CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments

"If you notice an abnormal situation, please contact an agent." Such announcements are commonplace in public spaces, highlighting a fundamental human trait: the ability to detect anomalies, situations that deviate from expectations. As Vision-Language Models (VLMs) are increasingly deployed in dynamic real-world settings, their ability to recognize and reason about uncommon or surprising situations is crucial for safe and efficient operation. Despite advances in multimodal learning, anomaly detection using VLMs remains underexplored. Existing benchmarks rely heavily on synthetic (Bitton-Guetta et. al., 2023) or domain-specific scenarios (Diers et. al., 2023, Fernando et. al., 2021), overlooking the diversity and complexity of real-life anomalies, leaving a critical gap in the evaluation of VLMs' anomaly detection capabilities.

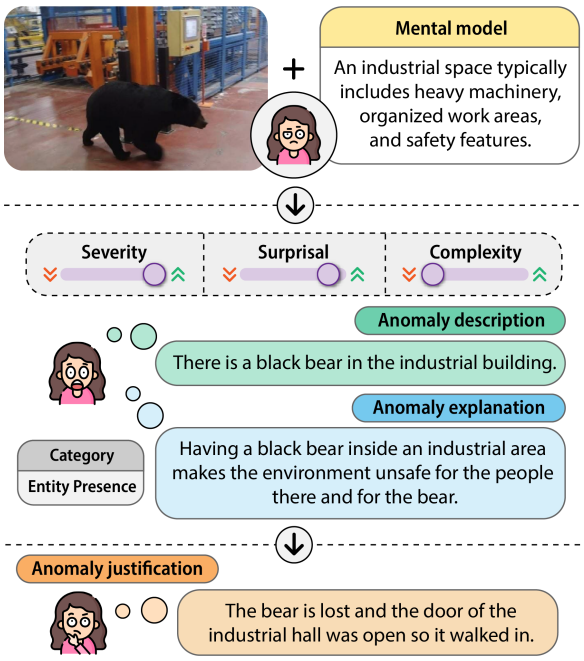

Figure 1: **CAVE Example:** a real-world image annotated with commonsense anomaly descriptions, explanations and justifications, as well as numerical features representing how humans perceive these anomalies.

Key Contributions: 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. Building on top of the cognitive science literature on how humans identify and understand anomalies, we present a multi-task anomaly understanding framework. We split the anomaly detection process into three open-ended tasks that align with human anomaly detection and sense-making processes: (a) Anomaly Description (AD), (b) Anomaly Explanation (AE), and (c) Anomaly Justification (AJ). We also categorize anomalies based on the type of visual reasoning required to identify them and further label them with three numerical features: severity, surprisal or rarity, and detection complexity based on cognitive science theories of anomaly perception (see Figure 1).

Dataset and Evaluation: CAVE contains 361 images (309 anomalous and 52 normal) encompassing 334 unique anomalies. We evaluate 8 state-of-the-art VLMs (5 open-source), on our CAVE benchmark, using both vanilla and advanced prompting strategies (*e.g.*, multi-step reasoning and self-consistency). For AD and AE, we employ GPT-4o as a judge (to pairwise compare each model output with the ground truth), given its agreement with human judgment on a small subset.

Key Results: Even the best performing model, GPT-4o, achieves only 56% F1 on AD, indicating VLMs' limited ability in anomaly detection. While models perform better on highly surprising (obvious) anomalies, they struggle with spatial reasoning and detecting pattern violations. For AE, models provide correct explanations >80% once the correct AD is provided. Although models provide correct justifications given the (ground truth) AD and AE, AJ still remains a challenge, with human evaluations displaying that model justifications often lack creativity or contextual relevance.

Conclusion: CAVE reveals fundamental limitations in current VLMs' perception and reasoning abilities when operating in real-world conditions. By combining real-world complexity, task diversity and cognitive grounding, we provide a robust platform to advance research in commonsense visual anomaly understanding.