

# Algorithmic Conservatism in Street-Level GeoAI: Evaluating Multimodal LLMs under a Bounded Visual Auditing Protocol

Apoorv Agrawal\*, Francesco Pilla\*, Anna Mölter\*

\* University College Dublin (UCD), Dublin, Ireland  
apoorv.agrawal@ucdconnect.ie, francesco.pilla@ucd.ie, anna.molter@ucd.ie

## Abstract

Multimodal large language models (LLMs) enable structured visual judgement to be incorporated into street-level geospatial analysis. Using CPTED-based bus stop auditing as a testbed, this paper examines where multimodal LLMs are best positioned in street-level GeoAI and how they behave when visual evidence becomes ambiguous. Street view imagery (SVI) from Dublin bus infrastructure is evaluated across object-explicit, spatial-relational, and condition-assessment tasks under a bounded visual auditing protocol that restricts unsupported inference and requires direct visual grounding. Results indicate the strongest agreement and stability for object-explicit tasks, while more interpretive tasks remain stable but exhibit cautious, evidence-grounded non-assertion under ambiguity. This pattern of algorithmic conservatism is interpreted as an observed behavioural tendency emerging under task ambiguity and protocol constraints, rather than as an intrinsic model property, with implications for safety-related auditing. The paper contributes a task taxonomy, a bounded visual auditing protocol, and a structured disagreement analysis for positioning multimodal LLMs within GeoAI workflows.

**Keywords:** Multimodal Large Language Models (LLMs), Street View Imagery (SVI), Algorithmic Conservatism

## 1. Introduction

Street view imagery (SVI) has become an important GeoAI data source for analysing urban environments at scale. Combined with computer vision (CV), SVI enables cost-efficient measurement of built-environment features and reduces reliance on labour-intensive field

---

Published in “Proceedings of the 1st International Conference on Geospatial Artificial Intelligence (GeoAI 2026) – Oral Presentation Papers”, edited by Haosheng Huang and Nico Van de Weghe, GeoAI 2026, 3-6 June 2026, Ghent, Belgium.

This contribution underwent single-blind peer review based on the extended abstract.

surveys (Biljecki and Ito 2021; Dai et al. 2025). Existing applications use SVI to detect, classify, and segment urban features, making visual auditing more scalable. However, street-level GeoAI tasks are not uniform. Some features can be assessed through direct visual observation, while others require interpretive judgement under partial, perspectival, or ambiguous street-level evidence. These tasks are harder to reduce to object detection because they depend on visibility, spatial arrangement, condition, and contextual interpretation. For such indicators, a purely detection-based framing may be insufficient. In this second category, multimodal large language models (LLMs) may add value not as replacements for CV, but as bounded visual reasoners within hybrid auditing workflows (Malekzadeh et al. 2025). Their role is to support structured judgments from visible cues when object-level recognition alone is insufficient. This positions the model as a structured rater rather than a universal classifier or predictor.

Against this background, this paper examines where multimodal LLMs fit within street-level GeoAI workflows, using Crime Prevention Through Environmental Design (CPTED)-based bus stop auditing as a testbed. CPTED is a design-based approach that links built-environment features, including visibility, access control, maintenance, and activity support, to crime prevention and perceived safety (Cozens and Love 2015). The study forms part of a broader research project developing an SVI-based CPTED index for bus stop environments through a hybrid workflow combining CV-based infrastructure detection and GPT-assisted indicator extraction. The aim is not to validate the CPTED index, but to use a balanced set of street-level tasks to examine where multimodal LLMs are useful in GeoAI. The paper addresses two research questions:

1. Where are multimodal LLMs best positioned within street-level GeoAI workflows?
2. How do multimodal LLMs behave in street-level visual judgement tasks with higher interpretive ambiguity under a bounded visual auditing protocol?

## 2. Study Design and Evaluation

### 2.1. Task Taxonomy and CPTED Indicators

Street-level visual auditing tasks differ in the type and complexity of judgment they require. Accordingly, this paper organises the evaluation around a task taxonomy, grouping indicators into three categories: object-explicit, spatial-relational, and condition-assessment tasks. Rather than organising the indicators by CPTED principles, this paper uses them as the empirical testbed for examining task-dependent multimodal LLM behaviour within street-level GeoAI. The study draws on the Google Street View Static API (Google 2025) imagery from 120 bus stops in the Dublin bus network. From a broader CPTED index of 25 indicators across six principles, six indicators were selected as test cases for single-image evaluation. *Table 1* presents the task taxonomy, corresponding CPTED indicators, and their expected suitability for conventional CV and multimodal LLM-based assessment.

### 2.2. Bounded Visual Auditing Protocol

The model was evaluated under a bounded visual auditing protocol designed to restrict unsupported inference across single-view SVIs. Using OpenAI’s GPT-4o-mini model (OpenAI 2024), each image was paired with a structured prompt built around five elements: *global*

*scoping rules* to maintain a consistent audit role across images and indicators; *fixed indicator definitions* to ensure that each indicator was interpreted consistently across all images; *exclusion criteria* to clarify what should not count as evidence for a given task, reducing over-inclusive interpretations; *non-speculation instructions* to prevent the model from inferring features beyond what was visually supported; *direct-visibility constraints* to limit assessment to cues directly observable in the image, aligning judgement with image-bounded evidence. No retrieval-augmented generation (RAG) or external CPTED knowledge base was used; inference relied on zero-shot prompting with compact indicator-specific instructions. To reduce output variability and support repeatability, the model was run with the temperature set to 0.

Task type	Visual judgement demands	Example indicators used in this study	Typical CV suitability	Expected role of multimodal LLMs
Object-explicit	Involves discrete, visually bounded entities, although detection may still be affected by occlusion or image perspective	Shelter; Bench	High, though sometimes domain-dependent	Limited or complementary
Spatial-relational	Requires judgement of spatial arrangement, openness, visibility, or relational structure rather than a single object	Unobstructed sight lines; Defined entrances	Moderate	Useful for scene-level judgement
Condition-assessment	Requires interpretive assessment of maintenance, disorder, or environmental quality	Graffiti/vandalism/trash; Infrastructure condition	Moderate to low	Useful for interpretive visual judgement

**Table 1.** Task taxonomy for positioning multimodal LLMs within street-level visual auditing (*Source: Author*).

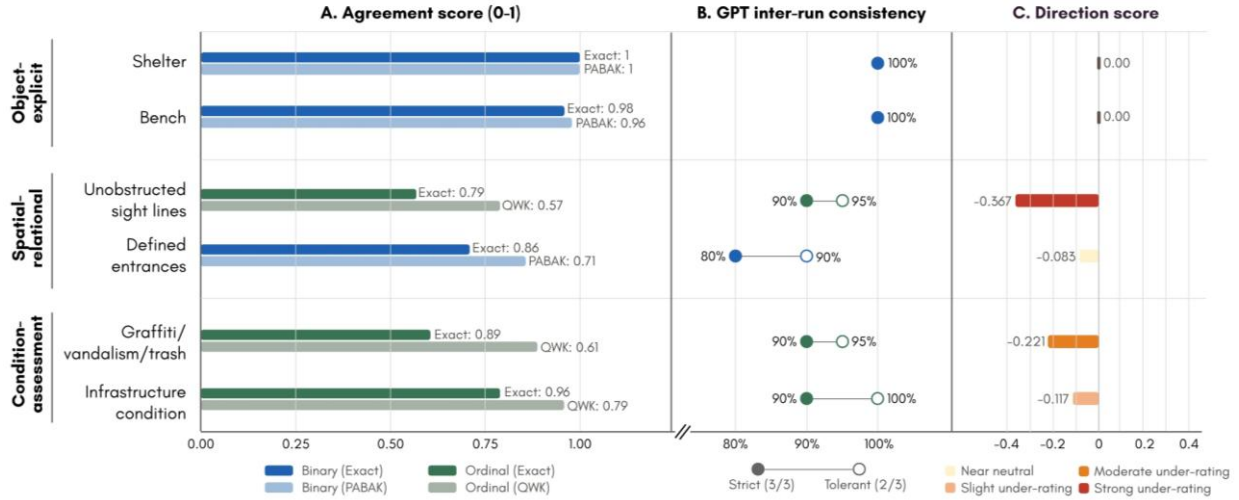
### 2.3. Metrics and Evaluation

Human-GPT evaluation was conducted at the indicator level and interpreted across the task taxonomy. For each indicator, ratings from two independent human raters were consolidated into one reference label, and GPT outputs were summarised by majority vote for comparison with it. Agreement was assessed using exact agreement for all indicators, PABAK for binary indicators, and quadratic-weighted kappa (QWK) for ordinal indicators. To assess stability, GPT inter-run consistency was measured as strict consistency when all three runs matched, and tolerant consistency when at least two runs matched. Disagreement direction was coded relative to the human reference using the coded response scale: GPT-higher/human-lower or GPT-lower/human-higher for mismatches, with exact agreement retained where labels matched.

## 3. Results

*Figure 1* summarises the indicator-level Human-GPT evaluation across the three task types, presenting agreement, GPT inter-run consistency, and disagreement direction. Negative direction scores indicate lower-scoring or more conservative GPT judgements relative to the consolidated human reference. The clearest alignment was observed for object-explicit tasks, where the visual target was discrete, directly visible, and closer to conventional CV detection. Across these tasks, agreement with the human reference was very high, with exact agreement ranging from 0.98 to 1.00 and PABAK ranging from 0.96 to 1.00. Strict consistency was also 100%, indicating that

the model reproduced the same judgment across repeated runs. Direction scores were neutral, showing no systematic tendency to score lower or higher than the human reference. Object-explicit tasks, therefore, represent the most stable and high-alignment task type in the audit, where model judgement is primarily based on recognising visible infrastructure.



**Figure 1.** Human-GPT evaluation across street-level visual auditing task types: (A) agreement score, (B) GPT inter-run consistency, and (C) disagreement score (*Source: Author*).

For spatial-relational and condition-assessment tasks, agreement remained meaningful but became more variable and task-dependent. The binary spatial-relational indicator achieved Exact = 0.86 and PABAK = 0.71, while the ordinal spatial-relational indicator showed Exact = 0.79 and QWK = 0.57. Condition-assessment tasks showed moderate-to-high agreement, with Exact ranging from 0.89 to 0.96 and QWK ranging from 0.61 to 0.79. Across both task types, strict consistency remained relatively high, ranging from 80% to 90%, while tolerant consistency ranged from 90% to 100%. This indicates that variation in agreement was driven less by erratic run-to-run behaviour than by the interpretive demands of assessing openness, visibility, maintenance, or disorder from street-level evidence.

Direction scores further distinguished these interpretive tasks. They were consistently negative, ranging from -0.083 to -0.367 for spatial-relational tasks and -0.117 to -0.221 for condition-assessment tasks. This indicates that disagreement was not randomly distributed; the model more often produced conservative (less affirmative) judgments when visual evidence was partial, ambiguous, or perspective-limited. Across task types, the results therefore show a shift from high-alignment object recognition to stable but more cautious interpretive judgement. This observed pattern is described as algorithmic conservatism: a recurrent tendency to avoid unsupported affirmative or higher-scoring judgements when street-level visual evidence is constrained or ambiguous.

## 4. Discussion and Conclusion

This paper contributes a task taxonomy for street-level visual judgement, evaluates multimodal LLMs under a bounded visual auditing protocol, and uses structured disagreement analysis to

examine how cautious non-assertion emerges under ambiguity. The findings support a task-sensitive understanding of multimodal LLMs in street-level GeoAI, showing that their behaviour varies with the task's demand for interpretive judgement. For object-explicit tasks, the high alignment observed in the results indicates a potential role for multimodal LLMs in low-training-data auditing contexts, particularly where locally variable infrastructure may not be well represented in task-specific pre-trained CV models.

For spatial-relational and condition-assessment tasks, the key finding is not simply reduced agreement, but the directional nature of disagreement. Under this protocol that restricted unsupported inference and required direct visual grounding, the model tended towards cautious non-assertion in ambiguous scenes. This study, therefore, interprets algorithmic conservatism not as an intrinsic property of the model, but as an observed behavioural pattern emerging from the interaction between task ambiguity and protocol design. This distinction is important in safety-related auditing: over-assertive outputs could falsely confirm the presence or quality of safety-relevant features, while conservative under-assertion may overlook cues that human raters infer from contextual knowledge. It is especially relevant for perception-oriented applications, where ambiguity is not merely noise but part of how people evaluate street-level environments.

Future perception-based validation, including pairwise safety-choice scenarios, may help calibrate this conservative behaviour by identifying when cautious non-assertion is appropriate, when prompts require refinement, and when task definitions should better reflect human perceptual judgement. Overall, the study positions multimodal LLMs as structured raters within bounded visual auditing workflows, where their value lies in task-sensitive judgement rather than universal classification or prediction.

## References

- Biljecki F, Ito K (2021) Street view imagery in urban analytics and GIS: a review. *Landscape and Urban Planning* 215:104217. doi:10.1016/j.landurbplan.2021.104217
- Cozens P, Love T (2015) A review and current status of crime prevention through environmental design (CPTED). *Journal of Planning Literature* 30(4):393-412. doi:10.1177/0885412215595440
- Dai Y, Liu L, Wang K, Li M, Yan X (2025) Using computer vision and street view images to assess bus stop amenities. *Computers, Environment and Urban Systems* 117:102254. doi:10.1016/j.compenvurbsys.2025.102254
- Google LLC (2025) Google Street View Static API. <https://developers.google.com/maps/documentation/streetview>.
- Malekzadeh M, Willberg E, Torkko J, Toivonen T (2025) Urban attractiveness according to ChatGPT: contrasting AI and human insights. *Computers, Environment and Urban Systems* 117:102243. doi:10.1016/j.compenvurbsys.2024.102243
- OpenAI (2024) GPT-4o mini. OpenAI, San Francisco, CA