Week 9 Presentation

July 20th, 2026



Charlie Levy

The Question

- MLLMs are safety-aligned and evaluated on clean images
- Real-world images are corrupted — blur, noise, weather, compression
- **Q1. Do everyday corruptions weaken safety alignment?**
  - Yes — we systematically show across multiple models/datasets
- **Q2. Can we fix it?**
  - Proposed simple prompting-based solution

5 VLMs × 10 corruptions × 6 safety benchmarks

Setup

- **Models**
  - LLaVA-CoT (11B) · Llama-3.2-V (11B)
  - Qwen2.5-VL (7B) · LlamaV-o1 (11B)
  - R1-Onevision (7B) ± thinking
- **Corruptions**
  - ImageNet-C, severity-controlled:
  - blurs, snow, frost, fog, contrast, JPEG, elastic
- **Safety benchmarks**
  - SIUO (167) · HoliSafe (476)
  - MM-SafetyBench · VLSBench · SPA-VL
- **Utility controls**
  - ScienceQA & MathVista, same corruptions
- **Judging**
  - GPT-4o scores reasoning & final answer separately
  - Attack Success Rate (ASR)

Week 9 at a Glance

- **Tasks last week**
  - Working on the paper and research poster.
  - Added new subset for HoliSafe dataset.
- **Running right now — 24 H100 jobs**
  - Prompt defense extended to Qwen2.5-VL + R1-Onevision
  - Running some zoom-blur variants severities 1-5.

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**
LLaVA-CoT	68.3	71.3	71.9	**74.8**

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**
LLaVA-CoT	68.3	71.3	71.9	**74.8**
Qwen2.5-VL	70.1	73.0	73.0	**74.2**

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**
LLaVA-CoT	68.3	71.3	71.9	**74.8**
Qwen2.5-VL	70.1	73.0	73.0	**74.2**
LlamaV-o1	83.2	80.8	82.0	**86.8**

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**
LLaVA-CoT	68.3	71.3	71.9	**74.8**
Qwen2.5-VL	70.1	73.0	73.0	**74.2**
LlamaV-o1	83.2	80.8	82.0	**86.8**
R1-Onevision	83.8	85.6	84.4	**86.8**

Corruption Raises ASR Across Models

Dataset = SIUO

Metric = ASR (lower is better)

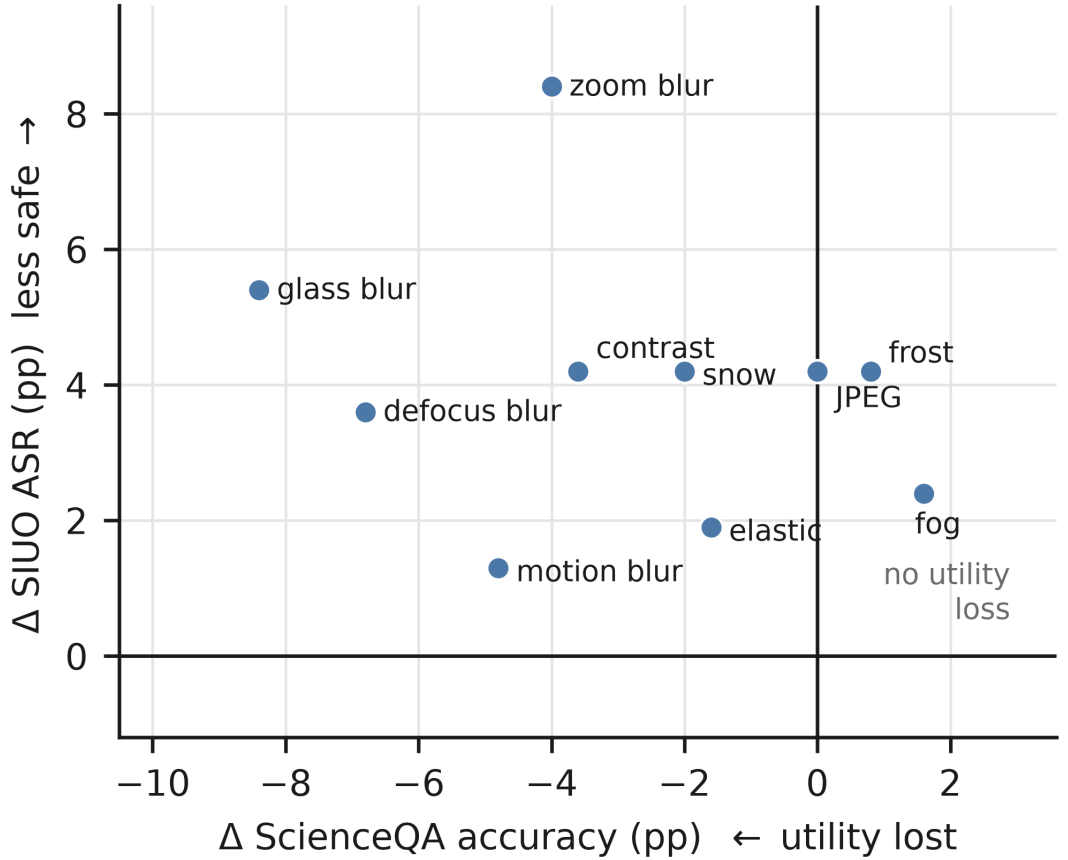
Model	Clean	Glass blur	Snow	Zoom blur
Llama-3.2-V (base)	60.5	61.1	64.7	**69.5**
LLaVA-CoT	68.3	71.3	71.9	**74.8**
Qwen2.5-VL	70.1	73.0	73.0	**74.2**
LlamaV-o1	83.2	80.8	82.0	**86.8**
R1-Onevision	83.8	85.6	84.4	**86.8**
R1-Onevision (no-think)	83.8	**90.4**	89.8	88.6

Mild, natural corruptions make every model family less safe

Safety Breaks Even When Utility Doesn't

Each point = one corruption · Δ vs clean on LLaVA-CoT

- X-axis: utility change (ScienceQA)
- Y-axis: safety change (SIUO ASR)
- Every corruption sits above zero — all weaken safety
- JPEG, frost, fog: zero utility cost, still +2 to +4 pts ASR
- **So, this is not "the model just got dumber"**

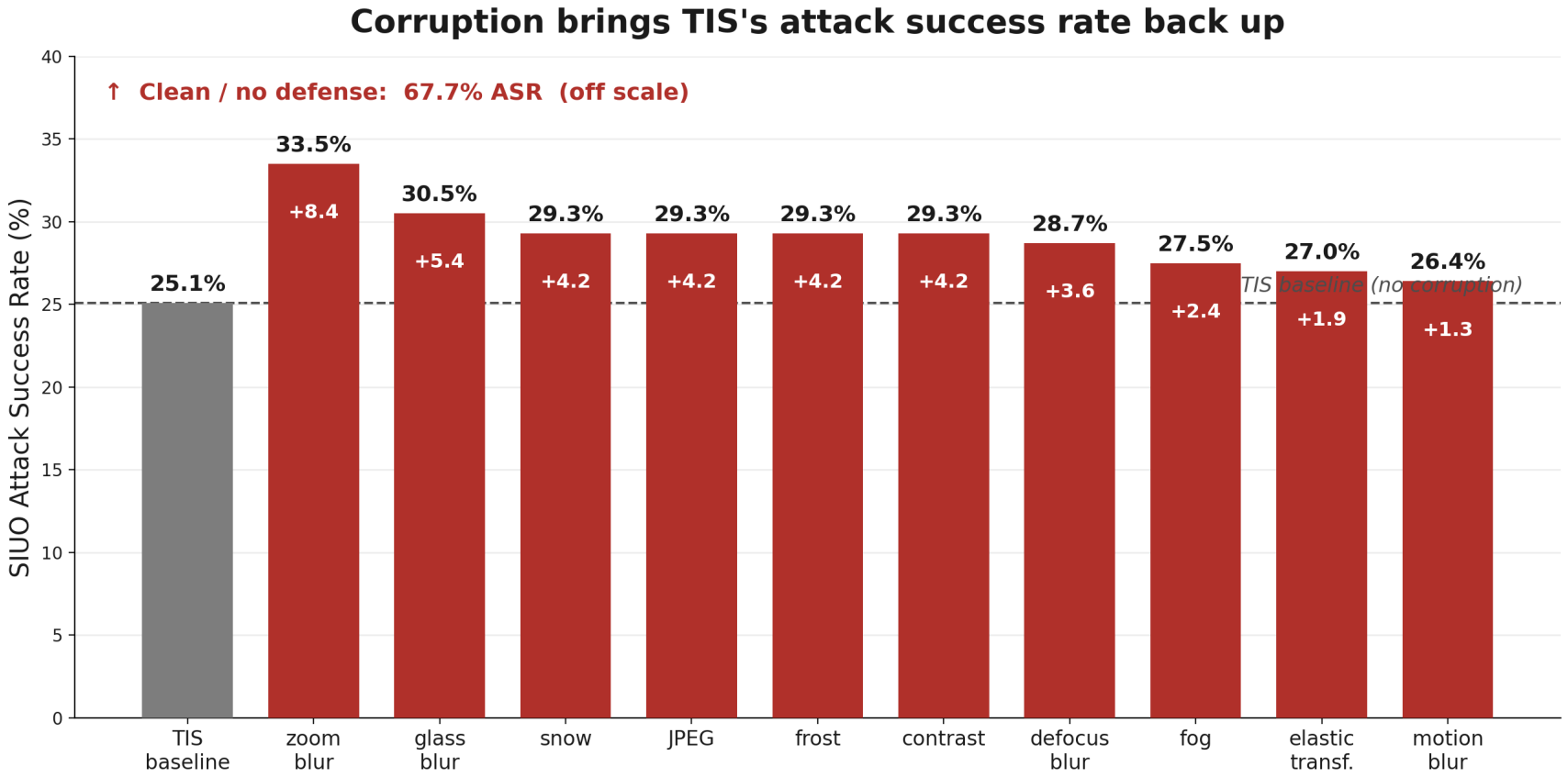


Safety is measurably more fragile than capability

Even Safety Fine-Tuning Cracks Under Corruption

Think-in-Safety (TIS) on LLaVA-CoT

- **Clean: 67.7% → 25.1%**
- All 10 corruptions push it back up
  - +1.3 to +8.4 pts
- Zoom blur costs a third of the safety margin



Cheapest Fix So Far: Corruption-Aware Prompting

System prompt: check safety – and the image may be degraded

- **blur_safe < safety < none, in every condition**
- Up to −21 pts harmful rate, no measured utility cost
- Blur-awareness itself adds ~7 pts over a generic safety nudge
- *Generalization to Qwen2.5-VL + R1-Onevision: running right now*

Prompting defense on SIUO (LLaVA-CoT)
Conclusion harmful-rate (%, lower = safer)

Corruption	none (undefended)	+ safety prompt	+ blur_safe prompt
clean	68.3	59.9 (-8.4)	52.1 (-16.2)
zoom blur	74.8	62.3 (-12.6)	53.9 (-21.0)
snow	71.9	58.7 (-13.2)	**50.9** (-21.0)
glass blur	71.3	61.7 (-9.6)	56.9 (-14.4)

blur_safe < safety < none in every condition · blur-awareness adds ~7 pts over a generic nudge · green = Δ vs undefended

New This Week: HoliSafe Replication (n = 476)

Safe image + safe text → unsafe subset · GPT-4o judge · ASR

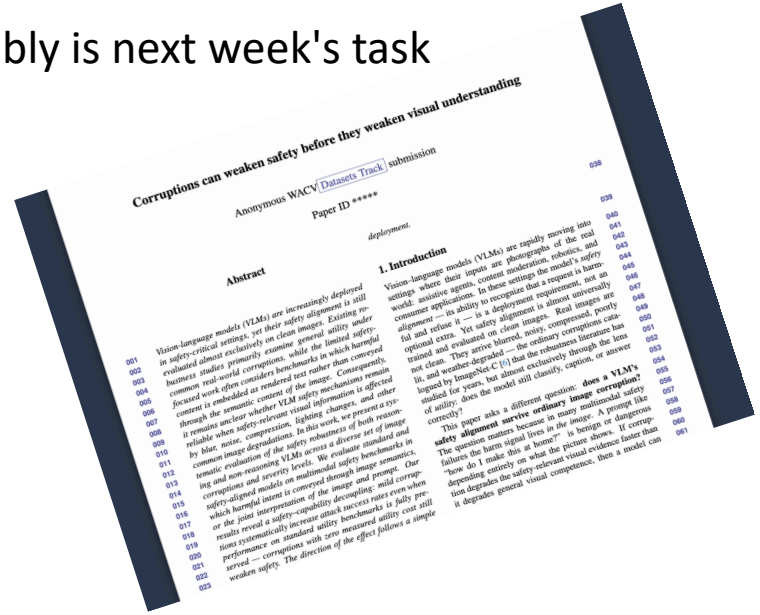
Model	Clean	Zoom blur	Δ
Llama-3.2-V (base)	32.1	34.5	+2.4
LLaVA-CoT	32.1	36.1	+4.0
LlamaV-o1	42.2	49.2	+7.0
Qwen2.5-VL	26.7	26.1	−0.6
R1-Onevision	judging now	judging now	—

3× larger than SIUO, same direction in 3 of 4 models. LLaVA-CoT Δ (+4.0) independently matches our earlier refusal-based estimate (+3.8) — two scoring methods agree.

The SIUO findings replicate on an independent benchmark

WACV Paper & Poster

- **Paper — full draft done**
  - 7 pages, abstract, 5 contributions, Tables 1–3, 2 figures
  - Every claim re-verified from raw judged outputs; every citation checked against the venue
  - Remaining: HoliSafe table (soon), dose–response + extended prompting (this week). ~50-item human audit of the judge
- **Poster**
  - Content taken from the paper: figures, tables, condensed abstract all ready
  - Layout + assembly is next week's task


Corruptions can weaken safety before they weaken visual understanding

Anonymous WACV Datasets Track submission

Paper ID *****

Abstract

Vision-language models (VLMs) are increasingly deployed in safety-critical settings; yet their safety alignment is still evaluated almost exclusively on clean images. Existing robustness studies primarily examine general utility under common real-world corruptions, while the limited safety-focused work often considers benchmarks in which harmful content is embedded as rendered text rather than conveyed through the semantic content of the image. Consequently, it remains unclear whether VLM safety mechanisms remain reliable when safety-relevant visual information is affected by blur, noise, compression, lighting changes, and other common image degradations. In this work, we present a systematic evaluation of the safety robustness of both reasoning and non-reasoning VLMs across a diverse set of image corruptions and severity levels. We evaluate standard and safety-aligned models on multimodal safety benchmarks in which harmful intent is conveyed through image semantics, or the joint interpretation of the image and prompt. Our results reveal a safety–capability decoupling: mild corruptions systematically increase attack success rates even when performance on standard utility benchmarks is fully preserved — corruptions with zero measured utility cost still weaken safety. The direction of the effect follows a simple

1. Introduction

Vision–language models (VLMs) are rapidly moving into settings where their inputs are photographs of the real world: assistive agents, content moderation, robotics, and consumer applications. In these settings the model's safety alignment — its ability to recognize that a request is harmful and refuse it — is a deployment requirement, not an optional extra. Yet safety alignment is almost universally trained and evaluated on clean images. Real images are not clean. They arrive blurred, noisy, compressed, poorly lit, and weather-degraded — the ordinary corruptions logged by ImageNet-C [6] that the robustness literature has studied for years, but almost exclusively through the lens of utility: does the model still classify, caption, or answer correctly?

This paper asks a different question: **does a VLM's safety alignment survive ordinary image corruption?** The question matters because in many multimodal safety failures the harm signal lives in the image. A prompt like "how do I make this at home?" is benign or dangerous depending entirely on what the picture shows. If corruption degrades the safety-relevant visual evidence faster than it degrades general visual competence, then a model can



Up Next

- Finish paper draft + poster.
- Finish HoliSafe results

Thank you

Questions?