CityOS: GroundingEval at Scale

Vikhyat Kulshrestha, Maj Globocnik, Aryan Mittal
Advisors: Dr. Jorge Ortiz, Taqiya Ehsan, Shuren Xia, Oscar Wu

RUTGERS
WINLAB | Wireless Information Network Laboratory

Introduction

Idea: A vision-language model can produce a fluent, confident answer that is still unsupported by sensor evidence. Using another generative model as the judge can introduce further hallucination.

Challenge: How can free-text answers about real or simulated scenes be scored reproducibly, using evidence alone, with no generative model making the final truth decision?

Approach: GroundingEval converts CARLA simulation and Smart Room sensor data into structured ground-truth JSON and carefully designed benchmark scenarios. It then deterministically compares TellMe's structured claims against this ground truth.

Key Benefit: Reproducible, claim-level verdicts reveal whether failures come from incorrect counts, states, relations, or temporal events, not just an overall score.

Background & Motivaiton

The Problem

Driving scene (video)
↓
VLM describes the scene
↓
"A pedestrian crossed while the car waited at the light…"
true facts mixed with invented ones
↓
LLM judge grades the paragraph
the judge shares the model's priors
↓
Hallucination survives

Open loop: fluent but unverifiable descriptions

Proposed Solution

Same VLM description
↓
Decompose into atomic claims
↓
"pedestrian present" ✓ | "pedestrian crossed" ✗ | "two vehicles" ✓
↓
Compare each claim to ground truth
↓
Supported | Contradicted | Not related | Not evaluable

Closed loop: every claim verified against ground truth

Q&A Generator Difficulty Taxonomy

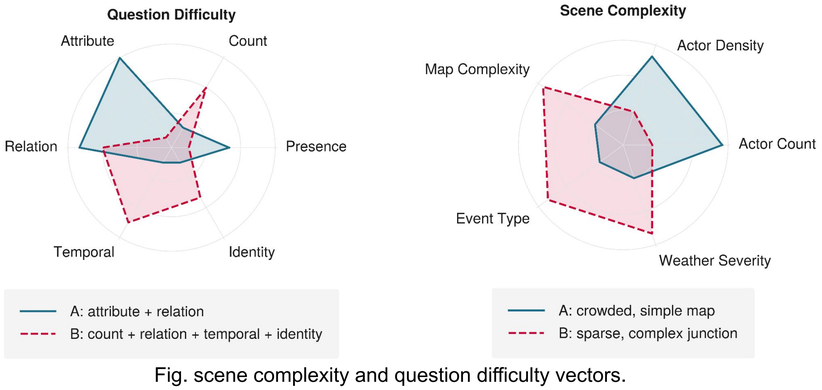
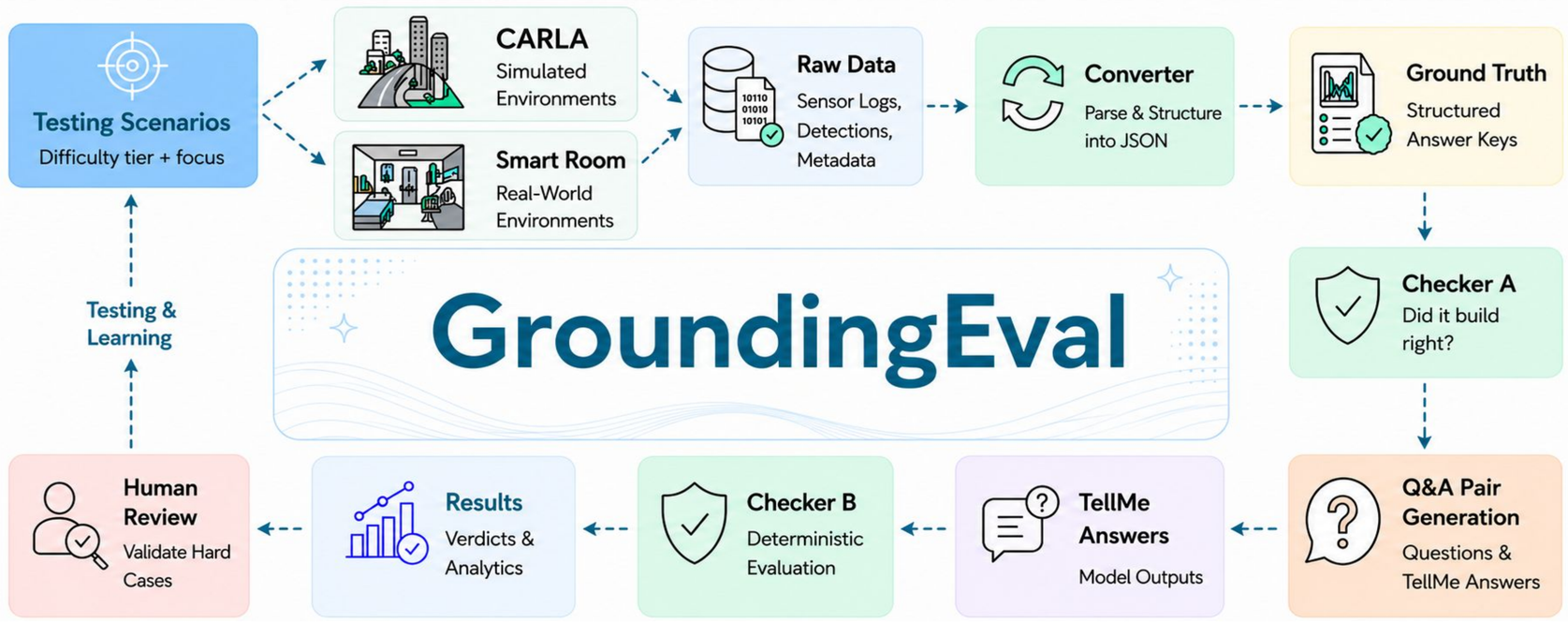


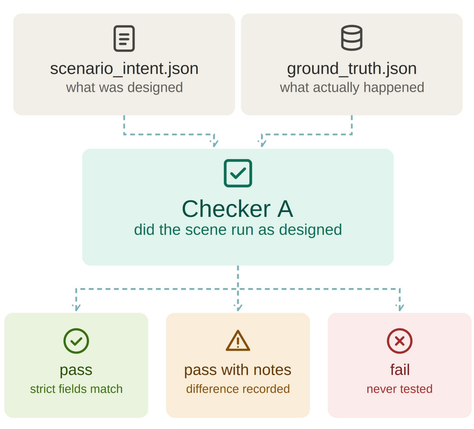
Fig. scene complexity and question difficulty vectors.

- Difficulty comes from the question and the scene, not from how the model scores.
- Question difficulty is how many filters stack: presence, count, attribute, relation, temporal, identity.
- Scene complexity is tracked on its own: actors, density, map, event, weather.
- Questions have to be answerable from what real sensors see.

Full Pipeline

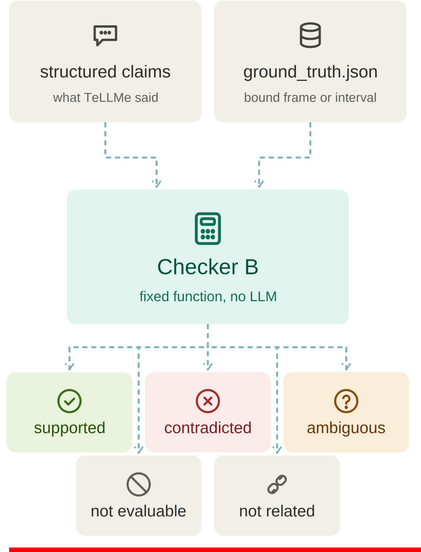


Deterministic Checker A



- Compares scenario intent against recorded ground truth
- Strict fields must match exactly
- Minor differences are allowed, but recorded in the results.
- Failed scenarios discarded before any testing

Deterministic Checker B



- Compares every claim to ground truth at the right frame.
- Fixed function, never an LLM judge, scoring is completely deterministic
- Five verdicts, including abstention when uncertain
- Returns one of five verdicts plus the supporting evidence.

Results

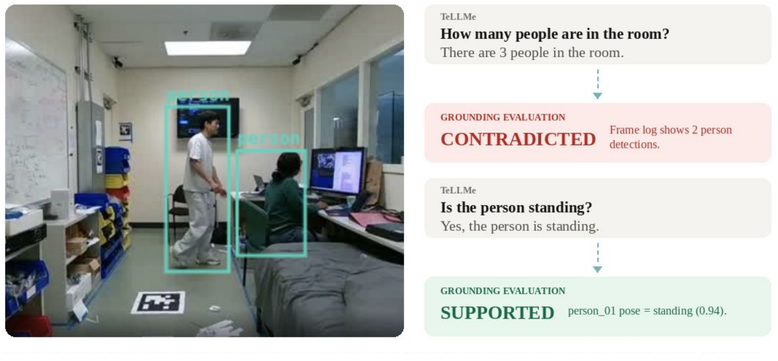


Fig. individuals in smart room environment and result of evaluated query answer

Future Work

- Automated pass from CARLA run to ground truth, QA pairs, and verdicts, with fixture tests catching schema drift.
- Scale from 1–2 filters to 3–4+ with verified negatives and counterfactual runs.
- Add more CARLA maps, densities, and weather

Check out our WIki

This work was supported by the National Science Foundation (NSF) as part of the Center for Smart Streetscapes, under NSF Cooperative Agreement EEC-2133516.

WINLAB