From Street Views to Urban Science: Discovering Road Safety Factors with Multimodal Large Language Models

Yihong Tang & Lijun Sun*
McGill University
yihong.tang@mail.mcgill.ca, lijun.sun@mcgill.ca

Ao Qu & Jinhua Zhao
Massachusetts Institute of Technology
{qua,jinhua}@mit.edu

Xujing Yu & Weipeng Deng & Jun Ma
The University of Hong Kong
xujingyu@connect.hku.hk, wpdeng@connect.hku.hk, junma@hku.hk

Abstract

Urban and transportation research has long sought to uncover statistically meaningful relationships between key variables and societal outcomes such as road safety, to generate actionable insights that guide the planning, development, and renewal of urban and transportation systems. However, traditional workflows face several key challenges: (1) reliance on human experts to propose hypotheses, which is time-consuming and prone to confirmation bias; (2) limited interpretability, particularly in deep learning approaches; and (3) underutilization of unstructured data that can encode critical urban context. Given these limitations, we propose a Multimodal Large Language Model (MLLM)-based approach for interpretable hypothesis inference, enabling the automated generation, evaluation, and refinement of hypotheses concerning urban context and road safety outcomes. Our method leverages MLLMs to craft safety-relevant questions for street view images (SVIs), extract interpretable embeddings from their responses, and apply them in regression-based statistical models. URBANX supports iterative hypothesis testing and refinement, guided by statistical evidence such as coefficient significance, thereby enabling rigorous scientific discovery of previously overlooked correlations between urban design and safety. Experimental evaluations on Manhattan street segments demonstrate that our approach outperforms pretrained deep learning models while offering full interpretability. Beyond road safety, URBANX can serve as a general-purpose framework for urban scientific discovery, extracting structured insights from unstructured urban data across diverse socioeconomic and environmental outcomes. This approach enhances model trustworthiness for policy applications and establishes a scalable, statistically grounded pathway for interpretable knowledge discovery in urban and transportation studies.

1 Introduction

Understanding how cities' physical structures shape societal outcomes is central to urban science. Across transportation, planning, and policy, researchers seek links between urban form and social indicators such as traffic safety (Yu et al., 2024), walkability (Ewing & Handy, 2009), equity (Guzman & Bocarejo, 2017), and environmental health (Majchrowska et al., 2022). A key challenge is discovering generalizable, interpretable factors that explain urban phenomena (Batty, 2024). However, urban form is complex and heterogeneous, with relevant information often stored in unstructured formats such as street-level imagery and visual cues (Biljecki & Ito, 2021), which are difficult to analyze using traditional methods.

*Corresponding author. Project: https://github.com/YihongT/UrbanX.git

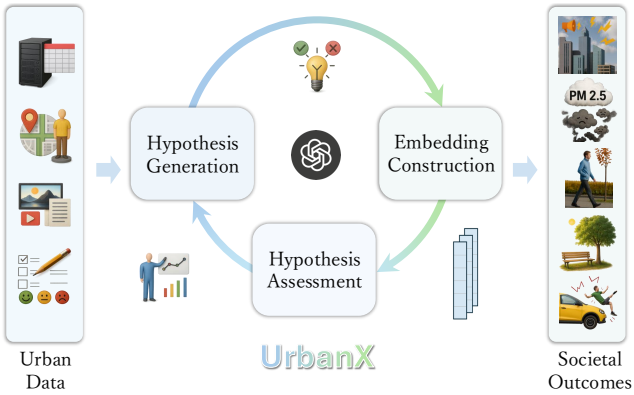

Figure 1: Real vs. synthetic mobility patterns.

Despite progress in urban analytics, identifying new interpretable factors remains challenging. Existing methods often rely on expert-defined features, black-box models, or handcrafted metrics (Xia et al., 2025), each with limitations. First, hypothesis generation is manual and prone to cognitive bias (Gettys & Fisher, 1979). Second, deep models usually lack interpretability. Third, unstructured data like SVIs are underused due to difficulties in extracting meaningful structure (Tang et al., 2025). These limitations hinder scalable and transparent urban analysis. Foundation models such as Large Language Models (LLMs) (Naveed et al., 2023) have transformed data-driven reasoning. Trained on large text corpora, LLMs perform flexible and context-aware inference (Wei et al., 2022). Recent extensions to visual inputs have produced Multimodal LLMs (MLLMs) (Wu et al., 2023), which jointly process images and text for tasks like scene interpretation and visual reasoning. MLLMs align visual content with language, allowing them to generate human-interpretable variables from raw imagery.

Given these, we present URBANX, a framework for hypothesis-driven urban discovery powered by MLLMs. As illustrated in Figure 1, URBANX treats machine learning as a collaborator in scientific inquiry. It iteratively generates hypotheses, derives variables from multimodal data, and evaluates their statistical relevance. Weak hypotheses are discarded, and new ones are proposed, gradually refining a set of interpretable, empirically supported factors. We apply URBANX to urban road safety, where interpretability is crucial. In a Manhattan case study, the framework uncovers novel visual variables from SVIs that correlate with crash rates. Our approach surpasses deep learning baselines such as ResNet and Vision Transformer, while maintaining transparency. Our contributions are:

- We frame scientific discovery in urban contexts as inference over a hypothesis space, enabling machines to generate, test, and refine hypotheses using available data.
- We propose using MLLMs as semantic engines that transform unstructured inputs, such as SVIs, into interpretable variables based on natural-language hypotheses.
- We design an interpretable, nonparametric, iterative framework that approximates the posterior over hypotheses, enabling scalable and statistically grounded discovery of novel urban factors.
- We demonstrate the effectiveness of our framework on road safety in Manhattan, where it discovers visual predictors of crash rates that outperform vision baselines while offering interpretable results. The framework generalizes to other domains in urban science.

2 Related Work

Understanding how urban form influences outcomes like public health, equity, and road safety is central to urban science and transportation research (Hall, 2012). Traditional approaches rely on statistical models that relate expert-defined variables to outcomes (Santamouris, 2013). In road safety, for example, street design, traffic calming, and pedestrian infrastructure have been linked to crash rates (Ewing & Dumbaugh, 2009). These methods, however, face several limitations. Hypothesis generation is often manual and based on intuition, making it slow, biased, and narrow in scope (Xia et al., 2025). This can restrict the discovery of less obvious relationships. While deep learning models can enhance prediction, they usually lack interpretability (Goodfellow et al., 2016), making it difficult to identify causal drivers or support policy decisions. Their opacity can reduce trust, particularly in critical applications (Benara et al., 2024). Another challenge lies in the underused unstructured data. SVIs contain rich visual details about the urban environment, such as infrastructure

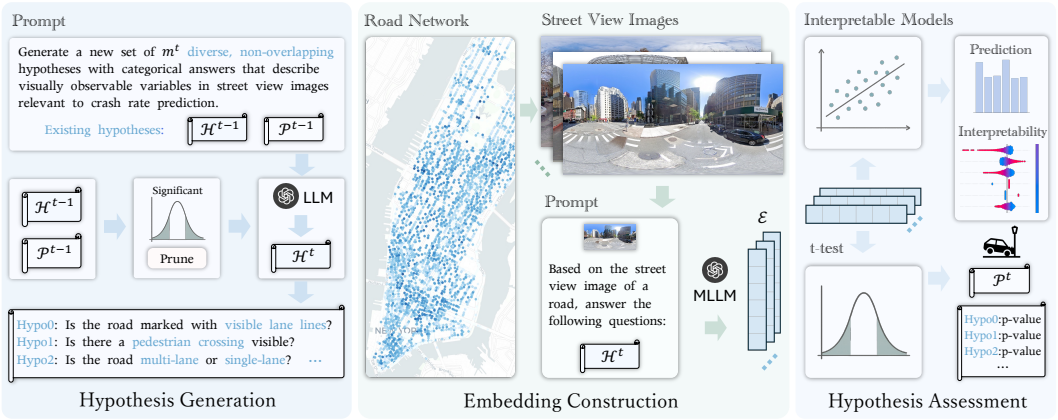

Figure 2: The URBANX framework consists of three iterative modules: (1) Hypothesis Generation using LLMs, (2) Embedding Construction via MLLM-based VQA on SVIs, and (3) Hypothesis Assessment using interpretable regression analysis.

condition and safety cues (Biljecki & Ito, 2021). Yet, their use in quantitative research is limited by issues of image consistency, variability, and the difficulty of extracting structured variables (Tang et al., 2024). Current methods often use general computer vision models that require extensive tuning and still may miss subtle, context-specific signals. Recent advances in AI-driven discovery offer promising directions. Some work uses LLMs for causal inference in urban settings (Xia et al., 2025), or combines LLMs with knowledge graphs for hypothesis generation in other domains (Lopez et al., 2025). These efforts show AI's potential to support scientific reasoning, but a transparent framework that generates and tests hypotheses directly from SVIs remains largely unexplored. Our work addresses this gap by using MLLMs to link urban visuals to road safety outcomes.

3 Methodology

Overview Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote a dataset of n SVIs x_i and their associated road-level crash rates $y_i \in \mathbb{R}$. We define a hypothesis space $\mathcal{H}$ comprising all natural-language queries that describe visually observable variables potentially related to road safety. Our objective is to uncover an optimal subset of hypotheses $\mathcal{H}^* = \{h_1, h_2, \ldots, h_k\} \subset \mathcal{H}$ that captures meaningful visual semantics from each SVI and enables interpretable, accurate prediction of y_i. We formalize this as a posterior mode estimation problem over the hypothesis space: $\mathcal{H}^* = \arg\max_{\mathcal{H}' \subseteq \mathcal{H}} P(\mathcal{H}' \mid \mathcal{D}) \propto P(\mathcal{D} \mid \mathcal{H}') \cdot P(\mathcal{H}')$, where $\mathcal{H}'$ is a candidate hypothesis subset. The likelihood $P(\mathcal{D} \mid \mathcal{H}')$ captures how well the hypothesis-derived variables explain variation in crash rates, typically assessed via a regression model. The prior $P(\mathcal{H}')$ encodes structural preferences over hypothesis subsets and is implicitly governed by the generative behavior of the MLLM. Each hypothesis $h_j \in \mathcal{H}^*$ corresponds to a semantically meaningful question with a categorical answer that could be inferred from an SVI using an MLLM. Applying these k hypotheses to each image x_i yields a k-dimensional interpretable embedding $\phi(x_i, \mathcal{H}^*) \in \mathbb{R}^k$, where each component reflects the MLLM's answer to the corresponding hypothesis. We denote the complete embedding matrix as $\mathcal{E} \in \mathbb{R}^{n \times k}$, where $e_i = \phi(x_i, \mathcal{H}^*)$ is the embedding vector for the i-th image. Since exact Bayesian inference over all subsets of $\mathcal{H}$ is intractable due to the large search space and unknown likelihoods, we use an approximate approach and cast the problem as nonparametric structure learning. Starting with an initial set $\mathcal{H}^0$ sampled from an LLM, we iteratively evaluate and refine it using a linear regression model. For each hypothesis h_j, we test the significance of its regression coefficient via a two-sided t-test under the null hypothesis that the coefficient is zero. This gives a p-value vector $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$, where each p_j represents the probability of observing the result under the null. Hypotheses with $p_j > \alpha$ (typically 0.05) are treated as statistically insignificant and removed. New hypotheses are generated to replace them, forming an iterative refinement loop. This procedure approximates Bayesian inference over

$\mathcal{H}$ using statistical evidence and LLM priors, in a nonparametric and data-driven way. An overview of the URBANX framework is shown in Figure 2.

Hypothesis Generation At each iteration t, the framework refines the hypothesis set $\mathcal{H}^{t-1}$ using statistical evidence derived from the previous assessment. For each hypothesis $h_j \in \mathcal{H}^{t-1}$, we compute a p-value p_j using a two-sided t-test on the coefficient estimated by a regression model, where the input variable is derived from the MLLM-inferred categorical responses to h_j across all SVIs. The detailed procedure for constructing hypothesis-driven embeddings is described in a later subsection. Hypotheses with $p_j > \alpha$ (typically $\alpha = 0.05$) are considered statistically insignificant. While the prompt for the LLM includes the full set of previous hypotheses $\mathcal{H}^{t-1}$ and their p-values $\mathcal{P}^{t-1}$, only m^t new hypotheses are generated, where m^t equals the number of pruned hypotheses. This maintains a fixed hypothesis set size while ensuring that each iteration incorporates empirical feedback into the generative process. Formally, the hypothesis generation step is given by: $\mathcal{H}^t \sim \text{LLM}\left(\text{Prompt}_{HypoGen}(\mathcal{H}^{t-1}, \mathcal{P}^{t-1}, m^t)\right)$, where m^t is the number of new hypotheses to generate. The prompt is constructed to elicit m^t diverse, categorical, and visually inferable questions that are relevant to crash prediction. By conditioning on statistically grounded examples, the LLM acts as a posterior-informed generator, implicitly sampling from a distribution biased toward hypotheses that are both semantically coherent and empirically promising. This design allows the system to balance exploration of new concepts with exploitation of previously validated structure, enabling effective refinement of the hypothesis space over time. An illustration of this process is shown in the left panel of Figure 2.

Embedding Construction To leverage the generated hypotheses $\mathcal{H}^t = \{h_1^t, h_2^t, \ldots, h_k^t\}$ for downstream modeling, we must transform their semantic content into structured, machine-interpretable representations. Traditional deep models rely on latent high-dimensional features extracted from images, which lack transparency and hinder hypothesis-driven analysis. In contrast, our goal is to construct a hypothesis-guided embedding that is transparently aligned with the semantics of each generated question. For each image x_i, we use an MLLM to answer all questions in $\mathcal{H}^t$ based on the visual content of the image. These categorical answers are then encoded into a k-dimensional vector $e_i^t \in \mathbb{R}^k$, where each element corresponds to the response to hypothesis h_j^t. Formally, we define: $e_i^t \sim \text{MLLM}\left(x_i, \text{Prompt}_{Embed}(\mathcal{H}^t)\right)$, where $\text{Prompt}_{Embed}(\mathcal{H}^t)$ denotes the prompt that queries the MLLM to answer all hypotheses in $\mathcal{H}^t$ based on the visual content of x_i. This embedding ensures full semantic traceability and supports interpretable downstream modeling.

This procedure yields a hypothesis-aligned, semantically interpretable embedding for each image, where each dimension has a well-defined linguistic meaning. It enables transparent variable construction while supporting statistical assessment and iterative refinement in subsequent stages. The embedding process is illustrated in the center panel of Figure 2.

Hypothesis Assessment After constructing semantically aligned embeddings $\mathcal{E}^t$ for all SVIs based on the current hypothesis set $\mathcal{H}^t$, the next step is to assess which hypotheses meaningfully explain variation in crash rates. Rather than focusing solely on predictive accuracy, we adopt interpretable models that provide transparent, decomposable attribution of outcomes to individual hypotheses. This is particularly important in societal domains such as road safety, where policy decisions and public accountability require not only reliable predictions but also actionable explanations. Each input dimension in the model corresponds to a specific hypothesis, enabling us to quantify its effect and assess statistical significance. In our implementation, we use linear regression as the default method due to its analytical tractability and well-established inference procedures.

For each embedding dimension, we perform a two-sided t-test on its regression coefficient to test the null hypothesis that the coefficient is zero. This yields a p-value for each hypothesis h_j^t, denoted p_j^t, indicating the likelihood that the observed effect is due to chance. The resulting vector $\mathcal{P}^t = \{p_1^t, p_2^t, \ldots, p_k^t\}$ serves as the statistical evidence that guides the next

iteration: hypotheses with $p_j^t > \alpha$ (typically $\alpha = 0.05$) are deemed statistically insignificant and are replaced by new hypotheses in the next round. This assessment mechanism plays a dual role: it enables interpretability by quantifying the contribution of each hypothesis, and it drives hypothesis refinement by filtering out those that lack explanatory value. The right panel of Figure 2 illustrates this evaluation process.

Algorithm 1 Iterative Posterior Approximation

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$; number of total hypotheses k; number of iterations T; interpretable model $\mathcal{M}$

Ensure: Final hypothesis set $\mathcal{H}^T$ and embedding matrix $\mathcal{E}^T$

1: Initialize $\mathcal{H}^0 \sim \text{LLM}(\text{Prompt}_{HypoGen}(k))$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad \mathcal{H}^t \sim \text{LLM}(\text{Prompt}_{HypoGen}(\mathcal{H}^{t-1}, \mathcal{P}^{t-1}, m^t))$ ▷ `Hypothesis Generation`
4: $\quad \mathcal{E}^t = \{e_i^t = \text{MLLM}(x_i, \text{Prompt}_{Embed}(\mathcal{H}^t))\}_{i=1}^n$ ▷ `Embedding Construction`
5: $\quad \{\widehat{y}_i\}_{i=1}^n, \mathcal{P}^t \leftarrow \mathcal{M}(\mathcal{E}^t)$ ▷ `Hypothesis Assessment`
6: **end for**

Iterative Posterior Approximation The overall framework is executed through an iterative loop that approximates posterior inference over the hypothesis space by alternating between generation, embedding, and statistical assessment. This process reflects a structure-learning approach where hypothesis subsets are progressively refined based on empirical evidence. Unlike standard optimization methods such as gradient descent or expectation maximization, where the objective function is guaranteed to monotonically improve, our setting involves sampling from a nonparametric, LLM-driven space that lacks such guarantees. To mitigate the risk of degeneracy or performance collapse, we adopt a conservative update rule: new hypotheses $\mathcal{H}^t$ are only retained if they yield improved predictive performance on the validation set compared to the previous iteration.

Algorithm 1 outlines the overall iterative procedure. In each iteration, insignificant hypotheses from the previous set $\mathcal{H}^{t-1}$ are filtered based on their p-values $\mathcal{P}^{t-1}$. The remaining hypotheses serve as context for LLM-based generation of new candidates. The resulting hypothesis set $\mathcal{H}^t$ is then used to construct interpretable embeddings $\mathcal{E}^t$ via MLLM-based reasoning, which are subsequently used to train an interpretable model and evaluate statistical significance. This iterative process continues for a predefined number of iterations.

4 Experiments

Settings To evaluate the effectiveness of URBANX in supporting interpretable and data-driven discovery of urban road safety factors, we conduct experiments on road segments within Manhattan, New York City. This area provides a complex urban setting characterized by high traffic density, varied land use, and extensive open data resources. Our objective is to predict and explain segment-level crash risk based solely on visual inputs from street-view SVI, without the use of predefined variables or manual feature annotation.

Crash risk is quantified using a standardized crash rate Hou et al. (2020); Zeng et al. (2017); Yu et al. (2024), defined for each segment as:

$$CR_i = \frac{\text{No\_crash}_i}{AADT_i \times L_i \times \frac{365}{1{,}000{,}000}}, \tag{1}$$

where No_crash_i denotes the annual average number of reported crashes, $AADT_i$ represents the average annual daily traffic volume, and L_i is the length of the road segment in kilometers. This formulation adjusts for traffic exposure and segment size, and is widely adopted in transportation safety research to enable fair comparisons across road types and traffic conditions.

Crash records were sourced from NYC Open Data[1], and traffic volume data ($AADT$) was obtained from the New York State Department of Transportation[2]. Street-view imagery was collected using ArcGIS Wong & Lee (2005), with panoramic images sampled every 15 meters along road centerlines and retrieved via the Google Street View API[3]. After preprocessing and filtering, the dataset includes 16,000 images for training, 2,000 for validation, and 2,000 for testing.

For hypothesis generation, we use GPT-4o Hurst et al. (2024)[4]. To construct visual embeddings based on structured prompts, we use InternVL2.5-78B Chen et al. (2024)[5]. All multimodal models are deployed using LMDeploy Contributors (2023), which provides an efficient and reproducible framework for serving large-scale MLLMs.

To provide a comparative reference for evaluating our approach, we also compile a comprehensive set of 58 built environment features, drawn from five domains: (1) road design (e.g., width, highway indicator), (2) land use composition and entropy, (3) point-of-interest (POI) features, (4) traffic-related facilities (e.g., crossings, bus stops), and (5) visual indices derived from panoptic segmentation (e.g., vegetation, building, road proportions). These variables serve as a baseline representation of conventional urban form and are used in a post hoc analysis to assess the added value of the discovered hypotheses.

In our primary modeling pipeline, however, these handcrafted features are not included. Instead, we rely exclusively on hypotheses generated by the language model and their corresponding embeddings inferred by the MLLMs. This design ensures that predictive insights emerge solely from automatically discovered, visually interpretable patterns, allowing us to assess the capacity of URBANX to support transparent, scalable, and data-driven scientific discovery grounded in the visual environment.

Predictive Performance We first evaluate the predictive performance of our interpretable embedding framework by comparing it with conventional vision-based baselines. Figure 3 reports results across three standard metrics: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). For baselines, we use two representative pretrained image encoders: ResNet50, a widely adopted convolutional architecture, and ViT-Base (ViT-B/16), specifically the `vit_base_patch16_224` variant that segments each image into 16×16 patches and processes them with transformer blocks. These models are fine-tuned to predict crash rates directly from raw SVIs. We compare these against two variants of our framework that rely on interpretable embeddings constructed from MLLM responses: one using linear regression (LR) and another using LightGBM (LGBM) as the downstream predictor. Across all metrics, our method consistently outperforms the deep learning baselines while maintaining transparency and semantic interpretability. The LightGBM variant achieves the strongest overall results. These results demonstrate that the embeddings retain sufficient information to make accurate predictions while also enabling interpretability.

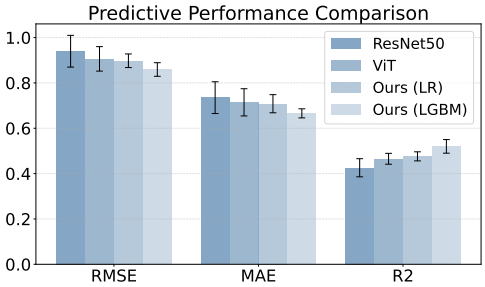

Figure 3: Performance comparison between ResNet, ViT, and our interpretable embedding-based models using linear regression (LR) and LightGBM (LGBM).

Discovered Factors A core objective of URBANX is to move beyond predictive accuracy and toward interpretable, data-driven discovery of visual factors that are often missing from conventional urban analytics. To evaluate the substantive relevance of these learned variables, we apply SHAP (SHapley Additive exPlanations) Lundberg & Lee (2017) to a

[1] https://opendata.cityofnewyork.us/
[2] https://dos.ny.gov/location/new-york-state-department-transportation
[3] https://developers.google.com/maps/documentation/streetview/overview
[4] https://platform.openai.com/docs/models
[5] https://huggingface.co/OpenGVLab/InternVL2_5-78B

regression model trained on both traditional built environment features and hypothesis-derived embeddings. This allows us to quantify the marginal contribution of each variable to the prediction of segment-level crash rates.

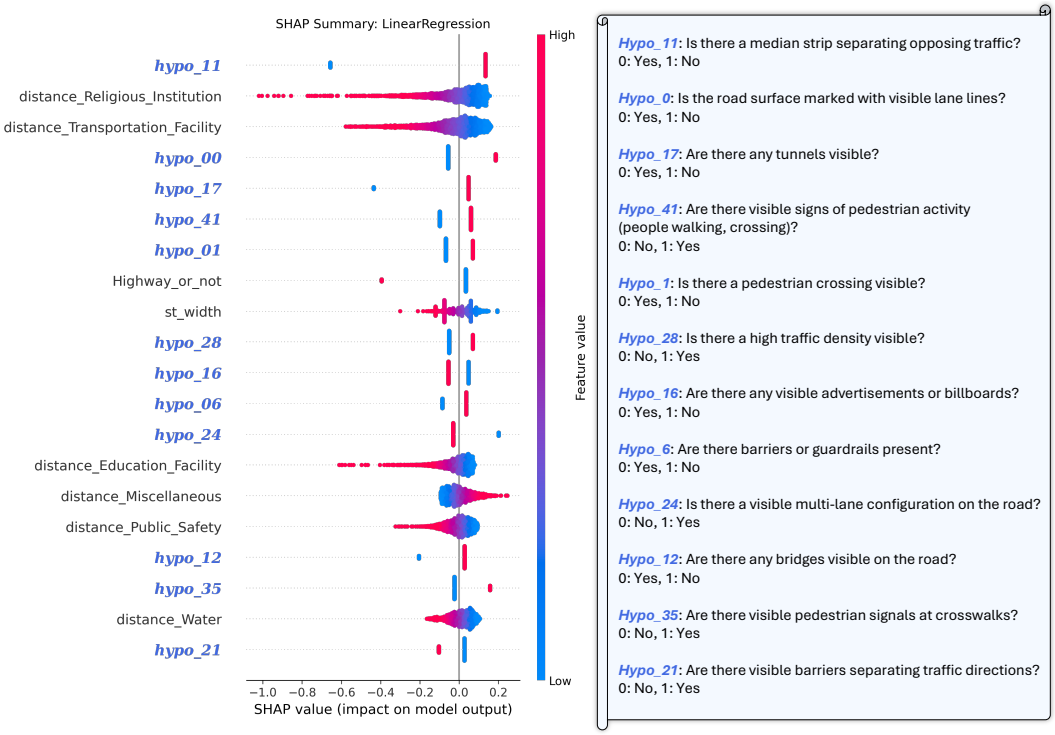

Figure 4: SHAP summary plot of the regression model with both traditional built environment variables and discovered hypotheses. The right panel maps the top hypothesis variables to their natural-language question meanings.

Figure 4 presents a top-20 ranked summary of both traditional (existing) built environment variables and the automatically discovered hypotheses. Remarkably, a majority of the top-ranked variables by explanatory power are generated by our LLM-based hypothesis pipeline. This highlights URBANX 's capacity to uncover impactful, interpretable factors that are not present in standard urban datasets, supporting its role as a scientific discovery tool rather than a black-box predictor. Many of the discovered hypotheses align with well-established urban safety principles, validating the model's ability to recover known but unstated domain knowledge. For example, `Hypo_11` ("Is there a median strip separating opposing traffic?") and `Hypo_0` ("Is the road surface marked with visible lane lines?") are both highly ranked and show negative SHAP contributions when absent, suggesting their presence is associated with lower crash risk. These align with conventional traffic engineering wisdom on lane separation and visual guidance.

At the same time, URBANX also surfaces more nuanced or less commonly considered factors. Several high-ranking hypotheses relate to pedestrian visibility and activity, such as `Hypo_1` (pedestrian crossing), `Hypo_41` (pedestrian presence), and `Hypo_35` (pedestrian signals). These factors may have complex and context-sensitive relationships with safety outcomes, underscoring the value of semantically grounded, hypothesis-level variables. In addition, URBANX identifies less conventional features that might escape manual enumeration. For instance, `Hypo_16` ("Are there any advertisements or billboards?") and `Hypo_6` ("Are there barriers or guardrails present?") point to visual distractions and physical protection measures that may subtly influence crash risk. These hypotheses extend the scope of interpretable modeling into environmental and perceptual dimensions that are often hard to encode using conventional GIS-based variables. Compared to traditional indicators such as street width or proximity to facilities, our hypotheses are more granular, semantically aligned,

and directly grounded in what is observable in urban space. This illustrates the advantage of URBANX in supporting structured discovery over unstructured inputs. Taken together, these results show that URBANX is capable of achieving competitive predictive accuracy and surfacing novel, interpretable factors that enrich the understanding of urban safety.

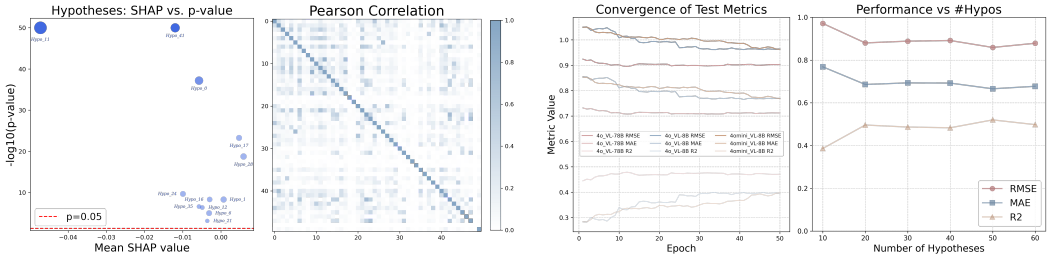

Figure 5: (Left) Variable quality analysis: SHAP vs. statistical significance and pairwise correlation. (Right) Robustness analysis: model performance across different settings.

Variable Quality, Robustness, and Practical Implications We assess the quality and robustness of URBANX's hypotheses based on significance, independence, and sensitivity to model and hypothesis settings. As shown in the left panel of Figure 5, we visualize each hypothesis by its average SHAP value and the negative log of its p-value from linear regression. This allows us to assess both predictive contribution and statistical significance. Hypotheses such as `Hypo_11`, `Hypo_41`, and `Hypo_0` score highly on both axes. These correspond to interpretable road safety features such as the presence of a median strip, pedestrian visibility, and clear lane markings. These results confirm the alignment between model-generated variables and established domain knowledge. The same panel also includes the pairwise Pearson correlation matrix, where low off-diagonal values indicate that the discovered variables are largely uncorrelated. This structural independence enhances interpretability and reduces the risk of multicollinearity.

The right panel of Figure 5 presents a robustness analysis. We vary the capacities of the LLM used for hypothesis generation and the MLLM used for answering those hypotheses. Larger MLLMs (such as InternVL2.5-78B) consistently produce better predictive accuracy and faster convergence than smaller ones (such as 8B), emphasizing the importance of visual reasoning capability. The size of the LLM has a smaller effect. GPT-4o leads to faster training convergence compared to GPT-4o-mini, but both eventually reach similar levels of performance. We also examine the impact of hypothesis count. Accuracy improves steadily up to around 50 hypotheses, after which performance plateaus or slightly declines. This reflects a tradeoff between representational richness and statistical noise.

5 Conclusion

In this paper, we presented URBANX, a framework that combines MLLMs with interpretable statistical modeling to automate scientific discovery from urban data. Taking road safety in the Manhattan area as a case study, URBANX formulates natural-language hypotheses, extracts semantically meaningful embeddings through visual question answering, and evaluates their significance using transparent regression models. Our experiments show that URBANX outperforms conventional deep learning approaches while uncovering novel, interpretable variables aligned with domain knowledge.

This work demonstrates a new paradigm for scientific discovery in urban research, one that integrates perception, language, and statistical reasoning in a unified pipeline. The generality of URBANX enables broad applicability to other domains such as walkability, equity, and environmental quality, where unstructured data possesses rich information and model interpretability are central. Future work may extend this approach to dynamic data, integrate causal inference, and benefit from ongoing advances in the alignment and efficiency of foundation models. By rethinking machine learning as a tool for interpretable, data-driven reasoning, URBANX offers a scalable foundation for MLLM hypothesis-driven urban science and beyond.

References

Michael Batty. *The computable city: histories, technologies, stories, predictions*. MIT Press, 2024.

Vinamra Benara, Chandan Singh, John X Morris, Richard J Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking llms questions. *Advances in neural information processing systems*, 37: 124137, 2024.

Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning*, 215:104217, 2021.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023.

Reid Ewing and Eric Dumbaugh. The built environment and traffic safety: a review of empirical evidence. *Journal of Planning Literature*, 23(4):347–367, 2009.

Reid Ewing and Susan Handy. Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban design*, 14(1):65–84, 2009.

Charles F Gettys and Stanley D Fisher. Hypothesis plausibility and hypothesis generation. *Organizational behavior and human performance*, 24(1):93–110, 1979.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Luis A Guzman and Juan P Bocarejo. Urban form and spatial urban equity in bogota, colombia. *Transportation research procedia*, 25:4491–4506, 2017.

Randolph Hall. *Handbook of transportation science*, volume 23. Springer Science & Business Media, 2012.

Qinzhong Hou, Xiaoyan Huo, and Junqiang Leng. A correlated random parameters tobit model to analyze the safety effects and temporal instability of factors affecting crash rates. *Accident Analysis & Prevention*, 134:105326, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Vanessa Lopez, Lam Hoang, Marcos Martinez-Galindo, Raúl Fernández-Díaz, Marco Luca Sbodio, Rodrigo Ordonez-Hurtado, Mykhaylo Zayats, Natasha Mulligan, and Joao Bettencourt-Silva. Enhancing foundation models for scientific discovery via multimodal knowledge graph representations. *Journal of Web Semantics*, 84:100845, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Sylwia Majchrowska, Agnieszka Mikołajczyk, Maria Ferlin, Zuzanna Klawikowska, Marta A Plantykow, Arkadiusz Kwasigroch, and Karol Majek. Deep learning-based waste detection in natural and urban environments. *Waste Management*, 138:274–284, 2022.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Matheos Santamouris. *Energy and climate in the urban built environment*. Routledge, 2013.

Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, et al. Itinera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1413–1432, 2024.

Yihong Tang, Menglin Kong, and Lijun Sun. Large language models for data synthesis. *arXiv preprint arXiv:2505.14752*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

WSD Wong and Jay Lee. *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. Wiley, 2005.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

Yutong Xia, Ao Qu, Yunhan Zheng, Yihong Tang, Dingyi Zhuang, Yuxuan Liang, Shenhao Wang, Cathy Wu, Lijun Sun, Roger Zimmermann, and Jinhua Zhao. Reimagining urban science: Scaling causal inference with large language models. *arXiv preprint arXiv:2504.12345*, 2025.

Xujing Yu, Jun Ma, Yihong Tang, Tianren Yang, and Feifeng Jiang. Can we trust our eyes? interpreting the misperception of road safety from street view images and deep learning. *Accident Analysis & Prevention*, 197:107455, 2024.

Qiang Zeng, Huiying Wen, Helai Huang, Xin Pei, and SC Wong. A multivariate random-parameters tobit model for analyzing highway crash rates by injury severity. *Accident Analysis & Prevention*, 99:184–191, 2017.