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

Yihong Tang Ao Qu Xujing Yu Weipeng Deng Jun Ma Jinhua Zhao Lijun Sun ✉
McGill University Massachusetts Institute of Technology The University of Hong Kong
yihong.tang@mail.mcgill.ca qua@mit.edu lijun.sun@mcgill.ca

## 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.

## KEYWORDS

Urban Science, Road Safety, Street View Imagery, Multimodal Large Language Models, Hypothesis Inference, Scientific Discovery

## 1 INTRODUCTION

Understanding how the physical structure of cities shapes societal outcomes is a foundational objective in urban science. Across fields such as transportation, planning, and public policy, researchers have long aimed to identify statistically meaningful links between urban environments and key social indicators, including traffic safety [46], walkability [11], equity [15], and environmental health [27]. A central challenge in this endeavor is the pursuit of scientific discovery: finding interpretable and generalizable factors that explain urban phenomena and support data-driven decision-making [4]. Yet, this process is often hindered by the complexity and heterogeneity of urban form. Much of the relevant information exists in unstructured formats, such as street-level imagery, architectural designs, and visual cues that capture human perceptions of space [6]. These modalities are difficult to quantify and analyze using traditional feature engineering or structured data pipelines.


**Figure 1: UrbanX: an interpretable, MLLM-powered framework for hypothesis-driven urban scientific discovery.**

Although recent advances have enhanced the use of data in urban research, existing methodological pipelines still face fundamental challenges when it comes to discovering new, interpretable factors from complex urban environments. Traditional approaches typically rely on expert-curated variables, black box predictive models, or handcrafted metrics to study specific urban dimensions [19, 35, 44]. These strategies face three key limitations. First, hypothesis generation is a manual and cognitively intensive process that depends heavily on prior knowledge and is vulnerable to confirmation bias [12]. Second, while deep learning methods offer strong predictive capabilities, they often rely on latent representations, limiting their interpretability and scientific value. Third, unstructured urban data, particularly SVIs, remains underutilized as a source of meaningful variables. These data contain rich contextual cues about urban and social space, yet current methods struggle to extract structured insights from them [34]. Together, these limitations constrain our ability to explore the urban hypothesis space at scale and with transparency.

✉Corresponding Author. Project: https://github.com/YihongT/UrbanX.git

Meanwhile, the emergence of foundation models, particularly Large Language Models (LLMs) [29], has transformed the landscape of data-driven reasoning. These models are trained on large-scale corpora of natural language and are capable of understanding, generating, and reasoning with text in flexible and context-aware ways [40]. Building on this progress, recent developments have extended the capabilities of LLMs to include visual and other modalities' inputs, giving rise to Multimodal Large Language Models (MLLMs) [43, 49]. These models jointly process text and images, enabling them to perform tasks such as visual question answering, open-ended scene interpretation, and multimodal reasoning with minimal supervision. MLLMs can extract semantically meaningful concepts from visual scenes and align them with natural language in ways that reflect human-like understanding. Rather than serving only as tools for visual perception or image classification, MLLMs can be used as semantic engines that generate interpretable, human-aligned variables directly from raw visual data. This capability opens a new avenue for scalable, transparent, and cognitively grounded urban scientific discovery.

Building on this insight, we propose a new framework, UrbanX, for interpretable, hypothesis-driven scientific discovery in urban domains, powered by MLLMs. As shown in Figure 1, rather than treating machine learning as a black box predictor, we reframe it as a collaborative agent in the scientific process, one that can generate candidate hypotheses, operationalize them into interpretable variables from structured and unstructured data, and evaluate their statistical relevance to real-world outcomes. UrbanX introduces an iterative structure: at each step, the model formulates hypotheses as natural language queries, extracts semantically meaningful features by applying MLLMs to unstructured inputs, and assesses their explanatory power using interpretable statistical models. Hypotheses with weak statistical support are pruned, and new ones are proposed, allowing the system to gradually converge on a compact set of variables that are human-aligned and empirically grounded.

We instantiate this framework in the domain of urban road safety, a high-impact area where interpretability is essential for actionable insight, and where the discovery of new, semantically meaningful factors can directly inform planning and policy. We demonstrate the effectiveness of UrbanX on a Manhattan case study, where it discovers novel visual variables from SVIs that exhibit significant correlations with crash rates. Our approach achieves predictive performance superior to pretrained deep learning baselines such as ResNet [17] and Vision Transformer (ViT) [9], while preserving full transparency through interpretable variable construction and attribution. Our work makes the following contributions:

- We formalize scientific discovery in urban domains as an inference problem over a hypothesis space, where each hypothesis is a natural-language conjecture linking urban form to societal outcomes. This framing enables machines to generate, test, and refine hypotheses directly from structured and unstructured data, offering a scalable foundation for data-driven urban science.
- We propose a novel use of MLLMs as semantic engines that translate unstructured inputs (e.g., SVIs) into structured, interpretable variables through natural-language hypotheses. This approach bridges perception and statistical modeling within a unified, human-interpretable framework and holds broad applicability across urban and transportation research.
- We develop a nonparametric, interpretable framework for hypothesis inference, formulated as an iterative posterior approximation over a hypothesis space. At each iteration, the framework generates new hypotheses, constructs semantically aligned embeddings, and evaluates variable significance using interpretable statistical models. This process enables scalable, statistically grounded, and transparent discovery of novel urban factors while reducing the human effort involved in the scientific discovery process.
- We apply our framework to study road safety in the Manhattan area and demonstrate its ability to uncover novel, interpretable visual variables that significantly correlate with crash rates. Our approach outperforms strong vision baselines in predictive performance while maintaining interpretability through hypothesis-level attribution. Beyond this case study, our approach serves as a general-purpose framework for scientific discovery in urban domains, with the potential to reveal structured insights from unstructured data across a wide range of socioeconomic and environmental outcomes.

## 2 RELATED WORK

### 2.1 Urban Scientific Discovery

The pursuit of understanding how urban form influences societal outcomes, such as public health, equity, and road safety, is a cornerstone of urban science and transportation research [2, 16]. Traditionally, this has involved developing statistical models to find correlations between expert-defined built environment variables and specific outcomes [33]. For instance, in road safety, studies have long linked street design elements, traffic calming measures, and infrastructure for pedestrians and cyclists to crash frequencies and severity [10, 30, 46].

However, these conventional workflows face significant hurdles in uncovering novel, interpretable insights from the complex urban milieu. A primary challenge is the manual and intuition-driven nature of hypothesis generation, which is often slow, susceptible to researchers' confirmation biases, and may overlook unconventional relationships [44]. This reliance on pre-existing knowledge or limited observations can constrain the breadth of scientific inquiry.

Furthermore, while advanced machine learning, particularly deep learning, has shown promise in predictive tasks using urban data, such models often function as "black boxes" [13]. Their internal representations are typically opaque, making it difficult to understand which specific factors drive predictions or to extract actionable, causal insights for urban planning and policy-making. This lack of transparency can hinder trust and adoption, especially in high-stakes decisions [5].

Another critical limitation is the underutilization of rich, unstructured data sources. SVIs encapsulate vast amounts of visual information about the urban environment, from infrastructure quality to perceived safety cues [6]. Yet, their integration into quantitative analysis is hampered by challenges in image acquisition consistency, quality, spatial-temporal variability, and the difficulty of systematically extracting meaningful, structured variables [37]. Existing efforts to automate feature extraction from SVIs often rely on standard computer vision models that may not capture the nuanced, context-specific attributes relevant to complex societal outcomes without significant task-specific fine-tuning or annotation.

Recent explorations into AI-driven scientific discovery have begun to address some of these issues. For example, frameworks are emerging that use large language models for causal inference in urban contexts [44] or to assist in generating hypotheses in other scientific fields by leveraging knowledge graphs alongside LLMs [25]. These works highlight a growing recognition of AI's potential to augment the discovery process, yet a dedicated, interpretable framework for hypothesis inference directly from SVIs remains an open area. Our work seeks to bridge this gap by leveraging MLLMs to systematically generate and test interpretable hypotheses about the urban environment's impact on road safety.

## 2.2 Multimodal Large Language Models

The advent of Large Language Models (LLMs) has significantly advanced capabilities in natural language understanding, generation, and reasoning [1, 41, 47]. Building upon this foundation, Multimodal Large Language Models (MLLMs) have emerged, extending these powerful reasoning abilities to encompass multiple modalities, most notably vision and language [36, 45, 49]. These models are designed to jointly process and interpret information from textual descriptions and visual inputs, such as images or videos.

Typical MLLM architectures integrate a pre-trained vision encoder with a pre-trained LLM [39]. A crucial component is the vision-language connector module, which projects visual features into a space compatible with the LLM's word embeddings. This connector can range from a simple linear projection layer, as seen in early models like LLaVA [24], which can perform instruction-aware visual feature extraction targeted by textual queries [23]. Training MLLMs often involves a multi-stage process: an initial pre-training phase to align visual and language representations using image-text datasets, followed by fine-tuning on multimodal instruction-following datasets to enhance their ability to perform specific tasks and engage in dialogue [31].

MLLMs have demonstrated remarkable capabilities across a wide range of tasks, including visual question answering (VQA), image captioning, multimodal dialogue, and complex visual reasoning [3, 38]. They can generate nuanced textual descriptions of images, answer questions about visual content, and follow instructions that require grounding language in visual information. This ability to extract semantically meaningful concepts from visual scenes and align them with natural language is central to their potential. Recent research also explores techniques like Optimal Transport to achieve more interpretable semantic alignment between modalities, allowing for insights into the MLLM's reasoning process by visualizing how visual and textual elements correspond [21]. This is particularly relevant for applications requiring trustworthiness.

The application of MLLMs to automated scientific discovery is a burgeoning field, with studies exploring their use for generating novel research ideas, designing experiments, and even assisting in writing scientific papers [14]. In urban contexts, vision-language models have been used for tasks such as function inference from street-level imagery [20]. However, the dominant focus in automated scientific discovery has often been on the novelty or efficiency of hypothesis generation, rather than on the interpretability of the generated hypotheses or the variables used, especially when derived from complex visual data in specific domains like urban science. Our framework distinctively proposes using MLLMs not just as predictors or general-purpose reasoners, but as semantic engines to derive interpretable, human-aligned variables directly from unstructured visual data (SVIs) for statistically rigorous hypothesis inference concerning road safety.

# 3 METHODOLOGY

## 3.1 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.

Bayesian inference over all possible hypothesis subsets is computationally infeasible due to the combinatorial size of $\mathcal{H}$ and the lack of a tractable likelihood model. Instead, we adopt an approximate inference strategy and frame the task as a nonparametric structure learning problem. Starting from an initial hypothesis set $\mathcal{H}^0$ sampled from an LLM, we iteratively refine the set by evaluating each hypothesis using a linear regression model. For each hypothesis $h_j$, we assess the statistical significance of its corresponding regression coefficient via a two-sided $t$-test under the null hypothesis that the coefficient equals zero. This yields a $p$-value vector $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$, where each $p_j$ quantifies the probability of observing the estimated coefficient under the null. Hypotheses with $p_j > \alpha$ (typically 0.05) are considered statistically insignificant and are discarded. New hypotheses are generated to replace them, forming an iterative process. This process constitutes a data-driven, nonparametric approximation to Bayesian posterior inference over the hypothesis space, guided by statistical evidence and LLM priors. An overview of the UrbanX framework is illustrated in Figure 2.

## 3.2 Hypothesis Generation

A key challenge in data-driven urban science is formulating meaningful and testable hypotheses that connect observable environmental variables to societal outcomes. In conventional workflows, this process relies heavily on human intuition, expert-defined variables, and domain-specific heuristics, creating scalability, objectivity, and scope bottlenecks. To overcome these limitations, we propose LLMs

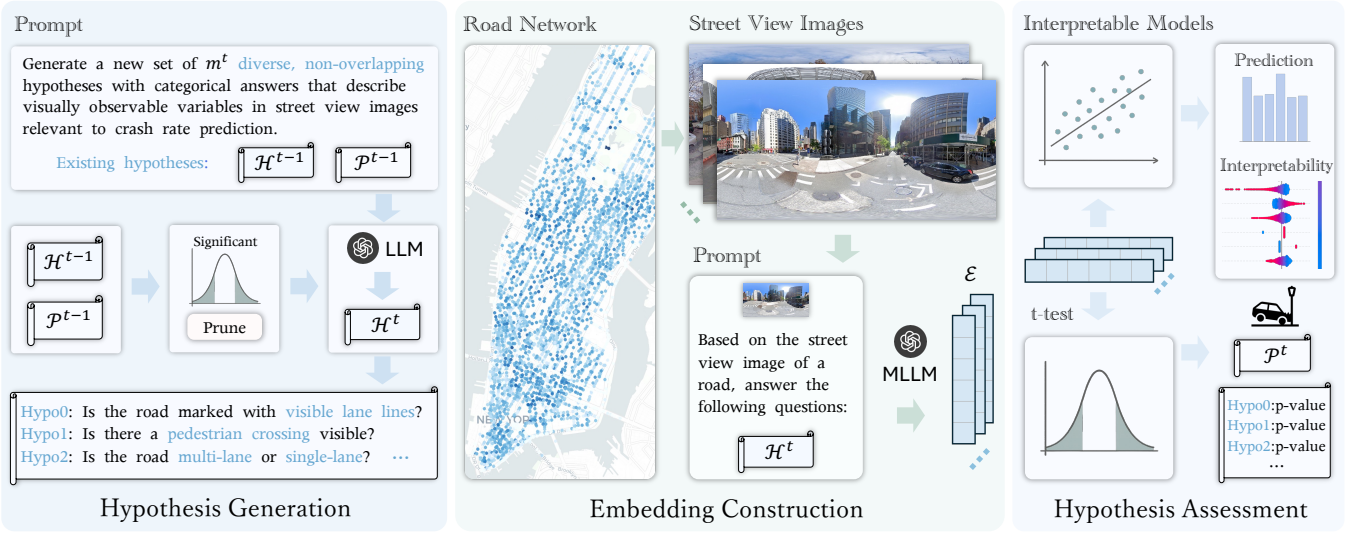



**Figure 2: The UrbanX framework consists of three iterative modules: (1) Hypothesis Generation using LLMs, (2) Embedding Construction via MLLM-based VQA on SVIs, and (3) Hypothesis Assessment using interpretable regression analysis.**

as cognitive engines that can explore and articulate semantically rich, visually grounded hypotheses about urban safety.

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:

$$\mathcal{H}^t \sim \text{LLM}\left(\text{Prompt}_{HypoGen}(\mathcal{H}^{t-1}, \mathcal{P}^{t-1}, m^t)\right), \tag{1}$$

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. The full construction of the prompt used in this stage is detailed in Appendix A.

This iterative, LLM-in-the-loop design ensures that hypothesis generation is continuously shaped by empirical evidence, fostering the discovery of novel but statistically grounded variables. The use of in-context prompting enables controlled and diverse exploration of the hypothesis space, avoiding redundancy and incorporating feedback from previous evaluations. The generation process also supports a nonparametric Bayesian interpretation: the LLM defines a flexible, data-informed prior over hypothesis space, while statistical assessment provides approximate posterior guidance. This synergy supports a principled and interpretable refinement of the variable space over successive iterations. An illustration of this process is shown in the left panel of Figure 2.

## 3.3 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:

$$e_i^t \sim \text{MLLM}\left(x_i, \text{Prompt}_{Embed}(\mathcal{H}^t)\right), \tag{2}$$

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.

### 3.4 Hypothesis Assessment

Given the hypothesis-aligned embedding matrix $\mathcal{E}^t \in \mathbb{R}^{n\times k}$ constructed from the current hypothesis set $\mathcal{H}^t$, the next step is to evaluate the empirical relevance of each hypothesis in explaining variation in segment-level crash rates. Rather than optimizing for predictive performance alone, our objective is to support transparent, interpretable modeling that enables attribution of outcomes to individual hypotheses. This is particularly important in transportation and urban policy contexts, where analytical traceability and explanatory clarity are essential for decision-making and public accountability. To this end, we adopt linear regression as the primary modeling framework. Each column of $\mathcal{E}^t$ corresponds to a hypothesis-specific variable derived from categorical responses generated by the MLLM. We fit a linear model of the form: $y_i = \beta_0 + \sum_{j=1}^{k} \beta_j e_{ij}^t + \varepsilon_i$, where $y_i$ is the observed crash rate for SVI $x_i$, $e_{ij}^t$ is the value of the $j$-th embedding dimension for the $i$-th SVI, $\beta_j$ is the corresponding regression coefficient, and $\varepsilon_i$ is an independent error term assumed to be normally distributed.

We then apply a two-sided $t$-test to each coefficient $\beta_j$ to assess the null hypothesis that $\beta_j = 0$, using standard errors estimated from the fitted model. This yields a $p$-value $p_j^t$ that quantifies the likelihood that the observed effect could arise under the null. Collectively, the vector $\mathcal{P}^t = \{p_1^t, p_2^t, \ldots, p_k^t\}$ summarizes the statistical significance of each hypothesis in $\mathcal{H}^t$. Hypotheses with $p_j^t > \alpha$ (typically $\alpha = 0.05$) are considered statistically insignificant and are pruned in the next round. The number of such hypotheses, $m^t$, determines how many new hypotheses are generated in the subsequent iteration. Fitting the regression model $\mathcal{M}$ to the embedding matrix $\mathcal{E}^t$ yields both fitted outcomes and statistical estimates of hypothesis relevance:

$$\{\widehat{y}_i\}_{i=1}^n,\ \{\widehat{\beta}_j,\ p_j^t\}_{j=1}^k \leftarrow \mathcal{M}(\mathcal{E}^t), \tag{3}$$

where $\mathcal{M}$ denotes the linear regression model applied to the embedding matrix $\mathcal{E}^t$.

This assessment step plays two complementary roles within the overall framework. First, it provides a quantitative basis for interpreting the influence of each hypothesis on crash outcomes, enabling transparent attribution and variable importance comparisons. Second, it functions as a mechanism for iterative hypothesis refinement, systematically pruning low-utility hypotheses and informing the next round of LLM-based generation. The right panel of Figure 2 illustrates this process within the broader iterative pipeline.

### 3.5 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 [32] or expectation maximization [28], 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** 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 $\mathcal{E}^T$

```
Initialize H^0 ~ LLM(Prompt_HypoGen(k))
for t = 1, 2, ..., T do
    H^t ~ LLM(Prompt_HypoGen(H^{t-1}, P^{t-1}, m^t))
    E^t = {e_i^t = MLLM(x_i, Prompt_Embed(H^t))}_{i=1}^n
    {ŷ_i}_{i=1}^n, P^t ← M(E^t)
end for
```

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 until convergence or a predefined number of iterations.

## 4 EXPERIMENTS

### 4.1 Settings

Our study focuses on road segments in Manhattan, New York City, using crash records, traffic volume data, and street-view imagery from 2013 to 2019. For each road segment, we compute the crash rate following the works [18, 46, 48] as: $CR_i = \frac{\text{No\_crash}_i}{AADT_i \times L_i \times \frac{365}{1{,}000{,}000}}$, where $\text{No\_crash}_i$ denotes the average annual number of crashes, $AADT_i$ is the average annual daily traffic, and $L_i$ is the segment length in kilometers. Crash records were obtained from NYC Open Data, and *AADT* data was sourced from the New York State Department of Transportation. We sampled SVIs at 15-meter intervals along road centerlines using ArcGIS [42], retrieving images via the Google Street View API. After filtering and processing, 16,000, 2,000, and 2,000 panoramic SVIs were used for training, validation, and testing, respectively. Unless otherwise specified, we use GPT-4o [22] as the LLM for hypothesis generation and InternVL2.5-78B [7] as the MLLM for answering hypotheses over images during embedding construction. We deploy MLLMs using LMDeploy [8], an optimized inference engine for serving MLLMs efficiently.

To support downstream evaluation and comparison, we also compile 58 conventional built environment variables from public sources. These include five categories: (1) road attributes (e.g., width, highway indicator), (2) land use (six category proportions and entropy), (3) point-of-interest (POI) features (density and distance for 13 POI types), (4) traffic-related facilities (e.g., crossings, bus stops, junctions), and (5) visual indices derived from panoptic segmentation of SVIs (e.g., proportions of sky, road, building, and vegetation pixels). A full list and description of these variables is provided in Table 1 in Appendix B. These features are aggregated from multiple sources, including NYC Open Data, PLUTO, CommonPlace, Geofabrik, and the Google Street View API.

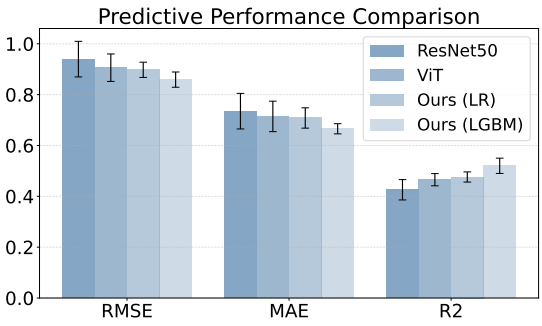



**Figure 3: Performance comparison between ResNet, ViT, and our interpretable embedding-based models using linear regression (LR) and LightGBM (LGBM).**

In our primary modeling pipeline, only the LLM-generated hypotheses and MLLM-derived interpretable embeddings are used for prediction. Built environment variables are used exclusively in post hoc SHAP analysis to compare their relative explanatory power against discovered hypotheses.

## 4.2 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.

To visualize the spatial quality of our predictions, Figure 4 presents the predicted and actual crash rates across Manhattan at the road segment level. The predicted map closely mirrors the true spatial distribution of risk, capturing key hotspots such as lower Manhattan and the Midtown corridor. To ensure comprehensive spatial coverage and generalizability, predictions are generated using a five-fold cross-validation setup, where each segment is held out once as test data. The aggregated predictions thus represent out-of-sample estimates across the entire study area. This spatial fidelity highlights the reliability of our interpretable model for real-world urban safety applications.

## 4.3 Discovered Factors

A central goal of UrbanX is not just to predict societal outcomes, but to enable interpretable, data-driven discovery of unstructured or previously overlooked urban factors. To assess whether our framework successfully identifies meaningful visual variables, we examine the learned regression model using SHAP (SHapley Additive exPlanations) [26] analysis, which quantifies each variable's marginal contribution to the prediction.

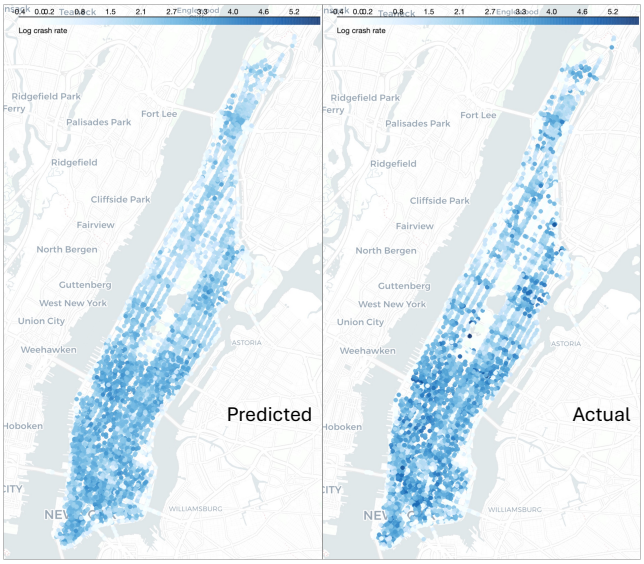



**Figure 4: Spatial distribution of predicted (left) vs. actual (right) crash rates (log) across Manhattan road segments.**

Figure 5 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

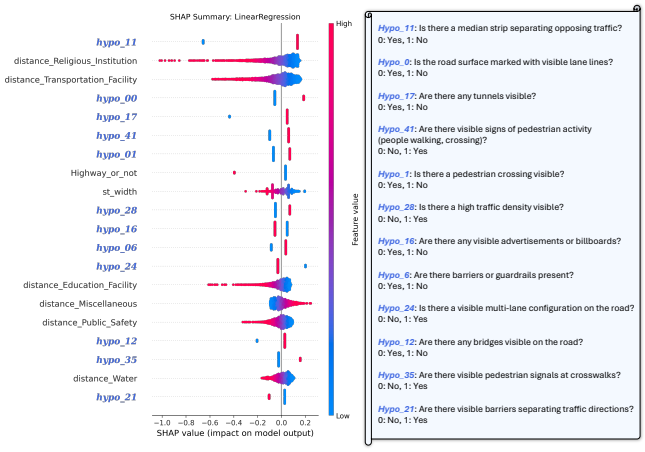



**Figure 5: SHAP summary plot of the regression model with both traditional built environment variables and discovered hypotheses. The right panel maps the top hypothesis variables to their natural-language question meanings.**

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 unique advantage of UrbanX in supporting structured discovery over unstructured inputs. A detailed list of all identified hypotheses and their corresponding analyses is provided in Appendix C.

Taken together, these results show that UrbanX is capable not only of achieving competitive predictive accuracy but also of surfacing novel, interpretable factors that enrich the scientific understanding of urban safety. Its ability to propose, test, and validate hypotheses from raw street imagery that without manual annotation or expert-defined features, highlighting the potential of MLLM-powered frameworks to transform how we conduct research in urban and transportation science. The results reveal that many of the top-ranked variables in terms of predictive contribution come from LLM-generated hypotheses, underscoring the framework's ability to discover meaningful and interpretable features beyond standard urban design indicators.

## 4.4 Variable Significance and Independence

To verify that the hypotheses uncovered by UrbanX are not only predictive but also statistically robust and non-redundant, we perform two complementary analyses: significance versus contribution, and pairwise correlation, as shown in Figure 6.

The left panel plots each hypothesis by its average SHAP value (x-axis) and the negative base-10 logarithm of its $p$-value from linear regression (y-axis). This joint visualization enables simultaneous assessment of feature importance and statistical significance. Hypotheses located in the upper-left region of the plot are both highly predictive and statistically significant. Notably, `Hypo_11`, `Hypo_41`, and `Hypo_0` stand out as dominant factors, exhibiting both high SHAP contributions and extremely low $p$-values. These questions correspond to well-established road safety indicators, whether there is a median strip separating traffic, whether visible pedestrian activity is present, and whether lane markings are detectable, offering domain-consistent evidence for their relevance.

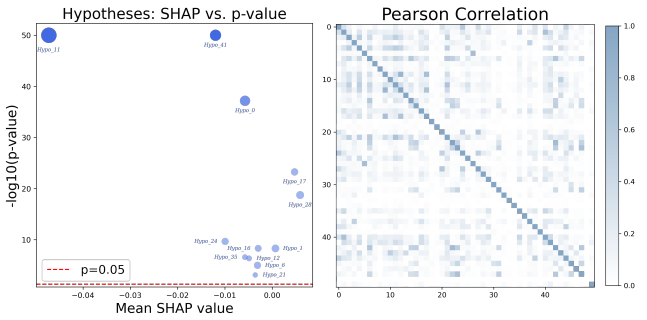



**Figure 6: (Left) Each hypothesis is plotted by its average SHAP value and $-\log_{10}(p\text{-value})$ from regression. Variables in the top left are highly significant and predictive. (Right) Pearson correlation matrix between hypotheses, showing low pairwise correlation and structural independence.**

The right panel presents the pairwise Pearson correlation matrix among all hypothesis-derived variables, based on their categorical values across SVIs. The mostly light-toned off-diagonal entries reflect generally low correlations, indicating that the learned variables capture complementary aspects of the visual environment rather than redundant or collinear signals. This structural independence further supports the interpretability and modularity of the learned embedding space, reducing concerns about multicollinearity or semantic overlap.

Together, these findings demonstrate that UrbanX is capable of discovering statistically grounded, interpretable, and non-redundant variables that not only improve predictive performance but also advance scientific understanding of urban safety phenomena.

## 4.5 Robustness and Validity

We assess the robustness of UrbanX by varying three key factors: (1) the capacity of the language model (LLM) used for hypothesis generation, (2) the capacity of the vision-language model (MLLM) used for embedding construction, and (3) the number of hypotheses used to generate interpretable features. Figure 7 presents results that illustrate how each of these choices affects convergence speed, predictive accuracy, and model stability.

*Effect of model capacity.* The left panel compares the convergence of test performance over 50 training epochs using different combinations of LLMs and MLLMs. Specifically, we consider GPT-4o and GPT-4o-mini as the hypothesis generators, paired with InternVL2.5 models of 8B and 78B parameters for embedding construction. Larger MLLMs (78B) consistently yield better predictive performance and faster convergence, underscoring the critical role of visual reasoning capacity in answering hypothesis queries from SVIs. For example, with InternVL2.5-78B, models converge in fewer than 10 epochs, while smaller models (8B) exhibit slower and noisier learning curves. Language model size also affects convergence, though to a lesser extent. GPT-4o achieves faster improvements than GPT-4o-mini, but both converge to similar final performance. This suggests that for categorical hypothesis generation, smaller LLMs are sufficient, though larger models may enhance efficiency by producing more immediately useful hypotheses.

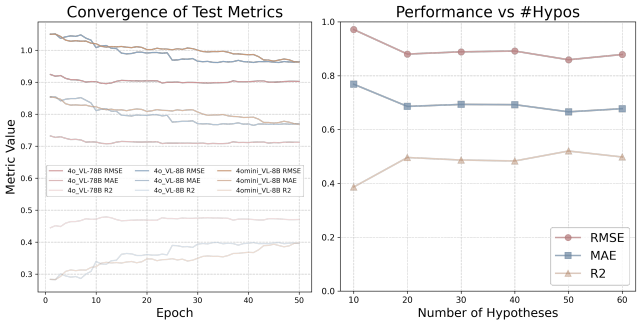



**Figure 7: Robustness analysis. (Left) Convergence of test metrics across different LLM and MLLM configurations. High-capacity MLLMs (78B) yield better and faster convergence. (Right) Performance as a function of the number of hypotheses used. Optimal performance is observed at 50 hypotheses.**

*Effect of hypothesis-set size.* The right panel examines performance as a function of the number of hypotheses used to construct interpretable embeddings. Performance improves steadily up to around 50 hypotheses, with diminishing returns and slight degradation beyond that point. This pattern reflects a balance between semantic expressiveness and statistical noise: too few hypotheses limit the model's representational capacity, while too many may introduce redundancy, spurious correlations, or overfitting. The model achieves optimal RMSE, MAE, and $R^2$ values when embedding dimensionality is neither overly constrained nor saturated.

*Implications for practice.* These experiments demonstrate that UrbanX is robust to reasonable changes in foundation-model capacity and to the choice of hypothesis-set size. In resource-constrained settings, an issue frequently faced by transportation agencies, one can adopt a smaller language model without a large performance penalty, provided that a sufficiently capable vision–language model is available. The sensitivity analysis on hypothesis cardinality also offers guidance for practitioners: start with a moderate set (40–60 items), monitor statistical significance during training, and prune or augment as needed. The results confirm that the proposed framework delivers stable, interpretable, and policy-relevant insights across a range of computational budgets and modelling choices.

## 4.6 Discussion

UrbanX demonstrates the potential of applying large language MLLMs for interpretable, automated hypothesis discovery in the context of urban safety analysis. The results show that the generated hypotheses not only match or surpass the predictive power of traditional computer vision models but also provide clear semantic insights aligned with established urban design principles. Moreover, UrbanX enables the discovery of novel, human-interpretable variables that may have been overlooked in prior literature, offering a scalable approach to data-driven scientific discovery.

This work implicitly builds upon a set of assumptions that reflect a shift in how machine learning can be used for knowledge generation. First, we assume that MLLMs, when queried appropriately, can reliably interpret and respond to natural-language hypotheses about complex visual scenes. This assumption effectively treats the MLLM as a proxy for human visual judgment, capable of semantically parsing urban environments in a consistent and informative way. A detailed analysis of MLLM answer quality is provided in Appendix D. Second, we assume that LLMs possess a rational understanding of societal constructs such as safety, risk, and infrastructure, and can leverage this latent knowledge to propose plausible hypotheses. These assumptions align with a broader philosophical view of foundation models not merely as function approximators, but as *cognitive instruments*, tools that can externalize latent human reasoning in scalable and programmable ways.

From this perspective, our framework is more than a predictive pipeline, it is a machine-in-the-loop system for structured discovery. It operationalizes a new epistemic loop: language models propose interpretable, theory-aligned variables; MLLMs extract structured representations from raw perceptual data; and statistical models evaluate and refine the space of explanatory factors. While not infallible, this loop offers a novel approach to bridging data, semantics, and scientific reasoning. Another key implication of this work is that hypothesis generation and refinement, traditionally limited by expert intuition and manual feature engineering, can be guided by LLMs in a statistically grounded loop. This offers a path toward semi-automated scientific workflows where human and machine jointly explore high-dimensional, unstructured data spaces.

Notably, our reliance on foundation models introduces limitations. The correctness of our results depends on the alignment and reliability of the underlying MLLMs and LLMs. Errors in visual understanding or gaps in commonsense reasoning may lead to spurious or irrelevant hypotheses. Moreover, the iterative nature of our approach, while principled, incurs significant computational overhead due to repeated prompting and inference. However, we believe these constraints are temporary. As foundation models continue to improve in efficiency, alignment, and accessibility, the feasibility of such machine-guided discovery frameworks will continue to grow.

## 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


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Michele Acuto, Susan Parnell, and Karen C Seto. 2018. Building a global urban science. *Nature Sustainability* 1, 1 (2018), 2–4.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

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

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

[6] Filip Biljecki and Koichi Ito. 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning* 215 (2021), 104217.

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

[8] LMDeploy Contributors. 2023. LMDeploy: A Toolkit for Compressing, Deploying, and Serving LLM. https://github.com/InternLM/lmdeploy.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

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

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

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

[13] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[14] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

[15] Luis A Guzman and Juan P Bocarejo. 2017. Urban form and spatial urban equity in Bogota, Colombia. *Transportation research procedia* 25 (2017), 4491–4506.

[16] Randolph Hall. 2012. *Handbook of transportation science*. Vol. 23. Springer Science & Business Media.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

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

[19] Yijia Hu, Long Chen, and Zhan Zhao. 2024. How does street environment affect pedestrian crash risks? A link-level analysis using street view image-based pedestrian exposure measurement. *Accident Analysis & Prevention* 205 (2024), 107682.

[20] Weiming Huang, Jing Wang, and Gao Cong. 2024. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science* 38, 7 (2024), 1414–1442.

[21] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).

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

[23] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *Comput. Surveys* 57, 8 (2025), 1–36.

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

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

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

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

[28] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, 6 (1996), 47–60.

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

[30] Ao Qu, Yihong Tang, and Wei Ma. 2023. Adversarial attacks on deep reinforcement learning-based traffic signal control systems with colluding vehicles. *ACM Transactions on Intelligent Systems and Technology* 14, 6 (2023), 1–22.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[32] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).

[33] Matheos Santamouris. 2013. *Energy and climate in the urban built environment*. Routledge.

[34] Yihong Tang, Menglin Kong, and Lijun Sun. 2025. Large Language Models for Data Synthesis. *arXiv preprint arXiv:2505.14752* (2025).

[35] Yihong Tang, Ao Qu, Andy HF Chow, William HK Lam, Sze Chun Wong, and Wei Ma. 2022. Domain adversarial spatial-temporal network: A transferable framework for short-term traffic forecasting across cities. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 1905–1915.

[36] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2024. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv preprint arXiv:2410.16162* (2024).

[37] Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, et al. 2024. 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*. 1413–1432.

[38] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186* (2025).

[39] Zhaokai Wang, Xizhou Zhu, Xue Yang, Gen Luo, Hao Li, Changyao Tian, Wenhan Dou, Junqi Ge, Lewei Lu, Yu Qiao, et al. 2025. Parameter-Inverted Image Pyramid Networks for Visual Perception and Multimodal Understanding. *arXiv preprint arXiv:2501.07783* (2025).

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

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

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

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

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

[45] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.

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

[47] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837* (2025).

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

[49] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).

# A PROMPT DESIGN AND SAMPLING STRATEGY

## A.1 Hypothesis Prompting Strategy.

In our implementation, we adopt an exploration-exploitation strategy for generating new hypotheses using Large Language Models. At each iteration $t$, a new set of candidate hypotheses $\mathcal{H}^t$ is sampled by prompting the LLM with one of two designed templates:

**HYPO_EXPLOIT_PROMPT**

```
Generate a new set of {n} diverse and non-overlapping binary classification questions that capture interpretable features from street view images relevant to
predicting crash rates.

Existing questions: ```{hypo t-1}```

Design requirements:
- Avoid redundancy and ensure each question is unique.
- Prioritize features with clear relevance to road safety.
- Questions must be answerable based solely on the visual content of the street view image.
- Each question must have exactly two answer choices.
- The answer options must be mutually exclusive and collectively exhaustive.
- Avoid vague, ambiguous, or overly subjective formulations.
- Focus on concrete visual cues (e.g., road structure, visibility, signage, obstructions) that could improve crash prediction.

First, provide a short reasoning paragraph analyzing the strengths and weaknesses of the existing questions with respect to safety prediction. Suggest new
aspects that could be captured to improve performance.

Then return a valid Python dictionary formatted as JSON:
{{
    "reasoning": "short reasoning text.",
    "Question 1": {{"0": "Answer option A", "1": "Answer option B"}},
    ...
}}

The response must be valid JSON (parseable by `json.loads`) and should include only the JSON dictionary, without additional commentary.
```

An exploitation prompt (see HYPO_EXPLOIT_PROMPT) that conditions on the currently retained hypothesis set $\mathcal{H}^{t-1}$ and their statistical significance $\mathcal{P}^{t-1}$. This prompt encourages refinement and expansion of hypotheses with known predictive value, anchoring the generation process to empirically validated ideas.

**HYPO_EXPLORE_PROMPT**

```
Generate {n} unique binary classification questions that explore the broad visual, social, and contextual dimensions of street view imagery.

Design requirements:
- Avoid repetition and ensure semantic diversity.
- Questions must be clearly answerable from a single street view image.
- Each question must have exactly two answer choices.
- The answer options must be mutually exclusive and collectively exhaustive.
- Avoid overly ambiguous, speculative, or subjective content.

Consider features beyond the obvious, such as socioeconomic signals, environmental quality, cultural context, artistic elements, or neighborhood character. Let
curiosity guide you to ask unconventional, thought-provoking questions that reveal hidden or surprising correlations.

Return a valid Python dictionary formatted as JSON:
{{
    "reasoning": "brief explanation of how the questions were conceived.",
    "Question 1": {{"0": "Answer option A", "1": "Answer option B"}},
    ...
}}

The response must be valid JSON (parseable by `json.loads`) and should include only the JSON dictionary, without additional commentary.
```

An exploration prompt (see HYPO_EXPLORE_PROMPT) that deliberately encourages broader semantic coverage, open-ended question generation, and inclusion of unconventional or underexplored visual features. This promotes diversity and mitigates local optima.

To balance these goals, we apply a stochastic control mechanism: at each iteration, with a fixed probability $p_{\text{explore}}$ (default 0.1), the exploration prompt is selected; otherwise, the exploitation prompt is used. This simple sampling scheme mirrors common strategies in reinforcement learning and approximate inference, ensuring both local exploitation and global search over the hypothesis space.

The final prompt used in each iteration is automatically constructed based on the current retained hypotheses and plugged into the appropriate template. Each LLM response is parsed as structured JSON and incorporated into the next round of variable construction and statistical evaluation.

This approach enables interpretable and data-driven exploration of the hypothesis space while maintaining relevance and control through prior feedback and statistical validation.

## A.2 Embedding Prompting Strategy

For embedding construction, we use Multimodal Large Language Models (MLLMs) to answer each generated hypothesis based on the visual content of a street view image. We define two templates for prompting the MLLM: a single-question version for sequential evaluation, and a batched version for more efficient parallel processing.

**EMB_PROMPT**

```
Based on the street view image of a road, answer the following question:
```{hypo t}```

Return only an integer indicating the option number and nothing else
```

The single-question prompt (EMB_PROMPT) is used to infer the answer to one hypothesis at a time. It specifies the question and options, instructing the MLLM to return only the index of the chosen option as an integer.

**EMB_BATCH_PROMPT**

```
Based on the street view image of a road, answer the following questions:
```{hypo t}```

example response format (should be replaced by your answer to the questions):
["0", "1", ...]

Return only a list of {n} integers indicating the option numbers, the response should be parsable with `json.loads` and nothing else
```

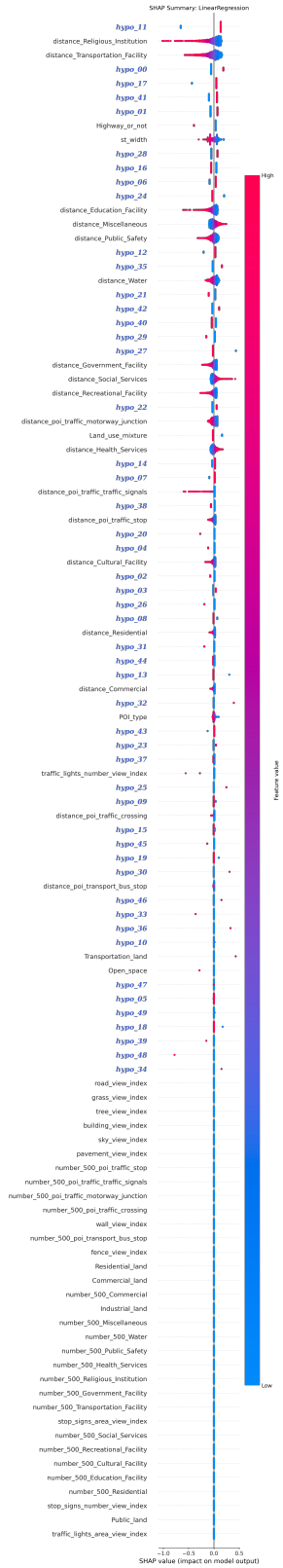
The batch prompt (EMB_BATCH_PROMPT) enables simultaneous evaluation of multiple hypotheses. It presents all questions at once and expects a list of integers corresponding to the chosen answer for each. Unless otherwise specified, we use the batch prompt for all experiments, as we find that it does not compromise VQA quality. This structured prompting strategy ensures semantic traceability, data efficiency, and ease of downstream statistical modeling.

## B SUMMARY OF BUILT ENVIRONMENT VARIABLES

This section summarizes the 58 built environment variables used in the study. These variables are derived from multiple data sources, including Google Street View images, New York City open data, land use polygons, OpenStreetMap, and POI (Point of Interest) datasets. The variables are categorized into five main classes: View Indices, Road Attributes, Land Use, Points of Interest (POI), and Traffic-related Facilities.

**Table 1: Summary of Built Environment Variables**

| Category | Variables | Description |
|---|---|---|
| **View Indices (12)** | road_view_index, pavement_view_index, sky_view_index, building_view_index, tree_view_index, grass_view_index, fence_view_index, wall_view_index, traffic_lights_area_view_index, stop_signs_area_view_index, traffic_lights_number_view_index, stop_signs_number_view_index | Proportion of image pixels representing each element (e.g., sky, pavement, building, vegetation, traffic signs), extracted from Google Street View images. |
| **Road Attributes (2)** | Highway_or_not, st_width | Binary indicator of whether the road is a highway, and the width of the road in meters. |
| **Land Use (7)** | Residential_land, Commercial_land, Industrial_land, Transportation_land, Public_land, Open_space, Land_use_mixture | Area of six land use types within a buffer around the road segment, and a land use mixture index capturing the diversity of land use types. |
| **POI (27)** | number_500_Residential, number_500_Education_Facility, number_500_Cultural_Facility, number_500_Recreational_Facility, number_500_Social_Services, number_500_Transportation_Facility, number_500_Commercial, number_500_Government_Facility, number_500_Religious_Institution, number_500_Health_Services, number_500_Public_Safety, number_500_Water, number_500_Miscellaneous, distance_Residential, distance_Education_Facility, distance_Cultural_Facility, distance_Recreational_Facility, distance_Social_Services, distance_Transportation_Facility, distance_Commercial, distance_Government_Facility, distance_Religious_Institution, distance_Health_Services, distance_Public_Safety, distance_Water, distance_Miscellaneous, POI_type | Number of POIs of 13 types within a 500-meter buffer, distances to the nearest POI of each type, and the primary POI type. |
| **Traffic-related Facilities (10)** | number_500_transport_bus_stop, number_500_traffic_stop, number_500_traffic_crossing, number_500_traffic_motorway_junction, number_500_traffic_traffic_signals, distance_transport_bus_stop, distance_traffic_stop, distance_traffic_crossing, distance_traffic_motorway_junction, distance_traffic_traffic_signals | Number and distance of various traffic-related facilities (e.g., bus stops, crossings, traffic lights, junctions) within 500 meters of the road segment. |




*Hypo_0*: Is the road surface marked with visible lane lines? 0: Yes, 1: No

*Hypo_1*: Is there a pedestrian crossing visible? 0: Yes, 1: No

*Hypo_2*: Are there any parked vehicles on the road? 0: Yes, 1: No

*Hypo_3*: Are there any trees or shrubs nearby that might obstruct visibility? 0: Yes, 1: No

*Hypo_4*: Is there a sidewalk present for pedestrians? 0: Yes, 1: No

*Hypo_5*: Is there signage indicating a school zone? 0: Yes, 1: No

*Hypo_6*: Are there barriers or guardrails present? 0: Yes, 1: No

*Hypo_7*: Is there a bus stop visible? 0: Yes, 1: No

*Hypo_8*: Are there any cyclists visible on the road? 0: Yes, 1: No

*Hypo_9*: Are there any pedestrian overpasses or underpasses? 0: Yes, 1: No

*Hypo_10*: Is the area heavily urbanized? 0: Yes, 1: No

*Hypo_11*: Is there a median strip separating opposing traffic? 0: Yes, 1: No

*Hypo_12*: Are there any bridges visible on the road? 0: Yes, 1: No

*Hypo_13*: Is there a visible incline or decline on the road? 0: Yes, 1: No

*Hypo_14*: Are there pedestrian barriers on the sidewalk? 0: Yes, 1: No

*Hypo_15*: Are there any visible traffic cones? 0: Yes, 1: No

*Hypo_16*: Are there any visible advertisements or billboards? 0: Yes, 1: No

*Hypo_17*: Are there any tunnels visible? 0: Yes, 1: No

*Hypo_18*: Is there any snow or ice visible on the road? 0: Yes, 1: No

*Hypo_19*: Is there a visible emergency vehicle access? 0: Yes, 1: No

*Hypo_20*: Is the road multi-lane or single-lane? 0: Multi-lane, 1: Single-lane

*Hypo_21*: Are there visible barriers separating traffic directions? 0: No, 1: Yes

*Hypo_22*: Is there any visible construction work on the road or sidewalk? 0: No, 1: Yes

*Hypo_23*: Is there any visible obstruction affecting driver visibility? 0: No, 1: Yes

*Hypo_24*: Is there a visible multi-lane configuration on the road? 0: No, 1: Yes

*Hypo_25*: Is there a merging lane visible? 0: No, 1: Yes

*Hypo_26*: Is there a visible school bus stop? 0: No, 1: Yes

*Hypo_27*: What is the width of the road? 0: Narrow, 1: Wide

*Hypo_28*: Is there a high traffic density visible? 0: No, 1: Yes

*Hypo_29*: Is there a visible pedestrian refuge island? 0: No, 1: Yes

*Hypo_30*: Is there visible road debris that could impact vehicle control? 0: No debris visible, 1: Debris visible

*Hypo_31*: Are there any temporary traffic signs indicating detours or lane closures? 0: No temporary signs visible, 1: Temporary signs visible

*Hypo_32*: Are there any visible electronic billboards that might distract drivers? 0: No electronic billboards, 1: Electronic billboards visible

*Hypo_33*: Is the road surface visibly wet or dry? 0: Dry, 1: Wet

*Hypo_34*: Are there any visible surveillance cameras along the road? 0: No, 1: Yes

*Hypo_35*: Are there visible pedestrian signals at crosswalks? 0: No, 1: Yes

*Hypo_36*: Are any road intersections controlled with roundabouts? 0: No, 1: Yes

*Hypo_37*: Is there visible vegetation that could obscure road signs? 0: No, 1: Yes

*Hypo_38*: Is there a visible presence of dedicated turning lanes? 0: No, 1: Yes

*Hypo_39*: Are there visible speed bumps or traffic calming measures? 0: No, 1: Yes

*Hypo_40*: Is there a visible traffic signal at intersections? 0: No, 1: Yes

*Hypo_41*: Are there visible signs of pedestrian activity (people walking, crossing)? 0: No, 1: Yes

*Hypo_42*: Are there visible temporary barriers or detours? 0: No, 1: Yes

*Hypo_43*: Are there any visible sharp curves or winding sections on the road? 0: Yes, 1: No

*Hypo_44*: Are there visible reflective road markers for improved night visibility? 0: No, 1: Yes

*Hypo_45*: Is there visible debris or obstacles on the road that could affect driving? 0: No visible debris or obstacles, 1: Visible debris or obstacles

*Hypo_46*: Are there any visible signs indicating upcoming sharp turns or curves? 0: No visible signs, 1: Visible signs for sharp turns

*Hypo_47*: Is there visible evidence of recent road maintenance or repairs? 0: No evidence of recent maintenance, 1: Visible signs of maintenance or repairs

*Hypo_48*: Are there any visible speed enforcement measures, such as cameras or speed limiters? 0: No visible speed enforcement measures, 1: Visible speed enforcement measures

*Hypo_49*: Is there visible pedestrian activity crossing the street at non-designated areas? 0: No pedestrian activity at non-designated areas, 1: Pedestrian activity at non-designated areas

**Figure 8: Final set of 50 natural-language hypotheses retained by UrbanX after the iterative process.**

## C QUALITATIVE ANALYSIS OF THE FINAL HYPOTHESIS SET

The final iteration of UrbanX retains 50 hypotheses that together provide a multifaceted description of the street environment, the resulting hypotheses are shown in Figure 8. A close reading of the list reveals several encouraging patterns as well as a few limitations that suggest directions for future refinement.

**Breadth of physical design cues.** A substantial portion of the questions targets canonical elements of roadway geometry: median strips, lane markings, multi-lane configurations, and road width. These queries parallel variables that civil engineers traditionally collect through manual audits, yet here they arise automatically from the LLM without prior codification. Their presence confirms that the system can rediscover core safety factors in a purely data-driven manner.

**Attention to vulnerable users.** Another cluster of hypotheses concentrates on pedestrian infrastructure and activity, including crossings, signals, overpasses, and informal crossing behaviour. The model further probes cyclist visibility and the existence of school-zone signage. By elevating these human-centric factors to high SHAP ranks, the framework highlights elements often under-represented in purely geometric crash models, reinforcing its potential for policy-relevant discovery.

**Contextual and perceptual variables.** Several questions extend beyond traditional inventories to capture visual distraction (billboards, electronic signage), driver visibility (obstructions, vegetation), and transient obstacles (construction work, debris, snow or ice). Such perceptual cues are rarely present in structured GIS layers yet can be critical for real-world safety. Their emergence illustrates how image-based hypothesis generation can broaden the discourse on urban risk.

**Redundancy and granularity considerations.** A few hypotheses partly overlap in meaning. For instance, both the multi-lane question and the lane-width question address capacity, and both barrier-related items refer to physical separation. While some redundancy is expected in an open-ended search, it suggests an opportunity to introduce a post-generation clustering step that merges semantically similar queries and thereby yields a more parsimonious variable set.

**Ambiguity and data-imbalance.** Certain questions may be ambiguous in practice or rarely triggered in the Manhattan data. Examples include snow or ice on the roadway and visible emergency vehicle access, which occur infrequently in the imagery and hence contribute little statistical signal. Future iterations might incorporate an adaptive pruning rule that removes hypotheses with low prevalence or high annotation uncertainty.

**Categorical framing limitations.** All queries are currently coded as questions with categorical answers. While this choice simplifies MLLM inference and statistical testing, it can obscure gradations of exposure. Road width, traffic density, and billboard prominence are inherently continuous or ordinal. Extending UrbanX to multi-class or scalar responses would permit richer descriptions while retaining interpretability.

**Implications for urban research.** Taken together, the hypothesis set demonstrates that UrbanX can surface both established and novel factors spanning geometry, infrastructure, human activity, and perceptual context. Even with the noted redundancies and simplifications, the variables provide a transparent basis for scientific inquiry, enabling planners to trace crash-risk patterns back to concrete visual attributes. As MLLMs continue to improve, we anticipate the framework will yield even finer-grained hypotheses, opening new avenues for theory building in urban safety and beyond.

# D QUALITATIVE VERIFICATION OF MLLM ANSWERS


**Figure 9: Panoramic SVI used for VQA analysis.**

To evaluate the reliability of MLLM-derived answers in our framework, we conducted a manual audit on a representative panoramic SVI (Figure 9) using the final hypothesis set (see Figure 8). Each answer was cross-checked by an expert in transportation systems and labeled as *correct*, *partially correct*, or *incorrect*. The results are summarized in Table 2, with detailed error analysis provided in Table 3.

**Table 2: Manual audit of MLLM answers for the 50 retained hypotheses.**

| Outcome | # Hypotheses | Examples | Typical cause |
|---|---|---|---|
| Correct | 42 | lane lines, crosswalk, bus stop, median strip | clear visual cue |
| Partially correct | 3 | vegetation obstruction, traffic density | ambiguous threshold |
| Incorrect | 5 | snow/ice, pedestrian signal, speed camera | rare or small object |

*Overall accuracy.* As shown in Table 2, 42 of the 50 MLLM answers (84%) were fully correct. Three were partially correct and five were judged incorrect. Correct predictions typically involved clear, high-contrast visual elements such as lane markings, sidewalks, crosswalks, parked vehicles, and urban density, consistent with the top contributors identified by SHAP analysis in Section 4.3.

*Error profile.* A breakdown of the eight non-correct answers is shown in Table 3. Errors generally fall into three categories:

- **Ambiguous semantics.** Hypotheses involving subjective thresholds (e.g., visibility obstructions or what counts as an "advertisement") yielded borderline results due to interpretive ambiguity in both human and model judgment.
- **Rare or subtle visual cues.** Conditions such as snow, temporary detours, and speed cameras were either absent or too small to detect reliably at the given image resolution, resulting in hallucinated or missed detections.
- **Fine-grained infrastructure.** Elements like pedestrian refuge islands or painted turning lanes were sometimes misidentified, likely due to resolution constraints in the panoramic image.

*Key insight: resolution is a limiting factor.* Many of the observed errors, especially those involving subtle signage, infrastructure details, or rare objects, can be traced directly to insufficient resolution in the source image. While panoramic images provide comprehensive spatial coverage, they often downsample visual detail, making it hard for even a capable MLLM to reliably detect small features such as traffic cones, refuge islands, or painted turn arrows. Switching to higher-resolution imagery, zoomed crops, or targeted visual attention modules could address most of these failures with minimal design change.

*Impact on model integrity.* Importantly, none of the outright failures involved the high-salience geometric or pedestrian-safety hypotheses that dominate model attribution and prediction. This helps explain why UrbanX retains high predictive accuracy despite modest error rates in the long tail of hypotheses. The most informative variables are generally those with clear, resolvable visual signals, precisely the ones the MLLM gets right most often.

**Table 3: Detailed audit of hypotheses flagged as *partially correct* or *incorrect* for the case study in Figure 9.**

| Hypothesis (paraphrased) | Model Answer | Reason for Deviation |
|---|---|---|
| *Partially correct* | | |
| Are there trees or shrubs that *might* obstruct visibility? | Yes | Small street trees are present, but they do not noticeably block sight lines; the "obstruction" qualifier is subjective. |
| Are any advertisements or billboards visible? | Yes | A store fascia (*Hallmark*) is present, yet it is a storefront sign rather than a driver-facing billboard, making the classification debatable. |
| Are reflective road markers visible? | Yes | Lane arrows could contain reflective paint, but this cannot be verified from the daytime image; confidence is therefore partial. |
| *Incorrect* | | |
| Is there visible construction work on the road or sidewalk? | Yes | No construction activity or equipment is present in the scene. |
| Is there a pedestrian refuge island? | Yes | The intersection lacks any raised or painted refuge island. |
| Are pedestrian signals present at the crosswalk? | No | Standard pedestrian countdown signals are visible on the far-left mast arm. |
| Is there a dedicated turning lane? | No | Painted turn arrows are clearly marked in the foreground lane. |
| Are temporary barriers or detours visible? | Yes | No cones, barriers, or road-closure signs can be observed. |

*Practical refinement strategies.* To further improve reliability without sacrificing transparency, several targeted interventions can be introduced:

- **Prompt rephrasing.** Adding clarifying definitions or thresholds (e.g., "construction work must include cones or equipment") could help disambiguate borderline cases.
- **View augmentation.** Supplementing wide-angle views with higher-resolution zoom-ins or directional crops would boost recognition of small but critical features.
- **Response calibration.** Incorporating uncertainty scores or allowing abstention on low-confidence answers could help filter out hallucinated positives.

*Conclusion.* This case study confirms that MLLMs can reliably answer structured hypotheses about urban form in most settings, especially when the cues are large, unambiguous, and visually distinct. Remaining errors are interpretable, largely due to either visual resolution or semantic vagueness, and can be systematically mitigated. These findings strengthen confidence in UrbanX 's use of MLLM responses as interpretable, robust inputs for scientific analysis.