CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments

Rishika Bhagwatkar[1,2,3]*, **Syrielle Montariol**[1]*, **Angelika Romanou**[1], **Beatriz Borges**[1], **Irina Rish**[2,3], **Antoine Bosselut**[1]

[1]EPFL, [2]Mila - Quebec AI Institute [3]Université de Montréal
Correspondence: rishika.bhagwatkar@mila.quebec, syrielle.montariol@epfl.ch

Abstract

Humans can naturally identify, reason about, and explain anomalies in their environment. In computer vision, this long-standing challenge remains limited to industrial defects or unrealistic, synthetically generated anomalies, failing to capture the richness and unpredictability of real-world anomalies. In this work, we introduce CAVE, the first benchmark of real-world commonsense anomalies. CAVE supports three open-ended tasks: anomaly description, explanation, and justification; with fine-grained annotations categorizing anomalies based on their visual manifestations, their complexity, severity, and commonness. These annotations draw inspiration from cognitive science research on how humans identify and resolve anomalies, providing a comprehensive framework for evaluating Vision-Language Models (VLMs) in detecting and understanding anomalies. Our results show that state-of-the-art VLMs struggle with visual anomaly perception and commonsense reasoning, even with advanced prompting strategies. By offering a realistic and cognitively grounded benchmark, CAVE serves as a valuable resource for advancing research in anomaly detection and commonsense reasoning in VLMs.

1 Introduction

"If you notice an abnormal situation, please contact an agent." Such announcements are commonplace in public spaces, highlighting a fundamental human trait: the ability to detect anomalies—situations that deviate from expectations [27, 25]. As Vision-Language Models (VLMs) [30, 42, 1, 29] are increasingly deployed in dynamic real-world settings [24, 56], their ability to recognize and reason about uncommon or surprising situations is crucial for safe and efficient operation [39].

Despite advances in multimodal learning, anomaly detection using VLMs remains underexplored. Existing benchmarks focus on specific domains like industrial inspection [11, 13, 3, 54], medical diagnosis [15, 61] or video surveillance [46]. More recently, commonsense-oriented anomaly detection benchmarks have started to appear. They typi-

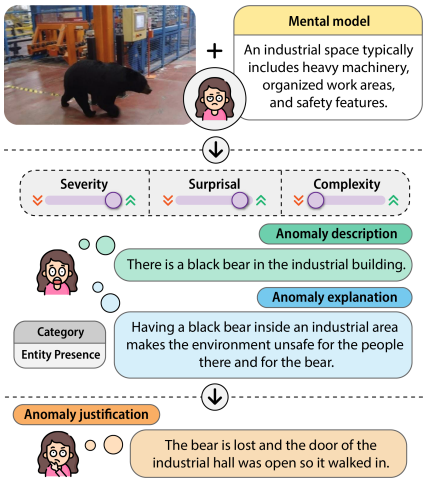


Figure 1: **CAVE Example:** a real-world image annotated with commonsense anomaly descriptions, explanations and justifications, as well as numerical features representing how humans perceive these anomalies.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

cally rely on synthetic image generation to create artificial scenarios [6, 31, 45, 5, 49]. Non-synthetic approaches rely on domain-specific datasets, such as understanding creative elements in advertisements [36] or detecting video game glitches [48]. As a result, existing benchmarks fail to capture the diversity, unpredictability, and realism of real-world anomalies, leaving a critical gap in the evaluation of VLMs' true anomaly detection capabilities.

In this work, we introduce **C**ommonsense **A**nomalies in **V**isual **E**nvironments (CAVE), the first visual anomaly benchmark curated from images captured from a human perspective, in real-life settings or as screenshots from smartphones and laptops.

Building on top of the cognitive science literature regarding the way humans identify and understand anomalies, we propose a **multi-task anomaly understanding framework**. We split the detection process into three open-ended tasks that align with human anomaly detection and sense-making processes: anomaly description, anomaly explanation, and anomaly justification. We also categorize anomalies based on the type of visual reasoning required to identify them (*e.g.*, spatial or attribute reasoning) and further label them with three numerical features: severity, surprisal or rarity, and detection complexity (see Figure 1).

This novel framework allows for systematic characterization and annotation of commonsense visual anomalies, allowing us to propose CAVE, a benchmark curated from Reddit comprising 361 images designed to evaluate VLMs' ability to detect and understand anomalies (Section 3). It captures a wide range of anomalies varying in visual manifestation, commonness, severity, and complexity. We evaluate 3 proprietary models and 5 open-source state-of-the-art models on CAVE, experimenting with 5 advanced prompting strategies (Section 4). We show that the best model, GPT-4o, only reaches 56% F1-score on anomaly detection with a multi-step reasoning strategy, highlighting significant room for improvement. We analyze VLMs' success and failure modes, finding that they perform better on surprising and severe anomalies but struggle with anomalies involving complex visual perception abilities, especially spatial reasoning and pattern detection; and that they lack the commonsense knowledge and reasoning ability to accurately identify anomalies.

2 Theoretical Framework

Leveraging cognitive science literature, we formalize how humans detect and understand anomalies into tasks. This framework guides our dataset creation process, model assessment, and analysis, enabling us to explore the alignment between human and machine processing of visual anomalies.

Perception of the anomaly. In this work, we define an anomaly not simply as a statistical rarity [17, 11, 43] but as a situation that disrupts an established pattern or expectation [27, 25]. This perspective underscores the key human ability to construct *mental models* of the world and identify deviations from these models [28]. The process of identifying this anomaly depends on three main characteristics:

- **Anomaly complexity.** Visually complex anomalies require greater cognitive resources for processing [14, 18, 47]. We leverage this formalism to assess the difficulty of detecting anomalies.
- **Anomaly severity.** Anomalies that signal immediate danger or high risk are more likely to be detected. Humans use both cognitive appraisal and emotional arousal to assess severity [44]. Hence, we operationalize severity by asking to what extent the anomaly requires immediate action.
- **Anomaly surprisal.** Surprise-based theories focus on how much an event updates prior beliefs (Bayesian) [23] or the amount of unexpected information it contains (Information-theoretic) [2]. We quantify surprisal with "How much does the situation deviate from expectations?".

We use these three formalizations to assess how humans perceive and detect anomalies. Similarly to [8], we posit that there are commonalities in the way humans and machines process visual information, and evaluate VLMs' ability to detect anomalies depending on these features.

Understanding of the anomaly. When detecting an anomaly, humans seek to understand it via three key steps. **(a) Description:** Identifying which elements violate expectations [27, 25]. **(b) Explanation:** Revisiting mental models to understand why the situation appears anomalous [26, 21], highlighting the model's understanding of the underlying commonsense knowledge. **(c) Justification:** Providing plausible explanations about the sequence of events that led to the anomalous situation, highlighting the model's sense-making ability [53, 62, 26]. Unlike synthetic datasets, with staged anomalies, each image in CAVE represents a real-world anomaly, naturally encouraging this multi-step interpretive process.

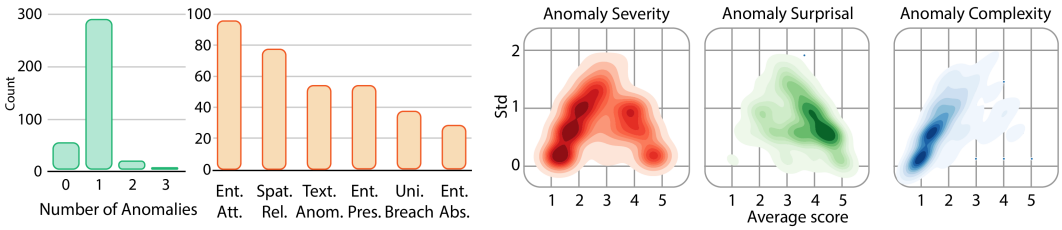


Figure 2: **CAVE statistics.** Distribution of the number of anomalies per image (left). Number of images in each anomaly category (middle). Density of severity, surprisal and complexity scores per average score and standard deviation (right).

Manifestation of the anomaly. Anomalies manifest in distinct visual ways. Inspired by MMBench [33], we categorize them into: **(a) Entity Presence/Absence**: Unexpected presence or absence of an object. **(b) Entity Attribute**: Anomalous object traits like shape, color, or usage. **(c) Spatial Relation**: Incorrect spatial positioning between objects. **(d) Uniformity Breach**: Breaks in expected visual patterns or regularities. **(e) Textual Anomaly**: Contradictory or surprising information in image text. This taxonomy complements the cognitive framework and supports our fine-grained benchmarking of VLM anomaly reasoning across varied contexts.

3 Dataset

Dataset Construction. To build CAVE, we curated real-world images from subreddits featuring unusual or surprising content. After filtering out unclear or inappropriate samples, we conducted a three-stage annotation process involving crowdworkers (via Amazon Mechanical Turk) and expert reviewers. Each anomaly was annotated with rich textual features (description, explanation, justification), categorized by manifestation type, and rated along severity, surprisal, and complexity. Please see Appendix C for full details on the data collection and annotation pipeline.

Figure 2 displays the distribution of these scores. The dataset is skewed toward visually simple anomalies, with severity showing moderate imbalance and surprisal tending toward more unexpected instances; the latter two having relatively high variance across annotators. A moderate but significant correlation exists between severity and surprisal, with a Spearman correlation of 0.52. This is consistent with the intuition that highly severe anomalies are typically rarer and hence more surprising.

Final dataset. CAVE consists of **309 anomalous and 52 normal images** for a total of 361 images. Images have up to 3 anomalies, totaling 334 anomalies. Overall, CAVE exhibits a rich diversity of anomalies as shown in Figure 2 and Figure 14.

Evaluation.

- **Anomaly Description (AD).** We evaluate whether model-generated descriptions match any ground-truth ADs via pairwise comparison. GPT-4o serves as an automatic judge (90% accuracy on human-labeled subset with prompt shown in Figure 33), scoring precision, recall, and F1 [34, 63, 35]. Matched pairs are labeled as True Positives (TP), unmatched ground-truth descriptions as False Negatives (FN), and unmatched model outputs as False Positives (FP).
- **Anomaly Explanation (AE).** Given a ground-truth AD, models are expected to generate plausible explanations. Again, GPT-4o is used for matching (89% agreement with humans, see judge prompt in Figure 34), comparing model and human explanations.
- **Anomaly Justification (AJ).** Justification quality is assessed along three criteria: (1) **Plausibility**—whether the justification makes sense for the anomaly; (2) **Relevance**—how well it aligns with the image context; and (3) **Creativity**—the depth and novelty of the reasoning, beyond generic or trivial explanations. Due to the possibility of having more than one correct anomaly justification and the subjectivity of these criteria, we rely entirely on human evaluation.[1] Using the same 50 TPs and FNs as in the AE task, three annotators compare each model-generated justification with the human one and rate it as better (+1), similar (0) or worse (-1).

[1]We experimented with LLM-as-a-judge for AJ, but observed low correlation with human assessments, particularly for creativity and plausibility. Hence, we prioritize reliability through human evaluation.

Model	AD						AE
	Vanilla	CoT	SoM	CoT + SoM	MS CoT	CoT + consist.	Vanilla
Llama3.2 90b	24.9	36.13 (+11.23)	28.00 (+3.10)	29.64 (+4.74)	32.19 (+7.29)	38.56 (+13.66)	85.22
LlavaOV 72b	27.3	27.12 (−0.18)	**43.21** (+15.91)	27.11 (−0.19)	29.38 (+2.08)	36.08 (+8.78)	85.22
InternVL2.5 38b	33.7	36.65 (+2.95)	37.79 (+4.09)	33.71 (+0.01)	32.42 (−1.28)	40.00 (+6.30)	84.24
QwenVL2.5 72b	35.7	32.92 (−2.78)	34.33 (−1.37)	29.13 (−6.57)	34.18 (−1.52)	34.32 (−1.38)	85.02
InternVL2.5 78b	36.7	39.06 (+2.36)	36.62 (−0.08)	37.55 (+0.85)	35.76 (−0.94)	39.88 (+3.18)	83.83
GPT-4o	**51.2**	**54.26** (+3.06)	40.70 (−10.50)	**45.05** (−6.15)	**56.64** (+5.44)	**53.69** (+2.49)	88.04
o1	46.0	49.76 (+3.76)	43.54 (−2.46)	41.55 (−4.45)	49.50 (+3.50)	52.78 (+6.78)	**90.96**
Claude	43.3	51.31 (+8.00)	34.66 (-8.65)	43.50 (+0.19)	51.31 (+8.00)	49.46 (+6.15)	80.54
Average	37.35	40.9 (+3.55)	37.35 (+0)	35.91 (-1.44)	40.18 (+2.82)	43.10 (+5.75)	84.67

Table 1: **AD and AE Results.** F1-scores on the Anomaly Description (AD) task using various prompting strategies (gains over vanilla in parentheses). AE results (last column) are based on the vanilla prompt only.

4 Experiments

We evaluate five open-source and three closed-source VLMs on the Anomaly Description (AD), Explanation (AE), and Justification (AJ) tasks. We focus primarily on AD, the core detection task, and briefly report AE and AJ results. Additional analyses, including breakdowns by anomaly type, score distributions, and cultural biases, are provided in Appendix G.5.

Anomaly Description. We prompt each model to describe anomalies in an image, then use GPT-4o as a judge to compare predicted descriptions against ground-truth. Using a vanilla prompt, GPT-4o achieves the best performance (51.2% F1) (Table 1). Qualitative analysis reveals two main failure modes: (i) **perception errors**: hallucinated or misidentified objects, and (ii) **reasoning errors**: misjudging contextually normal elements as anomalies. A breakdown of GPT-4o's false positives shows nearly half of them stem from reasoning issues (Table 5).

To improve performance, we test five advanced prompting strategies, including chain-of-thought (CoT), multi-step reasoning, and self-consistency. The best-performing strategy is CoT + self-consistency, improving F1 by +4.8% over vanilla on average (see details in Section D). However, improvements are limited and inconsistent across models, with some prompts introducing new errors due to noisy visual grounding.

Anomaly Explanation and Justification. For AE, all models achieve over 80% accuracy when explaining why a provided anomaly is anomalous, with slightly higher performance on anomalies they also successfully described. For AJ, we conduct a human evaluation of plausibility, relevance, and creativity. GPT-4o justifications are often reasonable when the anomaly is correctly detected, but tend to be simplistic and less creative than human responses (see Figure 12). Failures largely correlate with perception and reasoning errors in the AD task. Detailed AE/AJ examples are available in Appendix G.2 and G.3.

Analysis by anomaly category and numerical features. Using our taxonomy, we note that GPT-4o's FPs in AD involve hallucinating attribute, relation and textual anomalies (Figure 7; see classifier details in Appendix G.1). Textual anomalies are most frequently hallucinated but detected (56.28%) and explained (92.40%) best, uniformity anomalies are rarely hallucinated but detected (28.92%) and explained (88.28%) worst. Interestingly absence anomalies are harder to detect (28.2%) but easily explained (94.52%) once provided. Overall, *harder-to-detect categories are also harder to explain.*

5 Conclusion

We introduce CAVE, a multimodal benchmark of 334 visual anomalies in 361 images spanning seven tasks, designed to test VLMs' real-world anomaly detection and understanding. Leading proprietary and open-source models (>70B parameters) only score ∼56 % F1 on anomaly detection, highlighting significant room for improvement. While they perform better on anomalies seen as highly severe and surprising by humans, they struggle with anomalies that demand complex visual understanding, such as spatial reasoning and detection of pattern violations. Improving anomaly detection requires advances in both visual understanding and commonsense reasoning.

References

[1] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[2] P. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, 2010.

[3] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019.

[4] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019.

[5] N. Bitton-Guetta, Y. Bitton, J. Hessel, L. Schmidt, Y. Elovici, G. Stanovsky, and R. Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2616–2627, 2023.

[6] N. Bitton-Guetta, A. Slobodkin, A. Maimon, E. Habba, R. Rassin, Y. Bitton, I. Szpektor, A. Globerson, and Y. Elovici. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *ArXiv*, abs/2407.19474, 2024.

[7] D. Bogdoll, M. Nitsche, and J. M. Zollner. Anomaly detection in autonomous driving: A survey. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 4487–4498. IEEE, June 2022.

[8] D. I. Campbell, S. Rane, T. Giallanza, C. N. D. Sabbata, K. Ghods, A. Joshi, A. Ku, S. M. Frankland, T. L. Griffiths, J. D. Cohen, and T. W. Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[9] M. Cao, H. Tang, H. Zhao, H. Guo, J. Liu, G. Zhang, R. Liu, Q. Sun, I. Reid, and X. Liang. Physgame: Uncovering physical commonsense violations in gameplay videos. *arXiv preprint arXiv:2412.01800*, 2024.

[10] Y. Cao, X. Xu, J. Zhang, Y. Cheng, X. Huang, G. Pang, and W. Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *ArXiv*, abs/2401.16402, 2024.

[11] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[12] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[13] J. Diers and C. Pigorsch. A survey of methods for automated quality control based on images. *International Journal of Computer Vision*, 131(10):2553–2581, 2023.

[14] D. C. Donderi. Visual complexity: a review. *Psychological bulletin*, 132(1):73, 2006.

[15] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[16] S. G. Goto, Y. Ando, C. Huang, A. Yee, and R. S. Lewis. Cultural differences in the visual processing of meaning: Detecting incongruities between background and foreground objects using the n400. *Social cognitive and affective neuroscience*, 5(2-3):242–253, 2010.

[17] F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[18] Q. Guo and Y. Chen. The effects of visual complexity and task difficulty on the comprehensive cognitive efficiency of cluster separation tasks. *Behavioral Sciences*, 13(10):827, 2023.

[19] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[20] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[21] C. Heyes. Rethinking norm psychology. *Perspectives on Psychological Science*, 19(1):12–38, 2024.

[22] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017.

[23] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

[24] S. Jin, X. Jiang, J. Huang, L. Lu, and S. Lu. LLMs meet VLMs: Boost open vocabulary object detection with fine-grained descriptors. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] G. Klein. *Seeing what others don't: The remarkable ways we gain insights*. Public affairs, 2013.

[26] G. Klein, M. Jalaeian, R. R. Hoffman, and S. T. Mueller. The plausibility transition model for sensemaking. *Frontiers in psychology*, 14:1160132, 2023.

[27] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso. A data–frame theory of sensemaking. In *Expertise out of context*, pages 118–160. Psychology Press, 2007.

[28] G. Klein, R. Pliske, B. Crandall, and D. D. Woods. Problem detection. *Cognition, Technology & Work*, 7:14–28, 2005.

[29] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[30] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[31] J. Li, J. Mo, M. Vo, A. Sugimoto, and H. Nakayama. Nemo: Can multimodal llms identify attribute-modified objects? *arXiv preprint arXiv:2411.17794*, 2024.

[32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. yue Li, J. Yang, H. Su, J.-J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2023.

[33] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2025.

[34] Y. Liu, H. Zhou, Z. Guo, E. Shareghi, I. Vulić, A. Korhonen, and N. Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*, 2024.

[35] A. Liusie, P. Manakul, and M. Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics.

[36] S. Malakouti, A. Aghazadeh, A. Khandelwal, and A. Kovashka. Benchmarking vlms' reasoning about persuasive atypical images. *arXiv preprint arXiv:2409.10719*, 2024.

[37] A. Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024, 2024.

[38] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06, 2021.

[39] J. F. Mullen Jr, P. Goyal, R. Piramuthu, M. Johnston, D. Manocha, and R. Ghanadan. "don't forget to put the milk back!" dataset for enabling embodied agents to detect anomalous situations. *IEEE Robotics and Automation Letters*, 2024.

[40] J. Myung, N. Lee, Y. Zhou, J. Jin, R. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Perez-Almendros, A. A. Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

[41] S. Nayak, K. Jain, R. Awal, S. Reddy, S. Van Steenkiste, L. A. Hendricks, A. Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.

[42] OpenAI. Gpt-4o system card, 2024.

[43] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[44] T. Rabeyron and T. Loose. Anomalous experiences, trauma, and symbolization processes at the frontiers between psychoanalysis and cognitive neurosciences. *Frontiers in Psychology*, 6:1926, 2015.

[45] C. Roman and P. Meyer. Analysis of glyph and writing system similarities using Siamese neural networks. In R. Sprugnoli and M. Passarotti, editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 98–104, Torino, Italia, May 2024. ELRA and ICCL.

[46] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[47] Z. Sun and C. Firestone. Curious objects: How visual complexity guides attention and engagement. *Cognitive Science*, 45(4):e12933, 2021.

[48] M. R. Taesiri, T. Feng, C.-P. Bezemer, and A. Nguyen. Glitchbench: Can large multimodal models detect video game glitches? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22444–22455, 2024.

[49] Y. Tai, W. Fan, Z. Zhang, and Z. Liu. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27176–27185, 2024.

[50] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

[51] Z. Wang, G. Bingham, A. W. Yu, Q. V. Le, T. Luong, and G. Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer, 2024.

[52] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

[53] J. J. Williams, C. Walker, and T. Lombrozo. Explaining increases belief revision in the face of (many) anomalies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[54] G. Xie, J. Wang, J. Liu, J. Lyu, Y. Liu, C. Wang, F. Zheng, and Y. Jin. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*, 2024.

[55] G. Xu, P. Jin, H. Li, Y. Song, L. Sun, and L. Yuan. Llava-cot: Let vision language models reason step-by-step, 2025.

[56] Z. Xu, H.-T. L. Chiang, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, F. Xia, J. Hsu, J. Hoech, P. Florence, S. Kirmani, S. Singh, V. Sindhwani, C. Parada, C. Finn, P. Xu, S. Levine, and J. Tan. Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs. In *8th Annual Conference on Robot Learning*, 2024.

[57] Q. Yan, Y. Fan, H. Li, S. Jiang, Y. Zhao, X. Guan, C.-C. Kuo, and X. E. Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*, 2025.

[58] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[59] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023.

[60] A. Ye, S. Santy, J. D. Hwang, A. X. Zhang, and R. Krishna. Computer vision datasets and models exhibit cultural and linguistic diversity in perception. *arXiv preprint arXiv:2310.14356*, 2023.

[61] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv preprint arXiv:2003.12338*, 27(10.48550), 2020.

[62] P. Zhang and D. Soergel. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *Journal of the Association for Information Science and Technology*, 65(9):1733–1756, 2014.

[63] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[64] K. Zhou, E. Lai, W. B. A. Yeong, K. Mouratidis, and J. Jiang. ROME: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10185–10197, Singapore, Dec. 2023. Association for Computational Linguistics.

Appendix Table of Contents

A Research Tools **10**

B Related Works **10**

B.1 Vision Language Models 10

B.2 Anomaly detection benchmarks 10

B.3 Evaluation Methods 11

C Dataset Construction **12**

D Advanced Prompting Strategies **13**

E Human Annotations **14**

E.1 Data Collection and Filtering 14

E.2 Annotation round 1: Amazon Mechanical Turk 15

E.3 Annotation round 2: Expert annotation consolidation 16

E.4 Numerical features inter-rater agreement 18

E.5 Cultural representation & bias 18

F Prompts **19**

G Additional Results **19**

G.1 Anomaly Description 19

G.2 Anomaly Explanation 20

G.3 Anomaly Justification 21

G.4 Numerical features prediction 21

G.5 Analysis by numerical features 22

G.6 Analysis by anomaly category 23

G.7 Cultural bias assessment 25

H Failure examples **26**

A Research Tools

Compute details. We evaluated 5 open-source models: InternVL2.5 (38B et 78B parameters) [12], LlavaOneVision (72B) [30], QwenVL2.5 (72B) [58], and Llama 3.2 (90B) [37]. We use the PyTorch and Hugging Face Transformers implementations for all models examined in this work. Each model is publicly available on the Hugging Face Hub. Table 2 provides each model's corresponding Hugging Face identifier. All models are run in a zero-shot manner, with a temperature of 0, unless a self-consistency prompting strategy is used. Inference with the large models is done on 4 A100 80B GPUs for up to 3 hours for the full dataset.

Use of AI assistants. Portions of the code of this paper have been written with the support of a coding assistant (Copilot). All AI-generated codes were thoroughly verified. Portions of the paper were corrected using a writing assistant (Grammarly).

Model	Identifier
Open-source Models	
InternVL2.5 38B	`OpenGVLab/InternVL2_5-38B`
InternVL2.5 78B	`OpenGVLab/InternVL2_5-78B`
Qwen2.5-VL 72B	`Qwen/Qwen2.5-VL-72B-Instruct`
LlavaOneVision 72B	`llava-hf/llava-onevision-qwen2-72b-ov-hf`
Llama3.2 90B Vision	`meta-llama/Llama-3.2-90B-Vision`
Closed-source Models	
o1	`o1-2024-12-17`
GPT-4o	`gpt-4o-2024-11-20`
Claude	`claude-3-5-sonnet-20241022`

Table 2: **Models used.** Overview of the models considered in our study and their corresponding identifiers on the Hugging Face Hub.

B Related Works

B.1 Vision Language Models

Vision Language Models have made significant progress by integrating powerful vision encoders with LLMs. In most of the models considered in this work (Table 2), images are first processed by the vision encoder and then projected into the language model's embedding space [37, 58, 12]. These visual representations are fused with textual inputs and subsequently passed through the LLM. However, the overall performance of VLMs remains constrained by the capabilities of their vision encoders, particularly in capturing fine-grained visual details or handling out-of-distribution (OOD) images.

B.2 Anomaly detection benchmarks

Anomaly detection spans various modalities using specialized datasets, from industrial defect identification to autonomous driving [38, 3, 4, 7, 10]. Broadly, anomalies can be classified into *structural* (*e.g.*, physically detectable flaws or distortions in industrial inspections) and *semantic* (deviations at higher hierarchical levels, including the entity, relation, and frame levels) [10]. In this work, we focus on *semantic anomalies* that necessitate commonsense reasoning for detection and interpretation, hence, we emphasize prior works relevant to this domain.

Several recent multi-modal benchmarks have explored unusual, abstract, or commonsense-defying visual scenarios to evaluate the robustness of VLMs. Visual Riddles [6] introduces synthetically generated images, each depicting a unique situation and requiring commonsense to answer a question. WHOOPS [5] takes a broader approach, generating abnormal images across a wide range of scenarios using three diffusion models. Similar to our work, it extends beyond visual commonsense violations to include anomalies related to social norms, cultural knowledge, and celebrities. The main focus is on explanation generation and image captioning. HaloQuest [51] attempts to mitigate hallucination by collecting and generating unusual and abstract visual scenes along with VQA designed to trigger hallucinations and use them for VLM fine-tuning.

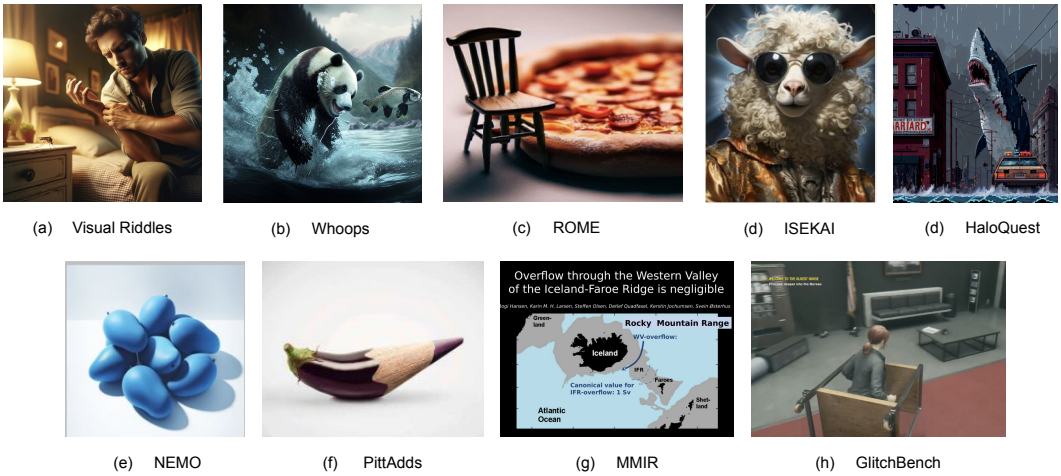


Figure 3: **Related Benchmark Examples.** Examples of images from related multimodal anomaly-detection benchmarks. More details about each benchmark are given in Section B and Table 3.

Complementing synthetic scenario generation, other benchmarks focus on systematically altering concrete object attributes and relationships to directly probe VLM reasoning. ROME (Reasoning Beyond Commonsense Knowledge) [64] explicitly modifies object attributes—such as color, shape, and size—and object relationships using DALL-E 2, creating images that defy commonsense expectations. Similarly, NEMO [31] investigates how VLMs recognize objects with uncommon properties, such as a blue mango. ISEKAI [49] explores a different approach by transferring real-world entities into an alternate world using diffusion models, introducing novel objects and entities and evaluating models on image-pair classification.

A separate line of research focuses on anomalies within structured visual styles, such as advertisements and video games. **(author?)** [36] leverage the PittAds dataset [22], which examines atypical visual elements in advertisements and defines specific tasks like multi-label atypicality classification, atypicality statement retrieval, and atypical object recognition. However, unlike open-ended benchmarks, these tasks constrain atypicality to a specific visual style. Similarly, MMIR [57] introduces a benchmark to assess VLMs' ability to detect and reason about semantic mismatches in webpages, presentation slides, and posters—focusing on images where performance is largely driven by OCR capabilities. In contrast, while CAVE also contains a category for such anomalies, it is limited to a subset of images with less amount of text.

Some recent benchmarks focus on leveraging non-photorealistic yet complex visual environments—such as video games—to evaluate anomaly detection and reasoning. GlitchBench [48] is a benchmark using unusual and glitched scenes from video games. Similar to ours, one of its strengths is the fact that, since it's not model-generated, there can be many distracting elements in the image, making the detection very challenging. Moreover, it's an open-ended benchmark that is also evaluated using LLMs as a judge. However, all the images are non-realistic and the anomalies defy commonsense. Similarly, PhysGame [9] benchmark models' ability to identify physical commonsense anomalies in gameplay videos.

B.3 Evaluation Methods

Across these benchmarks, evaluation typically relies on zero-shot testing on large-scale pretrained models to assess how well they generalize to rare or absurd scenarios without task-specific adaptation. A few studies, like WHOOPS and HaloQuest, also explore fine-tuning on a training subset to boost performance, illustrating how effectively VLMs adapt to OOD data. In our study, we focus exclusively on zero-shot evaluation, as most anomalies in CAVE are relatively easy for humans to identify (Figure 2 (right)), and the small size of our dataset makes fine-tuning impractical.

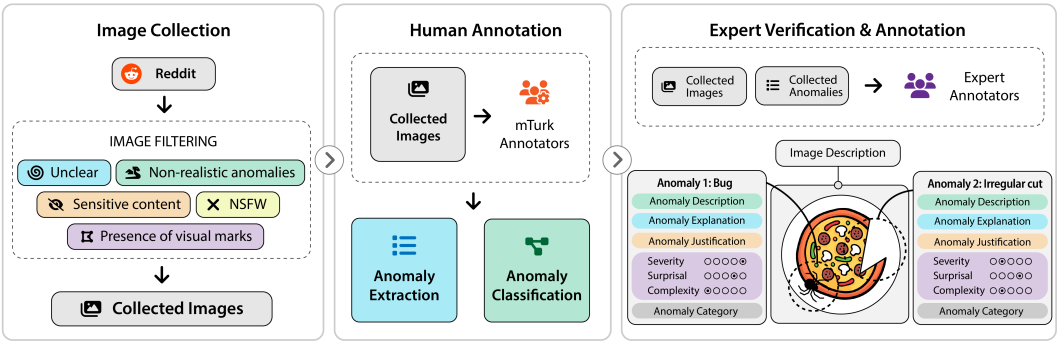


Figure 4: **An overview of CAVE data collection process. (1) Image Collection**: Images were sourced from the top 1,000 posts across various subreddits and filtered to ensure high-quality, safe data. **(2) Human Annotation**: Initial annotations were performed by Mechanical Turk workers, focusing on basic tasks such as anomaly descriptions and anomaly category identification. **(3)Expert Verification & Annotation**: A subsequent round of expert-driven annotation and verification ensured high-quality, consistent annotations across all six tasks, refining and validating the initial labels.

C Dataset Construction

Since our benchmark focuses on real-world, daily-life visual anomalies, our data collection process and annotation strategy are heavily human-centered. The dataset creation process is illustrated in Figure 4.

Data Collection. We collect images from four subreddits: `r/ocdtriggers`, `r/mildlyconfusing`, `r/mildlyinfuriating`, and `r/OSHA`. These subreddits specialize in content featuring unusual or uncommon situations, providing a rich source of real-life anomalies.

Data Filtering. We remove images that have unclear content, that contain non-realistic anomalies, and that contain NSFW content or content related to sensitive topics. We apply automatic and manual filters (see Section E.1 for details), and then annotate the remaining images through three annotation rounds.

Data Annotation. First, each image was reviewed by 5 annotators via Amazon Mechanical Turk. Annotators were asked whether the image was anomalous. If so, they were instructed to (i) describe and explain the anomaly in detail, (ii) describe what they expected instead, and (iii) categorize the anomaly.

Subsequently, expert annotators consolidated these initial annotations by validating and formalizing them along the following axes:

1. **Anomaly Description (AD)**: A visual description of the anomaly in the image.
2. **Anomaly Explanation (AE)**: An explanation of why it is anomalous.
3. **Anomaly Justification (AJ)**: A realistic and plausible explanation for how the anomaly might have occurred.
4. **Anomaly Category**: Category based on the anomaly manifestation taxonomy outlined in Section 2; the most frequent are anomalies about entity attributes, spatial relations, and textual anomalies (Figure 2).

Then, three annotators independently rated each anomaly along the 3 axes:

- **Anomaly Severity**: From 1 (does not require action; has no impact on functionality/safety) to 5 (requires immediate action).
- **Anomaly Surprisal**: From 1 (common, not very surprising; frequently observed in similar contexts) to 5 (extremely rare).
- **Anomaly Complexity**: From 1 (obvious and easy to notice) to 5 (very hard to detect or requires specific knowledge to identify).

Figure 2 displays the distribution of these scores. The dataset is skewed toward visually simple anomalies, with severity showing moderate imbalance and surprisal tending toward more unexpected instances ; the latter two having relatively high variance across annotators. A moderate but significant correlation exists between severity and surprisal, with a Spearman correlation of 0.52. This is consistent with the intuition that highly severe anomalies are typically rarer and therefore more surprising.

We measure the agreement between the 3 annotators (Table 4 in Section E.4). Spearman's Rank Correlation (0.65) and Krippendorff's Alpha (0.62) indicate moderate-to-strong agreement among annotators for severity, and weaker for surprisal, which is more subjective. Since complexity and –to a lesser extent–surprisal features have imbalanced distributions, we turn to the more adapted Gwet's AC2 [19], which shows a much higher agreement for the complexity score (0.76).

Final dataset. CAVE consists of **309 anomalous and 52 normal images** for a total of 361 images. Images have up to 3 anomalies, totaling 334 anomalies. Overall, CAVE exhibits a rich diversity of anomalies (see Figure 2 and Figure 14) across the dimensions of severity, surprisal, complexity and visual manifestation. Moreover, each anomaly is described through our comprehensive multi-task framework, which addresses anomaly detection, explanation, and justification.

Dataset	Anomaly Type		Dataset Size		Data source	Task			
	Real	Synthetic	#features	#Images		#Anomaly tasks	Y/N	multi	Open
Visual Riddles		✓	2	400	Text-to-Image models	1		✓	✓
WHOOPS		✓	4	500	Text-to-Image models	1		✓	✓
HaloQuest		✓	3	3,157	Text-to-Image models + Open Images dataset	1			✓
ROME		✓	1	1,563	ViComTe + ThingsNotWritten	1	✓		
NEMO		✓	1	900	Text-to-image models	1		✓	✓
ISEKAI		✓	1	1,498	Text-to-Image models	1			✓
PittAds		✓	1	3,928	Product ads & public service announcements	3		✓	
MMIR		✓	1	534	VisualWebArena, Zenodo	2		✓	✓
GlitchBench		✓	1	593	Game-Physics dataset + Unity + YouTube	1			✓
CAVE	✓		7	361	Reddit	3		✓	✓

Table 3: **Related Benchmarks.** Overview of multimodal reasoning benchmarks in images. Each benchmark is categorized based on the type of images it contains (real or synthetic), dataset scale (features per image and total number of images), generation method, and task involved (number of tasks related to anomaly, binary yes/no questions, multiple-choice VQA, and open-ended VQA).

D Advanced Prompting Strategies

(1) Chain-of-thought prompting (CoT) This strategy works by instructing models to "think step by step" before answering, breaking complex reasoning into explicit sequential steps [52]. See prompt in Figure 25.

(2) Set-of-Marks prompting (SoM) We incorporate object-level annotations and bounding boxes generated by Grounding DINO [32] to supplement the prompt with visual cues. Specifically, Grounding DINO identifies relevant regions in the image and provides precise bounding box coordinates, which serve as explicit visual references to guide the model's attention. Each bounding box is labeled with a number in the top-left corner, indicating the detected object. Following **(author?)** [59], we keep the textual prompt unchanged and instead replace the original images with versions that include these annotated boxes. As in the original work, the prompt does not explicitly mention the presence of bounding boxes. This strategy aims to reduce perceptual errors, such as hallucinations or counting mistakes, by focusing the model's attention on concrete visual entities [59]. The prompt used here is the vanilla inference prompt (see Figure 24.

(3) Combined CoT+SoM prompting This strategy integrates the step-by-step reasoning of CoT with visual cues of SoM. This hybrid approach first establishes precise visual references using bounding boxes, then builds logical reasoning chains based on these grounded elements, enabling both spatial understanding and logical inference. The prompt used is identical to the CoT inference prompt (see Figure 25), with the only change being the replacement of original images with versions containing bounding boxes.

(4) Multi-step CoT prompting Unlike standard CoT, this method decomposes the task into three sub-steps: (i) planning the reasoning process, (ii) identifying key visual elements, and (iii) generating anomaly descriptions based on these observations. Each sub-task is explicitly prompted, encouraging more organized and interpretable reasoning [55]. See prompt in Figure 26.

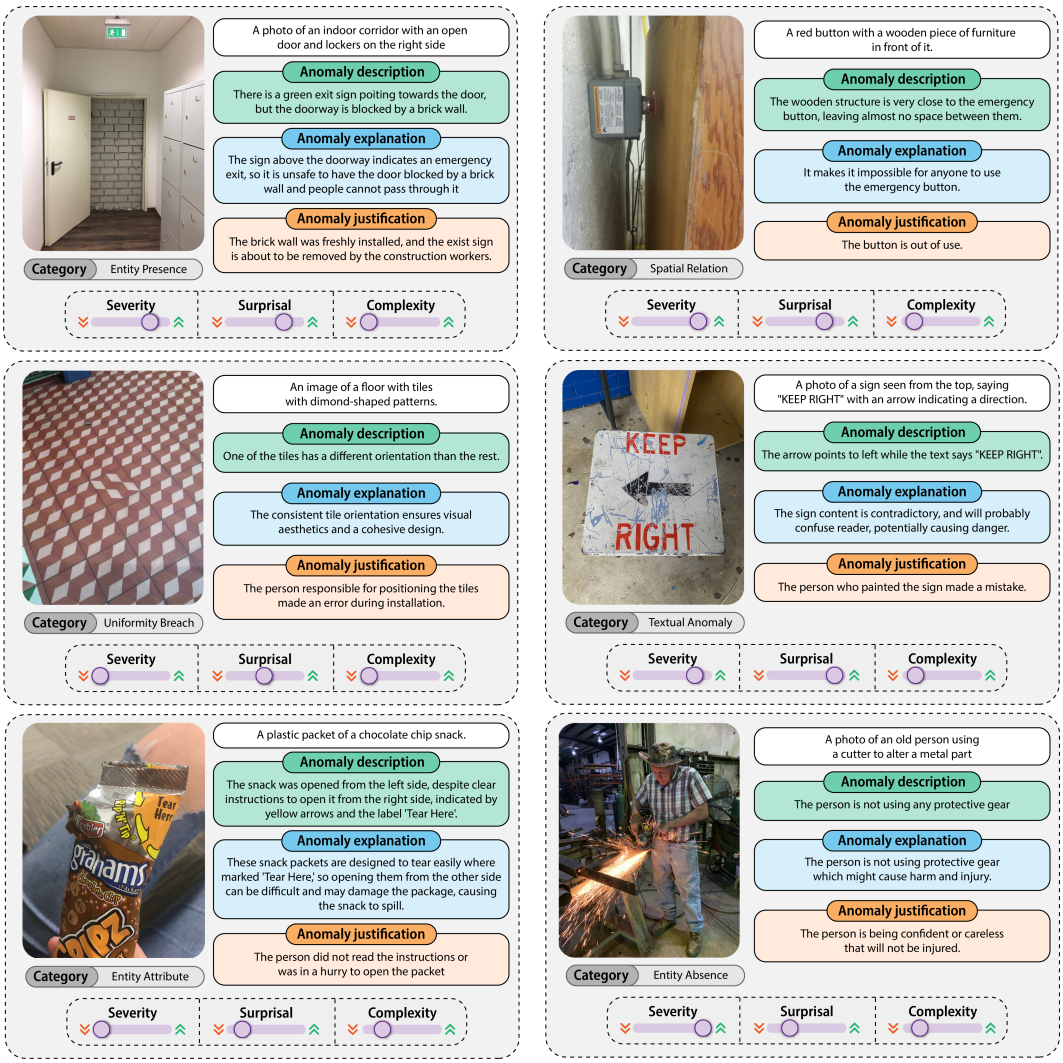


Figure 5: **Examples from CAVE.** Each image is accompanied by a human-provided image description, anomaly description, anomaly explanation, anomaly justification, anomaly manifestation category, and numerical features of severity, surprisal, and complexity scores, for each of the anomaly manifestation categories present in CAVE.

(5) CoT + Self-consistency prompting , In this strategy, the model is prompted multiple times (*e.g.*, three) using the CoT format with stochastic sampling (temperature = 0.5). The resulting outputs are then aggregated using a majority-vote mechanism: only anomalies mentioned in at least two of the three generations are retained. This technique reduces spurious detections by encouraging agreement across multiple reasoning paths, effectively filtering out unstable or hallucinated outputs [50]. See prompt in Figure 27.

E Human Annotations

E.1 Data Collection and Filtering

We scraped images from Reddit, focusing on four subreddits: `r/ocdtriggers`, `r/mildlyconfusing`, `r/mildlyinfuriating`, and `r/OSHA`. Using the PRAW[2] library, we downloaded the top 1,000 posts

[2] https://github.com/praw-dev/praw

from each subreddit. We kept only posts that contained images, and performed a first automated filtering, keeping only images above icon size.

We then manually filtered the remaining 1,725 images using the following criteria:

- Remove toxic, harmful, and not safe for work content.
- Remove image featuring unrealistic content.
- Remove images with annotations: text added on top of the image, circles, etc. When possible, we manually edited images that could be cropped to hide the annotations on the image.
- Remove images that are ambiguous or have unidentifiable content.

Many samples contain anomalies that were done on purpose; often for convenience, but sometimes as a joke. We keep these ones, as detecting the presence of a visual anomaly created on purpose for humoristic purposes, and understanding why it is anomalous, is part of the VLM abilities we want to probe.

E.2 Annotation round 1: Amazon Mechanical Turk

We used Amazon Mechanical Turk to obtain annotations for the Reddit images. To ensure high-quality annotations, we conducted a worker selection round, ultimately selecting 40 workers for the task. Workers were pre-screened using Amazon Mechanical Turk's automatic metrics with the following criteria: (a) HIT approval rate above 80%, (b) location in the United States, and (c) more than 1,000 approved HITs. Workers were compensated at a rate of 10 USD per hour, during the qualification and the annotation round. Each image received five annotations. We split the annotation into 3 rounds, allowing us to review the annotations between each round and provide feedback to the annotators when needed.

Below are the detailed instructions that were given to the annotators.

We need your help to identify and annotate anomalies in images. **An anomaly refers to anything that deviates from what most people consider standard, normal, or expected**. It can be an unusual element, action, or occurrence in an image that most people would find surprising or out of place. For example, bowls of soup accompanied by forks but no spoon would be considered an anomaly because a spoon is expected for eating soup. In contrast, a plant placed on a computer desk is not an anomaly, as most people wouldn't find it unusual.

Task Instructions:

1. **Presence of Anomaly:** Observe carefully the given image. Is there any anomalous element, according to the definition given above? Not all of the images necessarily have anomalies! You can right-click on the images and select *"Open in a new tab"* to zoom in.
2. **Description of Anomaly:** Describe the image and the anomaly in detail: What does the image show? What is abnormal or unexpected about it? Why is it considered an anomaly?
3. **Type of Anomaly: Select the type of anomaly** (an example for each type is given below):

- **Entity Presence:** Something is present in the image but shouldn't be there.
- **Entity Absence:** Something that should be present is missing.
- **Entity Attribute:** An object has an anomalous attribute such as *color, shape, label, orientation, or usage*.
- **Spatial Relation:** Something is incorrectly located or oriented relative to another element.
- **Uniformity Breach:** There is an unexpected or misplaced element in an ensemble that should be uniform or symmetrical.
- **Textual Anomaly:** The text in the image presents an unexpected, surprising, or illogical message.

You may choose more than one type of anomaly if applicable.

E.3 Annotation round 2: Expert annotation consolidation

Following the first round, we manually filtered out samples that were confusing for annotators. Our pool of expert annotators includes undergraduate degree holders, graduate students, and PhD students with a background in NLP.

Below are the detailed instructions that were given to the annotators.

Overview.

We are studying how well large vision-language models can identify anomalies that defy commonsense in images. Our goal is to assess their understanding of a situation, its severity, and potential solutions.
You will annotate anomalies visible in images. Each annotation form contains **5 images**. Each image has already been annotated by **4 to 5 workers** via MTurk, who answered the following questions:

1. **Is there an anomaly in this image?**
2. **If yes, they described:**
   (a) **Anomaly Description (AD):** Describe the image and the anomaly in detail: what does the image show, what is wrong about it, and why?
   (b) **Correct Version Description (CVD):** Describe what the correct version of the image would look like if the anomaly weren't present.

Definition of an Anomaly An anomaly is anything that **deviates from what most people consider standard, normal, or expected**. It can be an unusual element, action, or occurrence in an image that would seem surprising or out of place to most people.
Examples:

- A **bowl of soup served with a fork but no spoon** is an anomaly because a spoon is the expected utensil for soup.
- A **plant on a computer desk** is **not** an anomaly, as it is a common and expected item in such a setting.

Key Principle: Identifying an anomaly should rely **only on what is clearly visible in the image**—it should not require excessive assumptions about the situation.
Don't spend too much time on a single image. If you're unsure or confused about an image or an annotation, **skip it** and leave a note in the open field at the bottom of the page.

Instructions.

Workers often identified different anomalies in the same image. Your task is to consolidate their annotations into a structured format. You may input **up to 3 anomalies** per image. Most images contain only one anomaly. For each image, based on the workers' annotations, provide a final set of anomalies in the following format:

1. **Image Description**: Provide a short description of the image, without describing the anomaly. Include any useful context, such as whether the image is a *photo, screenshot, or illustration*, the *location*, etc.
2. **Anomaly Description (AD)**: Clearly describe the anomaly.
3. **Correct Version Description (CVD)**: Describe what the image would look like if the anomaly weren't present. **Do not** describe how to fix the anomaly—only describe the correct version as if it were normal.
4. **Anomaly Explanation (AE)**: Explain why it is anomalous. Avoid vague statements like "because it's abnormal." Instead, consider: *Why is the correct version expected? What makes the anomaly logically inconsistent or unexpected?*
5. **Anomaly Justification**: Provide a realistic and plausible explanation for how the anomaly might have occurred. Keep it concise (**max 2 sentences**). Example: If an object is blocking a door, a plausible justification might be: *"The door is not in use because it leads to an empty space."*
6. **Anomaly Severity (Does the anomaly require immediate action?)**
   - **1** = Does not require action; purely aesthetic or has no impact on functionality/safety. *Example: A small stain on a non-critical surface.*

- **3** = Moderately concerning; might cause inconvenience or minor inefficiencies but does not pose immediate risks. *Example: A misaligned sign that is still readable.*
- **5** = Requires immediate action; it could present a safety hazard, major malfunction, or significant risk. *Example: A worker using a circular saw without protection gear.*

7. **Anomaly Surprisal (How much does it deviate from expectations?)**
   - **1** = Common, not very surprising; frequently observed in similar contexts. *Example: A car parked in an inconvenient way.*
   - **3** = Unusual but not shocking; uncommon but plausible.
   - **5** = Extremely rare and highly surprising; would cause strong reactions (shock, confusion, amazement). *Example: A tree growing upside down from a roof.*
8. **Anomaly Complexity (How hard was this anomaly to detect?)**
   - **1** = Obvious and easy to notice; immediately stands out. *Example: A red apple in a pile of green apples.*
   - **3** = Requires some attention to notice; not the first thing seen but becomes clear after a few seconds. *Example: A misspelled word on a sign.*
   - **5** = Very hard to detect; blends into the environment or requires specific knowledge to identify. *Example: A minor defect in complex machinery.*

Guidelines:

In practice, you will reuse the MTurk annotations. Here are common situations you may encounter and how to handle them:

- **Same Anomaly from Different Workers.** If multiple workers describe the same anomaly, **merge their descriptions** into one clear and accurate version. **Two anomalies are the same if they have the same description and explanation.**
- **One Worker Describes Multiple Anomalies Jointly.** If a worker describes multiple anomalies together, **split them into separate entries** and fill in the necessary fields for each.
- **Invalid Anomaly.**
  - Does this truly qualify as an anomaly based on the definition?
  - Did the worker make **assumptions** about the situation that are not straightforward using the image alone?
  - Did the worker **misinterpret** the image?

  If invalid, **flag it and do not include it** in the consolidated list.
- **Unclear Anomaly Description.** If an anomaly is valid but **poorly described**, **rephrase it clearly** and complete the required fields (AD, AE, CVD, etc.).
- **Unclear or Incorrect Correct Version Description (CVD).** If a worker's CVD does not align with the anomaly or is poorly phrased, **rewrite it according to the guidelines**.
- **No Workers Found an Anomaly.** If no worker identified an anomaly, **check if you can spot an obvious one**. If not, **leave the fields empty**.
- **All Reported Anomalies Are Invalid.** If none of the workers' anomalies match the definition and you don't see any other valid anomaly, **leave everything empty**.

In practice:

For convenience, you can:

- Copy-paste the list of MTurk annotations to the side for easy reference.
- Open the image in full resolution in another window.
- Keep these instructions open in a separate tab.

LLM Usage:

- You can use a language model to check and correct the grammar of your annotations.
- **DO NOT upload or share the image with an LLM!**

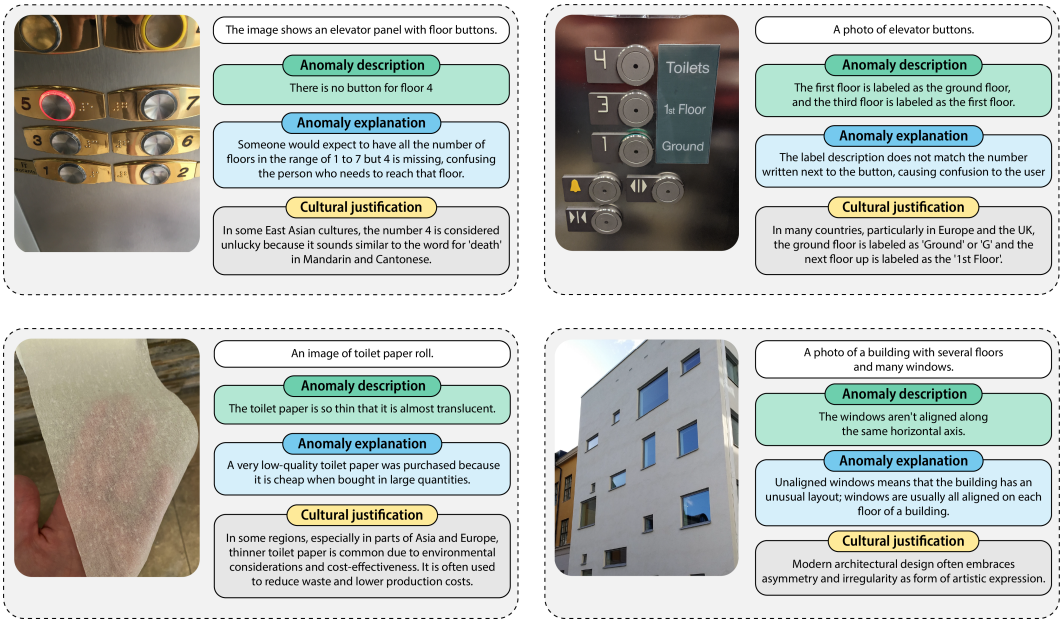


Figure 6: **Culture-specific examples of CAVE.** Examples of anomalies from CAVE annotated as Western-centric, along with culturally grounded justifications explaining why they should not be considered anomalies.

E.4 Numerical features inter-rater agreement

Each numerical feature – anomaly surprisal, complexity and relevance – is annotated by 3 people. We measure the agreement between the 3 annotators (table 4) using Spearman's Rank Correlation, Krippendorff's Alpha, and Gwet's AC2. Spearman's Rank Correlation (0.65) and Krippendorff's Alpha (0.62) indicate moderate-to-strong agreement among annotators for severity, and weaker for surprisal, which is more sujective. Since surprisal and complexity are imbalanced, we turn to Gwet's AC2 [19], a paradox-resistant agreement score, where the chance agreement is measured in a less distribution-sensitive fashion. We use quadratic weights, meaning that larger disagreements are exponentially more problematic than smaller ones. Indeed, likert-scale ratings with relatively subjective tasks such as here may lead to confusions between similar ratings (4 and 5, 1 and 2). Gwet's AC2 highlights a much higher agreement for the complexity score of 0.76, which is considered good [20].

	Spearman ρ	**Krippendorff α**	**Gwet AC2**
Severity	0.65	0.62	0.58
Surprisal	0.34	0.32	0.54
Complexity	0.28	0.23	0.76

Table 4: Inter-rater agreement for each numerical feature.

E.5 Cultural representation & bias

An anomaly is generally defined as a deviation from the norm. In this context, "norm" refers to a set of expectations commonly shared within a particular social or cultural group. Some of these norms are broadly universal, for example, adhering to safety standards to avoid hazardous situations, while others are culturally specific, such as the custom of wearing red at weddings in China [16, 40, 41]. As a result, interpretations of what constitutes an anomaly can differ significantly across cultural contexts, leading to situations that may appear ordinary to individuals from one background and anomalous to those from another [60].

To explore the extent to which cultural bias influences the perception of anomalies, we conducted an analysis of the CAVE dataset. Specifically, we examined whether a subset of visual anomalies presented in the dataset reflected culturally contingent interpretations. We selected a subset of 35

anomalies based on high variance (above 1.5 for each feature) in the numerical features obtained from annotator responses, as this variance suggests a lack of consensus that may be attributable to differing cultural perspectives. Among these, we identified four images containing anomalies that appeared Western-centric but would not be considered anomalous in other cultural contexts. In addition, from the full benchmark, we selected 20 examples reflecting personal biases, such as anomalies related to how individuals park their cars or behave in public spaces, as well as a set of universally recognized anomalies. For each of these 24 samples, we provided explanations of the relevant cultural context, where applicable, and updated the corresponding Anomaly Justification (AJ) annotations accordingly. Using these manually curated annotations as reference labels, we constructed a prompt to evaluate whether each anomaly aligned with specific cultural, religious, regional, or historical norms, and not with personal biases. This prompt was submitted to GPT-4o for analysis on the same subset. The model performed well, misclassifying only one instance: a train seat colored differently from the rest. While this was intended to reflect a "uniformity breach," the model interpreted it as a designated priority seat—an error likely due to contextual ambiguity.

We subsequently applied the same automatic bias assessment method to the entire CAVE dataset to verify the initial manual annotation. This broader analysis identified the same four anomalies that exhibited a Western-centric bias. These instances are presented in Figure 6, along with the model's culturally influenced anomaly justifications for each. This analysis indicates that while the majority of anomalies in the CAVE dataset are perceived as universally anomalous and actionable, a small number are influenced by culturally specific norms, particularly those aligned with Western perspectives. These findings underscore the importance of accounting for cultural variability in the development of robust and inclusive anomaly detection systems.

F Prompts

The prompts for six tasks, the automatic evaluation and the cultural assessment are listed below:

- Anomaly Description: Figure 24
- Anomaly Explanation: Figure 28
- Anomaly Justification: Figure 29
- Anomaly Severity: Figure 30
- Anomaly Surprisal: Figure 31
- Anomaly Complexity: Figure 32
- AD judge prompt: Figure 33
- AE judge prompt: Figure 34
- Cultural bias assessment prompt: Figure 36

G Additional Results

G.1 Anomaly Description

WE categorize all false positives (anomalies hallucinated by the VLM) into the different anomaly visual manifestation types (according to our taxonomy), by tuning a classifier of Anomaly Descriptions. We run the classifier on GPT-4o's false positives using the prompt given in Figure 35. Figure 7 shows that GPT-4o predominantly hallucinates anomalous entity attributes (*e.g.*, count, color), anomalous spatial relations, and textual anomalies (anomalies in the context of text seen in the image).

To further understand the limitations of the models, we perform a **qualitative error analysis**. We identify two main failure modes with the vanilla prompt. First, *perception errors*—hallucinations of missing or non-existent objects, miscounts, or incorrect spatial relations—arise from over-reliance on language priors and weak visual cues. For instance, in Figure 21, GPT-4o claims a chair is missing, despite all spots being filled. Second, *reasoning errors* occur when models flag contextually normal elements as anomalous due to faulty commonsense reasoning or limited commonsense knowledge. In Figure 18, QwenVL incorrectly marks a star next to the elevator button "1" as anomalous, overlooking its common use to denote the ground floor. Finally, some cases involve both *perception and reasoning errors*.

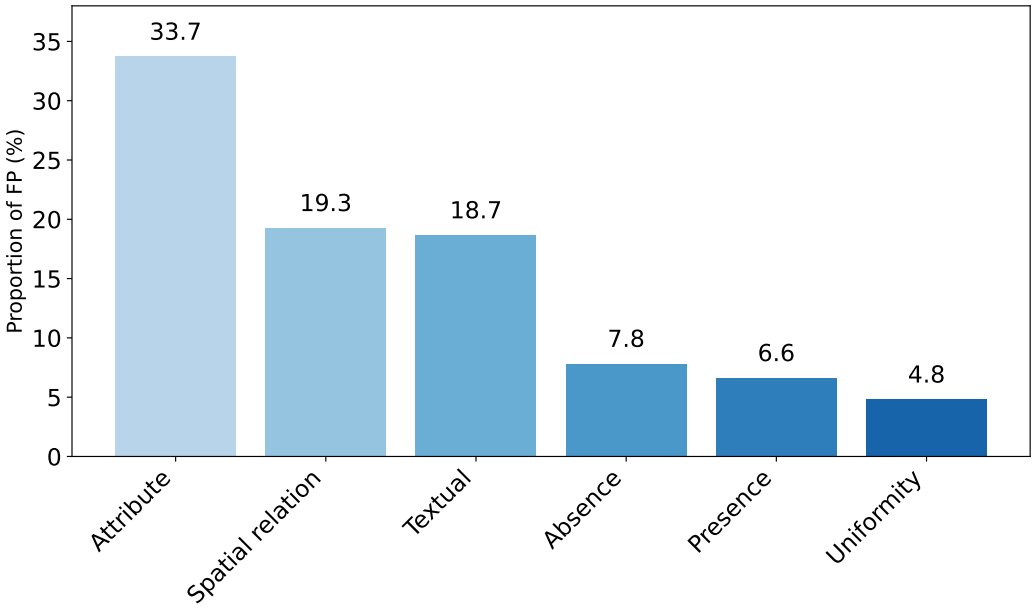


Figure 7: **FP classifier performance.** Anomaly category proportions in GPT-4o FP.

Additional examples of model errors can be found in Figures 18-23. We perform a manual classification to assign each GPT-4o FP (hallucinated anomaly) into one of these categories in Table 5, finding that around half of them are reasoning mistakes.

Prompt	Perception	Reasoning	Both	Count
Vanilla	44%	49%	7%	86
MS CoT	68%	32%	0	95

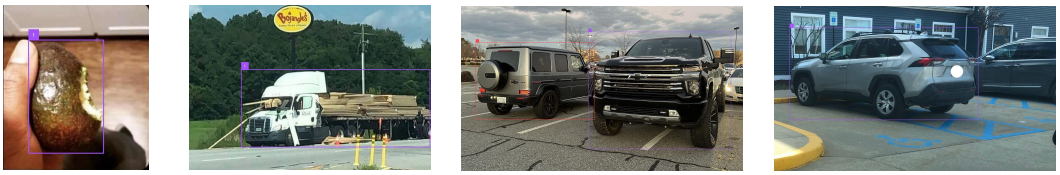
Table 5: **GPT-4o Qualitative Analysis.** Proportion of FP error analysis across prompting strategies as determined by human evaluation.

G.2 Anomaly Explanation

Each model's performance on TP and FN from the AD task is detailed in Table 6. Most of the models have higher performance on TP examples than FN.

Model	TP Acc. (%)	FN Acc. (%)
open-source models		
Llama3.2 90b	82.22	76.88
LlavaOV 72b	90.67	79.76
InternVL2.5 38b	84.26	84.21
QwenVL2.5 72b	87.39	82.64
InternVL2.5 78b	81.08	86.58
closed-source models		
GPT-4o	90.86	85.22
o1	93.02	88.89
Claude	87.10	73.97
Average	83.97	81.02

Table 6: **AE Results on TP vs FN.** AE Accuracy on TP vs FN for each model.



GPT-4o: A banana is unusually shaped like a phone receiver.

GT: The avocado was eaten from the outside

GPT-4o: A truck is parked on the roof of a building.

GT: The wooden planks on the cargo bed of the truck penetrated the driver's cabin and the windshield.

GPT-4o: A car is parked on top of another car in a parking lot.

GT: The big Chevrolet truck is crossing over into the parking spot on its right, taking up two parking spots.

GPT-4o: The car is parked on the roof of a house.

GT: The car is parked between the handicap spots, on the blue lines.

Figure 8: **Set-of-Marks images.** GPT-4o anomaly descriptions based on images with bounding boxes derived from GroundingDINO.

Prompting Strategy	TP	FP	FN	Precision	Recall	F1 Score
Vanilla	119.75	191.38	219.13	41.19	35.37	37.35
CoT	139.63 (+19.88)	159.00 (-32.38)	247.38 (+28.25)	47.95 (+6.76)	36.08 (+0.71)	40.90 (+3.55)
SoM	136.00 (16.25)	222.38 (+31.00)	251.00 (+31.88)	43.38 (+2.20)	35.14 (-0.23)	37.35 (+0.00)
CoT+SoM	123.50 (+3.75)	181.13 (-10.25)	263.50 (+44.38)	42.01 (+0.83)	31.91 (-3.46)	35.91 (-1.44)
MS CoT	144.50 (+24.75)	150.13 (-41.25)	242.50 (+23.38)	50.45 (+9.26)	33.88 (-1.49)	40.18 (+2.82)
Self-consistency	145.13 (+25.38)	141.75 (-49.63)	240.63 (+21.50)	51.72 (+10.53)	37.50 (+2.13)	43.10 (+5.75)

Table 7: **Overall anomaly detection performance.** Values in parentheses indicate deltas from the Vanilla baseline; green with for improvement, red for decline.

Prompting Strategy	Absence	Attribute	Presence	Relation	Textual	Uniformity
Vanilla	24.78	35.10	51.13	32.02	53.00	28.86
CoT	30.84 (+6.05)	39.13 (+4.03)	56.95 (+5.82)	35.62 (+3.60)	60.58 (+7.58)	30.56 (+1.71)
SoM	27.03 (+2.25)	33.58 (-1.52)	47.46 (-3.67)	30.57 (-1.45)	55.36 (+2.36)	25.95 (-2.91)
SoM+CoT	27.85 (+3.06)	33.20 (-1.91)	52.85 (+1.72)	30.74 (-1.28)	54.61 (+1.61)	28.38 (-0.48)
MS CoT	26.74 (+1.95)	39.90 (+4.79)	53.44 (+2.31)	33.63 (+1.61)	57.15 (+4.15)	27.35 (-1.51)
Self-consistency	31.97 (+7.19)	41.83 (+6.73)	56.12 (+4.99)	36.52 (+4.50)	56.97 (+3.97)	32.45 (+3.59)
Average	28.20	37.12	52.99	33.18	56.28	28.92
Rank	6	3	2	4	1	5

Table 8: **F1 scores per anomaly category.** Values in parentheses indicate deltas from the Vanilla baseline; green for improvement, red for decline.

G.3 Anomaly Justification

Figure 13 compares InternVL2.5 78B with human anomaly justifications.

G.4 Numerical features prediction

The last set of tasks of CAVE is the classification of the anomalies into ordinal features: surprisal, severity, and complexity, across a scale of 1 to 5. Echoing the inter-rater agreement that we computed between the 3 expert annotators on the surprisal, severity, and complexity scores, we measure the agreement between the human average score for each feature and the models' predictions of each score. The prompts used for these features can be found in section F.

GPT-4o and InternVL show high agreement with humans for severity (section G.4), with both models achieving strong agreement scores. Surprisal and complexity prediction are harder tasks for both models.

The analysis of complexity, severity, and surprisal scores across different anomaly categories has been shown in Figure 14. The severity scores indicate anomalies categorized under entity absence and presence tend to be perceived as more severe. Conversely, anomalies related to uniformity breaches are consistently viewed as less severe. Examining the complexity scores, we observe

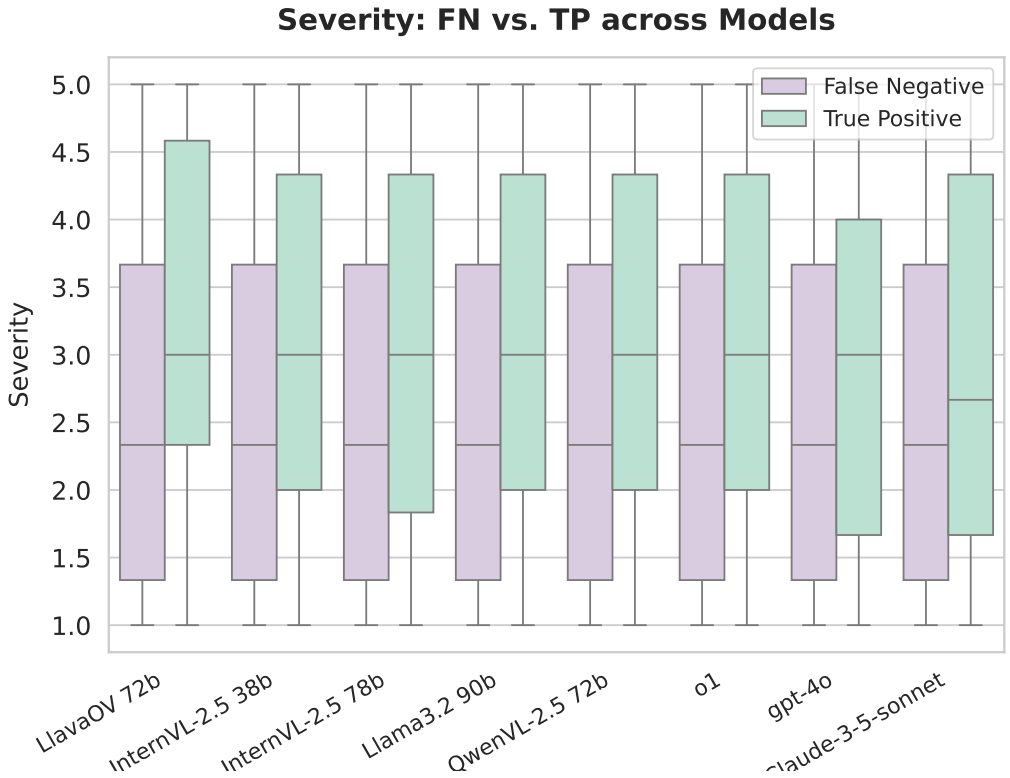


Figure 9: **Models' performance across severity feature.** Plot showing the deviation in the models' performance across different levels of anomaly severity for the anomaly description task. The results indicate that models perform well on less severe anomalies, while performance drops significantly for highly severe anomalies on average.

	GPT-4o		InternVL2.5 78b	
	ρ	**AC2**	ρ	**AC2**
Severity	0.78	0.79	0.75	0.77
Surprisal	0.49	0.81	0.28	0.24
Complexity	0.27	0.80	0.26	0.61

Table 9: **Numerical Feature Prediction.** Comparison of GPT-4o and InternVL2.5 78b prediction of Anomaly Severity, Surprisal and Complexity. We measure Gwet's AC2 and Spearman's ρ.

that categories like textual anomalies exhibit greater variability, suggesting diverse perceptions of complexity within annotators, whereas uniformity anomalies show lower complexity scores with minimal variance. The distribution of surprisal scores indicates that anomalies in the textual and presence categories consistently evoke stronger feelings of unexpectedness, while again, anomalies categorized as uniformity remain at lower surprise levels.

G.5 Analysis by numerical features

We analyze anomaly detection TPs and FNs across CAVE's three numerical features: severity, surprisal, and complexity (Figure 15). GPT-4o with vanilla prompt performs best on anomalies that are more surprising and less complex – *i.e.*, those humans found the most uncommon and easy to spot – while missed ones are often less severe, less surprising, and more complex. Other models show similar trends (Section G.4).

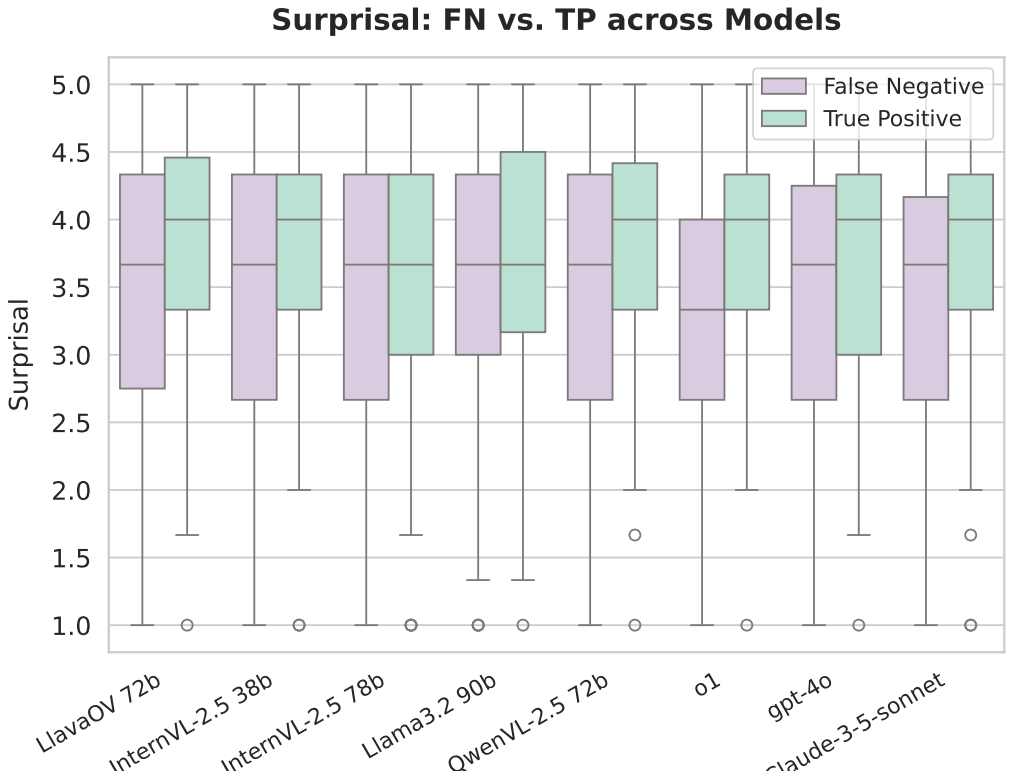


Figure 10: **Models' performance across surprisal feature.** Plot showing the deviation in the models' performance across different levels of anomaly surprisal for the anomaly description task. The results reveal that models perform well on high-surprisal anomalies but also exhibit more false positives for more surprising anomalies on average.

Model	Absence	Attribute	Presence	Relation	Textual	Uniformity
Open-source Models						
Llama3.2 90b	92.00	83.66	87.91	88.57	89.66	86.15
LlavaOV 72b	94.34	89.57	91.49	91.16	94.51	92.75
InternVL2.5 38b	94.34	90.91	90.32	87.32	93.33	91.18
QwenVL2.5 72b	94.34	90.91	91.49	90.41	90.91	87.88
InternVL2.5 78b	92.31	88.89	90.32	89.66	92.13	86.15
closed-source models						
GPT-4o	98.18	94.74	94.85	91.89	92.13	94.29
o1 2	96.30	94.1	96.97	94.04	95.65	91.18
Claude	94.34	87.90	92.47	88.11	90.91	76.67
Average	94.52	90.09	91.98	90.15	92.40	88.28
Rank	1	5	3	4	2	6

Table 10: **AE performance per category.** AE performance per anomaly category for vanilla inference prompt.

G.6 Analysis by anomaly category

Using our anomaly taxonomy (Section 2), we categorize GPT-4o's FPs in AD and find it most often hallucinates attribute, relation, and textual anomalies (Figure 7; see classifier details in Section G.1). Although textual anomalies are among the most frequently hallucinated, they are handled best, with

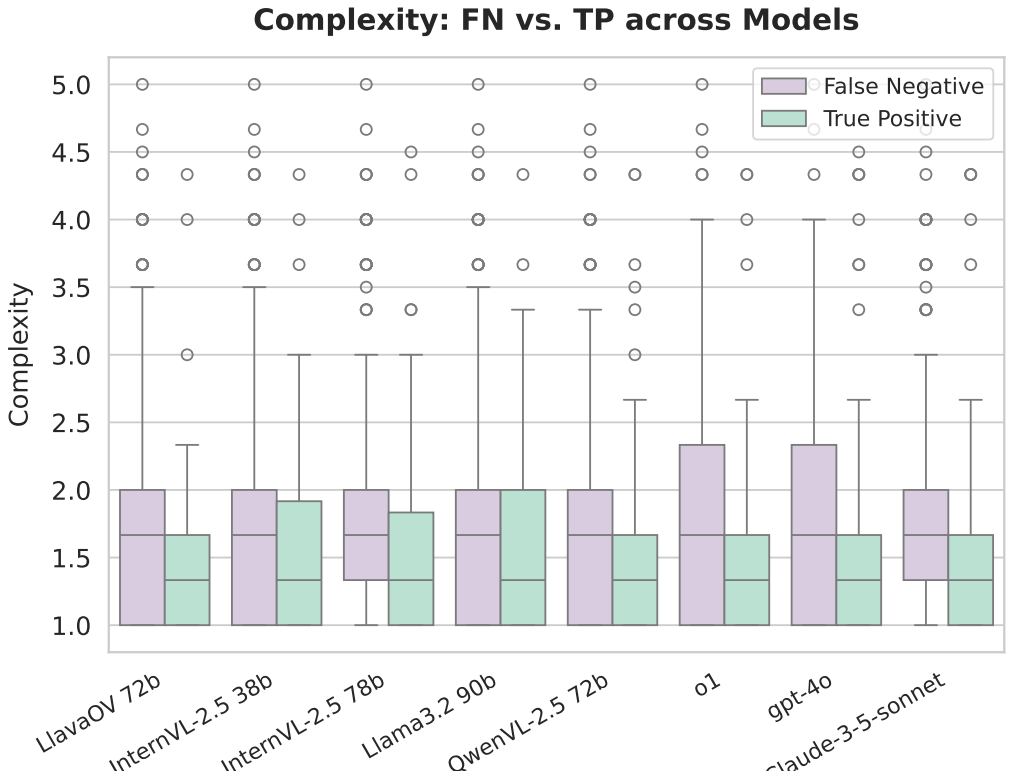


Figure 11: **Models' performance across complexity feature.** Plot showing the deviation in the models' performance across different levels of anomaly complexity for the anomaly description task. The results reveal that models perform well only on low-complex tasks but also exhibit false positives for much simpler anomalies on average.

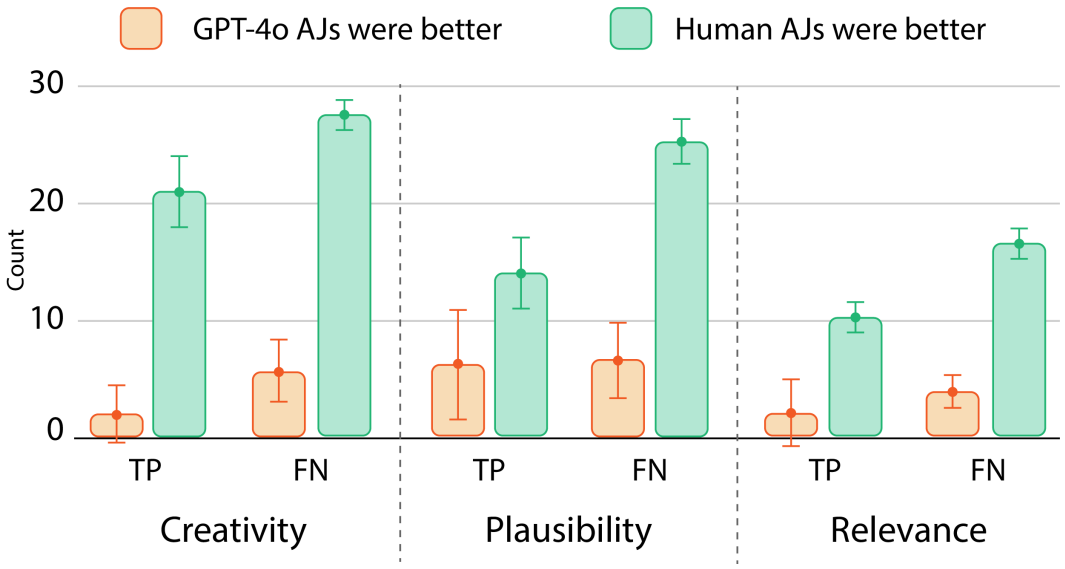


Figure 12: **AJ Results.** Comparison of GPT-4o vs. Human Anomaly Justifications.

top detection (56.28%) and strong explanation scores (92.40%) (See Appendix Table 8 and Table 10). In contrast, uniformity anomalies, which are rarely hallucinated, are the hardest to detect (28.92%) and explain (88.28%). Interestingly, absence anomalies show low detection (28.20) but the highest

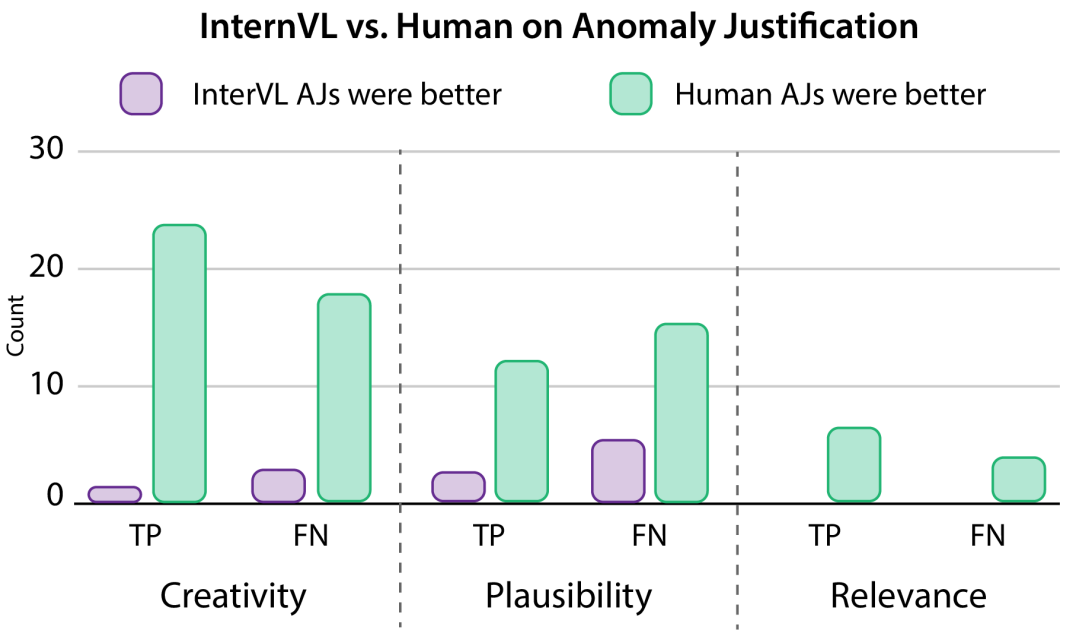


Figure 13: **Comparison of InternVL vs. Human Anomaly Justification.** Bars above the x-axis indicate cases where InternVL outperformed humans, while bars below indicate cases where InternVL underperformed. The 3 bars on the left are results over 50 False Negatives (FN), where the model failed to identify anomalies; the 3 bars on the right are over 50 True Positives (TP).

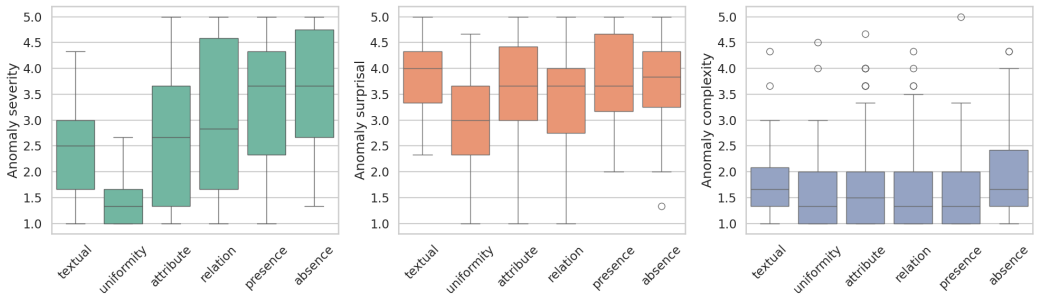


Figure 14: Distribution of anomaly scores across categories. The boxplots illustrate the distribution of complexity, severity, and surprisal scores across different anomaly categories, highlighting variations in human perception of anomalies.

explanation performance (94.52), suggesting models can reason well once the anomaly is identified. Overall, *harder-to-detect categories are also harder to explain*.

G.7 Cultural bias assessment

Considering the diversity of cultures and personal experiences, a situation may be perceived as anomalous in one cultural context while appearing entirely normal in another [16, 41, 60]. To investigate this phenomenon, we manually investigate which of the anomalies in CAVE reflect cultural biases. Our analysis shows that while the majority of anomalies are independent from cultural influence, a subset of four cases may reflect a Western-centric bias in their annotations. Notably, when GPT-4o is prompted with these images, it consistently identifies them as anomalies, suggesting an implicit alignment with Western cultural norms in the model's internal knowledge. Further details on the experimental setup and findings are provided in Section E.5 and the four cases are depicted in Figure 6.

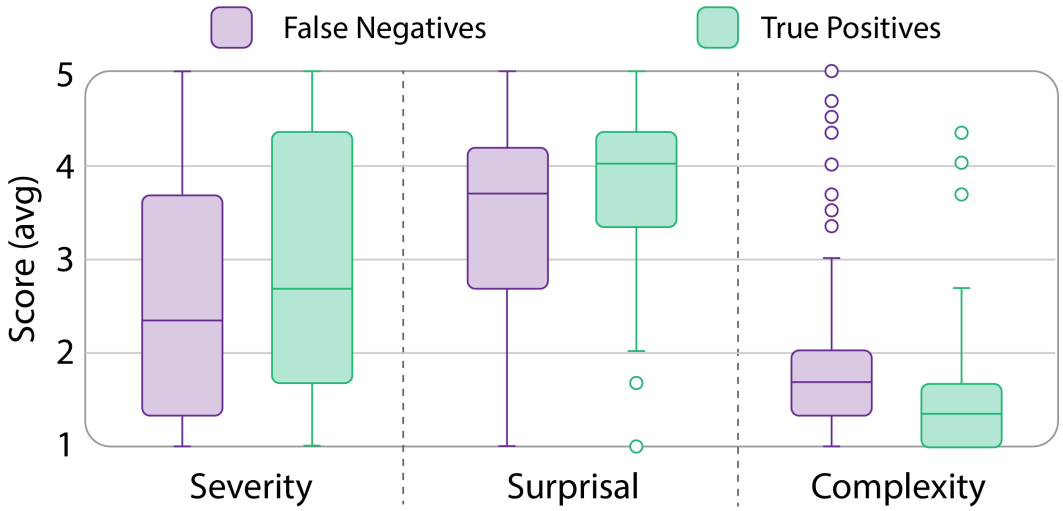


Figure 15: **Distribution of anomaly descriptions stratified by GPT-4o's TP vs FN** across severity, surprisal, and complexity scores.

H Failure examples

Examples of model predictions on images of CAVE are shown in Figure 18 to 23.

Examples of incorrect AE generations using GPT-4o are given in Figure 16. In the first image, the model generation mentions that the rusty wrench is a functional part of the control panel rather than recognizing it as an improvised obstruction to the switch usage. In the second image, it doesn't give reason for why the toilet bowl should not have objects. In the third image, it misidentifies the anomaly by describing the miscolored chair as white and entirely ignoring the missing chair.

Examples of incorrect AJ generations using GPT-4o are given in Figure 17. In the first image, the provided justification is highly implausible and lacks creativity. In the second example, the model makes a reasoning mistake when generating the AJ, assuming that 292 is a valid age displayed on a birthday cake. In the third example, the model makes a perception mistake, incorrectly describing the trash can as full or not properly open; the resulting AJ is plausible but incorrect given the image.

GT Anomaly Description	Human Explanation	GPT4o Explanation
A rusty wrench is bolted to a control panel to hold a switch in a permenant position.	The presence of the rusty wrench prevents the usage of the switch.	Using a rusty wrench to hold a switch in place on a control panel is unusual because it is an improvised and potentially unsafe method for securing electrical components, which should be done with proper tools and mechanisms.

GT Anomaly Description	Human Explanation	GPT4o Explanation
There is a soda can in the toilet bowl.	A toilet needs to be devoid of any objects in the bowl that might cause blockages for proper functioning.	A soda can is typically found in kitchens or dining areas, not in a toilet bowl, which is meant for sanitation purposes.

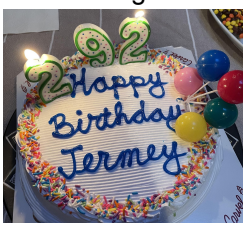
GT Anomaly Description	Human Explanation	GPT4o Explanation
(1) The second from left chair in the first row is in the wrong color. *(2)There is a chair missing on the second row from the back.*	The regularly alternating pattern is broken due to the second from left chair in the first row being red instead of white.	The second chair from the left in the first row is white, while the rest of the chairs in that row are red, disrupting the color pattern.

Figure 16: **GPT-4o Incorrect AE Generations.** Three examples from the AE task where GPT-4o incorrectly explains the anomaly. *Note:* In the rightmost image, two anomalies are present, but only the first one is considered in this example of a failed explanation.

Low plausibility and creativity

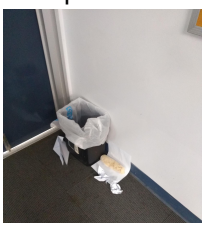
GT Anomaly Description	Human Justification	GPT4o Justification
The logo for the restaurant is off-center and sticks out of the side of the building	There used to be a second sign on the right, but it fell before the photo was taken	The logo may have been improperly installed due to lack of precise measurement or miscommunication during the installation process

Reasoning error



GT Anomaly Description	Human Justification	GPT4o Justification
The candles depict the age of “Jermey” as 292 years old	The candles have been placed as a joke and exaggerating Jermey’s age	The candles might have been arranged incorrectly, resulting in the age being displayed as 292 instead of the intended order, such as 29 or 92.

Perception error



GT Anomaly Description	Human Justification	GPT4o Justification
There is trash on the floor next to the trash can	A person threw the trash towards the bin and missed it	The trash can is full or was not opened properly, causing overflow and resulting in trash being left outside

Figure 17: **GPT-4o Incorrect AJ Generations.** Three examples from the AJ task where GPT-4o provides a poor or incorrect justification.

Example 1

GT AD: The button for floor number 2 is missing.
✗**o1:** The braille labels for 3 and 5 appear identical., The floor numbers are arranged out of a typical sequence.
✗**GPT-4o:** The button for floor 1 is labeled with a star instead of a number.
✗**LlavaOneVision:** The elevator button for the first floor is marked with a star instead of the number 1., The close button is red, which is unusual as it is typically found in a different color.
✗**InternVL:** The number 1 button is missing its label.
✗**QwenVL:** The button labeled '1' has a star symbol next to it, which is not present on other numbered buttons.

Figure 18:

Example 2

GT AD: The poster says forty carrots but there are only twenty eight illustrations of carrots.
✗**o1:** No output
✗**GPT-4o:** The word 'forty' is written with a carrot symbol replacing the letter 'i'.
✗**LlavaOneVision:** No output
✗**InternVL:** No output
✗**QwenVL:** No output

Figure 19:

Example 3

GT AD: One of the men is standing on the far end of the plank with no other support or safety gear.
✗**o1:** no output
✗**GPT-4o:** no output
✗**LlavaOneVision:** no output
✗**InternVL:** no output
✗**QwenVL:** no output

Figure 20:

Example 4

GTAD: There is a chair missing on the second row from the back.
✗**o1:** No output
✗**GPT-4o:** The seating arrangement is missing a seat in the bottom row, creating an empty space.
✗**LlavaOneVision:** No output
✗**InternVL:** No output
✗**QwenVL:** The seat at the bottom right corner appears to be partially cut off by the edge of the image.

Figure 21:

Example 5

GT AD: The vending machine is sponsored by Aquafina, a water brand, but there are no Aquafina water bottles being sold by the machine.
✗**o1:** No output
✓**GPT-4o:** The vending machine labeled 'Aquafina' contains a variety of drinks other than Aquafina water.
✗**LlavaOneVision:** No output
✗**InternVL:** A power cord is visible inside the vending machine.
✗**QwenVL:** No output

Figure 22:

Example 6

GT AD: The label on the water bottle is positioned upside down.
✓**GPT-4o:** The label on the water bottle is upside down.
✗**o1:** There is a small hole in the tile with brownish staining surrounding it.
✗**LlavaOneVision:** The water bottle is upside down., The water bottle is not properly sealed.
✓**InternVL:** The label on the water bottle is upside down.
✓textbfQwenVL: The water bottle label is upside down.

Figure 23:

Vanilla Anomaly Description (AD) Prompt

System: You are an advanced Vision-Language Model designed to extract and describe if any visible anomalies are present in images with precision and brevity.
User: Your task is to accurately describe anomalies observed in images in a concise manner. An anomaly is anything that deviates from what most people consider standard, normal, or expected. It can be an unusual element, action, or occurrence in an image that would seem surprising or out of place to most people.
Key principles for anomaly identification:
1. Base your descriptions solely on what is clearly visible in the image—no assumptions or external knowledge.
2. Descriptions must be objective and straightforward, without explanations, justifications, or subjective language.
Example Format:
If anomalies are present:
{
"anomaly_description_1": [Anomaly description],
...
}
If no anomalies are visible:
{}
Please just provide the description of the anomaly or return '{}' if there are no anomalies. **Do not provide any explanation or justification of why it is anomalous.**
Note:
- Limit each description to a maximum of two sentences.
- Focus on describing the anomaly itself.
- Do not focus on **why it is unusual or unexpected**.

Let's analyze the image and describe the visible anomalies:

Figure 24:

Chain-of-Thought (CoT) Prompt

System: You are an advanced Vision-Language Model designed to extract and describe if any visible anomalies are present in images with precision and brevity.
User: Your task is to carefully analyze the image, reason step-by-step through what is visible, and then describe any anomalies in a concise and objective manner.
An anomaly is anything that deviates from what most people consider standard, normal, or expected. It can be an unusual element, action, or occurrence in an image that would seem surprising or out of place to most people.
Key principles for anomaly identification: 1. Base your descriptions solely on what is clearly visible in the image—no assumptions or external knowledge. 2. Descriptions must be objective and straightforward, without explanations, justifications, or subjective language.
Example Format:
Let's think step by step: [Thinking steps]
If anomalies are present:
{
"anomaly_description_1": [Anomaly description],
...
}
If no anomalies are visible:
{ }
Note:

- Limit each description to a maximum of two sentences.
- Focus on describing the anomaly itself.
- Do not focus on **why it is unusual or unexpected**.

Let's analyze the image, think step by step and then describe the visible anomalies:

Figure 25:

Multi-step reasoning (MS CoT) Prompt

System: You are an advanced Vision-Language Model designed to extract and describe if any visible anomalies are present in images with precision and brevity.
User: Your task is to accurately describe anomalies observed in images in a concise manner. An anomaly is anything that deviates from what most people consider standard, normal, or expected. It can be an unusual element, action, or occurrence in an image that would seem surprising or out of place to most people.
Your goal is to carefully analyze the image using simple, structured reasoning, and describe any visible anomalies. Do not use external knowledge or assumptions — only what can be clearly seen in the image.
Use the following structure in your response:
1. **Planning**: Briefly explain the steps you will take to perform the task.
2. **Image Contents**: List the main elements visible in the image (e.g. objects, people, actions, text).
3. **Step-by-step reasoning**: Think through the image in a logical sequence to identify if anything looks unusual or out of place.
4. **Final Answer**: If anomalies are present:
{
"anomaly_description_1": [Anomaly description],
...
}
If no anomalies are visible:
{}
Note:
- Limit each description to a maximum of two sentences.
- Focus on describing the anomaly itself.
- Do not focus on **why it is unusual or unexpected**.

Let's begin by planning, then analyzing the image step by step, and finally reporting any anomalies found:

Figure 26:

Self-consistency ensembler Prompt

System: You are an advanced Vision-Language Model designed to extract and describe if any visible anomalies are present in images with precision and brevity.
User:
You are given three sets of anomaly descriptions for the same image:
1. [*Anomaly Descriptions from 1st inference*]
2. [*Anomaly Descriptions from 2nd inference*]
3. [*Anomaly Descriptions from 3rd inference*]
Your job is to identify the anomaly descriptions that are repeated — that is, descriptions that appear in at least twice. These may be worded slightly differently but must describe the same anomaly.
Do not make up any new descriptions. Ignore differences in phrasing if the meaning is clearly the same.
Return only the repeated anomaly descriptions in the given json format.
Only include those that appear atleast twice.
Example Output Format:
{
"anomaly_description_1": [Anomaly description],
...
}
Do not include any extra explanation.

Figure 27:

Anomaly Explanation (AE) Prompt

System: You are an advanced Vision Language Model designed to analyze anomalies seen in images.
User: You will be provided with an image along with a textual description of an anomaly present in the image. Explain why the anomaly is anomalous.
Anomaly: [Anomaly description and explanation]
Let's analyze the image and the anomaly first and then provide a concise response (max 2 sentences) in a json format:
{
"explanation": [Explain why it is unusual or unexpected]
}

Figure 28:

Anomaly Justification (AJ) Prompt

System: You are an advanced Vision Language Model designed to analyze anomalies seen in images.
User: You will be provided with an image and the description of an anomaly seen in the image. Provide a single, most realistic and plausible reason for how the anomaly happened.
Anomaly: [Anomaly description and explanation]
Let's analyse the image and the anomaly first and then provide a concise response (max 2 sentences) in a json format:
{
"justification": [Describe how the anomaly happened]
}

Figure 29:

Anomaly Severity Prompt

System: You are an advanced Vision-Language Model designed to analyze images and detect anomalies.
User: Assess the severity of the consequence of the anomaly on a scale from 1 to 5:

1. (No consequence): The anomaly is purely aesthetic or has no functional/safety impact. Example: A tile of a different color on the pavement.
2. (Low Concern)
3. (Moderate Concern): The anomaly may cause inconvenience or inefficiency but does not pose immediate risks. Example: A misaligned sign that is still readable.
4. (High Concern)
5. (Requires Immediate Action): The anomaly presents a safety hazard, major malfunction, or significant risk. Example: A worker using a circular saw without protection gear.

Inputs:
- Image: (Attached image)
- Anomaly Description: { }

Provide a severity rating in this format:
{
"severity": [Score between 1 and 5]
}

Figure 30:

Anomaly Surprisal Prompt

System: You are an advanced Vision-Language Model designed to analyze images and detect anomalies.
User: Assess how surprising or uncommon the anomaly is on a scale from 1 to 5:

1. (Common): Frequently observed in similar contexts; most people would not be surprised. Example: A car parked in an inconvenient way.
2. (Relatively Common)
3. (Average): Might raise curiosity but not shock. Example: A person eating spaghetti with chopsticks.
4. (Uncommon)
5. (Extremely Rare): Highly uncommon and surprising; most people have never seen it before. Example: A tree growing upside down from a roof.

Inputs:
- Image: (Attached image)
- Anomaly Description: { }

Provide a surprisal rating in this format:
{
"surprisal": [Score between 1 and 5]
}

Figure 31:

Anomaly Complexity Prompt

System: You are an advanced Vision-Language Model designed to analyze images and detect anomalies.
User: Assess how difficult it would be for a person to detect the anomaly on a scale from 1 to 5:

1. (Easy): Most people would notice the anomaly immediately without effort. Example: A red apple among green apples.
2. (Mild)
3. (Moderate): Requires some focus to identify but becomes clear after a few seconds. Example: A misspelled word on a sign.
4. (Difficult)
5. (Very difficult): Blends into the surroundings or demands specific knowledge to identify. Example: A contradiction in the screenshot of an email.

Inputs:
- Image: (Attached image)
- Anomaly Description: { }

Provide a complexity rating in this format:
{
"complexity": [Score between 1 and 5]
}

Figure 32:

Anomaly Description Evaluation Prompt

System: You are an advanced AI assistant designed to compare two descriptions of an anomaly in the image attached.
User: Compare the following two descriptions of an anomaly in an image. Judge whether they describe the same anomaly. If they match, respond with 'Yes' and briefly explain why. If they differ, respond with 'No' and provide a reason for the difference.
REFERENCE: [*Ground truth anomaly description*]
RESPONSE: [*Model-generated anomaly description*]

Figure 33:

Anomaly Explanation Evaluation Prompt

System: You are an advanced AI assistant designed to compare two explanations for a visual anomaly.
User: Determine whether the model explanation accurately reflects the core reasoning in the human annotation for why the given anomaly is considered unusual in the image.
The explanation does not need to match the human annotation word-for-word, but it should be logically aligned and refer to the same underlying cause.
Minor differences in wording are acceptable, but explanations that are unrelated or based on a different logic should be marked as incorrect.
Anomaly Description: [*Ground truth anomaly description*]
Human explanation: [*Human annotation*]
Model explanation: [*Model-generated anomaly explanation*]
If the explanations are unrelated or based on a different logic, answer 'False'.

Figure 34:

Anomaly Category Classification Prompt

System: You are an expert in classifying visual anomalies based on descriptions.
User: You are given a taxonomy of anomaly types:
1. Entity Presence –An object is present when it shouldn't be.
2.Entity Absence – An expected object is missing.
3. Entity Attribute – An object has an unusual attribute (color, shape, label, orientation, usage).
4. Spatial Relation – Objects are positioned or oriented incorrectly relative to one another.
5. Uniformity Breach – A disruption in an expected pattern or symmetry.
6. Textual Anomaly – The image contains text that is contradictory, unexpected, or illogical.
Given the following anomaly description, classify it into one of the five categories. Only respond with the category name.

Anomaly description: [*Model generated anomaly description*]

Figure 35:

Anomaly Cultural Analysis Prompt

User: You are a culturally-aware AI with expertise in global customs, social norms, and visual analysis. Based on the image, description, and noted anomaly:
Analyze the anomaly within its cultural context.
Determine if it aligns with any specific cultural, religious, regional, or historical norms.
If yes, identify the culture/region and explain why this is considered normal there.
If no, clearly state that and briefly explain why it does not align culturally.
Be objective, respectful, and avoid stereotypes. Consider that some anomalies may have universal meaning without cultural bias.
Respond as a dictionary with keys:
- cultural alignment: "yes" or "no"
- context: the relevant cultural norm that explains the anomaly (or null if none)
- justification: explanation why the anomaly is normal or not culturally aligned

Inputs:
- Image: (Attached image)
- Anomaly Description: [*Ground truth anomaly description*]

Figure 36: