Searchable.City: An Open-Vocabulary Semantic Atlas

Sean Hardesty Lewis
shl225@cornell.edu
Cornell Tech
New York, NY, USA

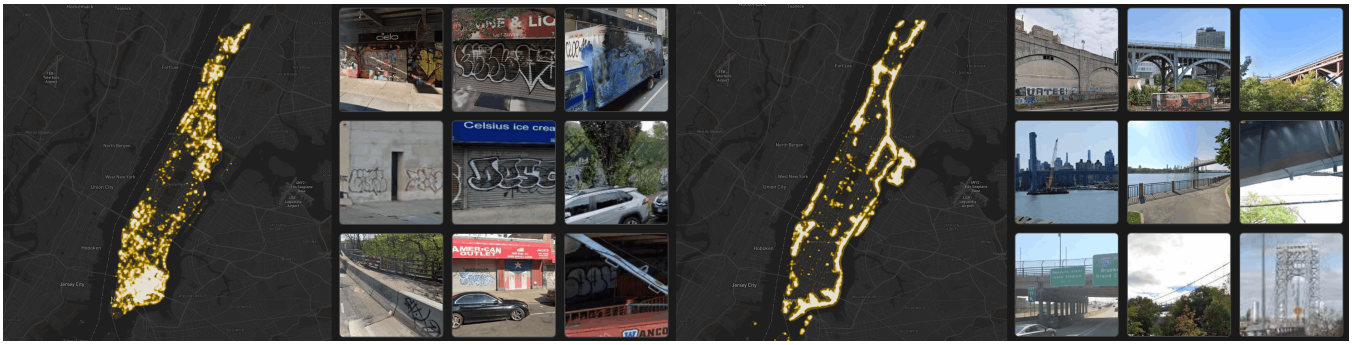


Figure 1: *Searchable.City* localizes visual patterns across city-scale street-level imagery. A query for graffiti (left) highlights dense corridor clusters, while bridges (right) traces Manhattan's perimeter and crossings. Each map view is paired with representative retrieved images (3×3).

Abstract

Maps excel at locations and are famously poor at surfaces: the fire escapes, awnings, scaffolds, shade, murals, and textures that make cities legible to residents rarely exist as database fields. We present *Searchable.City*, a pipeline that treats the city as an image corpus rather than only a registry of points of interest. By running a vision–language model (VLM) over millions of New York City streetscape images and indexing the resulting descriptions, we construct an *open-vocabulary semantic atlas*: a map that can be searched for meanings ("Chinese", "gothic", "construction") instead of addresses. We show how this translation, from city to caption to queryable field, reveals cultural gradients, architectural memory, and infrastructural churn, while also clarifying the limits of the street-view vantage. Open-vocabulary mapping is not only a technical interface but also a cultural instrument: it redraws neighborhood boundaries according to what the machine can name, and what it cannot see.

ACM Reference Format:
Sean Hardesty Lewis. 2026. Searchable.City: An Open-Vocabulary Semantic Atlas. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Emerging Technologies (SIGGRAPH '26), July 19–23, 2026, Los Angeles, CA, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

SIGGRAPH Emerging Technologies '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-XXXX-X/26/07
https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 Introduction

Current GIS systems represent the city as vectors and points, but they miss the semantic texture of urban life. A digital map knows where the pharmacy is, but rarely where the fire escape begins. We present *Searchable.City*, an interactive system that asks the city to describe itself. By processing millions of street-view images with a vision–language model (VLM), we build a queryable spatial index that lets users search the physical world for arbitrary visual concepts, such as "art deco" and "trash piles". This shifts the map from a navigational tool into a semantic search engine.

Prior work in urban computing has used street-view imagery mostly for scalar prediction, such as estimating perceived safety [Naik et al. 2014]. In parallel, open-vocabulary vision models such as CLIP [Radford et al. 2021] and SAM [Kirillov et al. 2023] enable recognition without a fixed label set. We bridge these strands by using instruction-tuned VLMs [Vasu et al. 2024] to generate a broad, open-ended lexicon of the streetscape, moving beyond pre-defined classes to capture the long tail of urban details.

2 Methodological Innovations

The core innovation of *Searchable.City* is a shift in the mapping primitive: mapping language onto geography, rather than only coordinates onto images. The pipeline begins by ingesting a large corpus of street-view imagery while retaining geospatial coordinates and camera heading. To handle the dense overlap of urban photography, we discretize the streetscape into directional "vantage cells" (Figure 2) so that queries align with what is visible from the street, not only with a top-down location.

For each vantage cell, a VLM generates a detailed descriptive caption. We decompose this caption into a multi-tag representation, producing a lexicon of thousands of descriptors. Unlike fixed-class

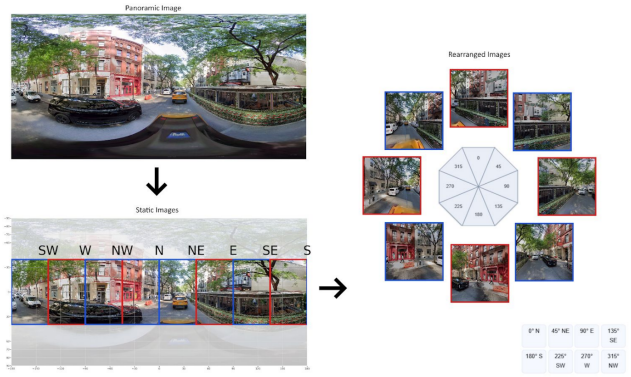
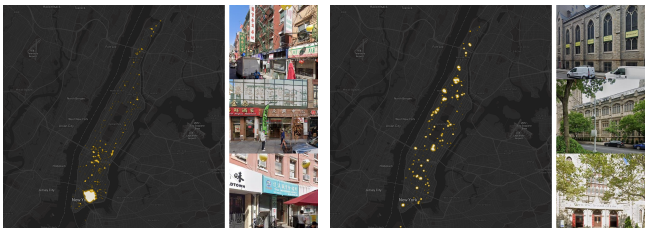


Figure 2: Vantage Cell Discretization: The streetscape is divided into directional slices based on available street-view metadata, creating spatial buckets for query aggregation.



(a) Query: "Chinese" **(b) Query: "Gothic"**

Figure 3: Open-vocabulary neighborhood delineation: the atlas outlines cultural and stylistic regions through repeated visual cues rather than explicit GIS layers.

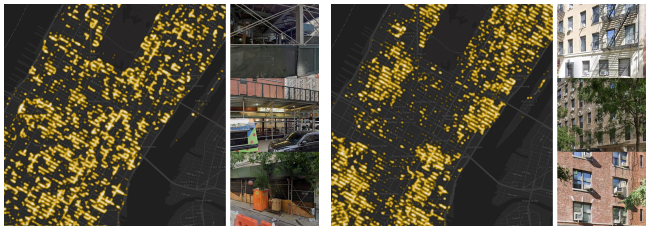
segmentation, this lets the system index ad-hoc concepts like "temporary scaffolding" or "art deco." Tags are then spatially aggregated into a probabilistic index. The interface works like a search engine: a user inputs a string, and the system returns a probability mass showing where that visual description clusters across the city.

3 Applications and Findings

The system enables new forms of urban reading for architects, sociologists, and planners. It can surface cultural gradients by mapping where stylistic markers (e.g., "gothic") or cultural signals (e.g., "Chinese") recur, producing neighborhood boundaries from visual density rather than administrative zip codes (Figure 3). It can also make infrastructure legible by indexing the city's churn: a query for "scaffolding" reveals the temporary city of renovation, while proxies like "air conditioning" suggest building age and retrofit status (Figure 4). Finally, the atlas functions as an everyday lexicon, letting users quantify the visual prevalence of unmapped objects, from subway vents to street art.

4 Challenges and Limitations

Searchable.City explicitly documents its limitations as part of the work's claim: the atlas is bound by the physics of the street-view camera and the semantics of the model. Occlusion turns objects into



(a) Query: "Scaffolding" **(b) Query: "Air conditioning"**

Figure 4: Two readings of infrastructure: (left) the temporary city of construction; (right) building retrofit signals.

absences; lighting collapses detail; and the camera's chosen routes omit courtyards, rooftops, and interior life. These limitations are not only technical error; they are cartographic arguments. Critical cartography reminds us that maps are made, not found [Harley 1989; Wood 2010]. In open-vocabulary mapping, the "made" quality becomes doubly visible: the map is produced by a chain of translations (world → image → caption → query), each of which can introduce bias.

5 Conclusion

Searchable.City asks a simple question with large consequences: what happens when a map becomes a search engine over images? The answer is an atlas of meaning: neighborhoods emerge from signage and style, renovation from scaffolding, and socioeconomic texture from retrofit artifacts. The project also exposes the politics of machine vision: the city that AI sees is constrained by where cameras go, what frames exclude, and which words models reach for first. Open-vocabulary cartography does not replace the map of coordinates; it complements it with a map of surfaces. If the twentieth century gave us navigation, the next may give us perception at scale: a city we can query not only for where, but for what.

Acknowledgments

We use imagery from Google Street View. © 2026 Google LLC, under fair use.

References

J. B. Harley. 1989. Deconstructing the Map. Cartographica, 26(2), 1–20.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643

Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César A. Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://openaccess.thecvf.com/content_cvpr_workshops_2014/W20/papers/Naik_Streetscore_-_Predicting_2014_CVPR_paper.pdf

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2024. FastVLM: Efficient Vision Encoding for Vision Language Models. arXiv:2412.13303 [cs.CV] https://arxiv.org/abs/2412.13303

Denis Wood. 2010. *Rethinking the Power of Maps.* Guilford Press.