# UrbanPhysicalDisorder-4K: Understanding Urban Perception via Counterfactuals and Street View Signs of Physical Disorder

Felipe Moreno-Vera
Fundação Getulio Vargas (FGV)
felipe.moreno@fgv.br

Andres De-la-Puente
Fundação Getulio Vargas (FGV)
andres.puente@fgv.br

Jorge Poco
Fundação Getulio Vargas (FGV)
jorge.poco@fgv.br

***Abstract*—This paper presents a novel framework for explainable urban safety perception analysis, utilizing counterfactual reasoning and providing human-readable interpretations. We leverage a collection of 3,659 street-level images annotated with perceptual safety scores. Unlike traditional segmentation approaches that return only general scene categories, we enrich the visual data with custom manual annotations of urban physical disorder elements (e.g., graffiti, broken infrastructure, overhead cables). Our goal is to classify safety perception from urban imagery and understand the causal impact of specific visual elements. To achieve this, we generate counterfactuals by adding or removing disorder-related elements within the scenes. Rather than relying on vector-based explanations, we translate these counterfactual edits into natural language using a large language model, yielding intuitive insights into how specific elements influence safety perception. Our findings indicate that a subset of disorder elements—particularly overhead cables, garbage, and structural damage—has the greatest impact on the perception of unsafe streets.**

***Index Terms*—Urban safety perception, Street view images, Counterfactuals, Large language models, Urban street disorder**

## I. Introduction

Perceived safety is a key component of urban perception, influencing how individuals experience and evaluate public spaces [1]. Beyond objective measures such as infrastructure quality or population density, the visual appearance of urban scenes plays a significant role in shaping subjective impressions of comfort, order, and safety [2], [3]. Elements such as overhead cables, deteriorating roads, graffiti, and broken windows contribute to a sense of physical disorder, which can negatively affect how urban environments are perceived [4], [5]. Current approaches to predicting urban perception often rely on semantic segmentation models trained on datasets such as Cityscapes or ADE20K, which label common urban elements, including roads, buildings, vehicles, and pedestrians [6]–[8]. These models have been used to encode images as visual features for classification and regression tasks, such as attractiveness (beauty), liveliness, or safety prediction [9], [10]. However, the segmented categories they cover tend to reflect the city's structural layout rather than its perceived quality or condition. As a result, these models overlook disorder-related visual cues that may play a more significant role in shaping subjective impressions of the urban environment.

Moreover, several studies have shown that disorder elements, such as the presence of graffiti, can influence how people perceive safety, cleanliness, and social order in a given area [4], [11], [12]. While not inherently criminal, graffiti is often associated with neglect and social disorganization, and its visibility has been linked to heightened perceptions of crime or risk [13], [14]. As a result, detecting and quantifying graffiti in street-level imagery has become a relevant task for both urban perception analysis and disorder mapping. Recent computer vision approaches have attempted to automate graffiti detection; yet, these efforts rarely integrate perception-based evaluation [15], [16]. Although graffiti is not the only physical disorder element, it is one of the most studied in urban perception literature and often serves as a key visual cue for perceived unsafety.In this work, we extend the existing perspective by incorporating a broader set of disorder-related elements through manual pixel-level annotations. The main contributions of this paper are as follows:

- (i) **Urban Street Physical Disorder (UrbanPD4k) dataset**: A curated set of 3,659 street-level images with manual pixel-level labels identifying 13 types of physical disorder elements.
- (ii) **Key visual elements of safety perception**: An analysis and comparison between the presence of non-disorder elements and disorder elements in urban perception.
- (iii) **Counterfactual and Human-readable interpretations**: Assess how specific disorder elements impact safety perception and identify minimal edits needed for perception shifts.

All **appendices**, supplemental materials, annotations, and code will be available here [1].

## II. Related Work

### A. Urban street disorder elements

Urban street disorder elements have been widely recognized as influential factors shaping how individuals perceive and evaluate urban environments. According to the "broken windows theory" [4], visible signs of disorder—such as graffiti, litter, and damaged infrastructure—can negatively impact perceptions of safety and social stability. Studies in environmental

[1] https://visualdslab.com/papers/UrbanPD4k

psychology and urban design have further shown that people often interpret such visual cues as indicators of neglect or decline, influencing feelings of comfort, trust, and aesthetic appreciation in urban settings [2], [5]. For instance, Nasar and Jones [17] found that participants rated neighborhoods with physical disorder significantly lower in perceived safety and pleasantness. More recent computational approaches have correlated physical disorder elements (graffiti) with the Human Development Index (HDI) and distinguished between graffiti as art versus vandalism in urban contexts [13]–[16]. Despite their relevance, such features are often missing or underrepresented in mainstream urban studies, limiting their utility in perception-centered modeling and analysis.

### B. Urban Perception, visual elements, and explanations

Recent advances in computer vision have enabled the use of segmentation-based representations to study how people perceive urban environments. By dividing street view imagery into object categories—such as buildings, roads, trees, and vehicles—researchers can extract interpretable visual features correlated with perception-based labels like safety, wealth, liveliness, and beauty [6], [7], [10], [18]–[20]. Several works have leveraged datasets like Cityscapes and ADE20K to identify links between specific object compositions and low perceived safety in urban scenes [8], [21], [22]. To further interpret model predictions, methods such as LIME and SHAP have been used to highlight the influence of visual elements on perception outcomes [23]–[28]. However, these efforts often rely on general-purpose segmentation datasets that lack categories related to physical disorders, potentially overlooking subtle yet impactful features—such as damaged infrastructure or overhead cables—that strongly affect how urban scenes are perceived.

**Position of the present work.** Incorporating 13 newly identified urban physical disorder elements into images and using counterfactuals to evaluate their impact on safe and unsafe perceptions offers a promising direction for more accurate and explainable perception predictions. Additionally, for further interpretation, we also incorporate large language models (LLMs) to generate human-readable explanations from counterfactual results, enhancing and enabling a higher-level understanding of how specific visual elements influence safety perception.

## III. Methodology

Our methodology comprises four stages: (i) *Image segmentation and manual annotations*, (ii) *Quantifying Urban perception*, (iii) *CounterFactual explanations*, and (iv) *Human-readable interpretations*.

### A. Image segmentation and manual annotations

We begin with 3,659 street-level scenes from Rio de Janeiro, each paired with a safety perception score provided by the PlacePulse 2.0 survey [9]. To extract structured visual information, we applied the OneFormer model pre-trained on ADE20K [29], which segments each image into 150 semantic classes. For interpretability, we regrouped these into nine higher-level *visual element* categories: **Sky**, **Human**, **Construction**, **Floor**, **Vegetation**, **City elements**, **Terrain vehicles**, **Body of water**, and **Other**, which includes all remaining classes and is excluded from the analysis. Note that some categories—such as water-related or specific construction elements—were not present in the selected imagery and were therefore not included in subsequent analyses (e.g., no occurrences of sea or waterfall). Table I summarizes the regrouped categories. For more details about the **Other** group, see Appendix A.

TABLE I
ADE20K CLASSES AND NEW GROUP-CATEGORIES (VISUAL ELEMENTS)

| Category | # classes | class name |
|---|---|---|
| Construction | 14 | wall, building, ceiling, house |
| | | fence, column, skyscraper, bridge. |
| | | bar, shack, tower, stadium |
| | | fountain, outside door |
| Floor | 7 | floor, road, sidewalk |
| | | ground, sand, path, land |
| Vegetation | 6 | tree, grass, plant, field |
| | | flower, palm |
| Terrain vehicle | 6 | car, bus, truck, van |
| | | motorcycle, bicycle |
| Body water | 5 | water, sea, river, waterfall, lake |
| City elements | 4 | signboard, streetlight |
| | | pole, stoplight |
| Sky | 1 | sky |
| Human | 1 | person |
| Other | 106 | not included |

In parallel, and drawing from psychological and sociological research on urban environments [4], [30] as well as Brazilian studies of urban disorder [13], [15], we identified 13 visual indicators of street disorder—such as overhead cables, graffiti, broken pavement, trash, the presence of homeless individuals, and police vehicles—that reflect both physical decay and social vulnerability. Figure 2 illustrates the process of adding manual annotations, resulting in new masks for the visual elements.

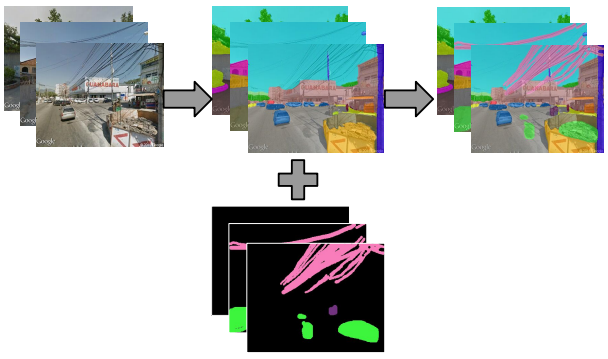

Fig. 1. An example of merging ADE20K segmentation masks and disorder elements manual annotations.

### B. Quantifying urban perception

MIT PlacePulse 2.0 dataset comprises approximately 1.22 million pairwise comparisons across 111,390 images from 56

cities, along with image IDs, geographic coordinates, and comparison outcomes. To preprocess the data and assign perceptual scores to each image, we apply the "strength of schedule" algorithm [31], which estimates a Q-score based on each image's win and loss rates using the following equation:

$$Award_i^k = \frac{1}{w_i^k} \sum_{j=1}^{n_1} \frac{w_i^k}{w_i^k + d_i^k + l_i^k} \tag{1}$$

$$Penalty_i^k = \frac{1}{l_i^k} \sum_{j=1}^{n_2} \frac{l_i^k}{w_i^k + d_i^k + l_i^k} \tag{2}$$

$$Q_i^k = \frac{10}{3}\left(\frac{w_i^k}{w_i^k + d_i^k + l_i^k} + Award_i^k - Penalty_i^k + 1\right) \tag{3}$$

In Equations 1 to 3, $w_i^k$, $d_i^k$, and $l_i^k$ denote the number of times image $i$ was selected as the winner, marked as equal, or marked as the loser in comparisons, respectively. The terms $n_1$ and $n_2$ represent the number of wins and losses for image $i$. The variable $Award$ refers to the average win rate against images that $i$ defeated, while $Penalty$ represents the average loss rate against images that defeated $i$. The final Q-score is scaled to a range from 0 to 10, where lower values indicate perceptions of low safety and higher values correspond to a high perceived safety [5]. In this study, we focus on images from the city of Rio de Janeiro, which were first scored for perceived safety and subsequently annotated with physical disorder elements.

### C. Counterfactuals explanations

Traditional predictive models can highlight correlations between visual features and perception scores, but often lack transparency and actionable insights [20], [26]. Counterfactual explanations address this gap by revealing how minimal visual changes—such as removing disorder elements (e.g., graffiti or overhead cables), or adding other visual elements (e.g, vegetation, construction)—could alter the safety perception.

More formally, we can define it as an optimization problem where the goal is to find a minimal change to the input image $x \in \mathcal{X}$ (through visual elements) that alters the model's prediction of perceived safety $f : \mathcal{X} \to \mathcal{Y}$, from the original outcome $y = f(x)$ to a desired outcome $y'$, using a $x' \in \mathcal{X}$ counterfactual input, subject to a cost function that measures the distance or difference between $x$ and $x'$ called $\mathcal{C}(x, x')$, and $\mathcal{F}$ a set of feasible edits (e.g., adding/removing specific elements). Here, we perform two experiments, counterfactuals based on the presence or absence of the element and the variation in element pixel ratios.

*1) Binary presence/absence:* The images are represented by binary visual feature vectors $x = x_1, x_2, \ldots, x_n$, where each $x_i \in 0, 1$ indicates the presence of a visual element (either present (1) or absent (0)). The goal is to flip as few elements $x_i$ as possible to reach a target prediction:

$$x' = \arg\min_{x' \in \mathcal{F}} \sum_{i=1}^{n} |x_i - x_i'| \quad \text{subject to} \quad f(x') \geq \tau \tag{4}$$

Where $\tau$ is a threshold safety score (e.g., to be classified as "safe").

*2) Pixel ratios:* The images are represented as pixel ratio vectors $x = x_1, x_2, \ldots, x_n$ where $x_i \in \mathbb{R} \mid 0 \leq x \leq 100$. The goal is to find the closest vector to change the prediction (e.g., from unsafe to safe):

$$x' = \arg\min_{x' \in \mathcal{F}} \mathcal{C}(x, x') + \lambda \cdot \mathcal{L}(f(x'), y') \tag{5}$$

Where $\lambda$ is the trade-off coefficient.

### D. Human-readable interpretations

Once $x'$ is obtained, the counterfactual difference $\Delta x = x' - x$ is mapped to textual statements:

$$\text{Interpretation} = \text{LLM}(\Delta x) \tag{6}$$

Where LLM is a large language model that converts added/removed elements into natural language (e.g., "Removing graffiti increased perceived safety").

*1) Binary presence/absence:* Following Equation 4, the difference $\Delta x_i = x_i' - x_i$ can be $\Delta x_i = 1$ if the element was added in the counterfactual, $\Delta x_i = -1$ if the element was removed in the counterfactual, or $\Delta x_i = 0$ if there is no change. Equation 6 can be rewritten as:

$$\text{Explanation} = \text{LLM}\left(\{i \mid \Delta x_i = 1\}, \{j \mid \Delta x_j = -1\}\right) \tag{7}$$

Where $\{i \mid \Delta x_i = 1\}$ is the set of visual elements that were added, and $\{j \mid \Delta x_j = -1\}$ is the set of visual elements that were removed.

*2) Pixel ratios:* Following Equation 5, the difference $\Delta x_i = x_i' - x_i$ can be a real number between -1 and 1. Equation 6 can be rewritten as:

$$\text{Explanation} = \text{LLM}\left(\{i \mid \Delta x_i > \epsilon\}, \{j \mid \Delta x_j < -\epsilon\}\right) \tag{8}$$

Where $\epsilon$ is a threshold to ignore small fluctuations, $\{i \mid \Delta x_i > \epsilon\}$ is the set of visual elements that had a noticeable increase in pixel area in the counterfactual image (e.g., tree coverage increased from 10% to 25%), and $\{j \mid \Delta x_j < -\epsilon\}$ is the set of visual elements that had a noticeable decrease in pixel area (e.g., graffiti coverage dropped from 15% to 2%).

## IV. Experiments & Discussions

Here, we discuss our results and insights obtained from the experiments conducted on street-level images from Rio de Janeiro. We evaluate how urban physical disorder (UDP) elements influence perceived safety.

### A. Urban Physical Disorder (UrbanPD4k) dataset

The UrbanPD4k dataset includes comprehensive manual annotations to support a fine-grained analysis of urban elements in street view imagery. Figure 2 summarizes the annotation statistics through three complementary visualizations. Subfigure (a) presents samples of the generated mask annotations across categories, highlighting the level of human involvement in the labeling process. Subfigure (b) shows the presence of each annotated element, providing insights into how frequently different components of urban physical disorder—such as cables, trash cans, garbage, and graffiti—appear across the dataset. Finally, subfigure (c) illustrates the pixel coverage of

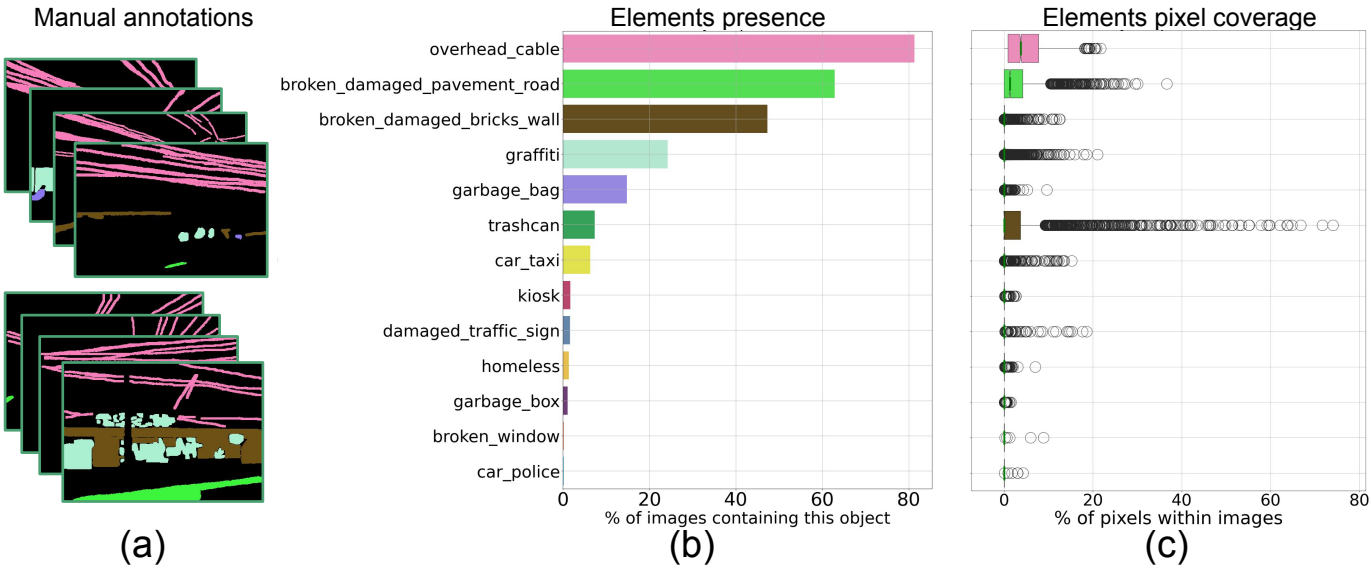


Fig. 2. Overview of the annotation statistics for the UrbanPD4k dataset. (a) Pixel-level annotation mask samples (b) Presence of annotated elements, indicating the frequency of each class in the dataset. (c) Pixel coverage of elements, showing the proportion of the image area occupied by each annotated class. Together, these visualizations highlight the diversity and balance of the annotated content within UrbanPD4k.

these elements, capturing their relative spatial dominance in the scenes. Together, these results reveal that while certain large-scale classes (e.g., buildings and roads) occupy most of the pixel area, smaller but semantically important elements (e.g., damaged walls, broken windows, homeless, and deteriorated traffic signs) appear less frequently but contribute significantly to the visual and functional diversity of the urban environment.

### *B. Physical Disorder Impacts on Urban Perception*

We conduct extensive experiments to investigate the following Research Questions (RQ):

- **RQ1**: Are disorder-related elements more predictive of low safety than general scene features?
- **RQ2**: What are the limitations of using the binary presence or absence of elements in subjective tasks, such as safety perception?
- **RQ3**: Which visual elements have the strongest causal influence on perceived safety and unsafety in urban environments?

*1) **Experimental settings**:* For the LLM experiments, we do not compare outputs from different LLMs, as previous work has already analyzed this and found that the GPT-4 series performs best for generating explanations [32], [33]. Therefore, we use OpenAI's GPT-4o-mini to create diverse and descriptive, human-readable textual interpretations of the counterfactual generations, with a focus on features relevant to urban perception. For counterfactual generations, we use Diverse Counterfactual Explanations (DiCE) [34] with a TensorFlow backend configuration. For all experiments, the dataset is split into 75% for training and validation and 25% for testing. Experiments are conducted on an NVIDIA RTX 3090 GPU with limited VRAM and half precision.

*2) **RQ1: Model performance comparison**:* We conduct experiments under two settings: using binary feature vectors and using pixel ratio values. Additionally, we compare the performance of four Random Forest classifiers trained with different feature configurations: (i) feature vectors composed of all ADE20K segmentation objects; (ii) feature vectors based on our new grouped categories; (iii) feature vectors combining ADE20K objects with our manually annotated disorder elements; and (iv) feature vectors combining our new grouped categories with the disorder annotations. For the labeling process, see Appendix C.

Table II reports the classification results. **We observe that grouping visual features and incorporating disorder elements (as in the best model) significantly improves the model's ability to identify unsafe and safe streets.** In particular, adding disorder elements boosts classification performance by 2–5% and 5–7% for ADE20K + disorder elements in binary and pixel ratios, respectively, and by 2–3% and 5–7% for Visual + disorder elements in binary and pixel ratios, respectively. These findings demonstrate that including relevant cues—such as urban disorder—provides a more informative representation of scene composition, thereby enhancing the model's ability to infer unsafe scenes. Additional details on the feature importance of each object for both binary and pixel ratio features are provided in Appendix E.

*3) **RQ2: SHAP explanations**:* We applied SHAP explanations [35] to two different feature representations of urban objects extracted via semantic segmentation: binary presence/absence and pixel ratio (i.e., the proportion of the image area). In the pixel ratio representation, features are continuous values indicating the percentage of the image occupied by each object class. This allows SHAP to compute contributions for all object classes, regardless of whether they dominate the

TABLE II
CLASSIFICATION RESULTS USING BINARY AND PIXEL RATIOS CONFIGURATIONS

| Setting | Categories | Perception | Metrics | | |
|---|---|---|---|---|---|
| | | | Precision | Recall | F-1 |
| Binary values | ADE20K (150 classes) | not safe | 0.59 | 0.65 | 0.62 |
| | | safe | 0.60 | 0.55 | 0.57 |
| | AE20K+Disorder (163 classes) | not safe | 0.64 | 0.65 | 0.65 |
| | | safe | 0.67 | 0.60 | 0.63 |
| | Visual elements (8 groups-classes) | not safe | 0.52 | 0.42 | 0.46 |
| | | safe | 0.51 | 0.61 | 0.55 |
| | Visual elements+Disorder (22 classes) | not safe | 0.66 | 0.71 | 0.67 |
| | | safe | 0.65 | 0.66 | 0.66 |
| Pixel ratios | ADE20K (150 classes) | not safe | 0.67 | 0.73 | 0.69 |
| | | safe | 0.71 | 0.63 | 0.67 |
| | AE20K+Disorder (163 classes) | not safe | 0.70 | 0.76 | 0.73 |
| | | safe | 0.73 | 0.68 | 0.70 |
| | Visual elements (8 groups-classes) | not safe | 0.69 | 0.72 | 0.70 |
| | | safe | 0.70 | 0.64 | 0.66 |
| | Visual elements+Disorder (22 classes) | not safe | 0.72 | 0.77 | **0.75** |
| | | safe | 0.75 | 0.69 | **0.72** |

image or are only marginally present. Appendix F-Figure 3 provides richer insights, capturing both positive and negative contributions of urban elements. Notably, disorder-related elements (e.g., trash, graffiti, broken sidewalks) are often assigned negative SHAP values, reflecting their association with lower perceived safety. In contrast, well-maintained or green elements typically contribute positively to the environment. Conversely, in the binary case, each object class is represented as a binary indicator (1 if present, 0 if absent). SHAP explanations under this setting only assign meaningful values to the objects present in the image. In contrast, absent objects receive near-zero attribution, as the model has no variation to assess their marginal contribution. As a result, SHAP outputs in this case are limited to a sparse subset of objects, which may hinder full interpretability. On the other hand, using pixel ratios allows us to gain deeper insights into objects. In both cases, SHAP explanations suggest that disorder-related features are strongly associated with perceptions of unsafety, with physical disorder elements having a negative impact on perceptions of safety.

*4)* ***RQ3: Counterfactuals explanations****:* From the SHAP results, we know that disorder elements are more strongly associated with "unsafe" predictions. However, we do not fully understand their causal influence. To better understand how specific visual elements influence perceived urban safety and to determine which elements can shift the perception of a street scene from "unsafe" to "safe" (or vice versa), we employ counterfactual explanations. The primary goal is to identify which features—whether represented as binary presence/absence or pixel ratio values—have the strongest causal impact on classification outcomes.

To quantify the causal importance of each feature, we generate 100 counterfactuals to change "unsafe" to "safe" samples using the DiceML library. We configure for $0$ and $1$ values to change in the binary case, and $[0-1]$ range values for pixel ratios. Appendix G-Figure 4 presents the number of feature changes per sample between the binary and pixel-ratio scenarios across the 100 generated counterfactuals for each sample. Features that are frequently modified are considered more causally influential, as they are often involved in the minimal transformations required to alter the model's decision. We observe that disorder-related elements, such as overhead cables, graffiti, damaged walls, trash cans, and visual elements like vegetation (plants, trees), people, and cars, exhibit the highest number of variations. This result suggests that changes in disorder-related factors are essential to determine whether a scene is "unsafe" or "safe".

### *C. Human-Readable Interpretations*

To enhance the interpretability of safety perception predictions and counterfactual suggestions, we employed a LLM to convert counterfactuals into natural language explanations. These edits were derived by identifying the semantic elements that most strongly influenced the model's predictions. Each counterfactual instance for interpretation was selected by analyzing object categories whose pixel ratios or binary presence significantly impacted the prediction outcome while remaining as close as possible to the original vector sample. The resulting vectors were then mapped to textual explanations via prompting. Following the recommendations of [32], [36], our prompt was structured to generate explanations in plain language for both cases. See Appendix D for prompt details.

We applied the prompt to the closest and least-modified counterfactual pair for each sample, aiming to obtain a natural language explanation of how the relative increase or addition, or decrease or removal, of certain visual elements may influence the perception of safety. Additionally, we define $\Delta$ safety as the difference in the probability of being classified as "safe" after applying the counterfactual. We select unsafe samples with both high and low prediction probabilities, corresponding to images that received a low safety score (between zero

TABLE III
LLM HUMAN-READABLE INTERPRETATIONS FOR COUNTERFACTUALS

| Image |  |  |  |
|---|---|---|---|
| $\Delta$ safety probability | 0.45 | 0.25 | 0.04 |
| Objects Removed | graffiti, garbage, damaged sidewalk | overhead cables, damaged road | garbage bags |
| Objects Added | trees, walls, fences | Trees, grass | tree, grass, road |
| LLM interpretation | It's better to remove graffiti from walls, repair brick walls, and avoid overhead cables | The overhead cables should be removed, and adding grass and trees along the road will help increase people's sense of safety. | This street is already in good condition, but by incorporating more greenery and eliminating garbage bags, it could be improved even further. |

and four). Table III shows the results of applying LLM-generated, human-readable interpretations for counterfactual explanations in both cases: pixel ratios (proportional coverage) and binary (presence/absence). We note that for the high-score (safe) image, changes are minimal; in contrast, for the middle- and low-score images, the changes (additions/removals of elements) are more significant, highlighting how visual content adjustments contribute more noticeably to perceived safety in less safe scenes.

## V. LIMITATIONS

### A. Subjectivity of Safety Scores

The safety scores used in our model are derived from crowdsourced perceptual judgments via the PlacePulse 2.0 platform. These scores reflect general public opinion rather than objective safety metrics or the views of local residents. The subjectivity inherent in perception data introduces noise and variability, which may affect model predictions and interpretations. A way to solve this is by incorporating local surveys, crime data, or demographic context to triangulate perception with lived experience and ground truth.

### B. Manual Annotation Scalability

Our analysis benefits from manually annotated physical disorder elements, which offer more targeted insights than general segmentation categories. However, the annotation process is time-consuming and subject to annotator bias. Scaling this approach to other cities or larger image sets would require significant effort or the adoption of semi-supervised labeling techniques.

## VI. ETHICAL IMPLICATIONS

The dataset used in this study consists of images obtained through the Google Street View API, which provides publicly accessible imagery under Google's Terms of Service. All usage of this data complies with the corresponding licensing and attribution requirements. Google Street View applies automatic blurring to faces and license plates, and we additionally verified that no unblurred identifiable elements were present in the images selected for annotation. Since the images originate from a platform specifically intended for public use and do not contain identifiable personal data, informed consent from individuals was not required. All manual annotations created for this study were limited exclusively to non-personal objects of interest relevant to the research. These annotations do not involve, derive, or expose any personal information. The UrbanDP4k dataset, including urban physical disorder annotations and geographic coordinates (latitude and longitude), will be available here [2].

## VII. CONCLUSIONS

This study introduces a novel approach to analyzing urban safety perceptions by combining semantic segmentation, manual annotation, counterfactual reasoning, and natural language explanations using LLMs. By analyzing 3,659 street-level scenes from Rio de Janeiro, we demonstrate how specific visual elements (e.g., graffiti, broken infrastructure, and overhead cables) contribute to perceptions of unsafety. We compare two encoding methods—binary presence/absence and pixel ratios—to assess their impact on safety judgments. Counterfactual analysis reveals that while both approaches identify influential elements, pixel ratios offer a more nuanced understanding of how subtle visual changes affect perception. By translating these insights into human-readable explanations, the study provides actionable guidance for urban planners and policymakers, highlighting the importance of considering physical disorder in urban safety assessments.

Finally, by translating visual edits into human-readable language, these explanations provide urban planners, policymakers, and designers with interpretable, scenario-based insights to support targeted interventions. These results highlight the value of integrating visual and textual reasoning for actionable

[2]https://visualdslab.com/papers/UrbanPD4k

urban analytics, showing that physical disorder—often overlooked in general-purpose segmentation—carries strong perceptual signals and can significantly shape safety perception.

## ACKNOWLEDGEMENT

This work was supported by the National Council for Scientific and Technological Development (CNPq, grant #311144/2022-5), Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State (FAPERJ, grants #E-26/204.593/2024 & #E-26/210.585/2025), São Paulo Research Foundation (FAPESP, grants #2021/07012-0 & #2024/05760-8), Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES, grant #88887.684234/2022-00), and the School of Applied Mathematics at Fundação Getulio Vargas (FGV).

## REFERENCES

[1] H. W. Schroeder and L. M. Anderson, "Perception of personal safety in urban recreation sites," *Journal of leisure research*, vol. 16, no. 2, pp. 178–194, 1984.

[2] R. S. Ulrich, "Visual landscapes and psychological well-being," *Landscape research*, vol. 4, no. 1, pp. 17–23, 1979.

[3] A. Rapoport and R. Hawkes, "The perception of urban complexity," *Journal of the American Institute of Planners*, vol. 36, no. 2, pp. 106–111, 1970.

[4] J. Q. Wilson and G. L. Kelling, "Broken windows," *Atlantic monthly*, vol. 249, no. 3, pp. 29–38, 1982.

[5] J. L. Nasar, "The evaluative image of the city," 1998.

[6] X. Zhang, S. Li, Y. Zhang, T. Yao, H. Han, L. Sun, and R. Stouffs, "Human-centric interpretable visual evaluation of urban street based on multimodal perception data: A case study of shanghai."

[7] Q. Cui, Y. Zhang, G. Yang, Y. Huang, and Y. Chen, "Analysing gender differences in the perceived safety from street view imagery," *International Journal of Applied Earth Observation and Geoinformation*, 2023.

[8] F. Moreno-Vera, B. Lavi, and J. Poco, "Urban perception: Can we understand why a street is safe?" in *Mexican International Conference on Artificial Intelligence*. Springer, 2021, pp. 277–288.

[9] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," in *ECCV*, 2016.

[10] Z. Ma, "Deep exploration of street view features for identifying urban vitality: A case study of qingdao city," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 123, p. 103476, 2023.

[11] K. Keizer, S. Lindenberg, and L. Steg, "The spreading of disorder," *Science (New York, N.Y.)*, vol. 322, pp. 1681–5, 12 2008.

[12] M. Albaik, "Influence of graffiti on people's perceptions of urban spaces in hashemi shamali, amman, jordon," *ISVS e-journal*, vol. 10, no. 7, pp. 68–90, 2023.

[13] A. M. A. Diniz and M. C. Stafford, "Graffiti and crime in belo horizonte, brazil: The broken promises of broken windows theory," *Applied Geography*, vol. 131, p. 102459, 2021.

[14] J. R. Alzate, M. S. Tabares, and P. Vallejo, "Graffiti and government in smart cities: a deep learning approach applied to medellin city, colombia," in *International Conference on Data Science, E-learning and Information Systems 2021*, 2021, pp. 160–165.

[15] B. Lavi., E. Tokuda., F. Moreno-Vera., L. Nonato., C. Silva., and J. Poco., "17k-graffiti: Spatial and crime data assessments in são paulo city," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SciTePress, 2022, pp. 968–975.

[16] E. K. Tokuda, C. T. Silva, and R. M. C. Jr., "Quantifying the presence of graffiti in urban environments," pp. 1–4, 2019.

[17] J. L. Nasar and K. M. M. Jones, "Landscapes of fear and stress," *Environment and Behavior*, vol. 29, pp. 291 – 323, 1997.

[18] Z. Li, P. Liu, J. Shi, and Y. Xing, "Research on street space quality combined with attention multi-task deep learning," *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*, pp. 434–441, 2021.

[19] F. Moreno-Vera, B. Lavi, and J. Poco, "Quantifying urban safety perception on street view images," in *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2021.

[20] J. Xu, Q. Xiong, Y. Jing, L. Xing, R. An, Z. Tong, Y. Liu, and Y. Liu, "Understanding the nonlinear effects of the street canyon characteristics on human perceptions with street view images," *Ecological Indicators*, vol. 154, p. 110756, 2023.

[21] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti, "Measuring human perceptions of a large-scale urban region using machine learning," *Landscape and Urban Planning*, vol. 180, pp. 148–160, 2018.

[22] X. Xu, W. Qiu, W. Li, X. Liu, Z. Zhang, X. Li, and D. Luo, "Associations between street-view perceptions and housing prices: Subjective vs. objective measures using computer vision and machine learning techniques," *Remote. Sens.*, vol. 14, p. 891, 2022.

[23] H. Ma, J. Li, and X. Ye, "Deep learning meets urban design: Assessing streetscape aesthetic and design quality through ai and cluster analysis," *Cities*, 2025.

[24] F. Moreno-Vera, B. Brandoli, and J. Poco, "What makes a place feel safe? analyzing street view images to identify relevant visual elements," in *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2024.

[25] C. Wu, Y. Liang, M. Zhao, M. Teng, Y. Han, and Y. Ye, "Perceiving the fine-scale urban poverty using street view images through a vision-language model," *Sustainable Cities and Society*, 2025.

[26] F. Moreno-Vera, "Understanding safety based on urban perception," in *International Conference on Intelligent Computing*. Springer, 2021, pp. 54–64.

[27] J. Luo, P. Liu, W. Xu, T. Zhao, and F. Biljecki, "A perception-powered urban digital twin to support human-centered urban planning and sustainable city development," *Cities*, 2025.

[28] F. Moreno-Vera and J. Poco, "Assessing urban environments with vision-language models: A comparative analysis of ai-generated ratings and human volunteer evaluations," in *2025 IEEE International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[29] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2989–2998.

[30] K. Lynch, "Reconsidering the image of the city," pp. 151–161, 1984.

[31] J. Park and M. Newman, "A network-based ranking system for us college football," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, pp. P10 014 – P10 014, 2005.

[32] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez, "Debating with more persuasive llms leads to more truthful answers," in *41st International Conference on Machine Learning*, 2024.

[33] A. Zytek, S. Pido, S. Alnegheimish, L. Berti-Équille, and K. Veeramachaneni, "Explingo: Explaining ai predictions using large language models," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 1197–1208.

[34] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," *2020 Conference on Fairness, Accountability, and Transparency*, 2019.

[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Neural Information Processing Systems*, 2017.

[36] Y. Wang, L. Zhou, Y. Wang, and Z. Peng, "Leveraging pretrained language models for enhanced entity matching: A comprehensive study of fine-tuning and prompt learning paradigms," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 1941221, 2024.

## APPENDIX

### A. Other Grouped Categories

TABLE I
ADE20K CLASSES AND GROUPS NOT INCLUDED

| Other Categories | # classes | class name |
|---|---|---|
| Indoor elements | 73 | bed, cabinet, door, table curtain, fan, crt screen, plate monitor, shower, ... |
| Outdoor elements | 11 | windowpane, grandstand, runway screen door, bench, stairway awning, poster, tent swimming pool, steps |
| Nature elements | 3 | mountain, rock, hill |
| Clothes elements | 2 | clothes, bag |
| Sea vehicle | 2 | boat, ship |
| Sea vehicle related | 1 | pier |
| Air vehicle | 1 | airplane |
| Terrain vehicle related | 1 | radiator |
| Animal | 1 | animal |
| Miscellaneous | 11 | blind, stage, basket food, trade, flowerpot, sculpture bulletin, glass, clock, flag |

### B. Images Comparisons to Scores

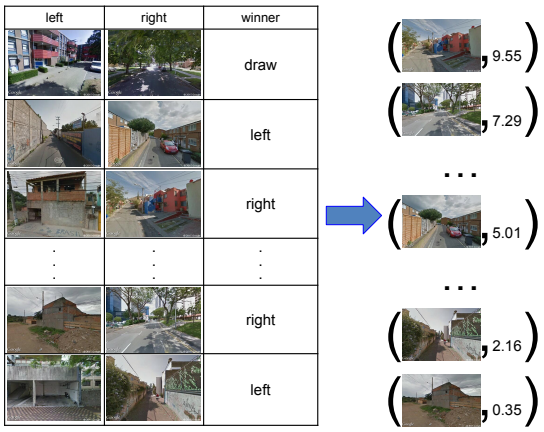


Fig. 1. An example of processing comparisons to perceptual scores using the "strengh of Schedule" algorithm.

### C. Labeling process

We use the score $Q_i^k$ calculated in Equation 3. We define class labels using a threshold based on the mean and standard deviation, using a $\delta = 0.8$:

$$y_{i,k} = \begin{cases} 1 & \text{if } Q_i^k > \mu^k + \delta\sigma^k\% \\ 0 & \text{if } Q_i^k < \mu^k - \delta\sigma^k\% \end{cases} \quad (1)$$

### D. Prompt

```
[System]: You are a person evaluating
street images based on their
visual appearance, tasked with
improving street safety perception.

[User]: Imagine describing how a street
scene might feel with visible changes
in certain elements.
We want you to describe which elements
should be removed/reduced and which
should be added/increased to improve
the safety perception.

[System]: These changes are derived
from comparing two representations
of the same scene:
original values and counterfactual values.

[User]: The counterfactuals give us
the following insights:

[Added or Increased Elements]:
{\\n.join(added) if added else 'None'}

[Removed or Decreased Elements]:
{\\n.join(removed) if removed else 'None'}
## End of Python code

Based on these changes, please write a
concise explanation (max 50 words)
describing how these changes might
affect your sense of safety.
```

### E. Feature importance

Figure 2 presents the computed feature importance. The results clearly indicate that the physical disorder elements play a significant role in shaping the model's predictions. The consistent importance of disorder-related elements suggests that they are especially informative in identifying environmental irregularities or anomalies that may signal unsafe conditions.

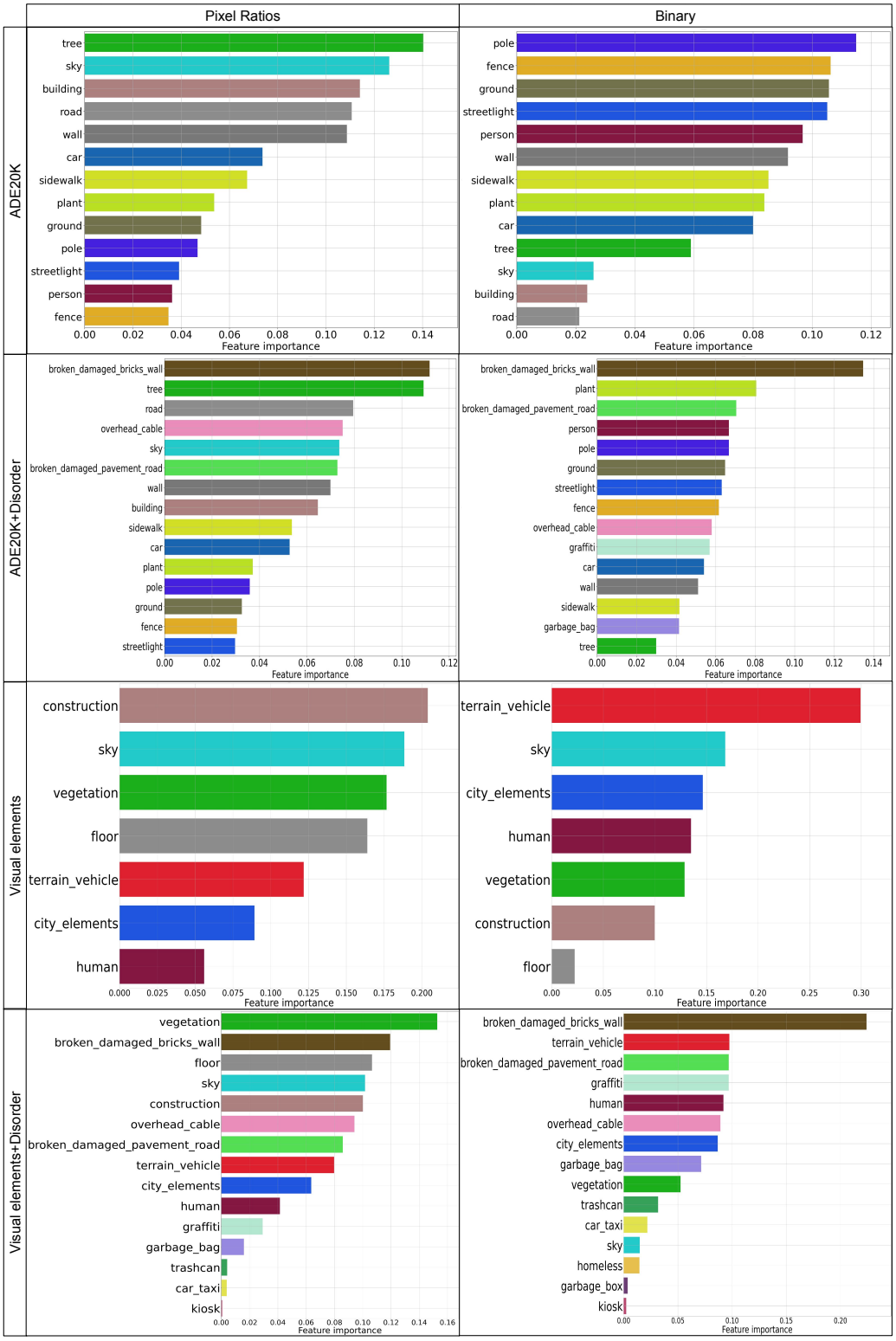


Fig. 2. Feature importance of the top 15 elements in all models evaluated.

### F. SHAP explanations

Figure 3 shows the SHAP explanations for both methods—pixel ratios (continuous values) and binary presence/absence values—indicating that pixel ratios provide more information and have a greater influence on model inferences. In contrast, binary values have a negligible impact and only on a subset of the features.

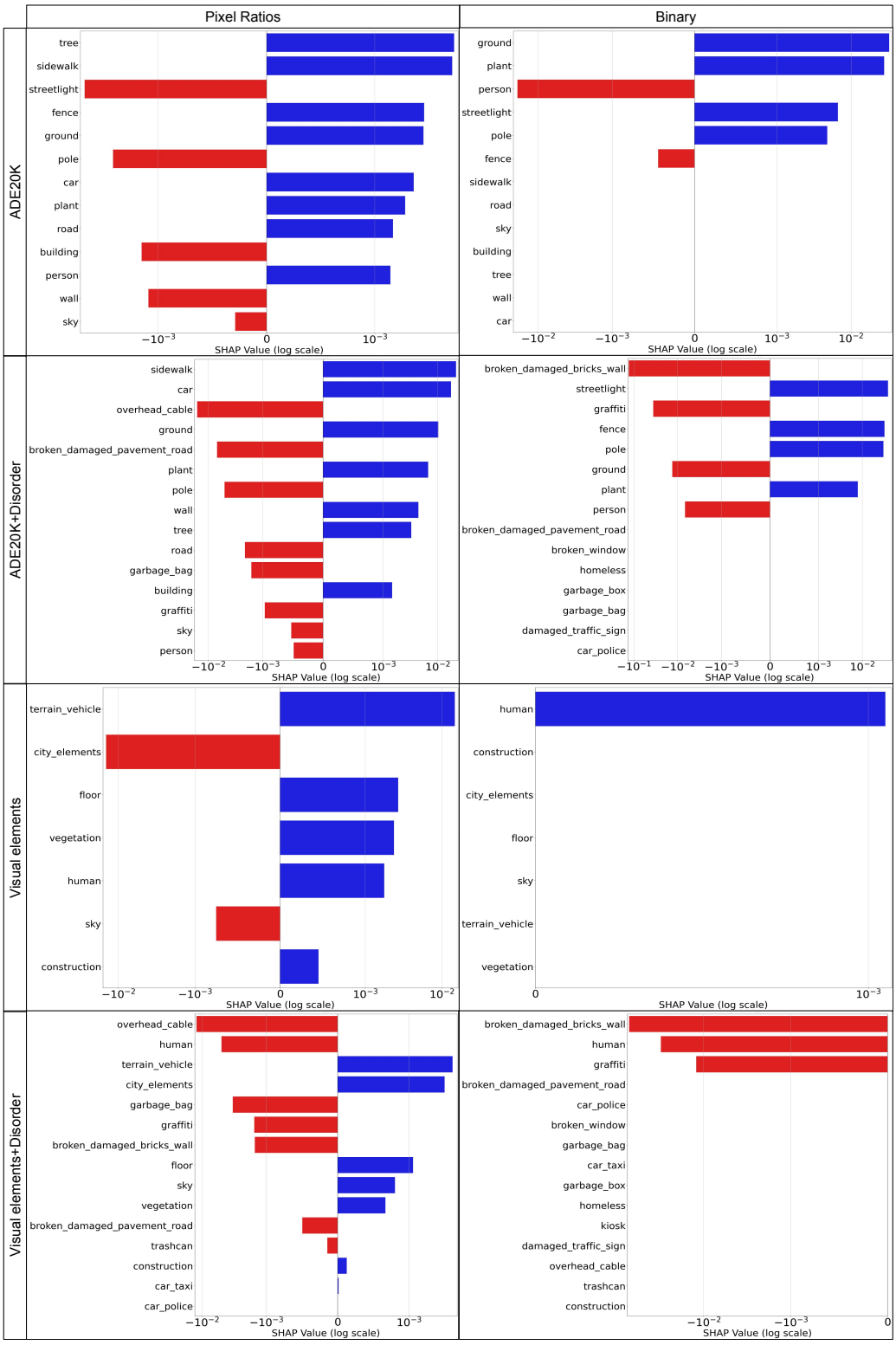


Fig. 3. SHAP values for both: binary presence and pixel ratios values.

## G. *Counterfactuals generations*

Figure 4 presents counterfactual generations for all unsafe samples. We compare binary and pixel ratio cases by measuring how often each object changes its value—presence or absence in the binary case, and percentage change in the pixel ratio case.

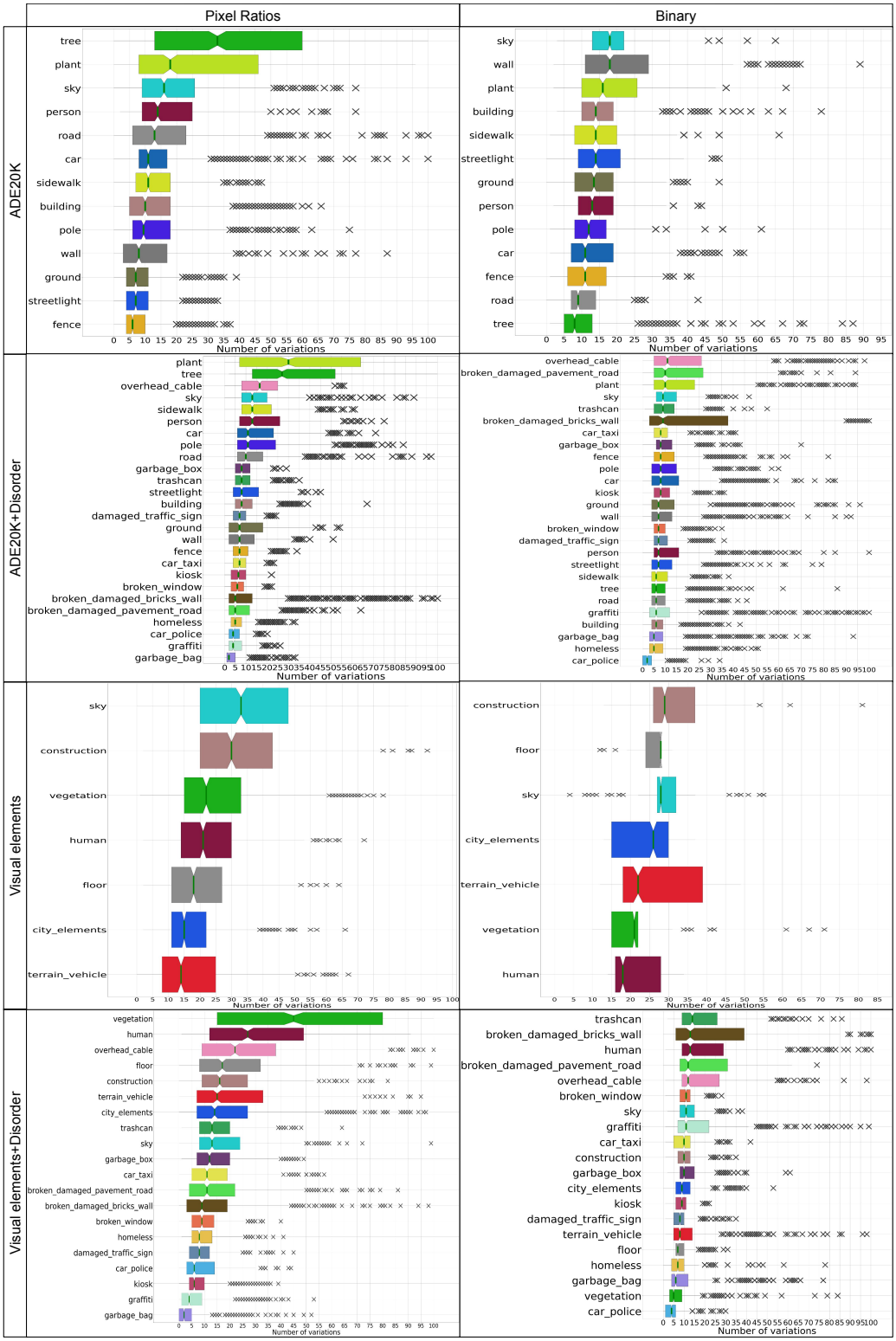


Fig. 4. Measure influence of elements by generating 100 Counterfactuals generations.