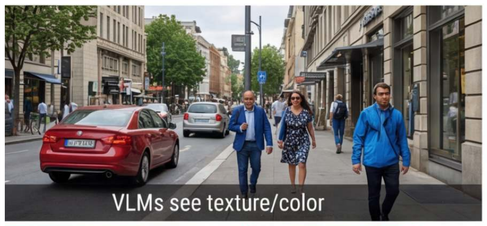
ThermEval: A Structured Benchmark for Evaluation of Vision-Language Models on Thermal Imagery

Ayush Shrivastava, Kirtan Gangani*, Laksh Jain*, Mayank Goel, Nipun Batra

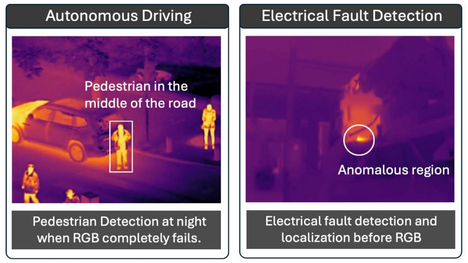
August 9–13, 2026
KDD 2026
Jeju, Korea

Motivation

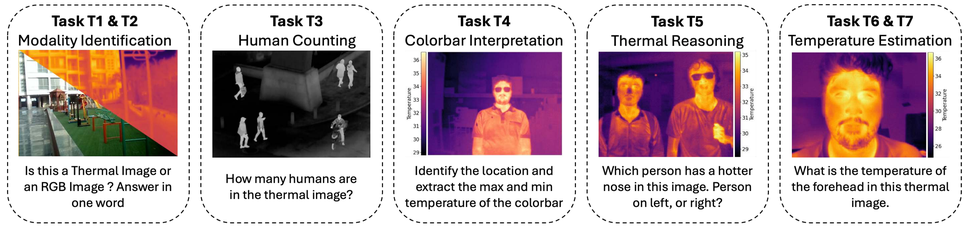


- **Thermal images encode temperature**, not color or texture.
- False-colored thermal images are only a visualization.
- Grounded Reasoning requires dense per-pixel temperature field.
- Can VLMs trained predominantly on RGB data reason about temperature-specific tasks in thermal images?

Applications



ThermEval Benchmark



7 **progressively challenging tasks** spanning **perception**, **reasoning**, and **temperature estimation**

70,000+ Thermal-VQA Pairs

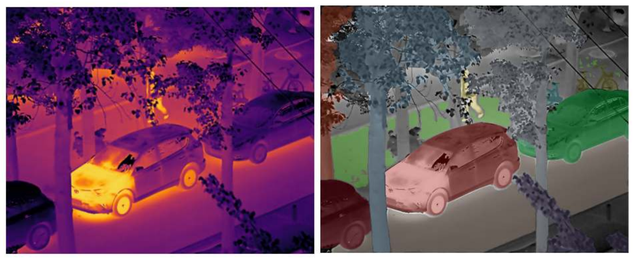
4 closed-source, 17 open-source and 3 chart specialised VLMs

Designed to uncover failure modes

ThermEval Dataset

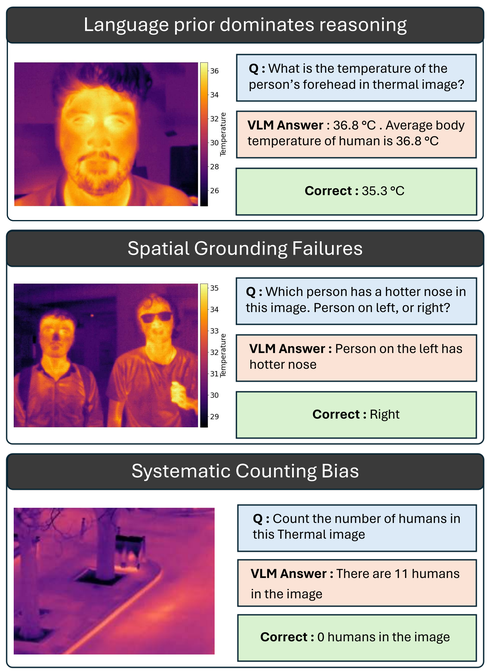
3,500 thermal images collected across **diverse indoor and outdoor scenes**

Dense **per-pixel temperature** maps



Thermal images containing bounding boxes and **segmentation masks across 20+ object categories**

Failure Modes



Findings

- **Failure Cascades :** Poor colorbar interpretation (T4) consistently leads to poor temperature estimation and thermal reasoning (T6–T7)
- **Bigger Models ≠ Better Thermal Understanding :** Failures persist from 0.3B to 235B models. The bottleneck is thermal grounding, not model size
- **Fine-Tuning Closes the Gap, Not the Problem :** Fine-tuning approaches human performance, but 1–2°C errors remain which is too large for safety-critical applications
- **Prompting Can't Replace Grounding :** Context helps recognition, but provides little or inconsistent improvement on thermal reasoning (T5–T7)

IIT Gandhinagar
Carnegie Mellon University
SUSTAINABILITY LAB
smashlab SMART SENSING FOR HUMANS

Github Kaggle Paper