FIRE360 : A Benchmark for Robust Perception and Episodic Memory in Degraded 360° Firefighting Video

Aditi Tiwari[1], Farzaneh Masoud[2], Dac Trong Nguyen[2], Jill Kraft[2], Heng Ji[1], Klara Nahrstedt[1]

[1]University of Illinois Urbana-Champaign, [2]Illinois Fire Service Institute

✨ NeurIPS 2025 Datasets & Benchmarks Track Spotlight ✨

Introduction

AI works in clean data. It breaks in chaos.

Fire360 tests if vision-language models can still **see, remember, and reason** when smoke blinds, heat warps, and fire reshapes reality.

Models scoring **97 % on clean data drop to 6 % in smoke.**

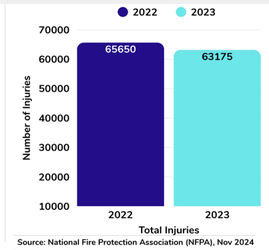
Can AI recognize melted equipment? Track objects through smoke? Remember identity across fire damage?

Current answer: No. Human-model gap: 54%.

Motivation

Over **63,000 injuries** were reported last year alone. Firefighting is one of the most high-risk, time-critical domains. Scenes push perception to its limits - smoke, heat, and motion everywhere.

360° vision captures what single cameras miss: the *full situational context* where reliability saves lives.

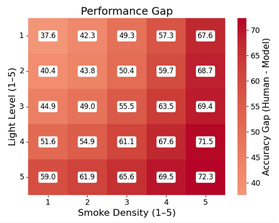


Reliability Gap - VQA under degradation

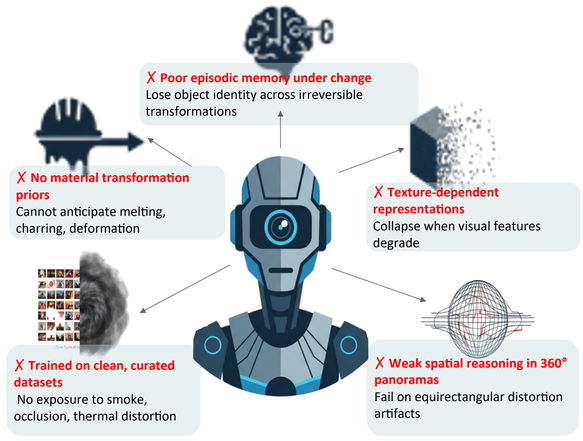
Humans ≈ 86 %

GPT-4o ≈ 27 %

Models collapse as smoke thickens and light fades.



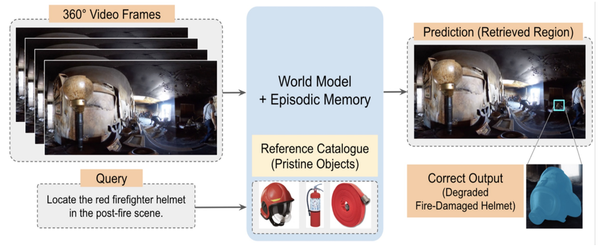
Failure Factors in Degraded Visual Understanding



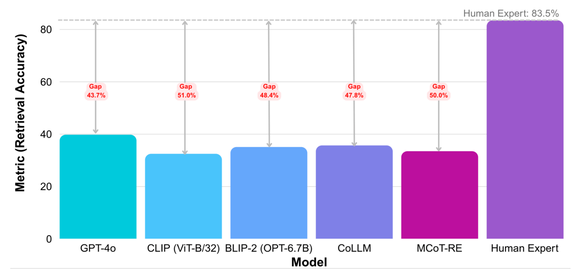
Fire360 Dataset

Transformed Object Retrieval (TOR)

Core Idea: Can AI recognize the same equipment **after fire destroys its appearance**?



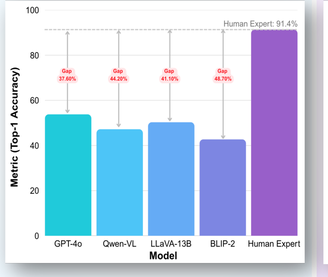
Gap reveals fundamental failures in transformation-invariant object recognition.



Benchmark tasks: Exposing failures across perception, reasoning, and memory

Task 1: Visual QA

Objective: Spatial reasoning across full 360° field-of-view under heavy degradation.

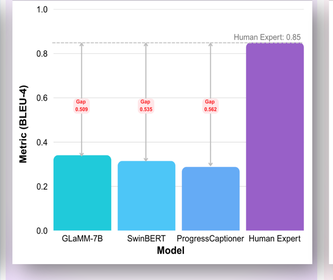


Why Models Fail?

Hallucinate objects in occluded regions. Accuracy drops to 6.1% in heavy smoke (vs. 81.3% for humans).

Task 2: Temporal Captioning

Objective: Generate grounded natural language descriptions of firefighter actions.

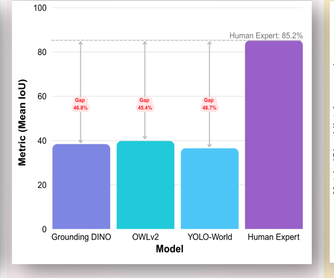


Why Models Fail?

Confuse visually similar but procedurally distinct actions. Generic captions lack domain specificity.

Task 3: Object Localization

Objective: Detect safety equipment (hoses, masks) under occlusion, thermal blur, and 360° distortion.

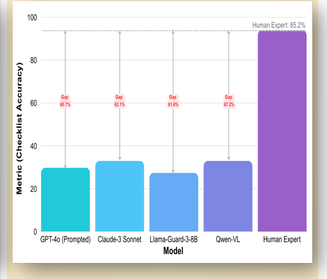


Why Models Fail?

IoU drops to 22.9% in low-visibility scenes. Cannot handle panoramic projection distortion.

Task 4: Safety Reasoning

Objective: Identify violations of standard firefighter safety protocols using expert-verified checklists.



Why Models Fail?

Cannot infer procedural violations from visual cues. Lack domain-specific protocol knowledge.

Why Models Fail?

Overreliance on surface texture: pipes misread as helmets, ladders as hoses

Sensitivity to occlusion: smoke- or debris-covered objects missed

No physics-aware material priors: plastic vs. metal confusion

Weak cross-scene correspondence under transformation

Research Directions Enabled

World models for safety-critical reasoning

- Predictive models of environmental state evolution
- Counterfactual reasoning: "What if smoke density increases?"
- Causal intervention for protocol violation detection

Episodic memory architecture

- Neural scene representations for cross-scene retrieval
- Memory-augmented transformers with degradation tokens
- Slot attention for object permanence under occlusion

Physics-informed vision models

- Integrate material science priors for transformation prediction
- Simulate melting points, charring patterns, thermal deformation
- Learn material-specific degradation trajectories