in Table 2 imply $2.409, consistent with that total. Observed mean cost per request ranges from $0.00007 to

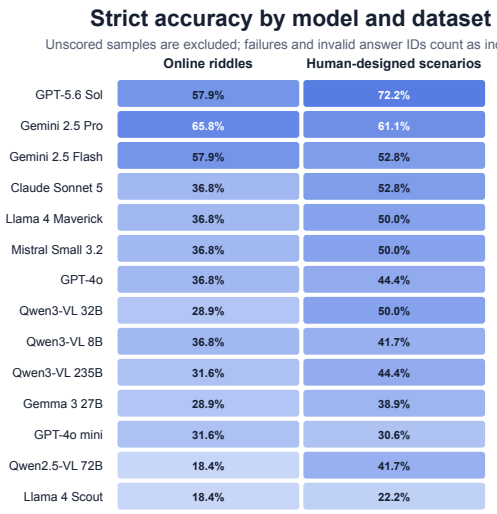


Figure 6: **Strict baseline accuracy by model and dataset source.** Source-specific performance varies across model families rather than producing a uniform change in difficulty across all evaluated LVLMs.

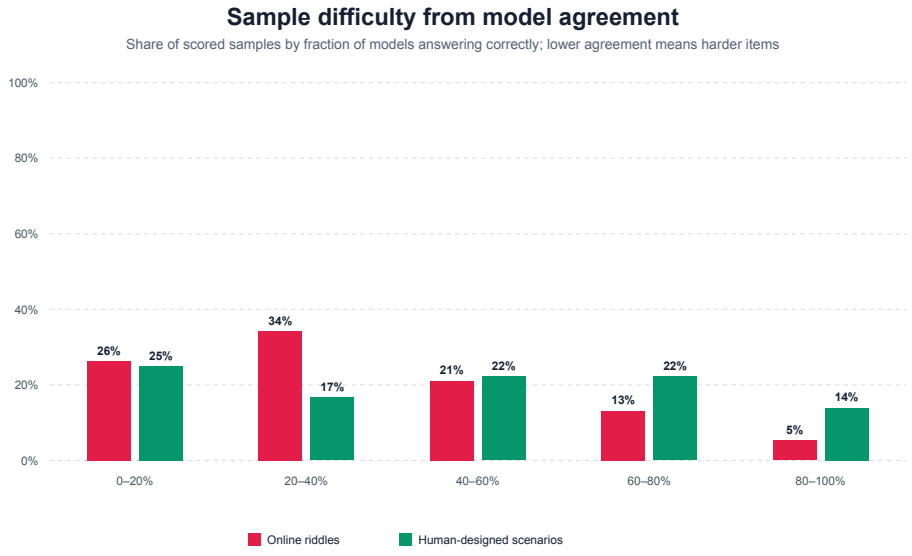


Figure 7: **Distribution of empirical riddle difficulty.** Difficulty is measured by the fraction of evaluated models answering a sample correctly. Lower-agreement bins therefore represent harder examples. Both sources contain a substantial low-agreement subset, with online riddles more concentrated in the harder bins.

$0.01027 and depends on provider route and evaluation date. For example, Mistral Small 3.2 and Llama 4 Maverick both reach 43.4% source-macro accuracy at different observed costs. These observations do not establish a general model-efficiency ranking.

Provider-reported token usage and latency are retained as operational metadata. Different tokenizers, image encodings, and reasoning-token accounting prevent raw token counts from serving as a uniform measure of multimodal computation.

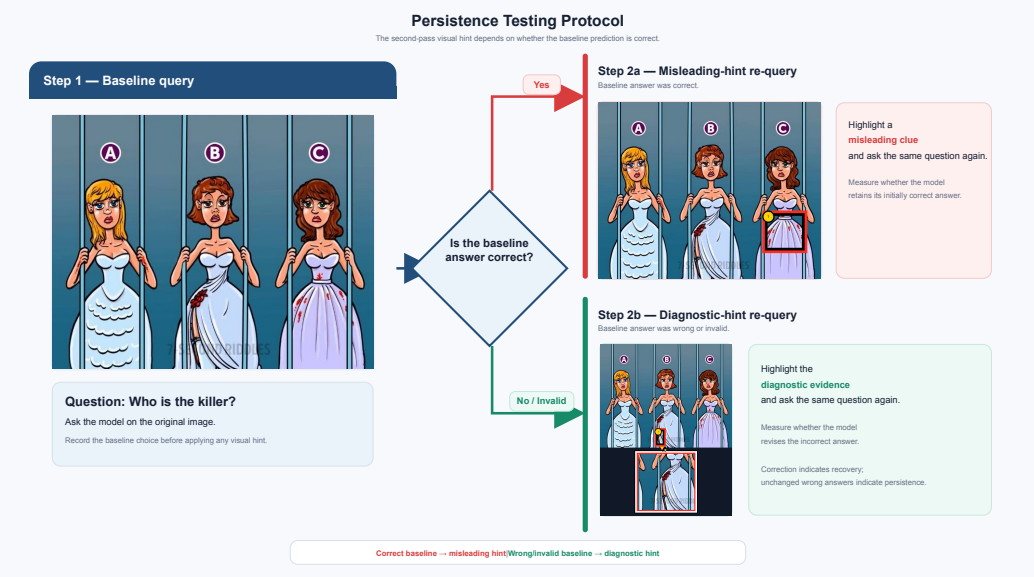


Figure 8: **Persistence-testing protocol.** Models are first evaluated on the original image. If the baseline prediction is correct, a misleading clue is highlighted and the same question is asked again to test resistance to deceptive evidence. If the baseline prediction is wrong or invalid, diagnostic evidence is highlighted and the model is re-queried to test whether it recovers the intended answer.

E.1 Accuracy and Observed Cost

Figure 9 summarizes the observed accuracy–cost operating points.

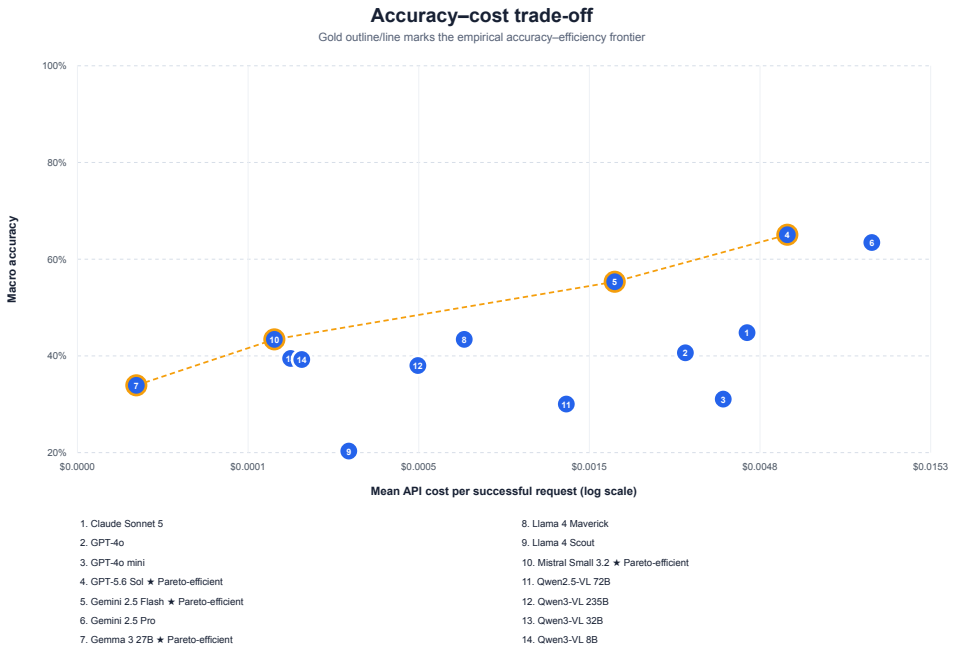


Figure 9: **Macro accuracy versus observed OpenRouter inference cost per request.** The horizontal axis is logarithmic because the observed costs span more than two orders of magnitude. Gold-outlined points and the dashed line indicate the empirical accuracy–cost frontier among the evaluated systems. Costs correspond to the OpenRouter provider routes used during these experiments.