StreetTransformer: A Large Scale Multimodal Benchmark for Change Detection in Urban Streets

Jon Atkins
University of California, Berkeley

Joao Rulff
New York University

Claudio Silva
New York University

Maryam Hosseini
University of California, Berkeley

Abstract

As the nodes of transportation networks, where modal and directional flows converge, intersections are inherently complex compositions of myriad infrastructure design choices over time. Yet, benchmarks that test machine or LLM understanding of street-level infrastructure and its longitudinal changes are scarce. We present STREETTRANSFORMER, a large scale multimodal corpus and benchmark of New York City streetscape change from 2006 to 2024. It links orthorectified aerial imagery, time-aware planimetric features, capital project records, and thousands of annotated civic design documents, covering about 47,000 intersections, 470,000 aerial snapshots, and more than 33,000 document pages with cross modal links to project records. As a benchmark, we define five tasks for LLMs in change- and feature-detection, document-to-image linking, and temporal-localization, with optional semantic segmentation masks. We find that GPT baselines detect change but struggle with temporal localization and reasoning across sources, which motivates task-specific training and ontology based evaluation.

1 Introduction

As the bedrock of urban form, city streets play an omnipresent role in civic life and are an input to countless models that describe urban systems. Annually, US local governments spend over $ 100 million on street capital reconstruction projects[10], including redesigning streets to improve safety, efficiency, comfort, or vibrancy. Yet, few quantitative methods attempt to comprehensively encapsulate the complexities of urban street design or its longitudinal changes. Existing data sources are plentiful but fragmented across modalities, siloed within jurisdictions, or temporally frozen, leaving no systematic scalable basis for representing the spatio-temporal evolution of modern streetscapes. Large-Language Models (LLMs) have shown promising performance in recent years, enabling "zero-shot" reasoning over diverse, unstructured data sources. But they have yet to prove effective in understanding the changing urban infrastructure.

We present STREETTRANSFORMER, a large-scale, multi-modal framework for representing and retrieving *longitudinal streetscape change* at the individual intersections level. Instantiated on New York City from 2006–2024, STREETTRANSFORMER integrates orthorectified aerial imagery, time-aware planimetric features, city project logs and thousands of unstructured design documents containing detailed diagrams scraped from agency archives. This integration produces a spatio-temporally grounded corpus covering more than $\sim$47,000 intersections and $\sim$470,000 biannual aerial snapshot images, and more than 33,000 pages of pdf documents, enriched with cross-modal links to 1,000's of capital project records and their design narratives.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Our contributions are threefold: 1) We introduce a novel, unified spatio-temporal schema that integrates disparate modalities enabling consistent fusion of imagery, features, records and documents at the intersection level. 2) We establish the first comprehensive benchmark on longitudinal urban streetscape infrastructure change, with tasks spanning detection, document grounding and temporal reasoning. 3) We demonstrate **Semantic Signal Extraction**, showing that segmentation-derived masks can amplify true infrastructural change by suppressing irrelevant noise and providing a novel evaluation axis for urban ML.

2 Related Work

Recent advances in multi-modal large language models (MLLMs) have opened new directions for analyzing urban environments by coupling visual inputs with natural language reasoning. Early studies applied vision–language models (VLMs) such as CLIP to score walkability from street-view imagery [4], while others explored generative evaluation of streetscapes or reviewed the role of LLMs in urban analysis [1, 7]. These works demonstrate both potential and brittleness: models can automate large-scale assessments, but outputs are often optimistic or generic unless carefully prompted. In Earth observation (EO), MLLMs have been tested on benchmarks such as VRSBench [5], SARLANG-1M [2], and GPT-4V evaluations [12], which show strong performance on captioning and landmark recognition but poor results on fine-grained tasks like change detection. Remote sensing image change captioning extends this direction, with datasets and models such as RSICCformer [6], Semantic-CC [13], SAT-Cap [11], and RB-SCD [9]. These benchmarks move beyond pixel differencing but remain centered on imagery and text, without aligning data from diverse sources such as project documents, operational features, and geospatial datasets. Our work advances this space by instantiating a multi-modal benchmark of *longitudinal streetscape change* on a novel dataset enabling tasks such as *semantic signal extraction*, document–image linking, and temporal localization, extending MLLM benchmarking into the domain of urban infrastructure.

3 STREETTRANSFORMER

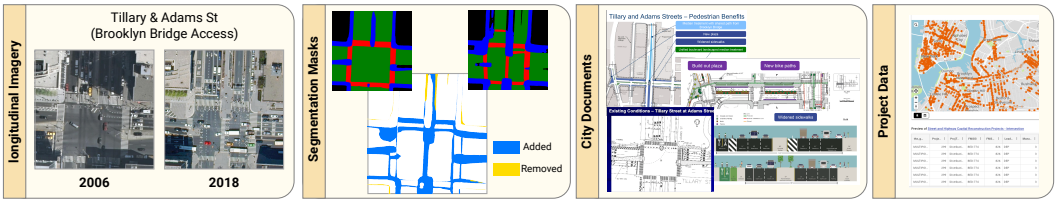

Figure 1: Data modalities representing New York City streetscape features.
STREETTRANSFORMER explores translation and coordination between each of the four.

Intersection Universe. We anchor our framework on the NYC LION street centerline dataset, extracting ∼47,000 consolidated intersection nodes as the unit of analysis. We extract stratified samples of ∼2000 of these locations related to each task ensuring balanced representation of completed projects on different times and scopes, and supplement with control of static and legacy project locations.

3.1 Multi-Modal Data Integration

To capture streetscape change, we assemble data across four modalities: imagery, city project data, planimetric features, and design documents; and reconcile them into a unified spatio-temporal schema.

Imagery: We collect aerial images (2006–2024) at zoom level 20 (15 cm resolution), producing ten temporal snapshots per intersection, stitch them into mosaics and crop them to standard bounding boxes around intersection centroids. We refer to them as Images. **City Projects**: From *NYC OpenData* we extract a geocoded log of *Capital Reconstruction Projects* since 2001 and match them to Locations within 100 ft. From this, we filter to transportation related reconstruction projects while filtering out minor changes like resurfacing or art installations. We refer to them as Projects.
Documents: Thousands of PDF documents (2007–2024) are scraped from the NYC DOT web

archive. OCR+LLM extraction identifies cross-streets, which are geocoded via the NYC Geoclient API [8] and matched to intersections. We refer to them as Docs. **Planimetric Differencing:**We use *Tile2Net* [3] to generate segmentation masks of sidewalks, crosswalks, and roads. By differencing masks across start–end year pairs, we obtain feature-specific change maps that isolate where and in which categories interventions occurred. **Cross-Modal Alignment:** All modalities are reconciled into a unified index keyed on (ℓ, t). For each intersection-year we store *state data* (Images, features), and for each year-pair we derive *change data* (intervening project, feature deltas, masks, linked documents).

3.2 Problem Formulation and Schema

Let $\mathcal{L}$ denote the set of locations $\mathcal{T}$ the set of years. For $\ell \in \mathcal{L}$, $t \in \mathcal{T}$, (t_1, t_2) with $t_1 < t_2$, define

$$S(\ell, t) = \big(I_{\ell,t}, F_{\ell,t}\big), \qquad \Delta(\ell; t_1, t_2) \quad = \big(\Delta F_{\ell;t_1\to t_2},\, M_{\ell;t_1\to t_2},\, \mathcal{P}_{\ell;t_1\to t_2},\, \mathcal{D}_{\ell;t_1\to t_2}\big).$$

Here $I_{\ell,t}$ is a georeferenced image and $F_{\ell,t}$ a standardized feature vector; ΔF are featural deltas, M are raster change masks (from planimetrics or Tile2Net), $\mathcal{P}$ matched city projects, and $\mathcal{D}$ grounded documents. This state/change schema resolves heterogeneous modalities into a common structure that supports both descriptive benchmarking and contrastive representation learning for interventions.

4 Evaluation Framework

The scale and heterogeneity of the dataset allow us to define a multi-faceted evaluation framework that goes beyond conventional pixel accuracy or retrieval precision vision benchmarks. Instead we design tasks built on conventions in the change detection and vision-language literature, tailored to unique challenges of urban streets.

1. Change Detection and Identification. We extend standard change detection benchmarks by evaluating how models can both identify *if* a redesign occurred and *what* features changed. Given paired before/after Images, Models must output a binary signal of change and a list of features that have changed among a set of features directly from the City's project dataset (see appendix).

2. Image and Document Summarization. A distinctive aspect of our benchmark is the integration of unstructured civic design documents (Docs). To test whether models can translate between Images and Docs, we introduce a captioning-style benchmark. Given image pairs, models generate a short description of the change given the *Feature* vocabulary. Models are given a similar task for a linked Document using same keywords. The outputs are embedded with CLIP and evaluated with RefCLIP Score, Cosine Similarity and human expert judgment. This task connects visual change detection to Engineering language suggesting future improvements through Document-annotated RAG.

3. Temporal Localization. Finally, we evaluate models' abilities to reason across longitudinal sequences. Given a set of chronological images, models must identify the year in which a change first appears. This task is scored by exact-year accuracy and mean absolute deviation in years, providing a measure of temporal reasoning in long-term infrastructure evolution.

4. Semantic Signal Augmentation. A key contribution of our framework is to move beyond raw imagery to novelly include segmentation masks as structured inputs for change analysis. Each location's Images are segmented into sidewalks, crosswalks, and roads using *Tile2Net* [3], which provides a stable foundation for longitudinal difference identification. We then rerun the tests from 1, 2 & 3 augmented by providing the model with the mask for each image and relevant prompting.

5 Results

We evaluate general-purpose vision–language models (*GPT-4o*, *GPT-5*) across our proposed tasks once with raw imagery and again augmented with segmentation masks. Across four tasks, we observe a consistent pattern: models can recognize that change has occurred, but struggle with more specific feature identification and reasoning tasks. They perform best in unstructured captioning tests, yet still, their outputs are inconsistent, generic, or confounded by noise. Models seem to perform worse on tasks involving more input files, and *GPT-5* did not outperform *GPT-4o* on this suite of evaluations. At the same time, incorporating segmentation-derived masks only moderately improves

robustness. However, a purely segmentation-based approach proves most effective underscoring Semantic Signal Extraction as a core contribution of our framework.

Table 1: Compact benchmark across tasks. Metrics differ by task: MCC for classification, MAD (yrs) for dating, RefCLIP for summarization, mIoU for segmentation. All values except MAD are fractions in [0,1].

Task	**Accuracy**	**Macro-F1**	**Task-specific score (Δ)**
Identifier (Image-only, GPT-4o)	0.054	0.002	MCC = 0.026
Identifier (+Mask, GPT-4o)	0.072	0.002	MCC = 0.037 (+0.011)
Classifier (Image-only, GPT-4o)	0.728	0.579	MCC = 0.182
Classifier (+Mask, GPT-4o)	0.737	0.614	MCC = 0.236 (+0.054)
Dater (Image-only, GPT-4o)	0.034	—	MAD = 3.85 yrs
Dater (+Mask, GPT-4o)	0.036	—	MAD = 3.57 yrs
Summarizer (Image-only, GPT-4o)	—	—	RefCLIP = 0.42
Summarizer (+Mask, GPT-4o)	—	—	RefCLIP = 0.47 (+0.05)
Segmentation (Tile2Net)	—	—	mIoU = 0.921, frac = 0.037

Table 1 summarizes results across all benchmark tasks, with each family reported in its most appropriate metric. Identifier and Classifier rows show Accuracy, Macro-F1, and MCC, where the Δ column highlights the effect of segmentation masks. Dater is reported as Accuracy alongside mean absolute deviation (MAD) in years, capturing temporal localization error. Summarizer rows use RefCLIP similarity to quantify alignment between generated captions and the controlled feature vocabulary. Segmentation is presented as mIoU, a weighted score dominated by large-area classes and the median fraction of pixels changed between years. All values except MAD are fractions in $[0, 1]$, consistent with convention.

Across a citywide corpus (≈47k intersections; ≈470k aerial crops), segmentation masks act as strong priors. In binary change detection (Classifier, N=923), GPT-4o gains Macro-F1 0.579→0.614 and MCC 0.182→0.236 (+0.054, ≈30% relative), with GPT-5 showing a similar lift (+0.040, ≈23%). The harder multi-class attribution (Identifier, 40–63 labels; $N \in \{864, 906, 99\}$) remains near chance, yet masks still raise Accuracy 0.054→0.072 and MCC 0.026→0.037 (+0.011). For temporal localization (Dater), masks reduce MAD from 3.85→3.57 years (-0.28 yr) on N=348/276 positives, sharpening dating even when exact hits are rare. Tile2Net itself reaches mIoU=0.921 with only 3.7% of pixels changing, still providing a stable and low-noise supervisory signal for detecting localized interventions. Together these results show structured cues improve robustness in detection and grounding, but fine-grained attribution and temporal reasoning remain brittle, highlighting the need for ontology-based retrieval and targeted fine-tuning to enable reliable reasoning about urban change.

6 Discussion and Conclusion

Our initial benchmarks reveal significant limitations in applying off-the-shelf LLMs to understand streetscape interventions. Models are overly optimistic and are confounded by irrelevant details, echoing findings in other domains where they struggle with specific, domain-grounded reasoning. In our setting, this brittleness underscores the need for domain-augmented improvement. Looking ahead, we see improved annotation, fine-tuning, and ontology-driven approaches improving model performance by helping to represent interventions within the vocabulary of those responsible for them. Additionally, retrieval-augmented generation (RAG) offers a promising complement, anchoring model reasoning in an authoritative local evidentiary context; planimetric change masks, project metadata, or documents. Moving forward, our goal is to grow **StreetTransformer** into a tool that translates between the temporospatially-linked modalities to build a structured model of streetscape evolution. This would allow researchers and practitioners to retrieve precedents across contexts, evaluate interventions, and design safer, more inclusive intersections. Early results show that general-purpose models struggle without augmentation, but the path forward; through fine-tuning, ontology-based reasoning, and RAG-enhanced grounding; offers an opportunity for ML to help improve cities through design.

References

[1] Chenyi Cai, Kosuke Kuriyama, Youlong Gu, Filip Biljecki, and Pieter Herthogs. Can a large language model assess urban design quality? evaluating walkability metrics across expertise levels. *arXiv preprint arXiv:2504.21040*, 2025. URL https://arxiv.org/abs/2504.21040.

[2] Z. Chen et al. Sarlang-1m: A large-scale vision-language dataset for synthetic aperture radar. *arXiv preprint arXiv:2504.03254*, 2025. URL https://arxiv.org/abs/2504.03254.

[3] Maryam Hosseini, Andres Sevtsuk, Fabio Miranda, Roberto M Cesar Jr, and Claudio T Silva. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101:101950, 2023.

[4] Youngok Kang, Jiyeon Kim, Jiyoung Park, and Jiyoon Lee. Assessment of perceived and physical walkability using street view images and deep learning technology. *ISPRS International Journal of Geo-Information*, 12(5):186, 2023.

[5] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*, 2024. URL https://arxiv.org/abs/2406.12384.

[6] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022.

[7] M. Malekzadeh et al. Urban visual appeal according to chatgpt: Contrasting ai and human perceptions. *arXiv preprint arXiv:2407.14268*, 2024. URL https://arxiv.org/abs/2407.14268.

[8] NYC OTI. Nyc geoclient api. https://maps.nyc.gov/geoclient/v2/doc, 2025. Accessed: 2025-08-29.

[9] Qingling Shu, Sibao Chen, Zhihui You, Wei Lu, Jin Tang, and Bin Luo. Rb-scd: A new benchmark for semantic change detection of roads and bridges in traffic scenes. *arXiv preprint arXiv:2505.13212*, 2025.

[10] Urban Institute. Highway and road expenditures. https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/state-and-local-backgrounders/highway-and-road-expenditures, 2024. ACcessed: 2025-11-23.

[11] Yuduo Wang, Weikang Yu, and Pedram Ghamisi. Change captioning in remote sensing: Evolution to sat-cap–a single-stage transformer approach. *arXiv preprint arXiv:2501.08114*, 2025.

[12] Chenhui Zhang and Sherrie Wang. Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7839–7849, 2024.

[13] Yongshuo Zhu, Lu Li, Keyan Chen, Chenyang Liu, Fugen Zhou, and Zhenwei Shi. Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.