

Searchable.City: An Open-Vocabulary Semantic Atlas

Sean Hardesty Lewis

end@mit.edu

Massachusetts Institute of Technology

Cambridge, MA, USA

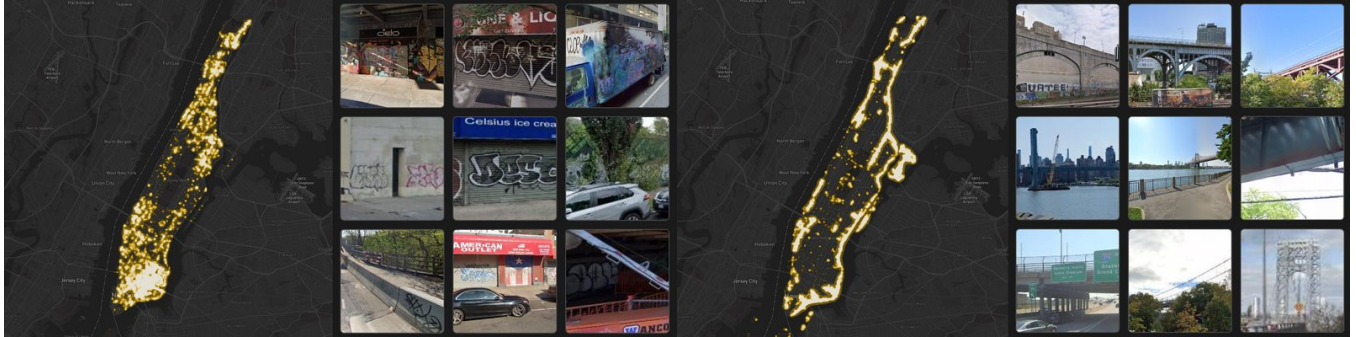


Figure 1: *Searchable.City* localizes visual patterns across city-scale street-level imagery. A query for graffiti (left) highlights dense corridor clusters, while bridges (right) traces Manhattan’s perimeter and crossings. Each map view is paired with representative retrieved images (3 × 3).

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 Emerging Technologies '26)*, July 19–23, 2026, Los Angeles, CA, USA. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3799826.3812433>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGGRAPH Emerging Technologies '26, July 19–23, 2026, Los Angeles, CA, USA

© 2026 ACM.

ACM ISBN 979-8-4007-2549-4/2026/07

<https://doi.org/10.1145/3799826.3812433>

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.

The contribution is not simply that a VLM can label urban images. The labels become a graphics primitive: a word becomes a spatial field that can be searched, compared, and projected. In this sense, *Searchable.City* treats language as a way to render the city.

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.

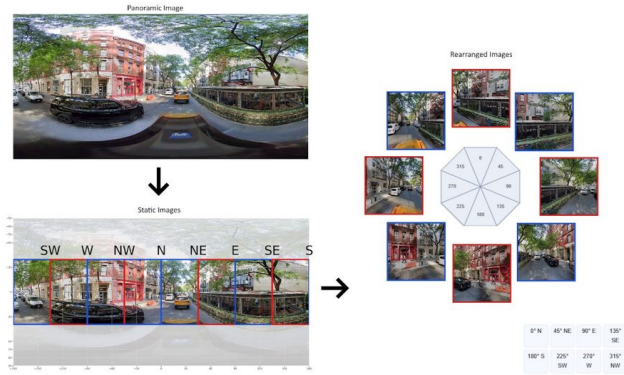


Figure 2: Vantage Cell Discretization: The streetscape is divided into directional slices based on available street-view metadata, creating spatial buckets for query aggregation.

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 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.

Unlike conventional GIS layers, the atlas is not organized around predefined object categories or administrative boundaries. Any phrase the captioning model can describe becomes a spatial query, allowing the same interface to retrieve architectural styles, cultural signage, temporary infrastructure, or informal street conditions.

The current implementation processes over 20 million street-view images from New York City imagery sources such as Mapillary. Each image retains coordinates, camera heading, and capture metadata before VLM captioning and semantic indexing. Query results are aggregated across directional vantage cells and rendered as continuous probability fields. Unlike conventional image retrieval, *Searchable.City* turns semantic matches into geographic layers that can be compared, projected, and interpreted at urban scale.

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. The same semantic layers can also leave the browser: in the installation below, query results become projected material on a physical city model.



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

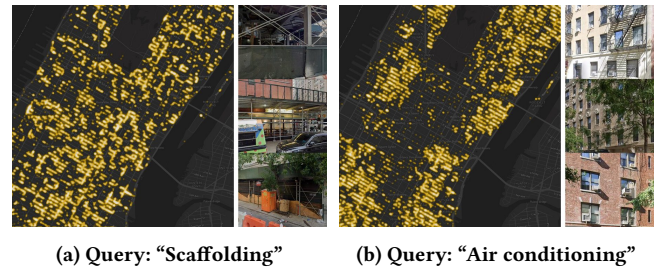


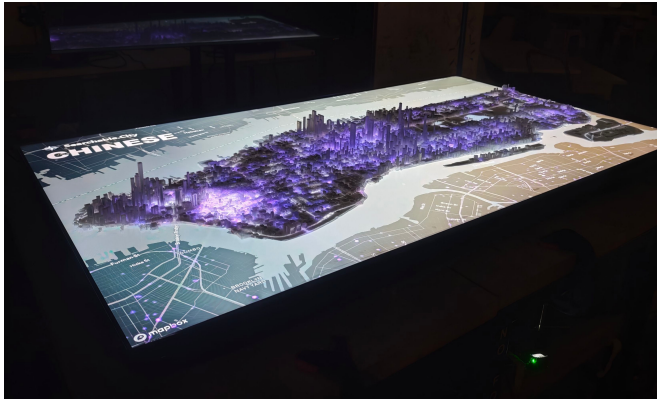
Figure 4: Two readings of infrastructure: (left) the temporary city of construction; (right) building retrofit signals.

These examples matter because the system was not trained to find Chinatown, Gothic architecture, facade work, or retrofit conditions as named categories. They appear through repeated visual descriptions attached to street-level views. A cultural query follows signage and storefronts; an architectural query follows ornament, facade form, and institutional massing; an infrastructure query follows the visible traces of repair and maintenance. Because the atlas is open-vocabulary, the same mechanism can be applied to almost any visible feature of the city, from sidewalk sheds and murals to pharmacies, bike lanes, flower boxes, and window air conditioners. The user is not limited to a prepared menu of map layers. Each query tests whether a phrase has a visible spatial footprint in the streetscape, and the atlas turns those repeated descriptions into searchable urban patterns.

The atlas also exposes the instability of language itself. Similar queries such as “gothic,” “cathedral,” and “ornate” can recover overlapping but non-identical spatial signatures, while culturally loaded phrases such as “wealthy” or “safe” reveal how results depend on visual evidence, model semantics, and user language rather than objective urban categories.

4 PHYSICAL INSTALLATION

The screen interface makes the city searchable; the physical interface makes that search collective. We built a projection-mapped installation around a three-dimensional model of Manhattan, using the model as a surface for semantic heatmaps (Figure 5). When a query such as “murals,” “bridges,” or “skyscrapers” is entered, the result appears both as a browser layer and as a projection across blocks, waterfronts, and corridors.



(a) Projection-mapped model.



(b) Close-up of a projected heatmap.

Figure 5: *Searchable.City* as a physical installation. The same semantic layers used in the browser are projected onto a three-dimensional model of Manhattan.

This format has precedents in tangible urban interfaces and augmented city models [Ishii et al. 2002; Noyman 2022]; here, the model tests what semantic search does when its output becomes public. On a screen, search is private. On the model, the same query becomes visible to a room at once, allowing viewers to point to clusters, compare them with retrieved street-level images, and debate whether the pattern is meaningful, partial, or misleading.

5 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 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.

6 CONCLUSION

Searchable.City asks a simple question with large consequences: what happens when a map becomes a search engine over images? The result is a new kind of interface: one where perception itself becomes spatially navigable, and where the visual world can be searched, compared, and interpreted through language. 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.

PROJECT AVAILABILITY

Searchable.City is publicly available at <https://searchable.city>, where users can explore interactive semantic atlases across multiple cities, including New York City and Los Angeles. Source code is available at <https://github.com/seanhlewis/searchablecity>.

REFERENCES

- J. B. Harley. 1989. Deconstructing the Map. *Cartographica*, 26(2), 1–20.
- Hiroshi Ishii, John Underkoffler, D. Chak, Ben Piper, Eran Ben-Joseph, Luke Yeung, and Zahra Kanji. 2002. Augmented Urban Planning Workbench: Overlaying Drawings, Physical Models and Digital Simulation. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Computer Society, 203–211. <https://doi.org/10.1109/ISMAR.2002.1115090>
- 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
- Ariel Noyman. 2022. *CityScope: An Urban Modeling and Simulation Platform*. Ph.D. dissertation, Massachusetts Institute of Technology.
- 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 Grascch, 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.