CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments

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

[1]EPFL, [2]MILA

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 visual anomalies. CAVE supports three open-ended tasks: anomaly description, explanation, and justification; with fine-grained annotations for visual grounding and 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. We 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. We release the code and benchmark on our project webpage.[1]

1 Introduction

"If you notice an abnormal situation, please contact an agent." Such announcements are commonplace in public spaces worldwide, highlighting a fundamental human trait: the ability to detect anomalies. Identifying uncommon situations, behaviors, and other elements that deviate noticeably from a norm is a natural and expected behavior for humans (Klein et al., 2007; Klein, 2013).

As Vision-Language Models' (VLMs) (Li et al., 2024a; OpenAI, 2024; Awadalla et al., 2023;
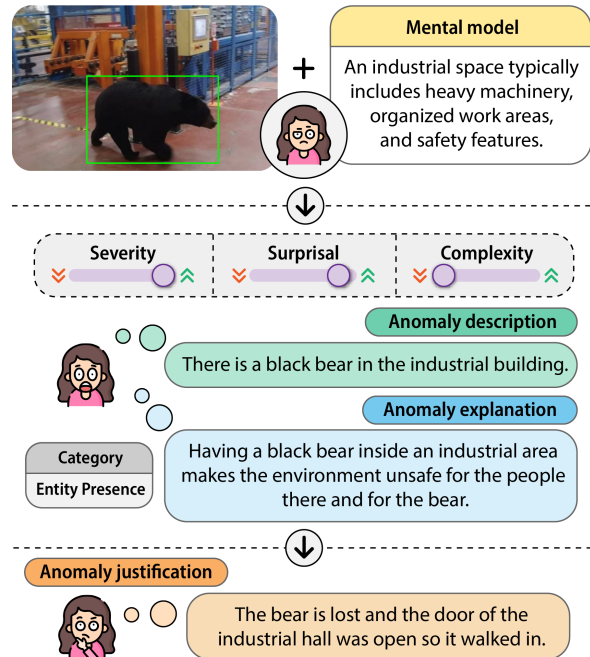

Figure 1: **CAVE Example.** CAVE contains images captured in real-world scenarios, annotated with commonsense anomalies, along with their textual explanations, justifications and bounding boxes around anomalous elements. It also includes numerical features representing how humans perceive these anomalies.

Laurençon et al., 2024) functionality broadens and deployment in real-world scenarios expands (Jin et al., 2024; Xu et al., 2024), so does their exposure to unexpected and novel situations. In this expanding landscape, their ability to differentiate between normal and anomalous situations is crucial for ensuring safe and efficient operation (Mullen Jr et al., 2024).

These models leverage the comprehensive world knowledge and reasoning capabilities of their Large Language Model (LLM) backbones (McKinzie et al., 2024; Liu et al., 2024a; Karamcheti et al., 2024; Laurençon et al., 2024), equipping them to handle a variety of tasks (Liu et al., 2024a; Caffagni et al., 2024). However, rare or uncommon situations are inherently underrepresented in training data, making

---

*Equal contribution.

[1] https://smontariol.github.io/cave-visual-anomalies/

it challenging for models to learn how to recognize them and to react to them. Moreover, VLMs tend to hallucinate image content toward the most probable interpretation, which directly conflicts with their ability to identify unexpected situations (Zhou et al., 2023).

While it is crucial to accurately evaluate VLM's ability to identify and understand anomalous situations, the scope of existing visual anomaly detection benchmarks is limited. In the literature, visual anomaly detection is mainly applied to specific domains such as industrial inspection (Chandola et al., 2009; Diers and Pigorsch, 2023; Bergmann et al., 2019a; Xie et al., 2024), medical diagnosis (Fernando et al., 2021; Zhang et al., 2020) or video surveillance (Sultani et al., 2018). More recently, commonsense-oriented anomaly detection benchmarks have started to appear. They typically rely on synthetic image generation to create artificial scenarios (Bitton-Guetta et al., 2024; Li et al., 2024b; Roman and Meyer, 2024; Bitton-Guetta et al., 2023; Tai et al., 2024). Non-synthetic approaches rely on domain-specific datasets, such as understanding creative elements in advertisements (Malakouti et al., 2024) or detecting video game glitches (Taesiri et al., 2024). 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 cognitive science literature, we pair each extracted anomalous image with annotations supporting three open-ended tasks that align with human anomaly detection and sense-making processes: anomaly description, anomaly explanation, and anomaly justification. Additionally, we support an anomaly localization task to evaluate the visual grounding capabilities of the models. 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: (a) how severe an anomaly is, (b) how surprising and uncommon it is, and (c) how complex it is to detect (see Figure 1). **Our main contributions and findings are as follows**:

- We propose an anomaly understanding framework that builds upon cognitive science literature regarding the way humans identify and understand anomalies (Section 2). We split the detection process into three sub-tasks formalized as open-ended visual question answering, and include a classification system based on visual manifestations and numerical attributes. This novel framework allows for systematic characterization and annotation of visual commonsense anomalies.

- We introduce 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 57% F1-score on anomaly detection with a multi-step reasoning strategy, highlighting significant room for improvement.

- We analyse 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.

2 Theoretical Framework

We leverage cognitive science literature to formalize the way humans detect and understand anomalies into a set of tasks. This framework guides our dataset creation process, model assessment, and analysis, allowing us to explore the alignment between human and machine processing of visual anomalies.

2.1 Perception of the anomaly

Anomaly detection focuses on identifying deviations from expected patterns (Klein et al., 2007; Klein, 2013). In this work, we define an anomaly not simply as a statistical rarity (Grubbs, 1969; Chandola et al., 2009; Pang et al., 2021) but as a situation that disrupts an established pattern or expectation. This perspective underscores the key human ability to construct *mental models* of the world and identify deviations from these models (Klein et al., 2005). Mental models are cognitive representations of the world, guiding information processing and anticipation of events (Borders et al., 2024). When an individual encounters something unexpected or surprising that disrupts their established mental model, it can be perceived as an anomaly (Klein et al., 2007); however, humans' accurate perception of situations is limited when doing rapid visual processing (Treisman and Schmidt, 1982; De Keyser

and Woods, 1990). The process of identifying this anomaly depends on three main characteristics.

Anomaly complexity. Complexity is often operationalized by evaluating perceptual features that influence how quickly and efficiently the brain detects anomalies in visual search tasks (Sun and Firestone, 2021). Visual anomalies are often salient stimuli that attract attention due to their deviation from expected patterns; more visually complex anomalies require greater cognitive resources for processing (Donderi, 2006; Guo and Chen, 2023). We leverage this formalization of complexity to assess the difficulty of detecting anomalies in CAVE.

Anomaly severity. Anomalies that signal immediate danger or high risk are more likely to be detected. Humans use both cognitive appraisal, *i.e.*, evaluating the potential consequences, and emotional arousal, such as fear and anxiety, to assess the severity of an anomaly (Rabeyron and Loose, 2015). Hence, we operationalize severity by asking to what extent the anomaly requires immediate action.

Anomaly surprisal. Surprise-based theories assess severity by how much an event updates prior beliefs (Bayesian Surprise) (Itti and Baldi, 2009) or the amount of unexpected information it contains (Information-theoretic Surprise) (Baldi and Itti, 2010). Prediction Error Theory measures surprisal by the magnitude of the discrepancy between expectations and reality, as well as the confidence in the original expectation (Friston, 2005). Following these concepts, we operationalize surprisal under the question "How much does the situation deviate from expectations?".

We use these three formalizations to quantify how humans perceive and detect an anomaly. Similarly to Campbell et al. (2024), we posit that there are commonalities in the way humans and machines process visual information, linking model visual processing limitations with human cognition constraints; and evaluate VLMs' ability to detect anomalies depending on these features.

2.2 Understanding of the anomaly

When a human detects an anomaly, the main underlying task is **anomaly description** (Klein et al., 2007; Klein, 2013): identifying and articulating what elements in the environment are inconsistent with expectations. The next step involves a reassessment of the observer's mental model: why does the situation appear anomalous (Klein et al., 2023)? In other words, why did these expectations exist in the first place (Heyes, 2024)? We define this intermediate step as **anomaly explanation**. It aims at assessing the

model's understanding of underlying commonsense knowledge on why the situation deviates from the norm.

Finally, in contrast to typical datasets that often contain artificially generated or staged anomalies, each image in CAVE represents a real-world scenario captured as an actual photograph or screenshot taken by an individual. These anomalies document real events, prompting the observer to naturally question, "How did this happen?" This leads to the final step, termed sense-making (Williams et al., 2012; Zhang and Soergel, 2014; Klein et al., 2023), which involves making hypotheses to make sense of the anomaly. To encapsulate this process, we define the **anomaly justification** task. It involves providing a plausible explanation for the anomaly by describing a sequence of events or circumstances that could have led to the scene.

2.3 Manifestation of the anomaly

The manifestation of an anomaly refers to the specific way in which it appears or deviates from the expected pattern in visual data. This categorization enables targeted benchmarking of VLMs against real-world challenges, ensuring that their anomaly detection capabilities generalize across diverse anomaly manifestations and complementing the cognitive aspect illustrated by the concepts of anomaly complexity, severity and surprisal.

Inspired by MMBench's taxonomy of visual reasoning types (Liu et al., 2025), we categorize the ways anomalies manifest in images as follows (see examples for each category in Appendix Figure 7).

- **Entity Presence/Absence**: An object is present in the image when it shouldn't be, or an expected object is missing.
- **Entity Attribute**: An object exhibits an anomalous attribute, such as an unusual color, shape, label, orientation, or usage.
- **Spatial Relation**: An object is incorrectly positioned or oriented relative to another specific object.
- **Uniformity Breach**: A disruption in an expected pattern, such as an out-of-place element in a uniform or symmetrical arrangement.
- **Textual Anomaly**: Some text in the image conveys an unexpected or contradictory message.

3 Dataset

We introduce CAVE, a vision-language benchmark which builds on our theoretical framework to evaluate the commonsense anomaly detection and understanding capabilities of VLMs.
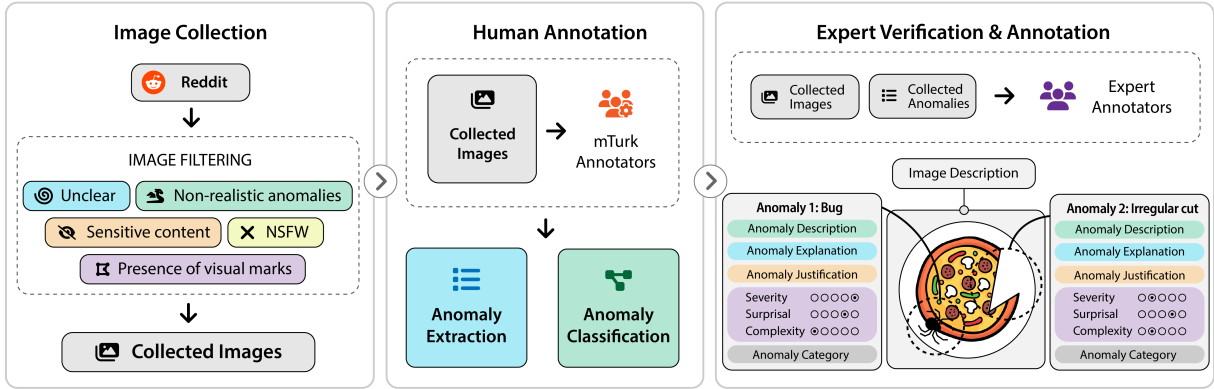
Figure 2: **An overview of CAVE data collection process. (1) Image Collection**: Images were sourced from the top 1,000 posts across various subreddits and filtered to ensure high-quality, safe data. **(2) Human Annotation**: Initial annotations were performed by Mechanical Turk workers, focusing on 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.

3.1 Dataset Construction

Since our benchmark focuses on real-world, daily-life visual anomalies, our data collection process and annotation strategy are strongly human-centered. The dataset creation process is illustrated in Figure 2.

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 or sensitive content. We apply automatic and manual filters (see Appendix E.1 for details), and then annotate the remaining images through two annotation rounds.

Data Annotation. First, each image was annotated by 5 annotators via Amazon Mechanical Turk. They were asked whether each 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 annotated a single bounding box per anomaly and consolidated the initial textual annotations by validating and formalizing them along the following axes:

1. **Anomaly Description (AD):** A textual description of the anomaly in the image.
2. **Anomaly Localization (AL):** The coordinates of the bounding box demarcating the anomaly.
3. **Anomaly Explanation (AE):** An explanation of why it is anomalous.
4. **Anomaly Justification (AJ):** A realistic and plausible explanation for how the anomaly might have occurred.
5. **Anomaly Category**: Category based on the anomaly manifestation taxonomy outlined in Section 2.3. Anomalies about entity attributes, spatial relations, and textual anomalies are the most frequent (Figure 3).

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 3 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, with 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 6 in Appendix 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
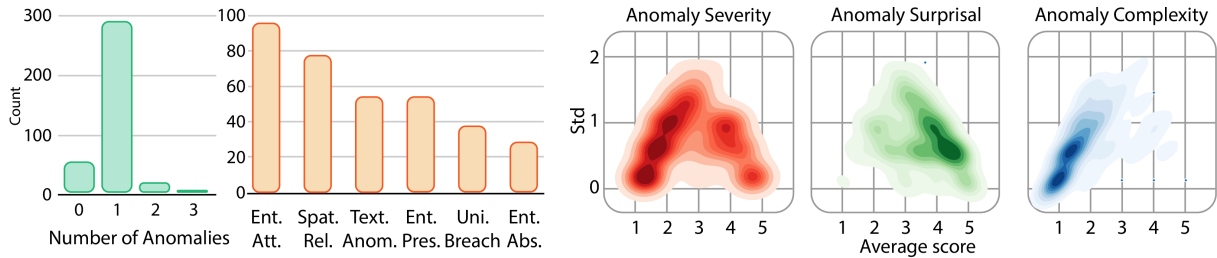
Figure 3: **CAVE statistics.** Distribution of the number of anomalies per image (left). Number of images in each anomaly category (middle). Density of severity, surprisal and complexity scores per average score and standard deviation (right).

distributions, we turn to the more adapted Gwet's AC2 (Gwet, 2008), which shows a much higher agreement for the complexity score (0.76).

In the case of anomaly localization, the bounding box annotations show that most anomalies occupy only a small area of the image, with an average of 24% (median area is 16%). This illustrates the difficulty of identifying small-sized anomalies in complex real-world scenes. Note that large anomaly areas are mostly associated with textual anomalies, where the entire text region is annotated.

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, each paired with a unique bounding box. Overall, CAVE exhibits a rich diversity of anomalies (see Figure 3 and Figure 15) 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.

3.2 Evaluation

Anomaly Description (AD). Since an image can contain multiple anomalies, we evaluate model predictions by performing systematic pairwise comparisons between each ground-truth anomaly description and each model-generated output. To enable scalable evaluation, we employ GPT-4o as a judge to assess whether two descriptions refer to the same anomaly (Liu et al., 2024b; Zheng et al., 2023; Liusie et al., 2024). This judge achieves 90% accuracy on 50 manually annotated pairs (GPT-4o vs. human), confirming its reliability (see judge prompt in Figure 34). Matched pairs are labeled as True Positives (TP), unmatched ground-truth descriptions as False Negatives (FN), and unmatched model outputs as False Positives (FP). We compute the precision, recall, and F1-score using these counts.

Anomaly Localization (AL). Given the ground-truth AD for each anomaly, we evaluate a model's ability to predict the correct bounding box by computing the Intersection over Union (IoU) score between the bounding box generated by the model and the ground-truth bounding box. The prompt is given in Figure 28.

Anomaly Explanation (AE). Each AD in CAVE is paired with a single explanation. We evaluate a model's ability to generate the correct explanation when provided with the ground-truth AD using an LLM judge, comparing model-generated and ground-truth explanations. To validate the judge, we manually label explanation pairs for 50 TPs and FNs from the AD task, achieving 89% accuracy on the label subset (see judge prompt in Figure 35).

Anomaly Justification (AJ). Since there can be more than one correct anomaly justification, we evaluate a justification quality 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 subjectivity of these criteria, we rely entirely on human evaluation.[2] 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, similar or worse. We report the average win rate according to the three annotators.

4 Experiments

We evaluate 5 state-of-the-art open-source models and 3 closed-source models (see model details in Appendix Table 3) on the AD task (Section 4.1), revealing limitations in the form of perception and reasoning errors. To further investigate these shortcomings, we analyze the models' commonsense

---

[2]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.

reasoning with our two complementary tasks: AE (Section 4.3) and AJ (Section 4.4). Then, we analyze the performance of models on the AD task against numerical features (Section 4.5) and visual manifestation (Section 4.6), identifying the most challenging aspects of anomaly detection.

4.1 Anomaly Description

Inference. We prompt each model to describe the anomalies in the input image and perform evaluation using LLM-as-a-judge. To ensure consistency and reduce evaluation bias, the prompts were carefully aligned with the instructions provided to human annotators (see prompt in Appendix F).

Vanilla Prompt Performance. Table 1 shows that using vanilla prompt, the best performance is achieved by GPT-4o with a 51.2% F1-Score.

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 understanding. 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*. 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 2 (first row), 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 2: **GPT-4o Qualitative Analysis.** Proportion of FP error analysis across prompting strategies as determined by human evaluation.

Advanced Prompting Strategies. These findings highlight the need for fine-grained visual and contextual reasoning, which general-purpose prompts fail to trigger. As fine-tuning is infeasible due to limited data, we explore five advanced prompting strategies. **(1) Chain-of-thought (CoT)** encourages models to generate explicit reasoning steps before answering (Wei et al., 2022); **(2) Set-of-marks (SoM)** leverages GroundingDINO (Liu et al., 2023) to generate object-level visual annotations to enhance visual grounding (Yang et al., 2023a); **(3) Combined CoT+SoM** integrates visual grounding annotations to the model's step-by-step reasoning; **(4) Multi-step CoT** guided models to plan reasoning steps, identify and describe key image elements, and reason step-by-step before answering (Xu et al., 2025); **(5) CoT + Self-consistency** generates three model outputs per image (using temperature=0.5) and applies consensus-based aggregation using the same model, considering only anomalies detected in at least two out of three generations, thereby reducing spurious detections (Wang et al., 2022).

As shown in Table 1, the five advanced prompting strategies lead to limited improvements over the vanilla baseline across all VLMs (see significance tests in Appendix Appendix G.1). CoT + self-consistency demonstrates the strongest overall performance gain (+4.83%). The highest absolute score remains at

56.64% with GPT-4o using multi-step reasoning prompting (MS CoT). The performance of many models degrades with the SoM and CoT+SoM prompting strategies; this is due to faulty spatial annotations introducing noise into the anomaly detection task (see example Figure 10). A more fine-grained analysis on TPs and FNs (Table 9 in appendix) shows that all strategies boost TPs over the vanilla baseline, especially self-consistency and MS CoT, while the number of FNs generally increases with advanced strategies. Notably, CoT, MS CoT, and Self-consistency sharply reduce FP, yielding fewer spurious detections.

Finally, qualitative analysis (Table 2) shows that for the MS CoT prompting strategy, the false positives predominantly shift toward perception errors (68% vs. 44% under vanilla prompting), with fewer reasoning mistakes (32% vs. 49%). Failure examples are shown in Appendix H. A more detailed analysis of performance across different anomaly categories is provided in Section 4.6.

4.2 Anomaly Localization

Inference. Given the ground truth anomaly description as input, we prompt the best-performing model, GPT-4o, to predict the bounding box coordinates corresponding to the anomaly mentioned in the AD. We then compute the Intersection-over-Union (IoU) score of the predicted bounding box with respect to the ground truth.

Results. Only 21.7% of the bounding boxes generated by GPT-4o achieved an IoU ≥ 0.10 with the ground truth. The predicted boxes generally under-cover the true anomalies: the mean ratio of predicted-to-ground-truth coverage is 0.69, with 75% of images exhibiting under-coverage (ratio <0.5). Our error analysis shows that localization accuracy drops significantly in cluttered scenes, and that GPT-4o often focuses on smaller subregions rather than the complete anomalous object or context, confirming the quantitative results.

These findings are consistent with prior work (Yang et al., 2023b; Ramachandran et al., 2025) demonstrating that naive prompting strategies are insufficient for reliable and precise localization with such models.

4.3 Anomaly Explanation

Inference. Building on our findings from the AD task, we examine how well models can explain–rather than identify–visual commonsense anomalies. For each sample, we prompt the model using the AE prompt in Appendix F, instructing it to explain why

a situation is anomalous, given the ground truth anomaly description.

Results. All VLMs achieve over 80% accuracy according to the LLM-as-a-judge evaluation (Table 1). To better understand the link between accurate anomaly description and correct anomaly explanation, we stratify the AE performance results by AD true positive and false negative (see Appendix Table 8). The highest-performing models are o1 (93.02% TP, 88.89% FN) and GPT-4o (90.86% TP, 85.22% FN); on average, there is a 3% performance gap between TP and FN. Hence, it is slightly easier for models to explain an anomaly that it was able to detect. Overall, this analysis allows us to disambiguate the model's ability to perceive the anomaly in the image from its internal knowledge and understanding of the anomaly. VLMs, despite often possessing the commonsense knowledge required to explain an anomaly, are unable to use that knowledge when performing visual processing. Some of the GPT-4o incorrect AE cases are illustrated in Appendix H, Figure 16.

4.4 Anomaly Justification

Inference. We provide the model with both the ground truth anomaly description and explanation alongside the image and ask it to generate a realistic and plausible justification for how the anomaly occurred (see the prompt in Appendix F).

Results. Figure 4 shows how many GPT-4o-generated justifications are better (orange bars) or worse (green bars) than human justifications for each criterion (plausibility, creativity and relevance), averaged over 3 annotators, along with standard deviation. Among the 50 FN and 50 TP samples, on average, fewer than 7 model-generated justifications outperform humans on any criterion. The model's justifications are less creative, plausible, and contextually relevant when it fails to identify the anomaly (FNs), in line with the AE results. This suggests that these harder cases require both stronger perception and deeper commonsense knowledge to generate plausible explanations. When the model successfully detects the anomaly (TPs), its justifications often resemble human explanations, but lack creativity. GPT-4o tends to favor simpler justifications, attributing anomalies to human forgetfulness, accidental errors, or machine failures, while humans often provide more imaginative explanations, sometimes at the expense of plausibility. This explains the few cases where the model scores higher in plausibility than humans. The model also often generated implausible and irrelevant
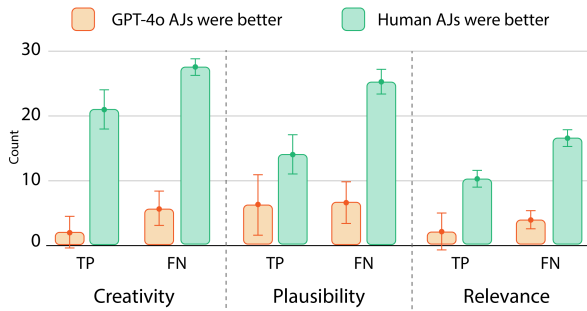
Figure 4: **AJ Results.** Comparison of GPT-4o vs. Human Anomaly Justifications.
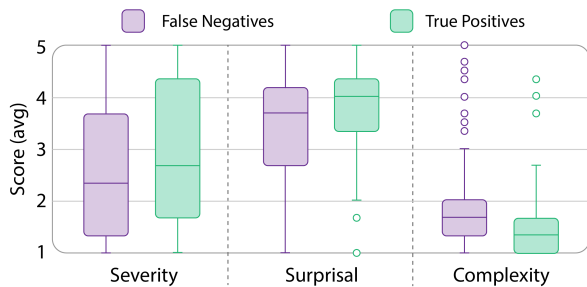

Figure 5: **Distribution of anomaly descriptions stratified by GPT-4o's TP vs FN** across severity, surprisal, and complexity scores.

FNs due to perception and reasoning errors (see examples in Appendix H, Figure 17). Figure 14 in Appendix G.3 presents results for InternVL, which follows a similar pattern across all features.

4.5 Analysis by numerical features

We analyze anomaly detection TPs and FNs across CAVE's three numerical features: severity, surprisal, and complexity (Figure 5). 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 (Appendix G.4).

4.6 Analysis by anomaly category

Using our anomaly taxonomy (Section 2.3), we categorize GPT-4o's FPs in AD and find it most often hallucinates *attribute, relation*, and *textual anomalies* (Figure 9; see classifier details in Appendix G.1). Although *textual anomalies* are among the most frequently hallucinated, they are handled best, with top detection (56.28%) and strong explanation scores (92.40%) (See Appendix Table 10 and Table 12). In contrast, *uniformity* anomalies, which are rarely hallucinated, are the hardest to detect (28.92%) and

explain (88.28%). Interestingly, anomalies on *object absence* show low detection (28.20) but the highest explanation performance (94.52), suggesting models can reason well once the anomaly is identified. Overall, harder-to-detect categories are also harder to explain.

4.7 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 (Goto et al., 2010; Nayak et al., 2024; Ye et al., 2023). To investigate this phenomenon, we manually annotate which of the anomalies in CAVE reflect cultural biases. Our analysis shows that while the majority of anomalies are independent from cultural influence, 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 Appendix E.5; the four culturally-biased samples are shown in Figure 8.

5 Related Works

Historically, image-based anomaly detection has focused on industrial defects such as surface flaws and structural issues (Mishra et al., 2021; Bergmann et al., 2019a,b)), tailored for industrial applications and relying on statistical anomaly detection methods. With the rise of VLMs, recent works explore their commonsense reasoning in rare or unusual situations. Many use *synthetic image generation* to create controlled anomalies (Bitton-Guetta et al., 2024; Wang et al., 2024; Tai et al., 2024; Li et al., 2024b; Bitton-Guetta et al., 2023; Zhou et al., 2023), allowing researchers to cover a wide range of scenarios, systematically control the degree and type of abnormality. These benchmarks typically feature violations of physics or logic, whereas our work uses real-world images – photographs and screenshots – with realistic, context-dependent, and varied anomalies. The *"in-the-wild"* nature of CAVE makes anomaly detection significantly more complex than these synthetic datasets, where anomalies are often clearly isolated; real-world anomalies may be subtle, contextually embedded, and require sophisticated perception and reasoning ability to be detected. Benchmarks using real-world unusual situations focus on specific image types, such as creative elements in advertisements (Malakouti et al., 2024)

or video game glitches (Taesiri et al., 2024; Cao et al., 2024a), with limited applicability to commonsense anomalies in real-world images. Additional details about related benchmarks can evaluations can be found in Appendix B, along with a visual comparison of the benchmarks' image types (Figure 6).

In parallel, tailored prompting strategies (Xu et al., 2025; Yang et al., 2023a) are increasingly designed to tackle vision-language tasks involving complex reasoning, *e.g.*, compositional (Thrush et al., 2022) or commonsense (Zellers et al., 2019; Bitton et al., 2022). In this work, we implement several advanced strategies to improve VLMs' perception and reasoning capabilities.

6 Conclusion and Future Work

We introduce CAVE, a multimodal benchmark of 334 visual anomalies in 361 images spanning eight tasks, designed to test VLMs' real-world anomaly detection and understanding. Leading proprietary and open-source models (>70B parameters) only score ∼57 % 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. Future research could explore fine-grained visual representations for capturing subtle patterns (*e.g.*, uniformity breach) and retrieval-based approaches that leverage large-scale image databases to provide situational commonsense knowledge often missing from existing knowledge sources.

7 Limitations

Dataset Size. Our dataset consists of 361 images and 334 anomalies, which may be considered small compared to large-scale vision-language benchmarks. However, this limitation is counterbalanced by the depth and quality of annotations across seven tasks, including three open-ended tasks (anomaly description, explanation, and justification), one visual-grounding task (anomaly localization) and four classification tasks (anomaly categorization, complexity, severity, and surprisal). These factors help mitigate the limited number of examples by offering fine-grained insights into VLMs' ability to address the tasks. Further, we provide a solid framework for future development and curation.

Dataset Bias. Cultural bias is an inherent challenge in anomaly detection, as what is perceived as anomalous in one culture may be considered normal in another. In our dataset, bias is present due to two main factors. First, image selection bias arises because all images are sourced from Reddit, a platform with a skewed user demographic that does not represent the full diversity of human experiences. Second, anomaly detection and description bias arises from the human annotators, despite the diversity of the backgrounds of the Amazon Mechanical Turk workers and expert annotators. Annotation bias occurs both in open-ended tasks (AD, AE and AJ) and numerical ratings—for example, the surprisal score (how uncommon an anomaly is) may vary based on an annotator's cultural background and personal habits. Moreover, our dataset remains entirely in English, which might further limit its cultural inclusiveness.

To tackle this issue, we implemented several measures.

Diverse annotation team: We employed a culturally diverse annotation team, with annotators of open-ended tasks coming from 4 different continents. Annotators of numerical features, which are key to represent the subjectivity of an anomaly, come from 3 different countries and cultures. This diversity helps provide a broader perspective on what constitutes an anomaly in different contexts.

Multiple annotations per image: We had 5 annotators per image during the initial annotation stage, highlight varying perspectives and provide a richer understanding of what different people may consider anomalous; these annotations were later consolidated into the final set of anomalies. Similarly, we had three raters per anomaly for numerical features; these annotations are released as-is, allowing future users of CAVE to exploit the knowledge stemming from the diversity in numerical scores for each sample.

Inclusive definition of anomaly: We encouraged our annotators to adopt an inclusive definition of anomalies, considering what the majority of people would find anomalous, beyond their own beliefs and expectations. Especially, we encouraged expert annotators to be mindful of potential biases that may influence their perceptions of anomalies, and to consider all original MTurk annotations with an open mind.

Transparency in the collection and annotation process: We provide transparent and comprehensive documentation for the dataset that explains the process followed for collecting and annotating images, allowing future CAVE users to be aware of the potential bias and coverage limitations of the benchmark.

Dataset Completeness. Our evaluation of anomaly detection relies on precision and recall, and makes the strong assumption that we have exhaustively identified anomalies in each image. This assumption is supported by the extensiveness of our annotation process, with five independent Amazon Mechanical Turk annotations per image and expert validation. Moreover, we excluded ambiguous images where the presence of an anomaly was uncertain or debatable, to minimize borderline cases.

Despite these efforts, it is still possible that some anomalies were overlooked. To account for this, we provide a detailed performance breakdown, reporting the number of false positives, true positives, and false negatives to analyze model behavior in a fine-grained manner.

Dataset Consistency. Like most datasets relying on human annotation, ours is subject to errors, subjectivity, and inconsistencies despite extensive efforts in validation and standardization. Differences in individual interpretation could introduce some inconsistencies in open-answer tasks and numerical ratings.

Model Evaluation. We employ LLM-based evaluation as an alternative to costly and time-consuming human assessments. While this enables scalability, it comes with the risk of biases or misjudgments from the LLMs themselves. To address this, we validate LLM-based scores against human annotations and conduct manual evaluations for the two most complex reasoning tasks: anomaly explanation and anomaly justification.

8 Acknowledgements

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019a. 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.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019b. 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.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. 2022. Winogavil: Gamified association benchmark to challenge vision-and-language models. *Advances in Neural Information Processing Systems*, 35:26549–26564.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. 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.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. *ArXiv*, abs/2407.19474.

Daniel Bogdoll, Maximilian Nitsche, and J. Marius Zollner. 2022. Anomaly detection in autonomous driving: A survey. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 4487–4498. IEEE.

Joseph Borders, Gary Klein, and Ron Besuijen. 2024. Mental model matrix: Implications for system design and training. *Journal of Cognitive Engineering and Decision Making*, 18(2):75–98.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi,

Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. 2024. 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*.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. 2024a. Physgame: Uncovering physical commonsense violations in gameplay videos. *arXiv preprint arXiv:2412.01800*.

Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. 2024b. A survey on visual anomaly detection: Challenge, approach, and prospect. *ArXiv*, abs/2401.16402.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

V De Keyser and DD Woods. 1990. Fixation errors: Failures to revise situation assessment in dynamic and risky systems. In *Systems Reliability Assessment: Proceedings of the Ispra Course held at the Escuela Tecnica Superior de Ingenieros Navales, Madrid, Spain, September 19–23, 1988 in collaboration with Universidad Politecnica de Madrid*, pages 231–251. Springer.

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

Don C Donderi. 2006. Visual complexity: a review. *Psychological bulletin*, 132(1):73.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2021. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37.

Karl Friston. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Sharon G Goto, Yumi Ando, Carol Huang, Alicia Yee, and Richard S Lewis. 2010. 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.

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

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

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

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

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

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306.

Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. 2024. LLMs meet VLMs: Boost open vocabulary object detection with fine-grained descriptors. In *The Twelfth International Conference on Learning Representations*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*.

Gary Klein. 2013. *Seeing what others don't: The remarkable ways we gain insights*. Public affairs.

Gary Klein, Mohammadreza Jalaeian, Robert R Hoffman, and Shane T Mueller. 2023. The plausibility transition model for sensemaking. *Frontiers in psychology*, 14:1160132.

Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*, pages 118–160. Psychology Press.

Gary Klein, Rebecca Pliske, Beth Crandall, and David D Woods. 2005. Problem detection. *Cognition, Technology & Work*, 7:14–28.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jiaxuan Li, Junwen Mo, MinhDuc Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024b. Nemo: Can multimodal llms identify attribute-modified objects? *arXiv preprint arXiv:2411.17794*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *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. Association for Computational Linguistics.

Sina Malakouti, Aysan Aghazadeh, Ashmit Khandelwal, and Adriana Kovashka. 2024. Benchmarking vlms' reasoning about persuasive atypical images. *arXiv preprint arXiv:2409.10719*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, and 1 others. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

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

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. 2021. Vt-adl: A vision transformer network for image anomaly detection and localization. *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06.

James F Mullen Jr, Prasoon Goyal, Robinson Piramuthu, Michael Johnston, Dinesh Manocha, and Reza Ghanadan. 2024. "don't forget to put the milk back!" dataset for enabling embodied agents to detect anomalous situations. *IEEE Robotics and Automation Letters*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

OpenAI. 2024. Gpt-4o system card.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38.

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

Rahul Ramachandran, Ali Garjani, Roman Bachmann, Andrei Atanov, Oğuzhan Fatih Kar, and Amir Zamir. 2025. How well does gpt-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks. *Preprint*, arXiv:2507.01955.

Claire Roman and Philippe Meyer. 2024. Analysis of glyph and writing system similarities using Siamese neural networks. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 98–104, Torino, Italia. ELRA and ICCL.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Zekun Sun and Chaz Firestone. 2021. Curious objects: How visual complexity guides attention and engagement. *Cognitive Science*, 45(4):e12933.

Mohammad Reza Taesiri, Tianjun Feng, Cor-Paul Bezemer, and Anh Nguyen. 2024. 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.

Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. 2024. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27176–27185.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Anne Treisman and Hilary Schmidt. 1982. Illusory conjunctions in the perception of objects. *Cognitive psychology*, 14(1):107–141.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Joseph Jay Williams, Caren Walker, and Tania Lombrozo. 2012. Explaining increases belief revision in the face of (many) anomalies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Jiayi Lyu, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. 2024. Im-iad: Industrial image anomaly detection benchmark in manufacturing. *IEEE Transactions on Cybernetics*.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Zhuo Xu, Hao-Tien Lewis Chiang, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, Fei Xia, Jasmine Hsu, Jonathan Hoech, Pete Florence, Sean Kirmani, Sumeet Singh, Vikas Sindhwani, Carolina Parada, Chelsea Finn, and 3 others. 2024. Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs. In *8th Annual Conference on Robot Learning*.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *arXiv preprint arXiv:2502.16033*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *Preprint*, arXiv:2310.11441.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *Preprint*, arXiv:2309.17421.

Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2023. Computer vision datasets and models exhibit cultural and linguistic diversity in perception. *arXiv preprint arXiv:2310.14356*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. 2020. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv preprint arXiv:2003.12338*, 27(10.48550).

Pengyi Zhang and Dagobert Soergel. 2014. 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. ROME: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10185–10197, Singapore. Association for Computational Linguistics.

Appendix Table of Contents

A Research Tools

Compute details. We evaluated 5 open-source models: InternVL2.5 (38B et 78B parameters) (Chen et al., 2024), LlavaOneVision (72B) (Li et al., 2024a), QwenVL2.5 (72B) (Yang et al., 2024), and Llama 3.2 (90B) (Meta, 2024). 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 3 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).

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 3), images are first processed by the vision encoder and then projected into the language model's embedding space (Meta, 2024; Yang et al., 2024; Chen et al., 2024). 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 (Mishra et al., 2021; Bergmann et al., 2019a,b; Bogdoll et al., 2022; Cao et al., 2024b). 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) (Cao et al., 2024b). 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 (Bitton-Guetta et al., 2024) introduces synthetically generated images, each depicting a unique situation and requiring commonsense to answer a question. WHOOPS (Bitton-Guetta et al., 2023) 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 (Wang et al., 2024) 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.

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) (Zhou et al., 2023) 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 (Li et al., 2024b) investigates how VLMs recognize objects with uncommon properties, such as a blue mango. ISEKAI (Tai et al., 2024) 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. Malakouti et al. (2024) leverage the PittAds dataset (Hussain et al., 2017), 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 (Yan et al., 2025) 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 (Taesiri et al., 2024) is a

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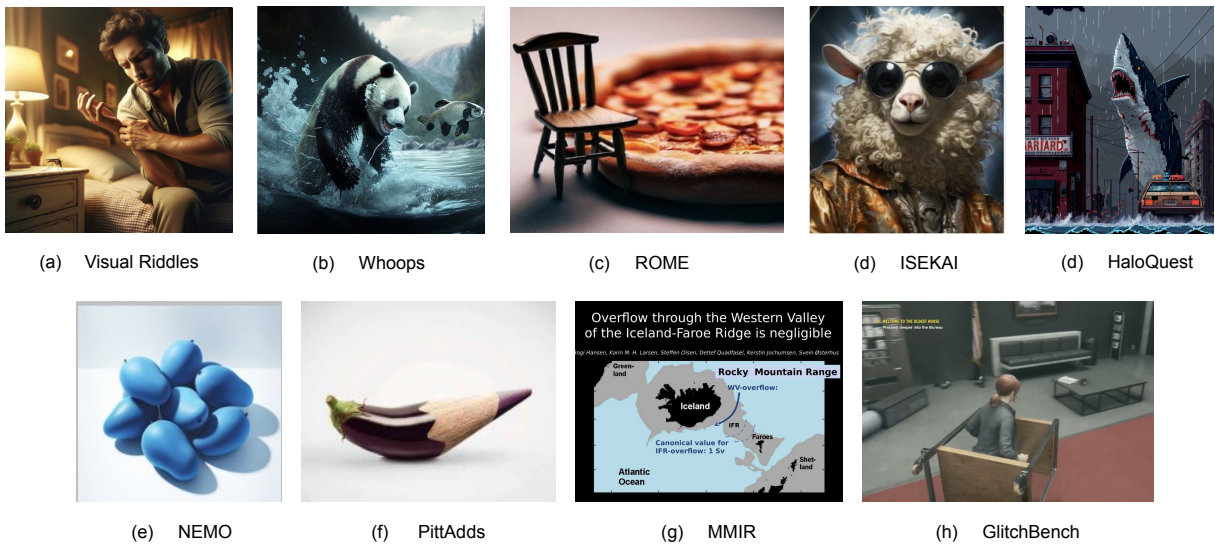

Figure 6: **Related Benchmark Examples.** Examples of images from related multimodal anomaly-detection benchmarks. More details about each benchmark are given in Appendix B and Table 4.

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 (Cao et al., 2024a) benchmark models' ability to identify physical commonsense anomalies in gameplay videos.

B.3 Taxonomy-level Comparison

We also provide a taxonomy-level comparison with synthetic datasets. Our taxonomy includes six categories (Section 2.3): Entity Presence, Entity Absence, Entity Attribute, Spatial Relation, Uniformity Breach

and Textual Anomaly. Below we briefly contrast these with existing synthetic benchmarks.

Novel Anomaly Manifestations: Real-world images naturally capture a broader and more open-ended set of anomalies. Entity Absence and Uniformity Breach (e.g., a misaligned row of products, one tile laid upside-down) arise organically in our data but are rarely or never seen in synthetic datasets, which typically only show alterations of a single entity (see Figure 6 for examples; taxonomies of datasets such as ROME (Zhou et al., 2023) and GlitchBench (Taesiri et al., 2024) only match with our Entity Attribute and Spatial Relation categories). Textual Anomalies are also usually absent from synthetic datasets, except MMIR (Yan et al., 2025).

Varied and Nuanced Anomalies: The richness of CAVE is not just that real images are noisier; rather,

the types of anomalies themselves are more varied and nuanced. Synthetic datasets are constrained by their generation rules: take an existing object or scene, alter it following a pre-defined template (e.g., change an object's color, swap animal species) to obtain an anomalous version. Given the limited set of alteration types, the resulting datasets are inherently limited in terms of anomaly diversity. In contrast, real data exposes new failure modes and contextual subtleties that template-driven methods cannot anticipate.

B.4 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 3 (right)), and the small size of our dataset makes fine-tuning impractical.

C Quantitative Comparison with Synthetic Benchmarks

We perform a direct comparison with two prior synthetic benchmarks: WHOOPS, which uses full-image generation to create anomalous scenes, and COCO-OOC, which introduces anomalies via image editing.

First, we evaluate WHOOPS, a benchmark closely aligned with our task formulation and commonsense-reasoning orientation. It provides ground truth anomaly descriptions, enabling the use of our full LLM-as-a-judge pipeline. Second, we sample 500 randomly selected anomalous images from COCO-OOC. Unlike WHOOPS, COCO-OOC modifies real COCO images by inserting localized anomalies while keeping the rest of the scene unaltered. However, the edits can introduce visual artifacts or noise that may affect anomaly detection. Since COCO-OOC lacks ground truth AD descriptions, we rely on manual evaluation to assess the correctness of the model-generated anomalies.

Across both benchmarks, we use identical evaluation prompts (GPT-4o with our AD prompt). We observe a substantial performance gap between CAVE and prior benchmarks as shown in Table 5.

This demonstrates that CAVE is significantly more challenging for state-of-the-art VLMs. The higher precision on WHOOPS reflects its simpler, single-object or uncluttered scenes, which reduce hallucinations; in contrast, CAVE's complex, real-world images introduce more distractors and plausible but subtle anomalies.

A manual analysis of false positives further underscores the difference: on WHOOPS, most FPs are perception errors, with only 4 reasoning errors. On CAVE, by contrast, reasoning errors account for about half of all FPs, indicating a much greater demand for contextual and commonsense reasoning (see Table 2 and Figure 18 to 23). Moreover, many of the WHOOPS FPs stem from artifacts of synthetic image generation (e.g., non-existent script, visual inconsistencies and missing elements). This leads to additional "valid" anomalies found by the model being marked as FPs, underestimating the model's performance.

We observed that the synthetic editing process (like in COCO-OOC) often results in unrealistic artifacts (e.g., partially added objects, such as a half-elephant). While this makes anomaly detection easier, it also limits ecological validity: models may succeed by flagging obvious artifacts rather than demonstrating true commonsense reasoning.

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 (Wei et al., 2022). See prompt in Figure 25.

(2) Set-of-Marks prompting (SoM) We incorporate object-level annotations and bounding boxes generated by Grounding DINO (Liu et al., 2023) 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 Yang et al. (2023a), 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 (Yang et al., 2023a). 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

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 4: **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).

Dataset	Precision	Recall	F1-score
WHOOPS	85.5	85	85.3
COCO-OOC	-	91	-
CAVE	**52.4**	**50**	**51.2**

Table 5: **Comparison with Synthetic Benchmarks.** We compare GPT-4o evaluation on WHOOPS and 500 samples of COCO-OOC. Note that COCO-OOC only has positive samples, hence we only report recall.

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 (Xu et al., 2025). See prompt in Figure 26.

(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 (Wang et al., 2022). 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[3] library, we downloaded the top 1,000 posts 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

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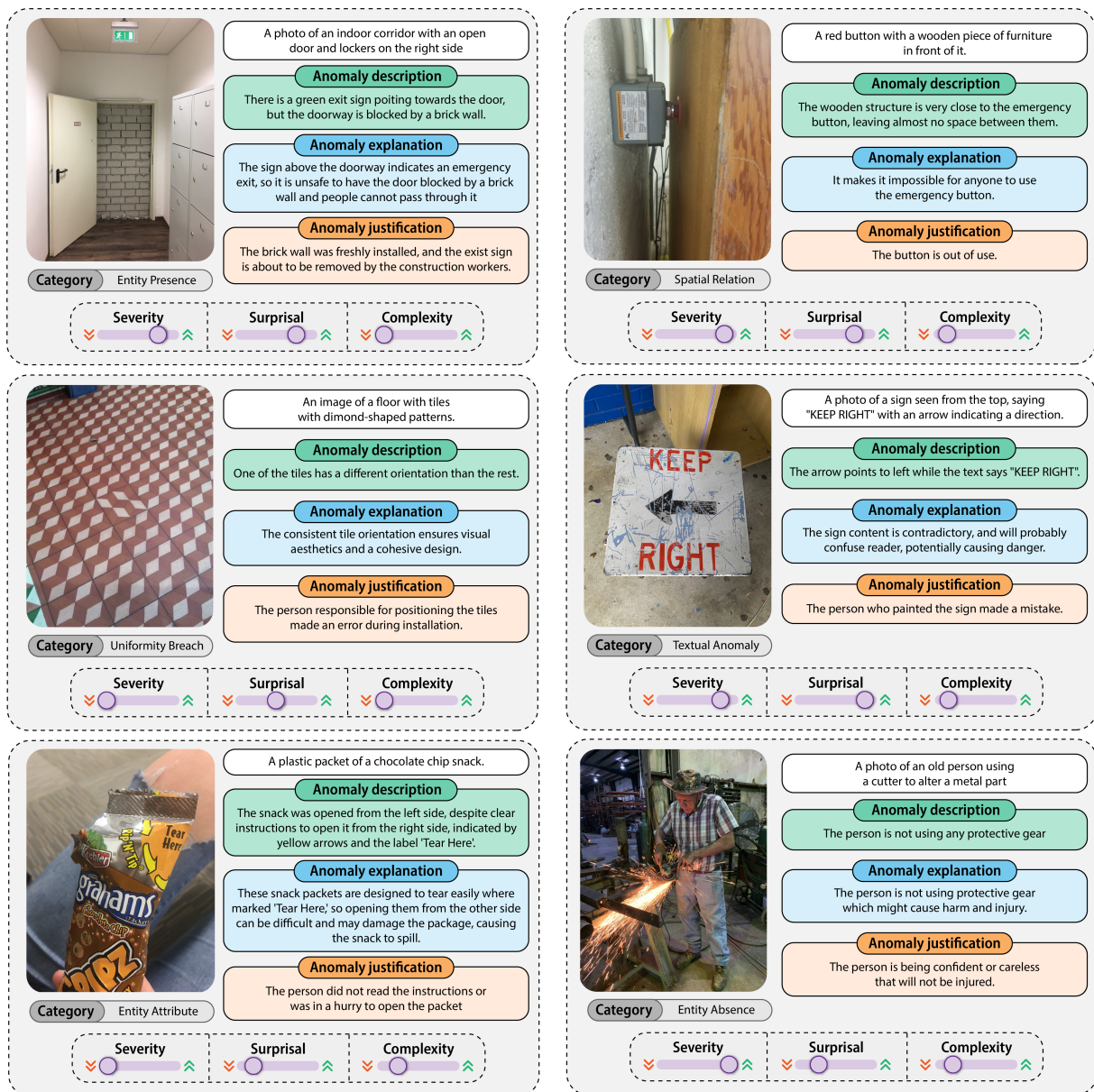
[3]https://github.com/praw-dev/praw

Figure 7: **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.

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!**

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 6) 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 (Gwet, 2008), 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 (Gwet, 2014).

	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 6: 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 (Goto et al., 2010; Myung et al., 2024; Nayak et al., 2024). 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 (Ye et al., 2023).

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
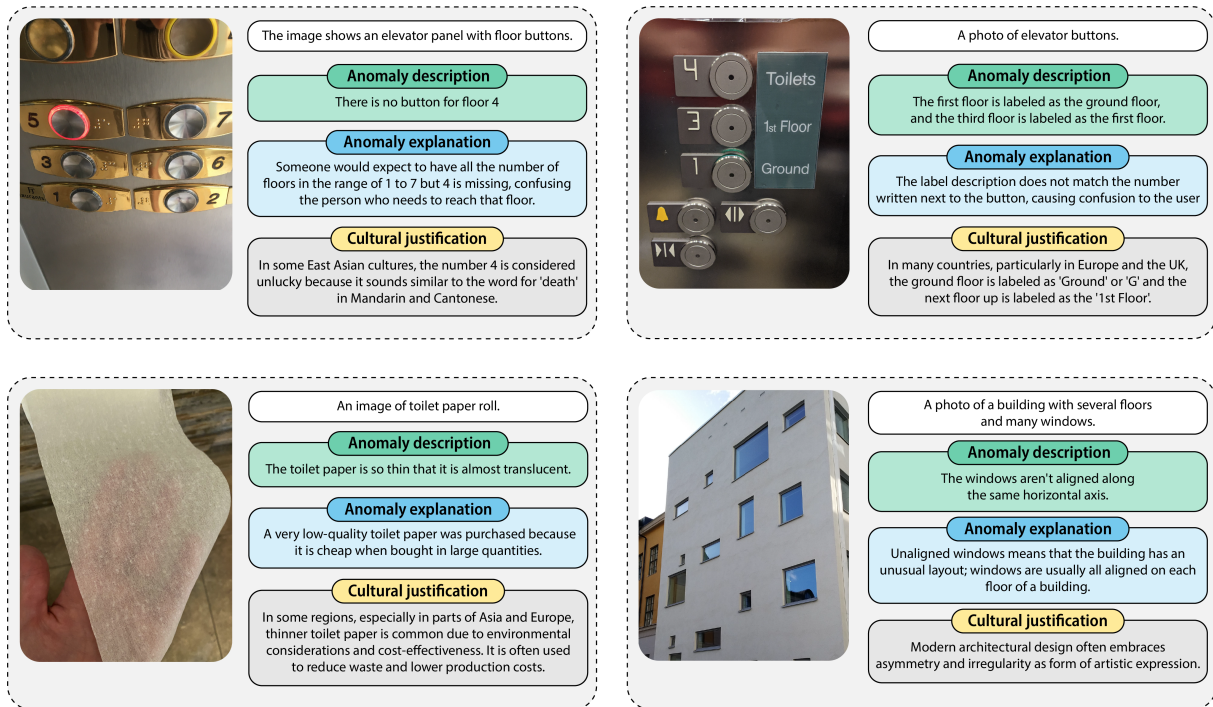
Figure 8: **Culture-specific examples of CAVE.** Examples of anomalies from CAVE annotated as Western-centric, along with culturally grounded justifications explaining why they should not be considered anomalies.

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 8, 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 29
- Anomaly Justification: Figure 30
- Anomaly Severity: Figure 31
- Anomaly Surprisal: Figure 32
- Anomaly Complexity: Figure 33
- AD judge prompt: Figure 34
- AE judge prompt: Figure 35
- Cultural bias assessment prompt: Figure 37

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 36. Figure 9 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).
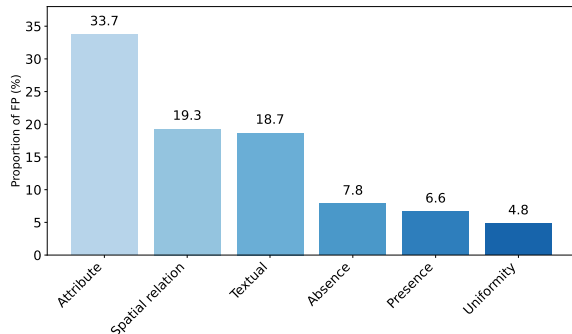

Figure 9: **FP classifier performance.** Anomaly category proportions in GPT-4o FP.

We also performed a bootstrap test on 5k different random samples of CAVE to assess significance for prompting strategies over the vanilla baseline. Table 7 reports the improvement in F1 scores for the 3 best prompting strategies compared to the vanilla prompt baseline, with 95% confidence intervals shown in brackets and corresponding p-values, in bold for statistically significant gains ($p \leq 0.05$). Overall, CoT with self-consistency yields significant gains for most models, whereas plain CoT and multi-step CoT reach significance for roughly half.

G.2 Anomaly Explanation

Each model's performance on TP and FN from the AD task is detailed in Table 8. Most of the models have higher performance on TP examples than FN.

G.3 Anomaly Justification

Figure 14 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 appendix F.

GPT-4o and InternVL show high agreement with humans for severity (appendix 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 15. 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 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 Judge Bias Evaluation

Since GPT-4o is both one of the evaluated models and the default judge in our LLM-as-a-judge evaluation pipeline for AD, this raises the possibility of bias in its favor. To assess this, we conducted an additional evaluation using Claude-3.5 as an independent judge. Specifically, Claude-3.5 was used to score outputs from both GPT-4o and the top-performing open-source model (InternVL2.5–78B) across several prompting strategies. This cross-model judgment setup allows us to quantify potential self-judging bias and validate the robustness of our conclusions.

Table 13 reports the delta (Claude judge score minus GPT-4o judge score) for each model and prompt. For most prompts, Claude assigns slightly higher scores than GPT-4o's self-judgment. Importantly, the deltas are of similar magnitude for both models, and the delta values for GPT-4o are positive, validating the robustness and absence of bias of our evaluation protocol.

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

Model	CoT	CoT + consist.	MS CoT
Llama3.2 90b	11.145 [0.059, 0.160] ***p=0.000***	13.553 [0.081, 0.188] ***p=0.000***	7.160 [0.020, 0.120] ***p=0.003***
LlavaOV 72b	-0.160 [-0.043, 0.040] *p=0.524*	8.780 [0.042, 0.133] ***p=0.000***	2.088 [-0.028, 0.067] *p=0.204*
InternVL2.5 38b	2.937 [-0.006, 0.065] *p=0.056*	6.284 [0.022, 0.103] ***p=0.002***	-1.330 [-0.054, 0.027] *p=0.732*
QwenVL2.5 72b	-2.845 [-0.068, 0.010] *p=0.925*	-1.468 [-0.053, 0.024] *p=0.764*	-1.609 [-0.057, 0.025] *p=0.773*
InternVL2.5 78b	2.411 [-0.018, 0.066] *p=0.125*	3.207 [-0.005, 0.069] ***p=0.044***	-0.876 [-0.049, 0.031] *p=0.659*
GPT-4o	3.113 [-0.006, 0.068] ***p=0.049***	2.543 [-0.010, 0.061] *p=0.081*	4.954 [0.010, 0.089] ***p=0.007***
Claude	8.051 [0.026, 0.133] ***p=0.001***	6.155 [0.015, 0.107] ***p=0.006***	16.384 [0.120, 0.206] ***p=0.000***

Table 7: **Effectiveness of prompting strategies via bootstrapping.** Each entry shows the improvement over vanilla AD prompt, 95% confidence interval, and p-value. Bold indicates statistically significant improvements ($p < 0.05$).

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

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

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.
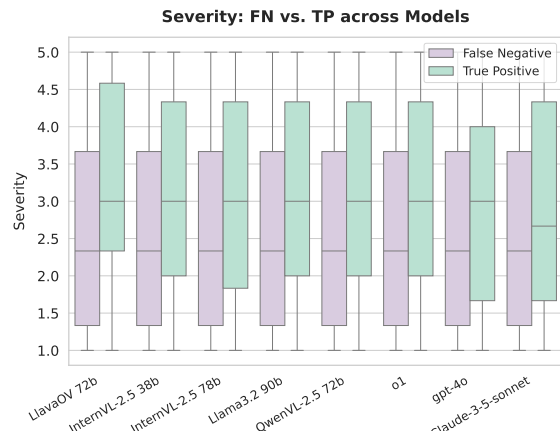

Figure 11: **Models' performance across severity feature.** Plot showing the deviation in the models' performance across different levels of anomaly severity for the anomaly description task. The results indicate that models perform well on less severe anomalies, while performance drops significantly for highly severe anomalies on average.

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 9: **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 10: **F1 scores per anomaly category.** Values in parentheses indicate deltas from the Vanilla baseline; green for improvement, red for decline.
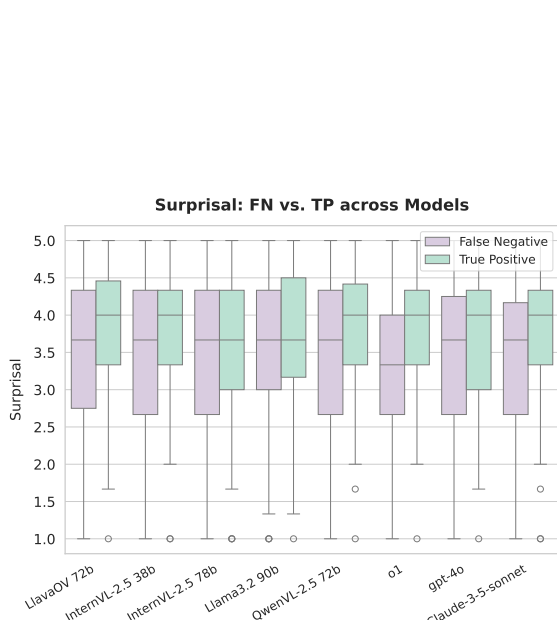

Figure 12: **Models' performance across surprisal feature.** Plot showing the deviation in the models' performance across different levels of anomaly surprisal for the anomaly description task. The results reveal that models perform well on high-surprisal anomalies but also exhibit more false positives for more surprising anomalies on average.
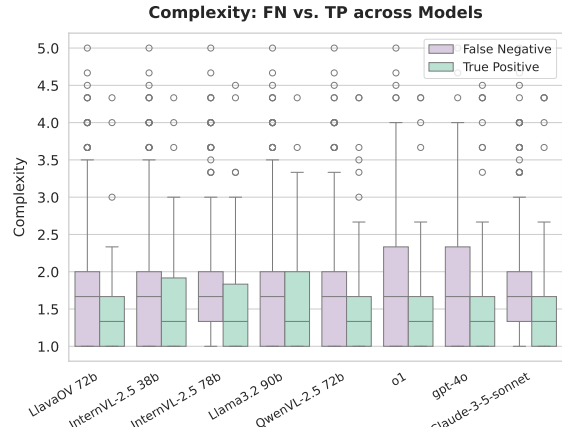

Figure 13: **Models' performance across complexity feature.** Plot showing the deviation in the models' performance across different levels of anomaly complexity for the anomaly description task. The results reveal that models perform well only on low-complex tasks but also exhibit false positives for much simpler anomalies on average.

	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 11: **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.

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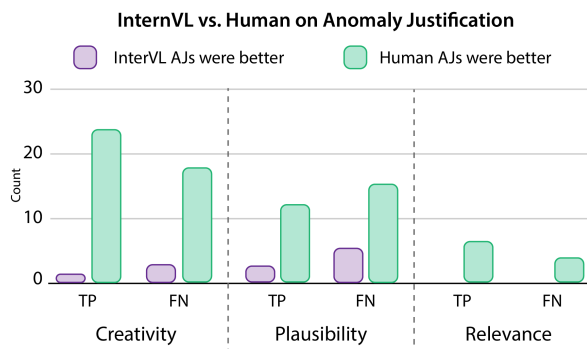

Figure 14: **Comparison of InternVL vs. Human Anomaly Justification.** Bars above the x-axis indicate cases where InternVL outperformed humans, while bars below indicate cases where InternVL underperformed. The 3 bars on the left are results over 50 False Negatives (FN), where the model failed to identify anomalies; the 3 bars on the right are over 50 True Positives (TP).

Prompt	GPT4o	InternVL2.5-78B
Vanilla	2.49	4.02
CoT	0.86	3.58
CoT + consist.	1.7	4.11
MS CoT	0	0
CoT + SoM	-1.37	1.11

Table 13: **Judge bias analysis using Claude-3.5 as an independent judge.** We report average score deltas (Claude minus GPT-4o). Similar positive deltas for GPT-4o and InternVL2.5–78B indicate no self-judging bias.
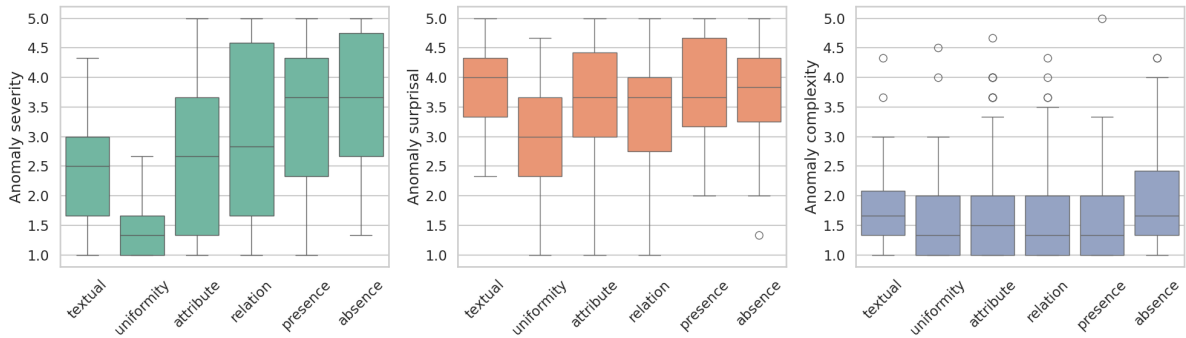
Figure 15: Distribution of anomaly scores across categories. The boxplots illustrate the distribution of complexity, severity, and surprisal scores across different anomaly categories, highlighting variations in human perception of anomalies.

GT Anomaly Description	GT Anomaly Description	GT Anomaly Description
A rusty wrench is bolted to a control panel to hold a switch in a permenant position.	There is a soda can in the toilet bowl.	(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.
Human Explanation	**Human Explanation**	**Human Explanation**
The presence of the rusty wrench prevents the usage of the switch.	A toilet needs to be devoid of any objects in the bowl that might cause blockages for proper functioning.	The regularly alternating pattern is broken due to the second from left chair in the first row being red instead of white.
GPT4o Explanation	**GPT4o Explanation**	**GPT4o Explanation**
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.	A soda can is typically found in kitchens or dining areas, not in a toilet bowl, which is meant for sanitation purposes.	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	Reasoning error	Perception error

GT Anomaly Description

The logo for the restaurant is off-center and sticks out of the side of the building
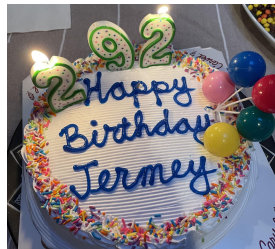
Human Justification

There used to be a second sign on the right, but it fell before the photo was taken

GPT4o Justification

The logo may have been improperly installed due to lack of precise measurement or miscommunication during the installation process

GT Anomaly Description

The candles depict the age of "Jermey" as 292 years old

Human Justification

The candles have been placed as a joke and exaggerating Jermey's age

GPT4o Justification

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.

GT Anomaly Description

There is trash on the floor next to the trash can

Human Justification

A person threw the trash towards the bin and missed it

GPT4o Justification

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 Localization Prompt

System: You are an advanced Vision-Language Model designed to locate the given anomaly description.

User: Your task is to find and localize the single anomaly described below in the given image.

Anomaly Description: {anomaly_description}

Image resolution: {width} x {height} pixels.

Please provide your output in JSON format exactly as follows:

{{

"box": ["x1": <top-left x>, "y1": <top-left y>,

"x2": <bottom-right x>, "y2": <bottom-right y>]

}}

Figure 28

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 29

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 30

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 31

27148

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 32

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 33

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 34

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 35

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 36

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 37