FUNDAÇÃO GETULIO VARGAS | EMAp ESCOLA DE MATEMÁTICA APLICADA

Exploring Urban Safety Perception with Vision-Language Models and Semantically Guided Image Counterfactuals

Felipe A. Moreno-Vera

Advisor: Prof. Jorge Poco Medina

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPESP FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

About me

M.Sc. Felipe A. Moreno-Vera

www.fmorenovr.com

Content

- Context & Motivation
- MAP — What do we know about this field?
- IDENTIFY — Which visual elements actually matter?
- TRANSLATE — What would make this street feel safer?
- SHOW — What would those changes look like?
- Conclusions

Context & Motivation

What is urban perception?

What is urban perception?

Would you walk here at night?

Would you rent an apartment on this street?

Does it feel taken care of?

Would you rather walk alone, or sit here and sigh by yourself?

What is urban perception?

Perception forms in milliseconds — before reasoning *(Lynch, 1960).*

Would you walk here at night?

Would you rent an apartment on

Would you rather walk alone, or sit here and sigh by yourself?

What is urban perception?

Urban perception

as the instant and automatic impression produced by a place from its visual appearance alone.

Cities are lived through perception

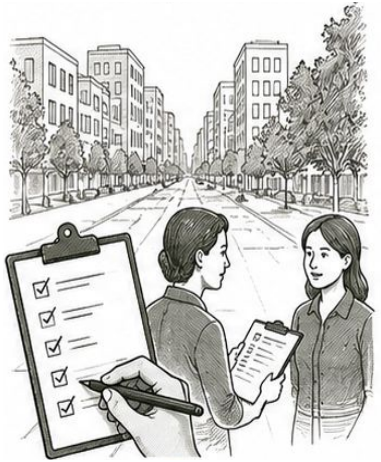
- **Visual elements give streets identity** — cities are experienced through what people see (Lynch, **1960**; Jacobs, **1961**; Rapoport, **1970**; Ulrich, **1979**)

- **Perceived disorder and neglected** signals **unsafety** (Wilson & Kelling, **1982;** Lynch, **1983**; Schroeder, **1984**)

- **Perceived lack of control** feeds fear and **insecurity in the city** (Kaplan, **1989**; Skogan, **1992**; Nasar, **1998**)

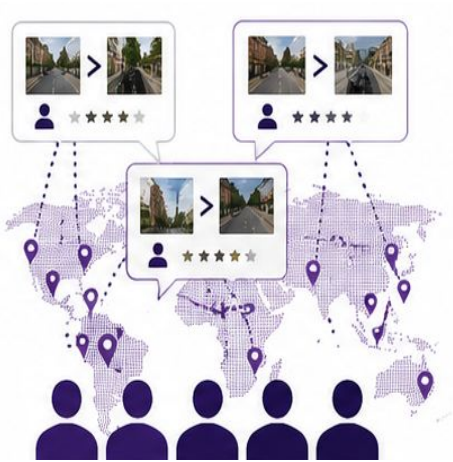
- **Perceiving greenery and maintenance** produces **safety** and invites use of space (Sampson, **2002**; Keizer, **2008**; Hinkle, **2009**)

Perception at city scale: How it evolved?



1960s-1990s
personal interviews

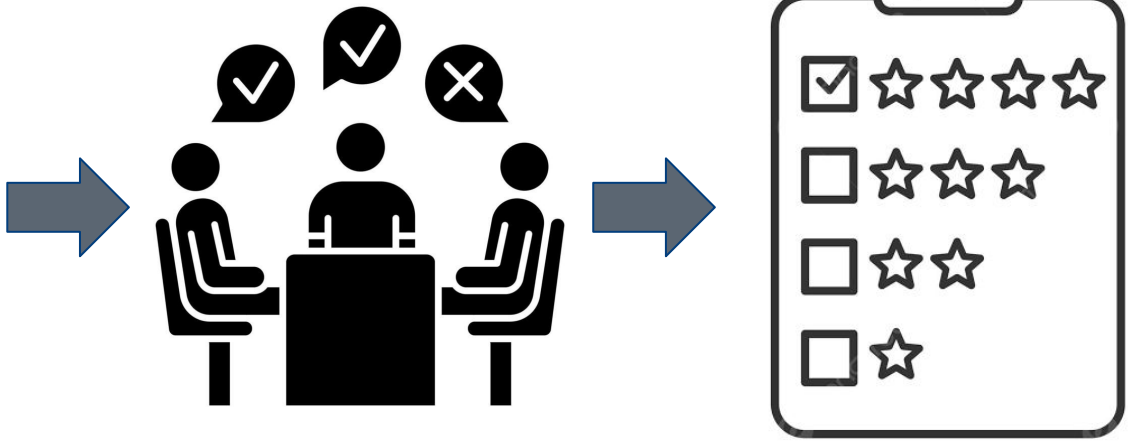
2000s-2010s
periodical surveys



2011-2016
Online crowdsourcings

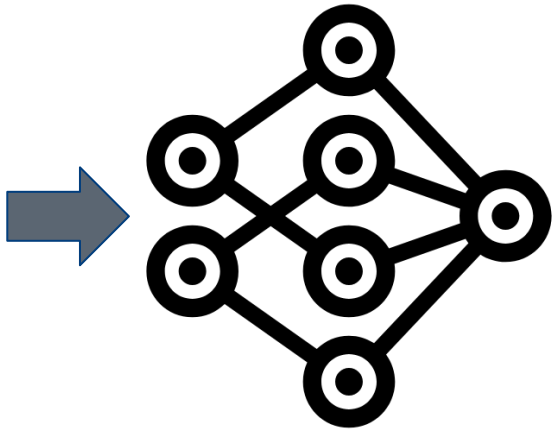
2017-2020s
Worldwide approach

Perception became assessable



— **“Is this place safe?”**

Perception became assessable — then predictable

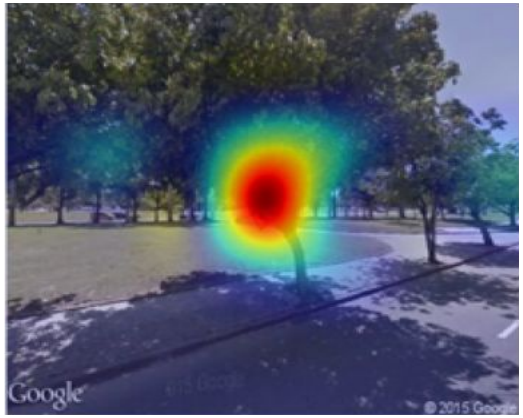


Predicted score: 7.8

- **2014** — Streetscore: first city-scale prediction of perceived safety from SVI
- **2016–2019** — Deep CNNs pre-trained on Place Pulse 2.0 predict perception across dozens of cities
- **2020s** — Vision-language and generative models can **predict, describe images — and even *modify*** — what a scene shows

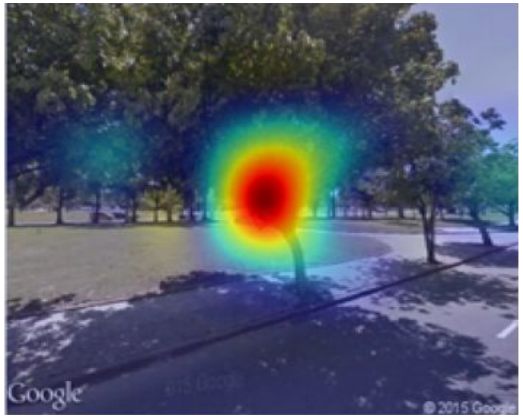
Perception became predictable — then explainable

Predicted score: 7.8



- Black box models
- **Gradient-based saliency maps (grad-cam)**
  — **“where a models looks”**

Perception became explainable — but not actionable



What would make this street feel different?

- Black box models
- Gradient-based saliency maps (e.g., grad-cam)
  — **“where a models looks”** but no **“what it is looking for”**

This thesis

Urban safety perception should be treated as a

grounded visual analytics problem

— not only a prediction problem.

Throughout: safety = PERCEIVED safety

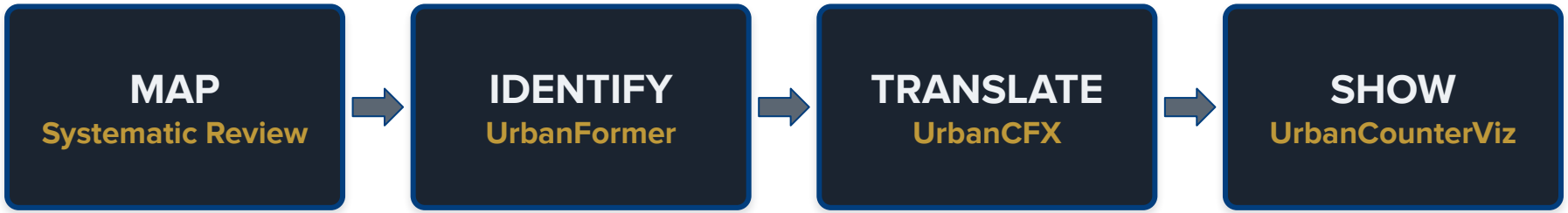
a human judgment from visual appearance — not crime rates, not objective risk, not driving safety

Goal

Investigates computational methods to generate ***hypotheses*** *about urban perception changes*

Overview

“What makes a street feel safe — and what visual evidence supports the prediction?”



- **MAP** the field
- **IDENTIFY** perceptually relevant elements
- **TRANSLATE** them into human-readable language
- **SHOW** the changes — visible and inspectable

Contributions

Contribution	Venue / status	Role
Systematic Literature Review (2,307 → 234 studies)	Submitted to TETCI journal	Methodological — the map
UrbanFormer — visual elements of perceived safety	Published in IEEE WI-IAT' 24	Core — identify
UrbanCFX — human-readable language	Published in IEEE BigData' 25	Core — translate
UrbanCounterViz — VA system + interpretable workflow	To be submitted to EuroVis' 27	Core — show (central)
UrbanVLM — VLM vs. human perception ratings	Published in IEEE IJCNN '25	Complementary — supporting

"Four contributions: one review that defines the map, three empirical works that walk it."

Complementary work informed the VLM scoring comparisons — it is not a thesis pillar.

Mapping the field

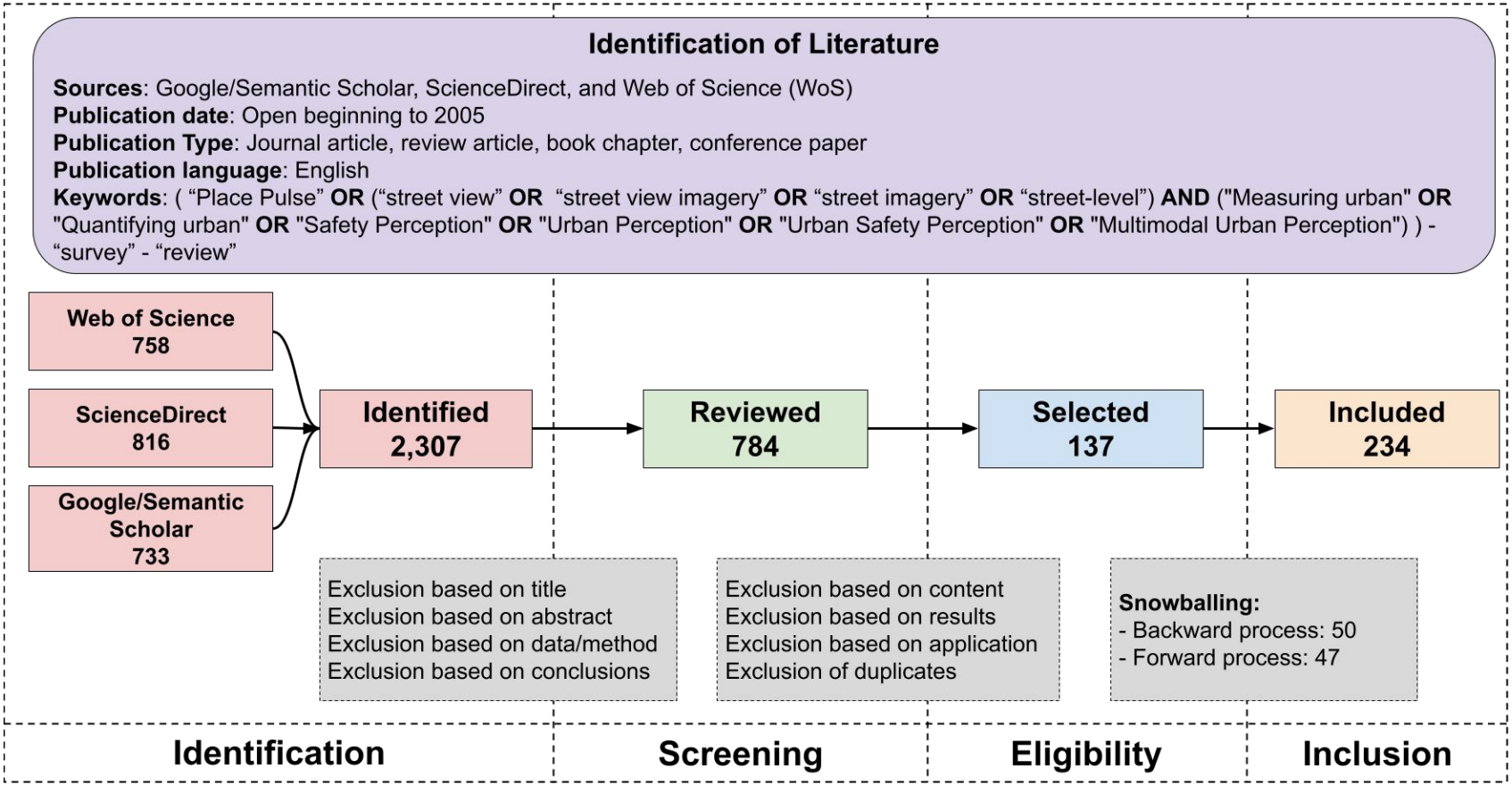
What do we know about this field? – We need to map it

- Systematic Review:
  – Find and analyze studies in urban perception research
  – Study trends
  – Identify gaps

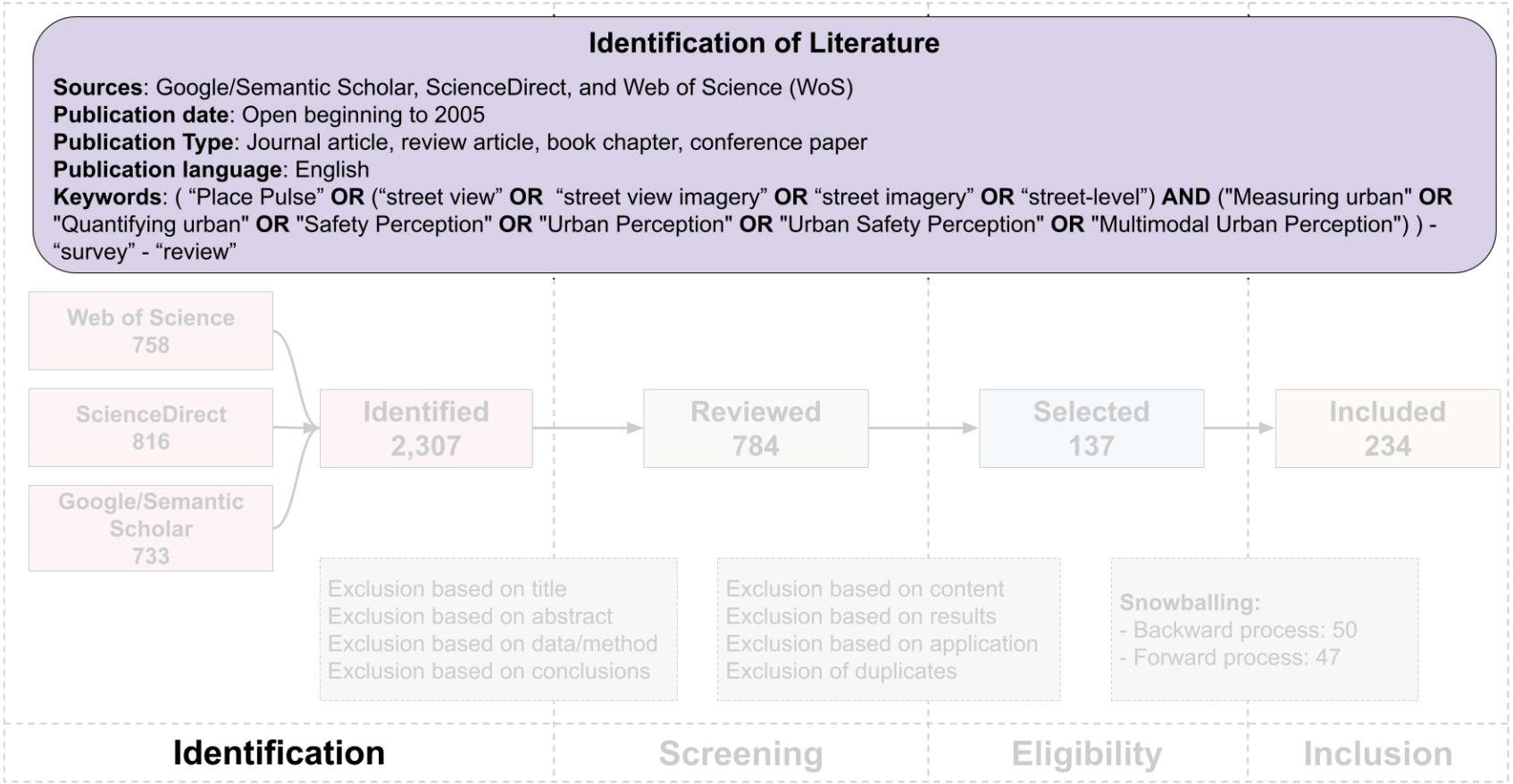
- Preferred Reporting Items for Systematic reviews and Meta-Analyses **(PRISMA)**
  – Identification, screening, eligibility, and inclusion

- **Snowballing**:
  – **Backward**: starts with the references of identified papers and works backward.
  – **Forward**: explores the papers that have cited the identified papers.

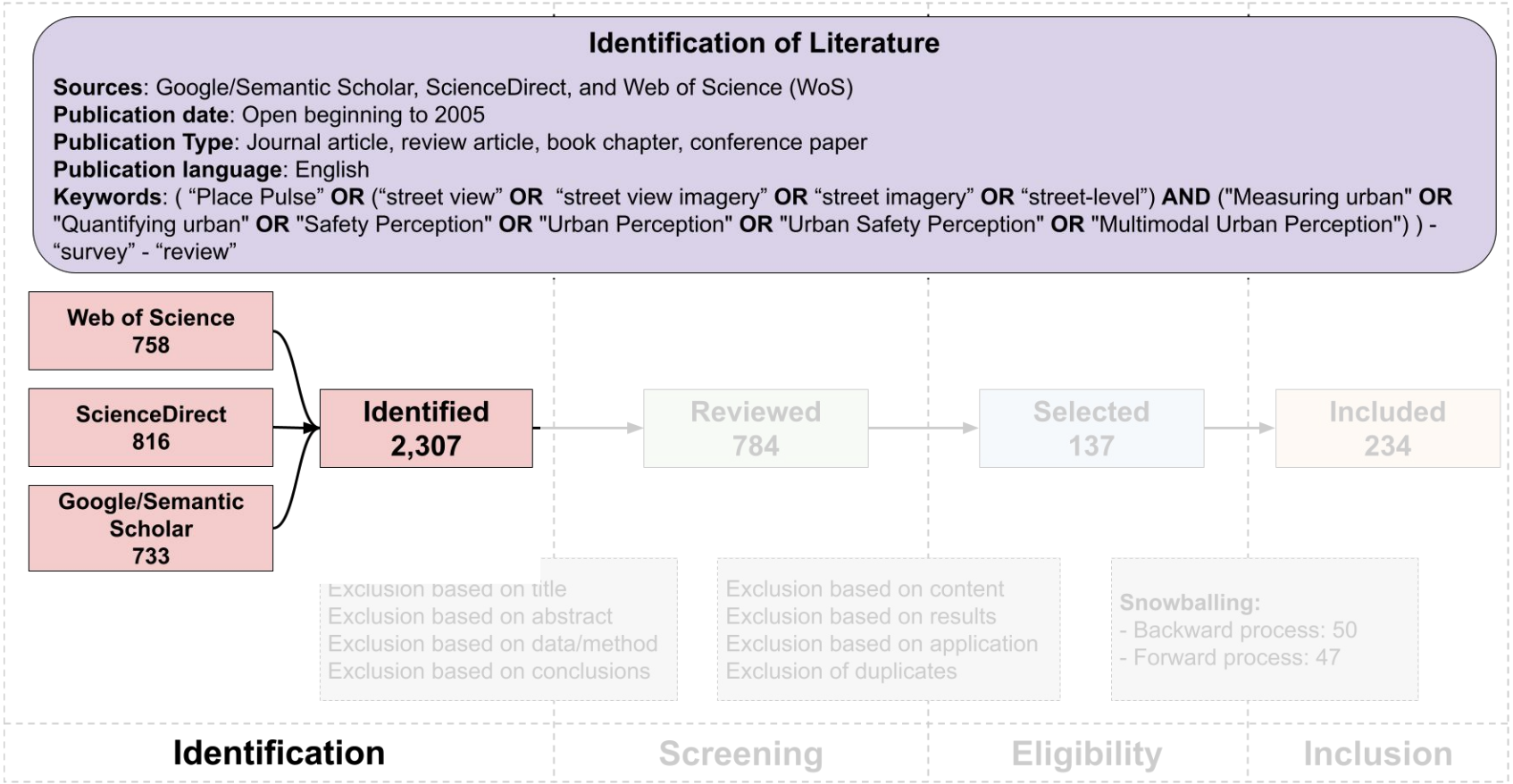
Mapping the field: PRISMA, 2005–2026



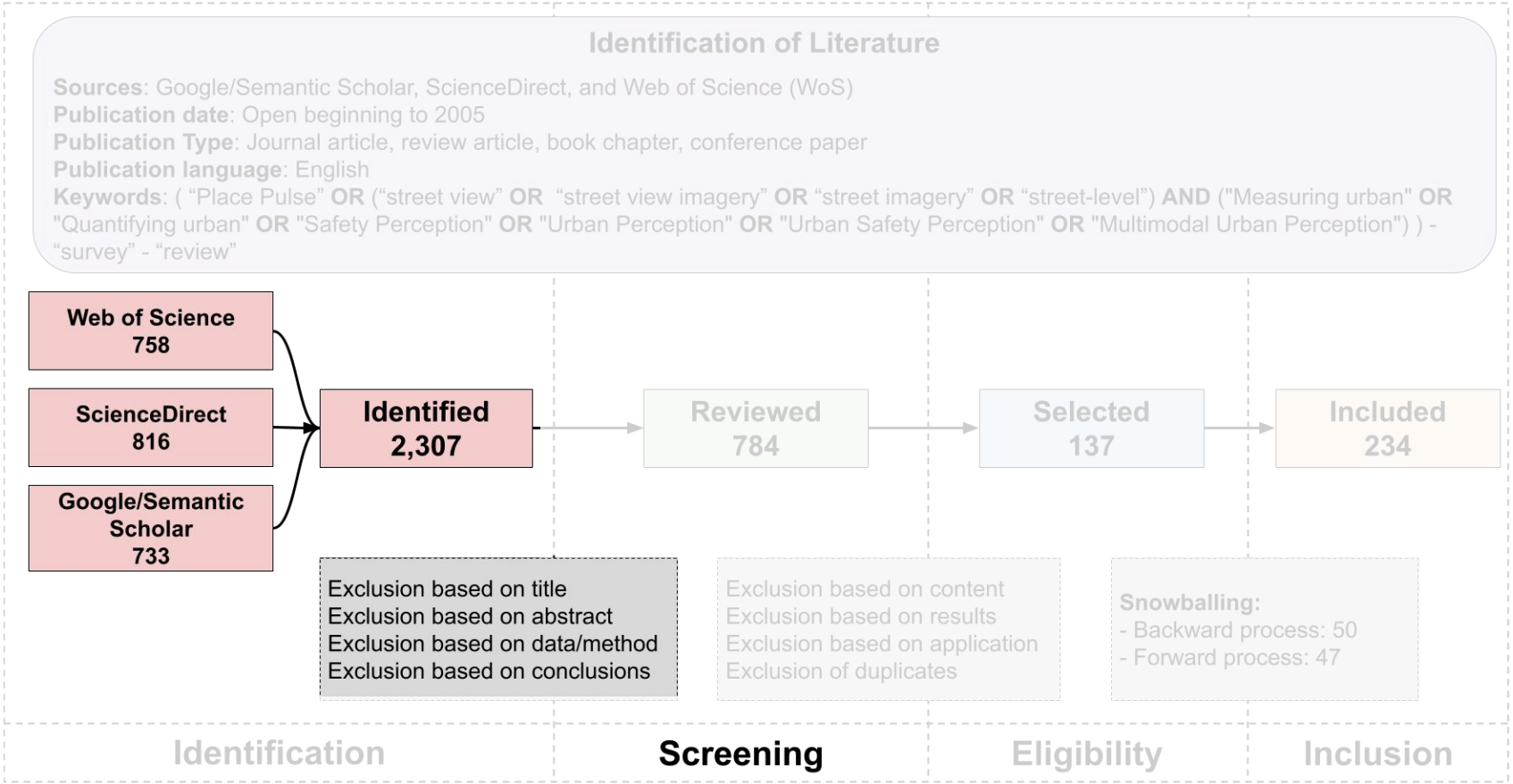
Mapping the field: PRISMA, 2005–2026



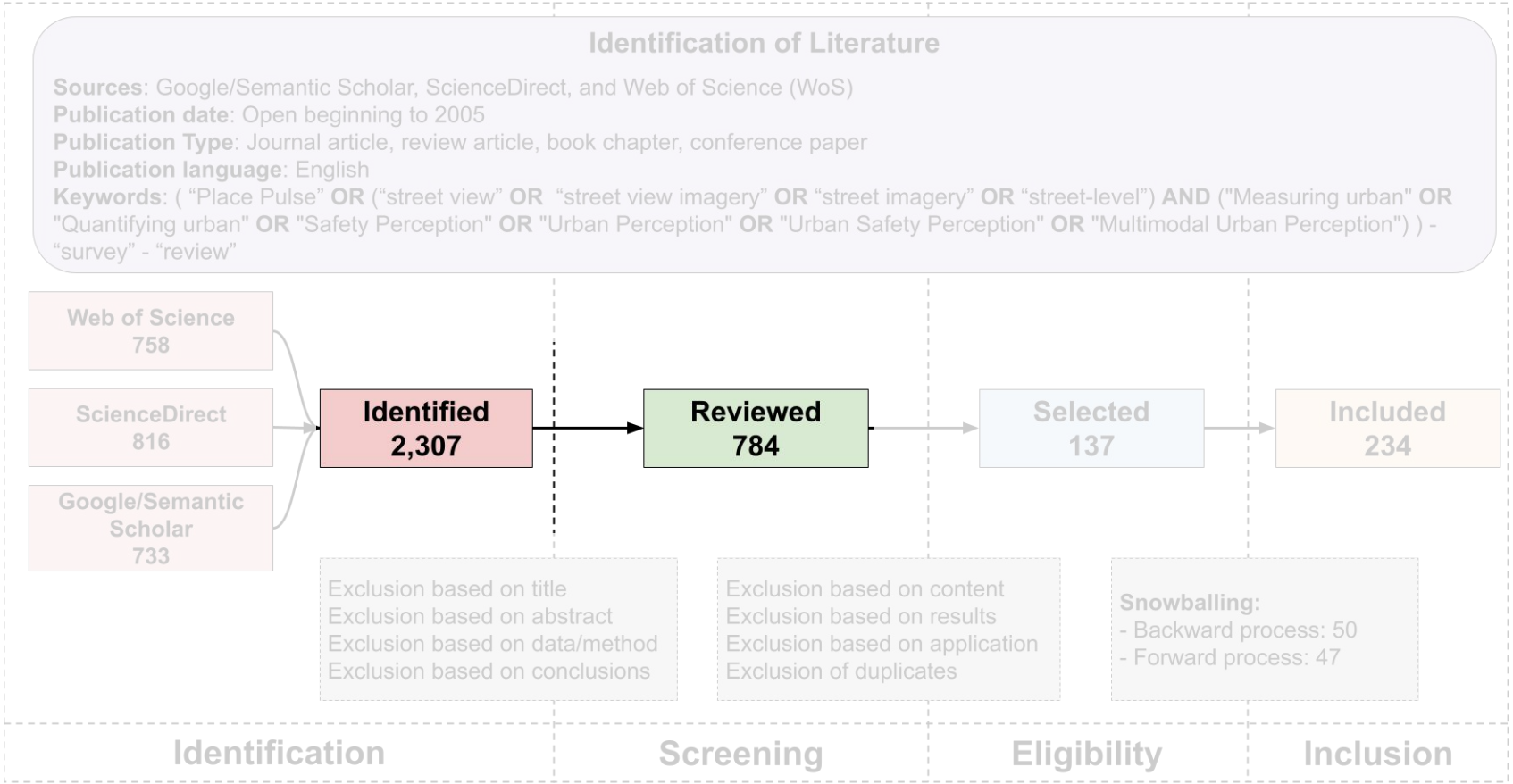
Mapping the field: PRISMA, 2005–2026



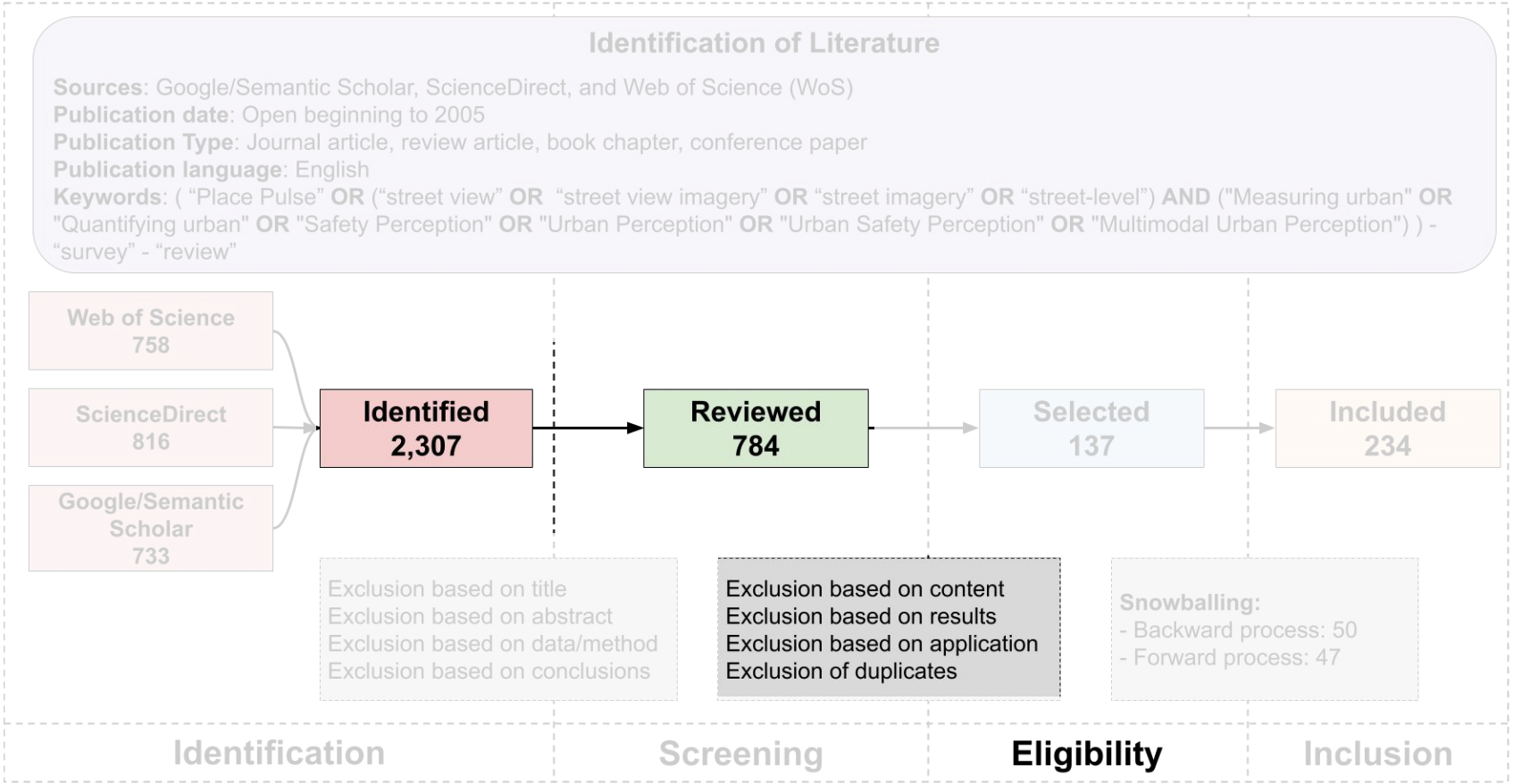
Mapping the field: PRISMA, 2005–2026



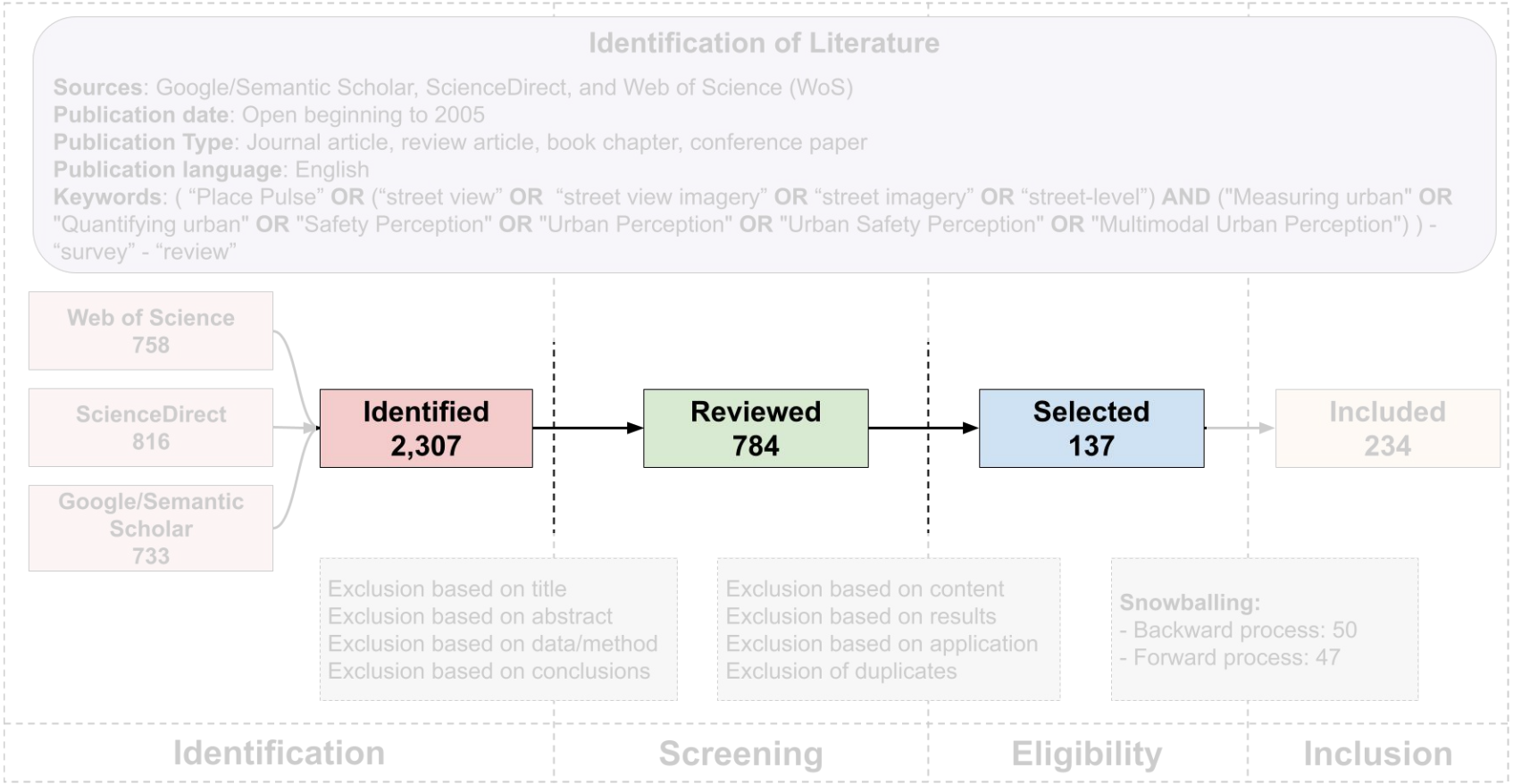
Mapping the field: PRISMA, 2005–2026



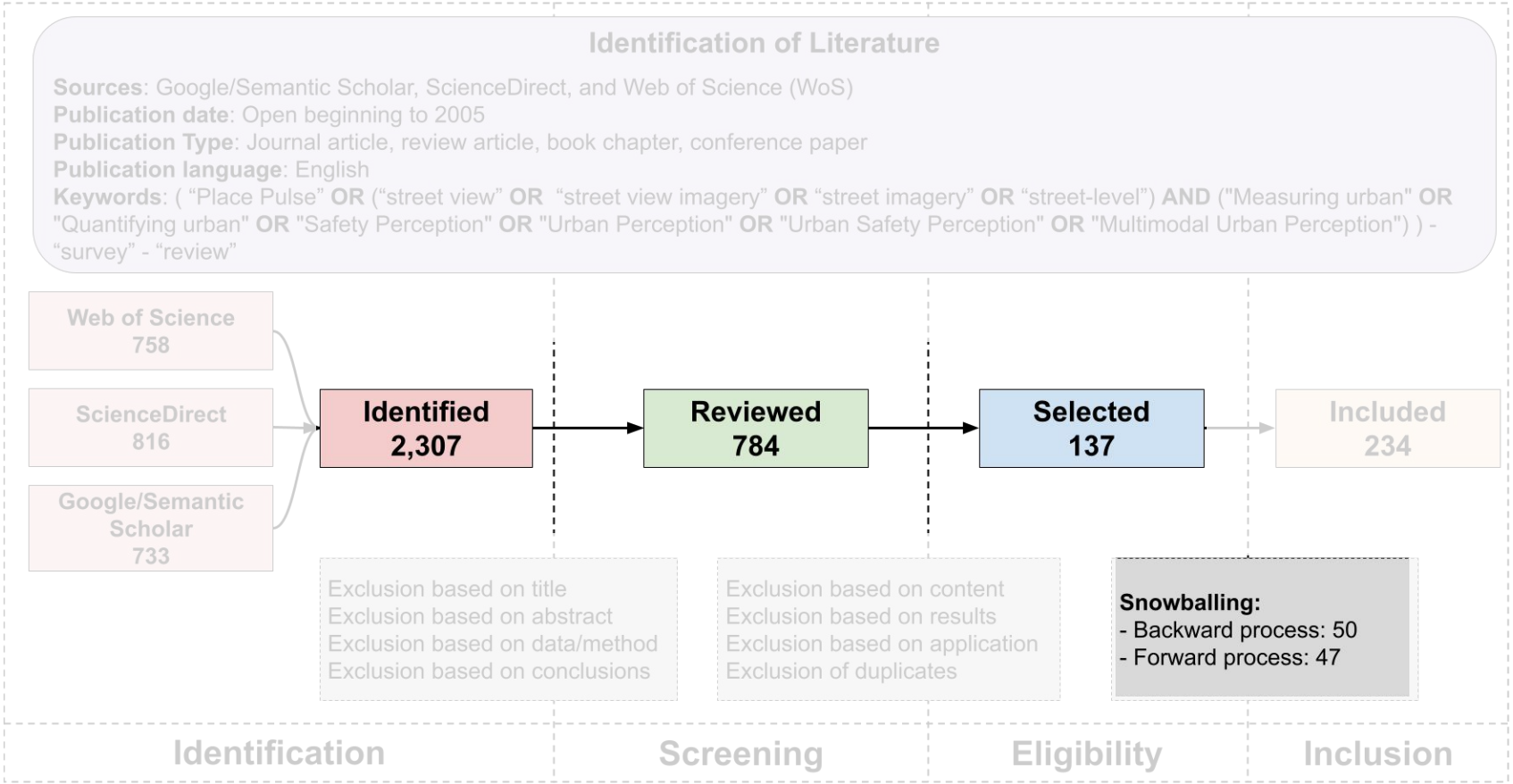
Mapping the field: PRISMA, 2005–2026



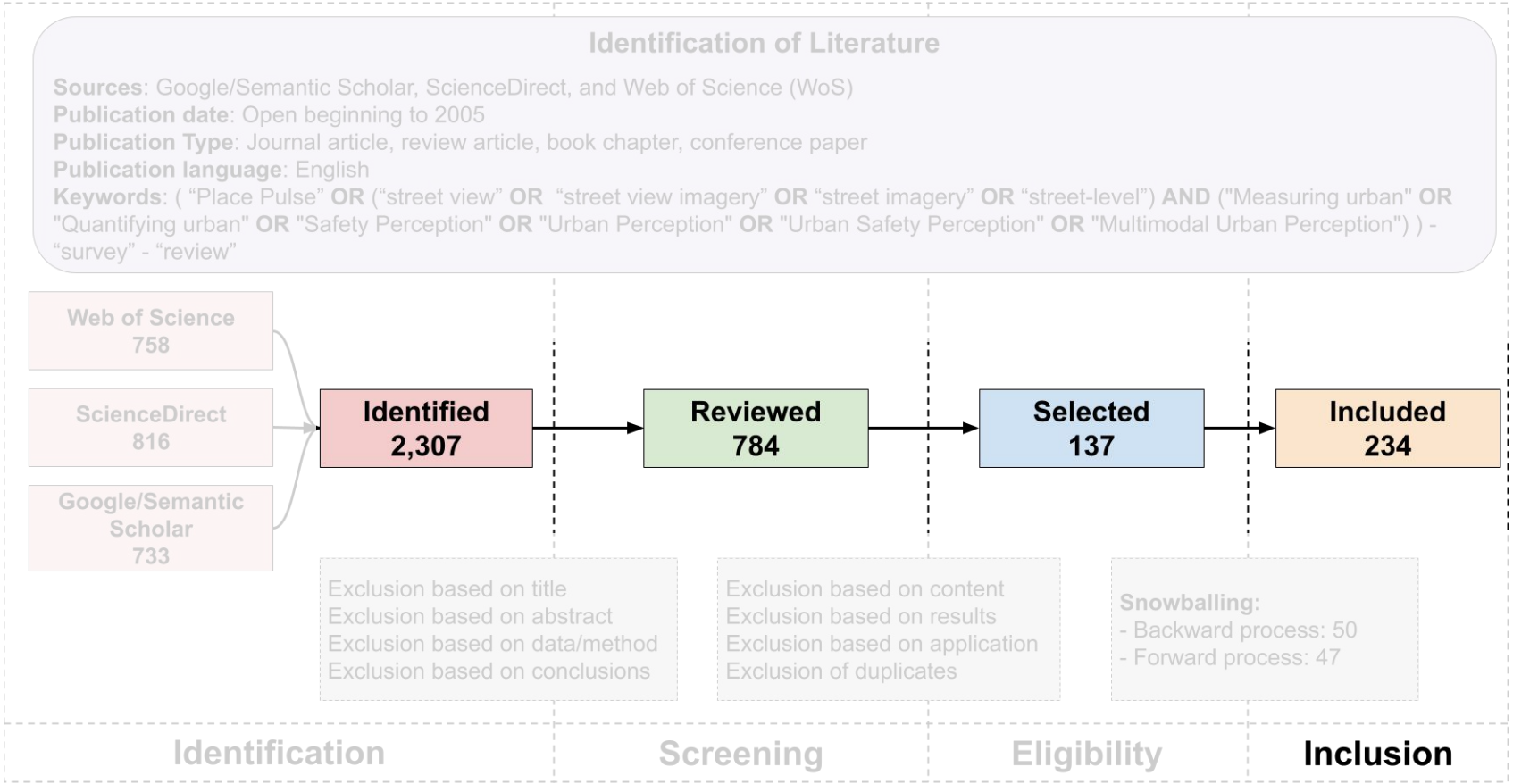
Mapping the field: PRISMA, 2005–2026



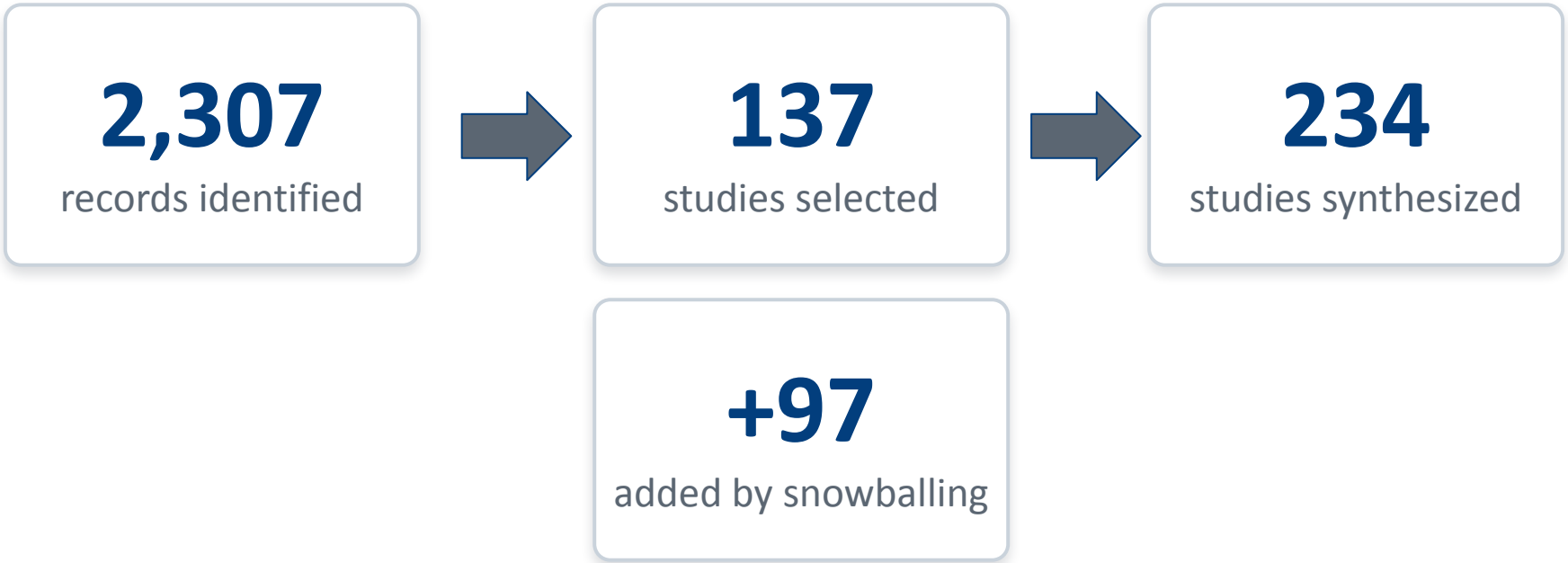
Mapping the field: PRISMA, 2005–2026



Mapping the field: PRISMA, 2005–2026



Studies synthesized



What 234 studies say about the field

Street View Imagery (SVI)

Perception Collection

- Pairwise comparisons
- Ratings (scales)

Learning Tasks

- Regression
- Classification
- Pairwise winner

XAI

- Gradients
- Perturbations

Generative AI

- VLM vs humans
- Image editions

What 234 studies say about the field

1. Find and analyze studies in urban perception research

2. Study trends

3. Identify gaps

1- Find and analyze studies

- Perception collection
- ML techniques
- Explainability methods
- Generative AI

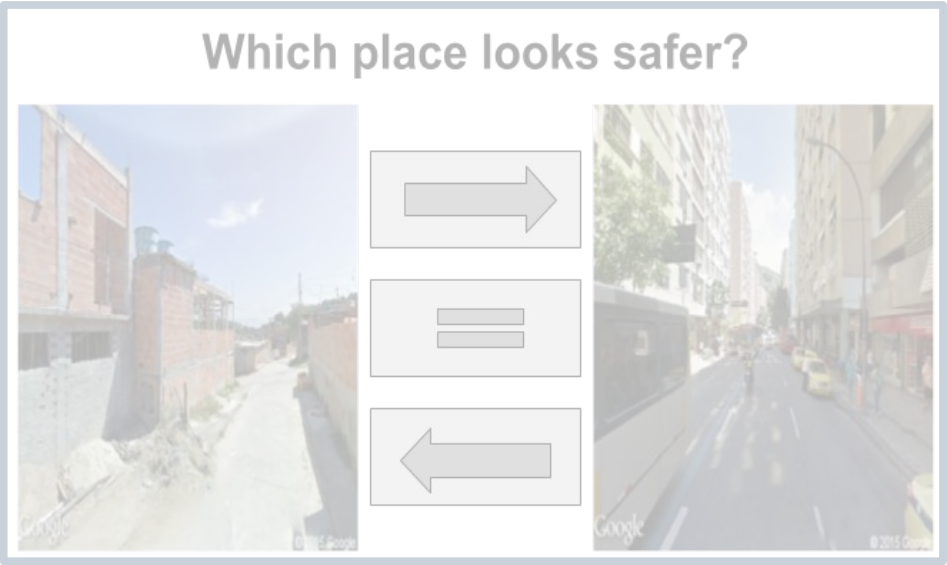
Perception collection

- Two main collection methods: pairwise comparisons and rating scales
- **Anonymized** information from volunteers, only their **choices** were recorded.
- Comparisons and rating votes are converted into scores.

Perception collection

- **Place pulse 2.0 dataset:** the most widely used dataset in this field.
- Two main methods to convert pairwise comparisons to scores:
  — Strength of schedule (recommended)
  — TrueSkill (limited, requires that all images have at least 35 comparisons to be stable).

Perception collection



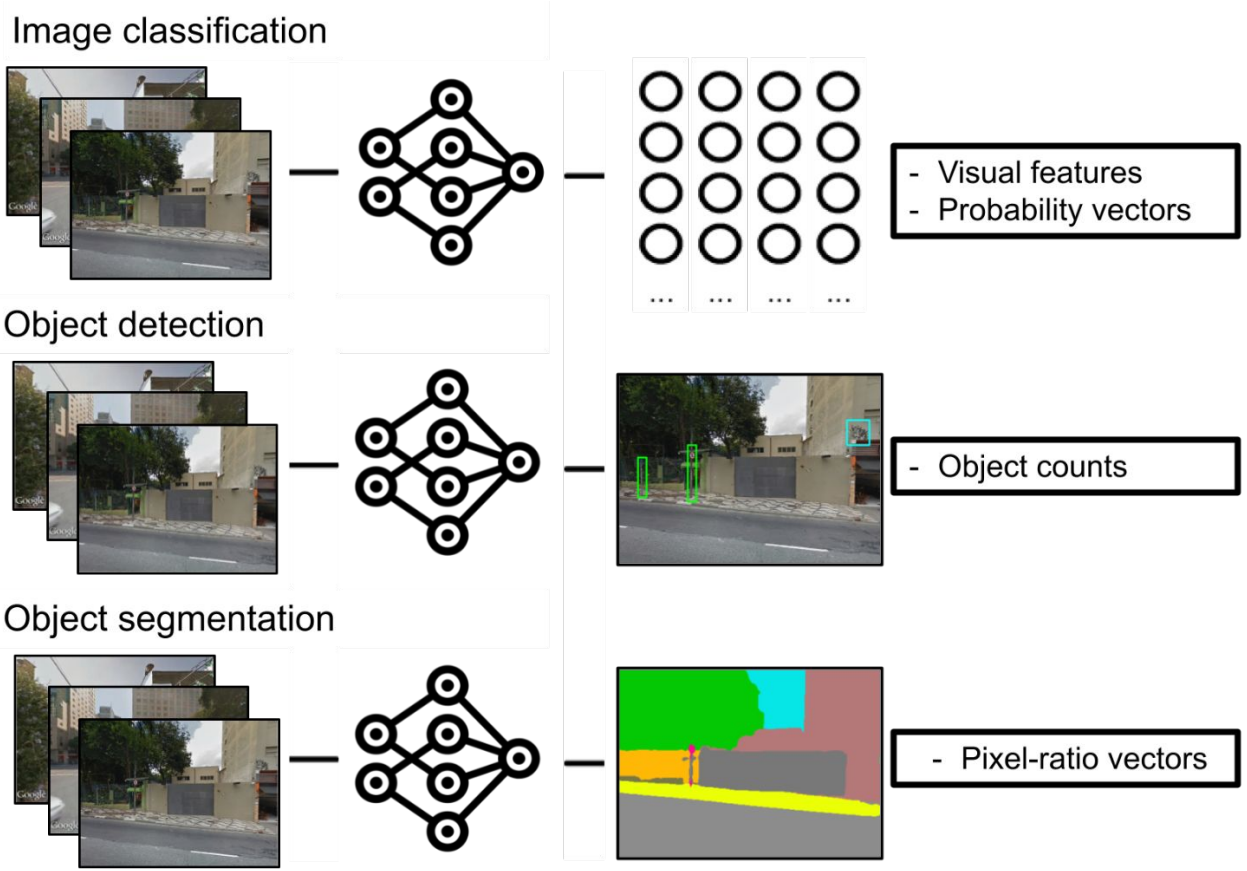
- **Likert rating scale datasets**
- Commonly used in single-city studies (mainly Chinese cities)
  — Scores taken as rating averages

ML techniques

- Feature extractions

- Learning tasks

Feature extractions



Learning tasks

Regression

( , 9.55)

( , 7.29)

Classification

( , safe)

( , unsafe)

Pairwise winner

left	right	winner
		draw
		left

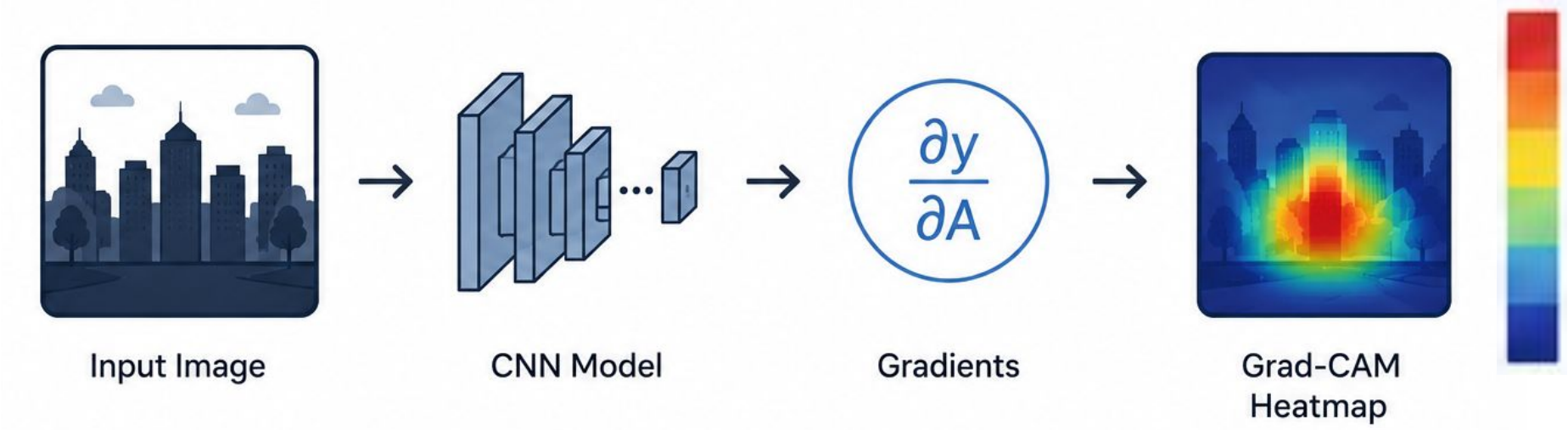
Siamese-based networks

XAI: explainability methods

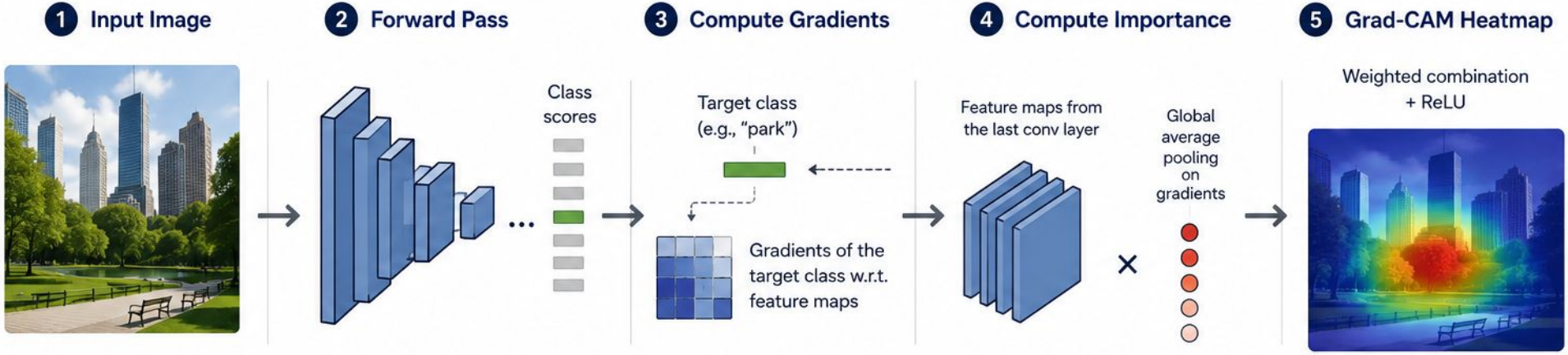
- Gradient-based attribution methods

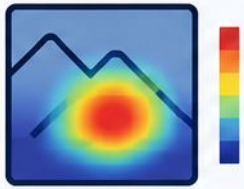
- Perturbation-based attribution methods

Gradient-based attribution methods



These method leverage gradients to understand how changing each spatial location in the feature maps would affect the target class score, highlighting the most influential regions





The heatmap highlights the regions in the image that most positively influence the prediction of the target class.

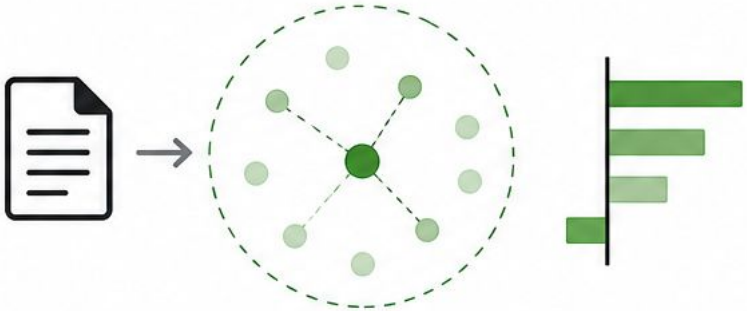
Class Discriminative: focuses on the importance for a specific class.

Visual: produces an intuitive localization map.

Model Agnostic (to architecture details): applicable to any CNN.

Perturbation-based attribution methods

LIME



Approximates the model locally around the instance using perturbed samples and a simple interpretable model.

SHAP

Uses game theory (Shapley values) to fairly and consistently assign each feature's contribution to the prediction.

Both model-agnostic methods explain a model's prediction for a single instance by estimating how much each feature contributes to the prediction.

LIME

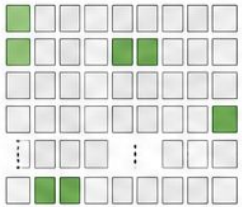
Local Interpretable Model-agnostic Explanations

1 Original instance

Park	=1
Trees	=1
Traffic	=2
Noise	=3
...	...

The instance we want to explain.

2 Perturb the instance



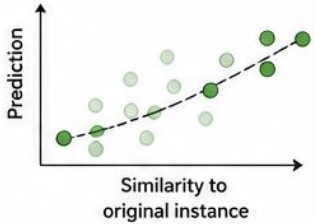
Generate many variations (perturbations) around the original instance.

3 Get model predictions

Prediction
0.82
0.64
0.21
0.78
...

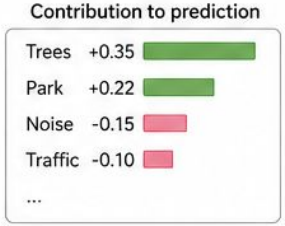
Obtain the model's prediction for each perturbed sample.

4 Fit local interpretable model



Train a simple, interpretable model (e.g., linear model) weighted by similarity.

5 Feature contributions (explanation)



The local model's coefficients indicate how each feature influences the prediction.

SHAP

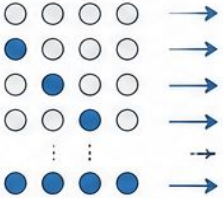
SHapley Additive exPlanations (game-theoretic approach)

1 Original instance

Park	=1
Trees	=1
Traffic	=2
Noise	=3
...	...

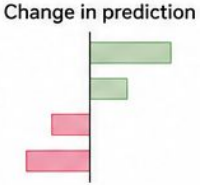
The instance we want to explain.

2 Consider all feature coalitions (subsets)



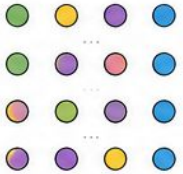
Evaluate the model with different combinations of features (with/without each feature).

3 Compute marginal contributions



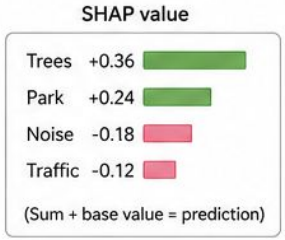
Measure how much adding a feature changes the prediction.

4 Average over all permutations



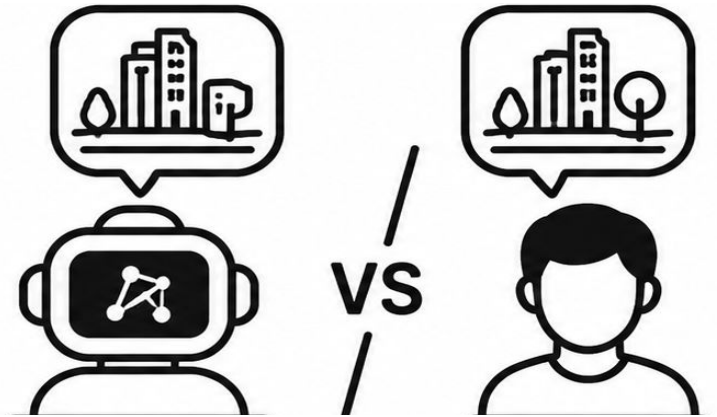
Compute the average marginal contribution across all possible feature orders.

5 SHAP values (explanation)



SHAP values quantify each feature's contribution fairly and consistently.

Generative AI



VLM vs Humans

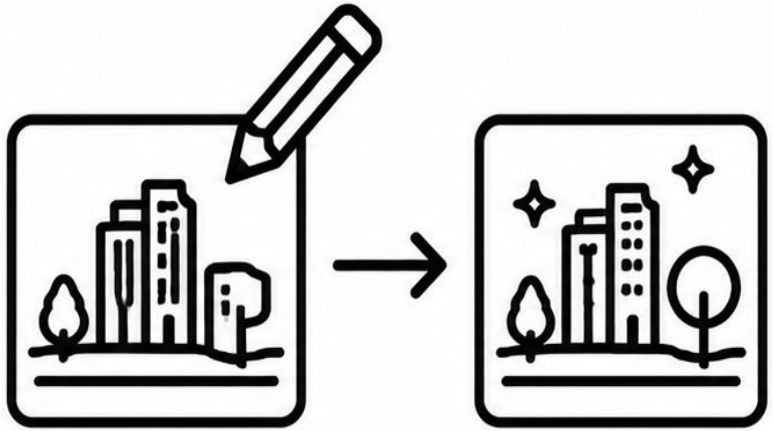
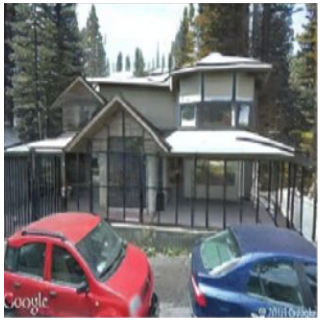


Image editing

Image Editing

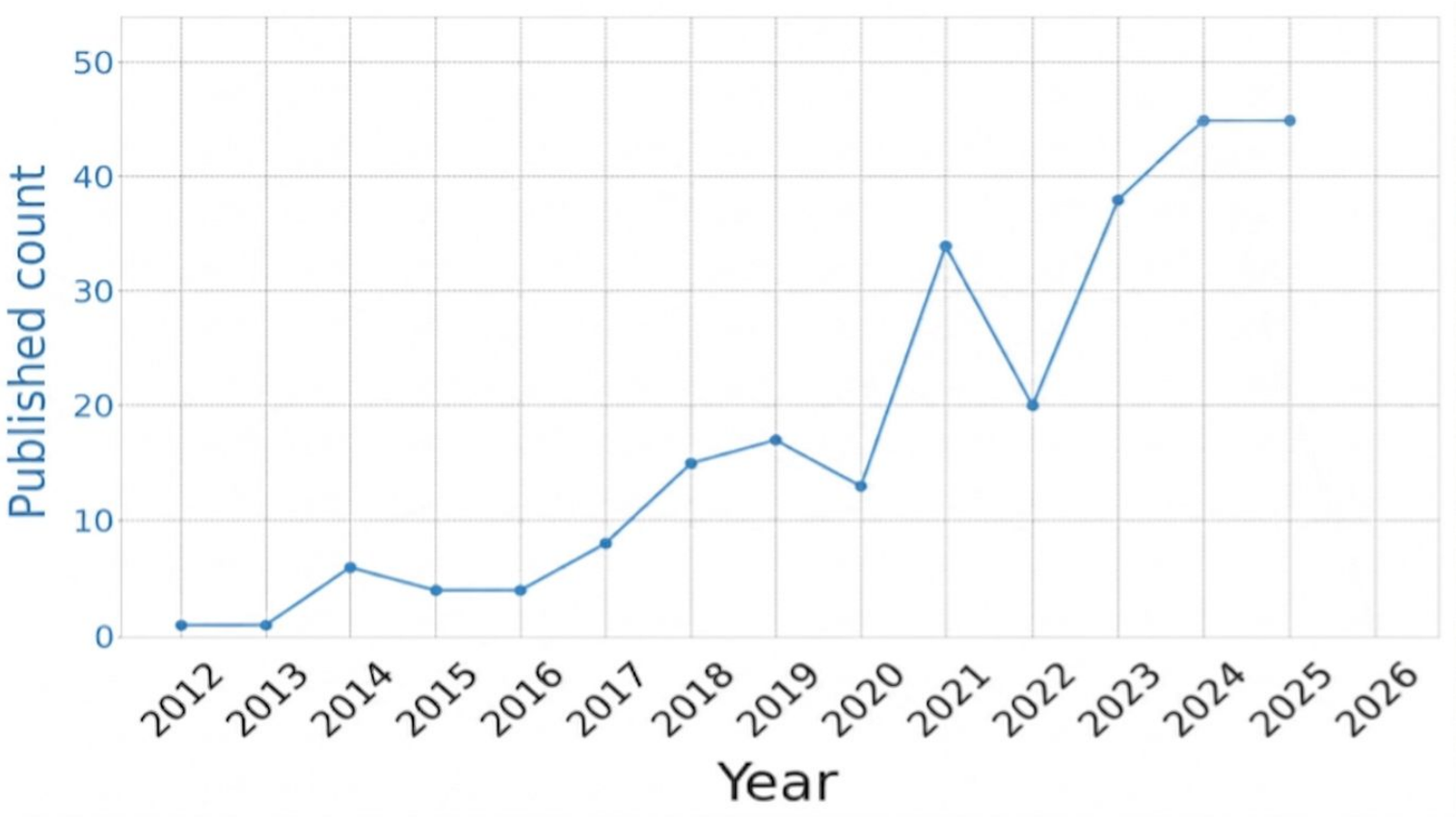

Distorted (law,2024)


Low quality (Tan,2025)

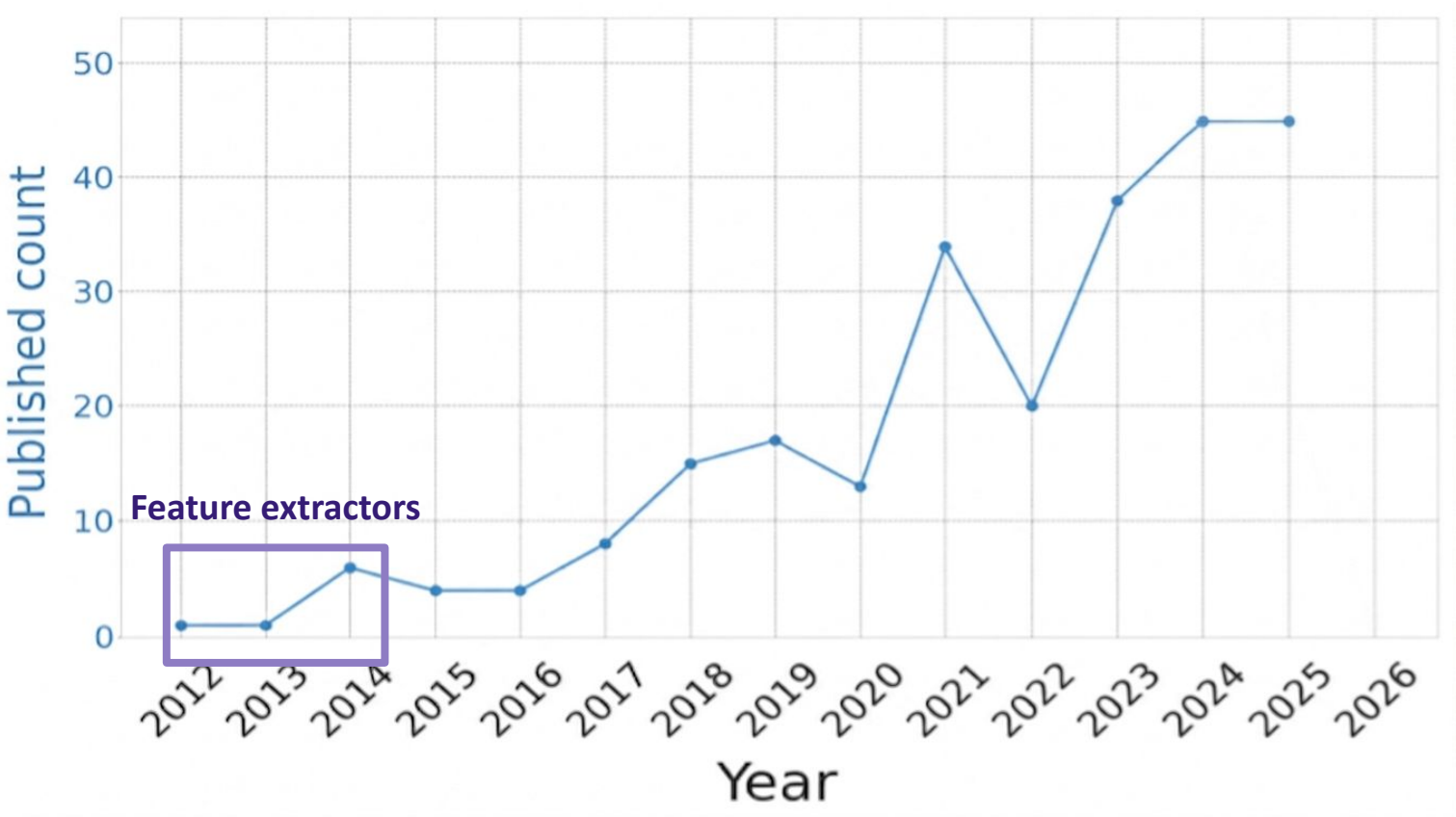



Loss context (Hu,2025)

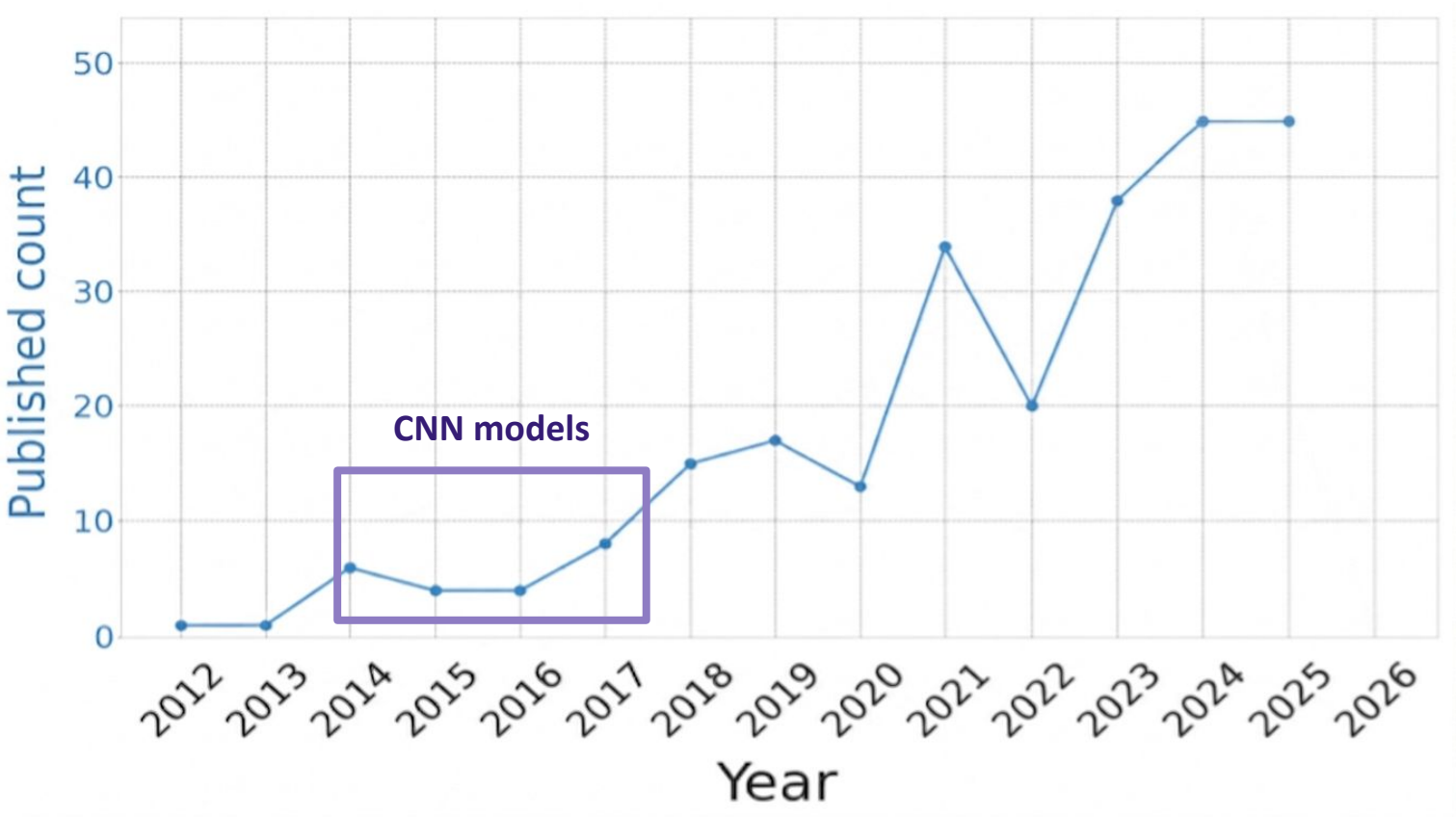
2- Study trends



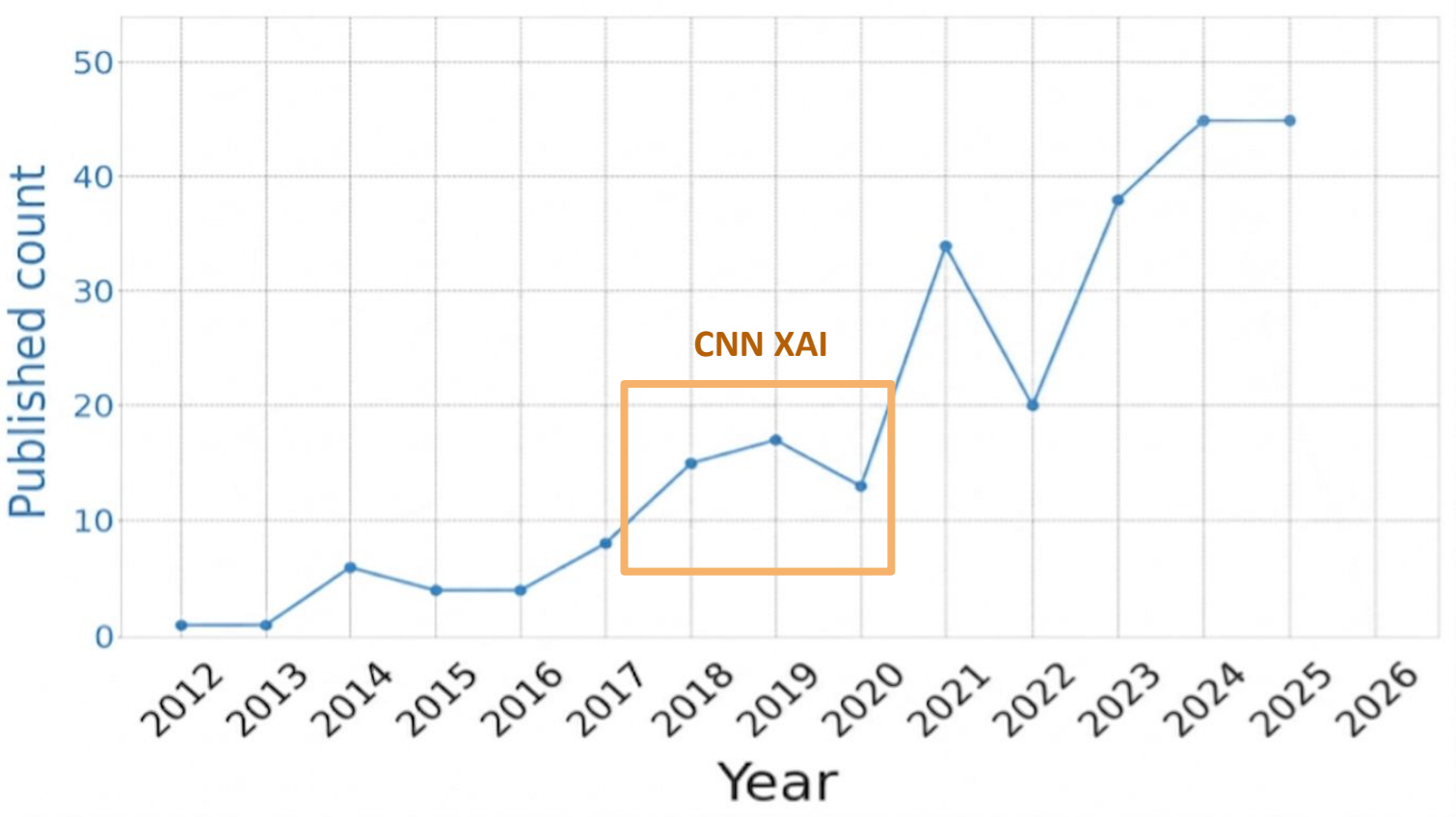
2- Study trends



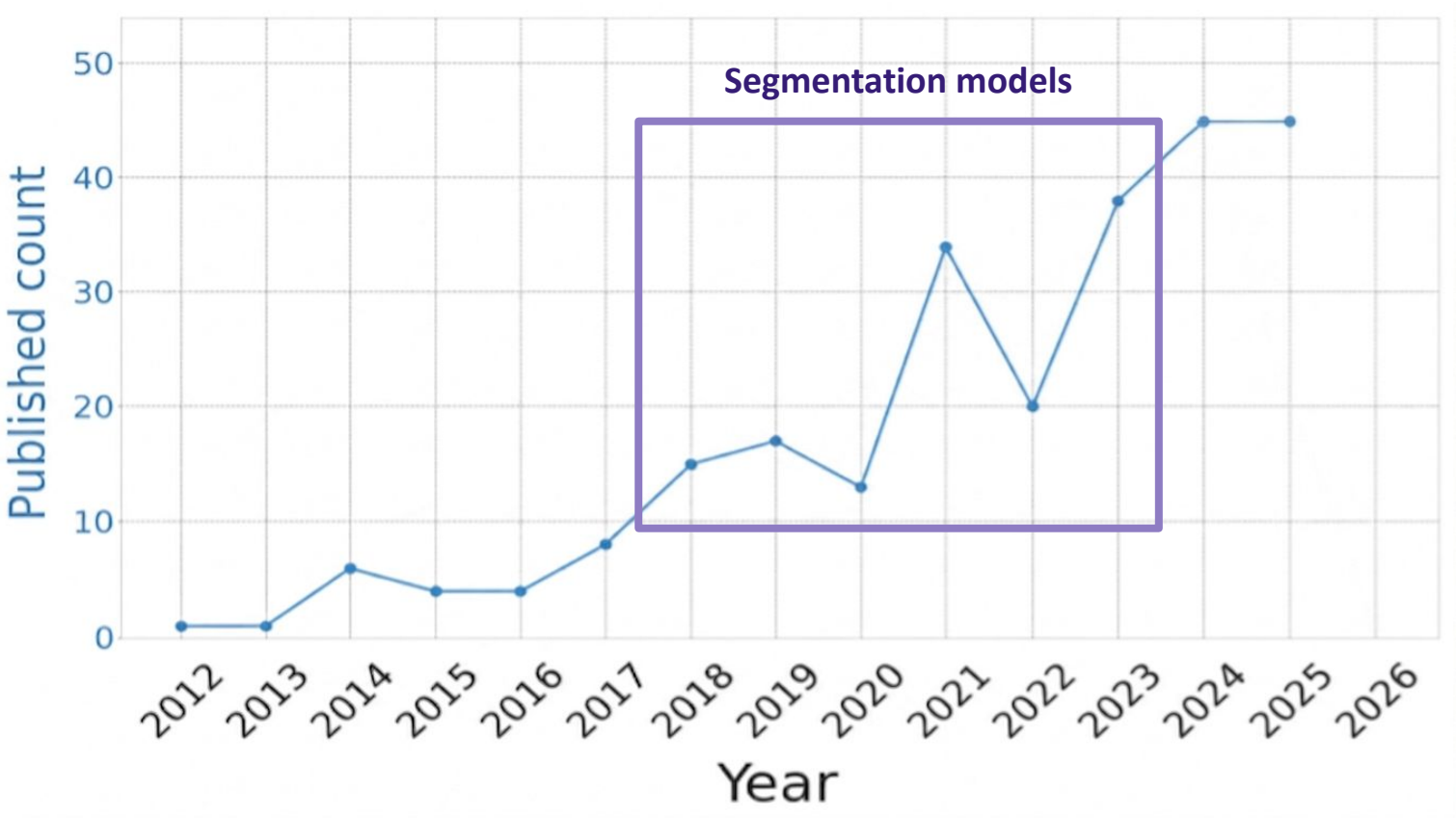
2- Study trends



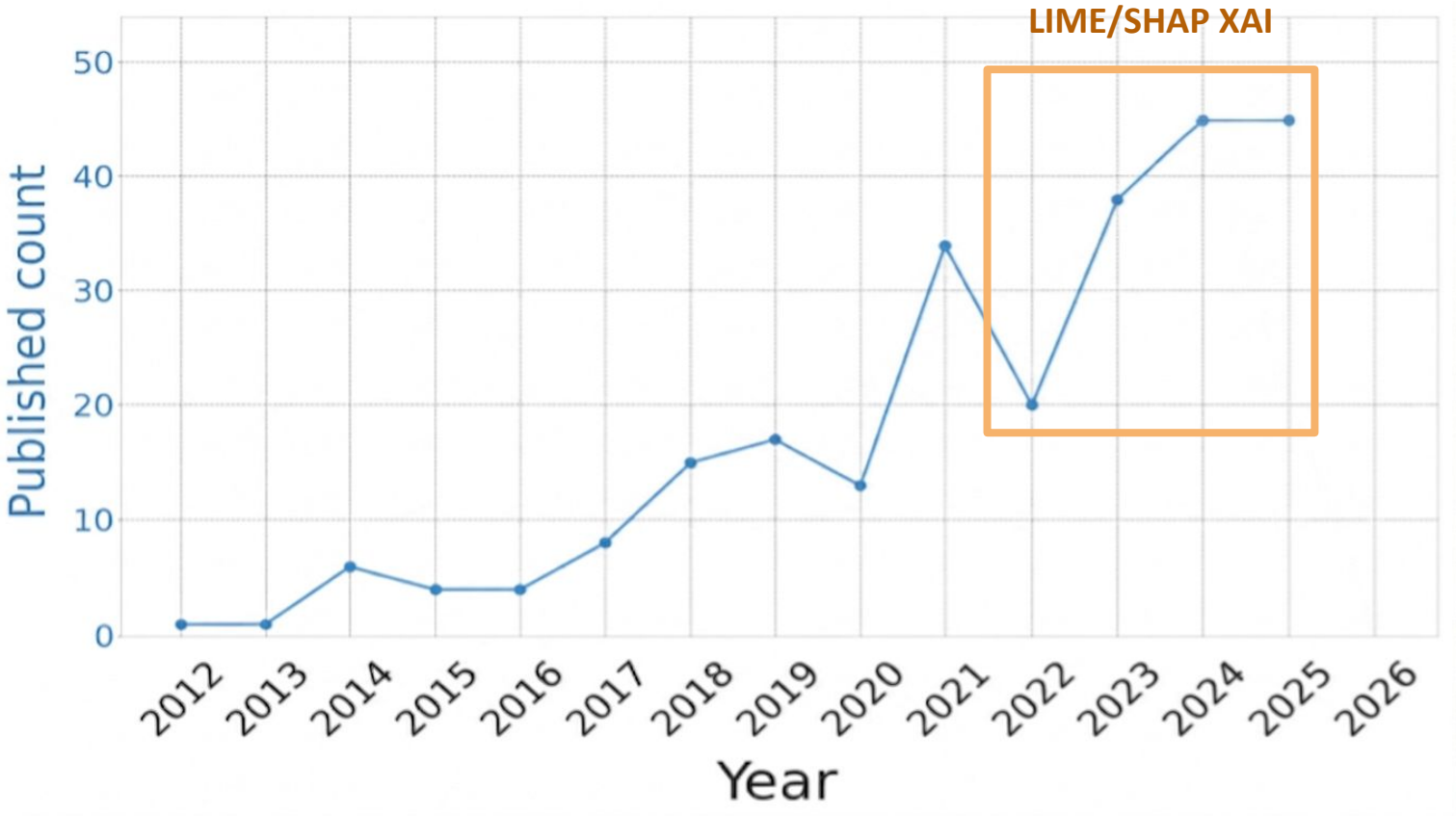
2- Study trends



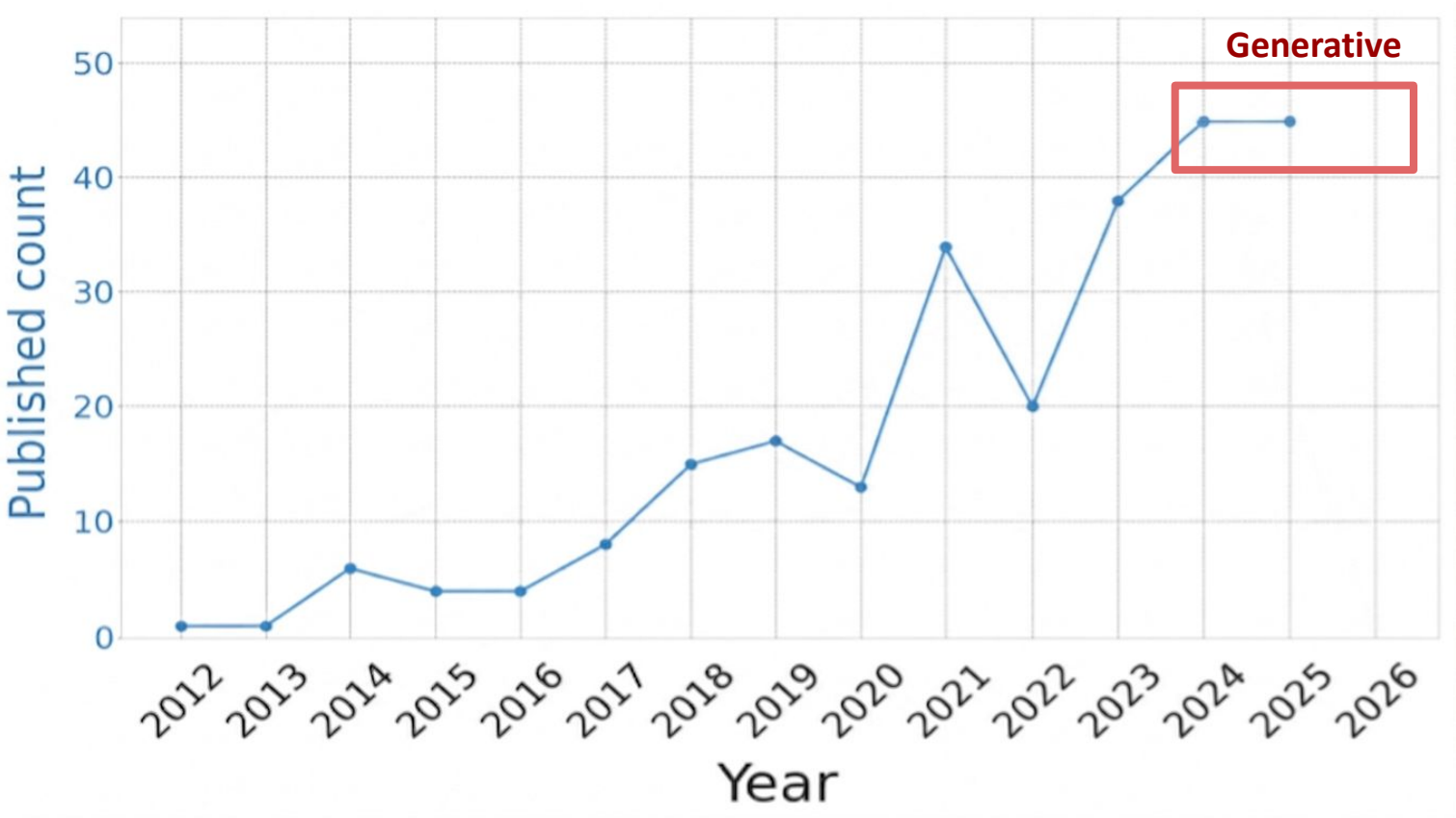
2- Study trends



2- Study trends



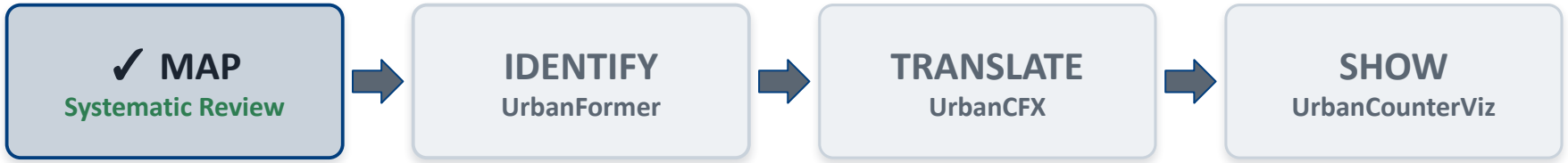
2- Study trends



Four gaps

1. Limited insight into visual factors
2. Static and non-readable explanations
3. Unrealistic generative edits
4. No interactive, human-in-the-loop tools

Three to go



This work was submitted to the *IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI)* Journal *2026*

Felipe Moreno-Vera and Jorge Poco

Identifying relevant visual elements

Question: which visual elements actually matter?

1. **Limited insight into visual factors**
   → IDENTIFY
2. **Static and non-readable explanations**
   → TRANSLATE
3. **Unrealistic generative edits**
   → SHOW
4. **No interactive, human-in-the-loop tools**
   → SHOW

Dataset

- Data: **Place Pulse 2.0**
  — 368,000+ **pairwise safety comparisons of street-view images**
  — 108,000+ street-view images from 56 cities
  — Strength of schedule method -> perceptual scores (Q-scores)

left	right	winner
		draw
. . .	. . .	. . .
		right
		left

Strength of schedule

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

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

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

* Park et. al., A network-based ranking system for us college football

Dataset: samples

High safety scores images

Low safety scores images

Previous approaches

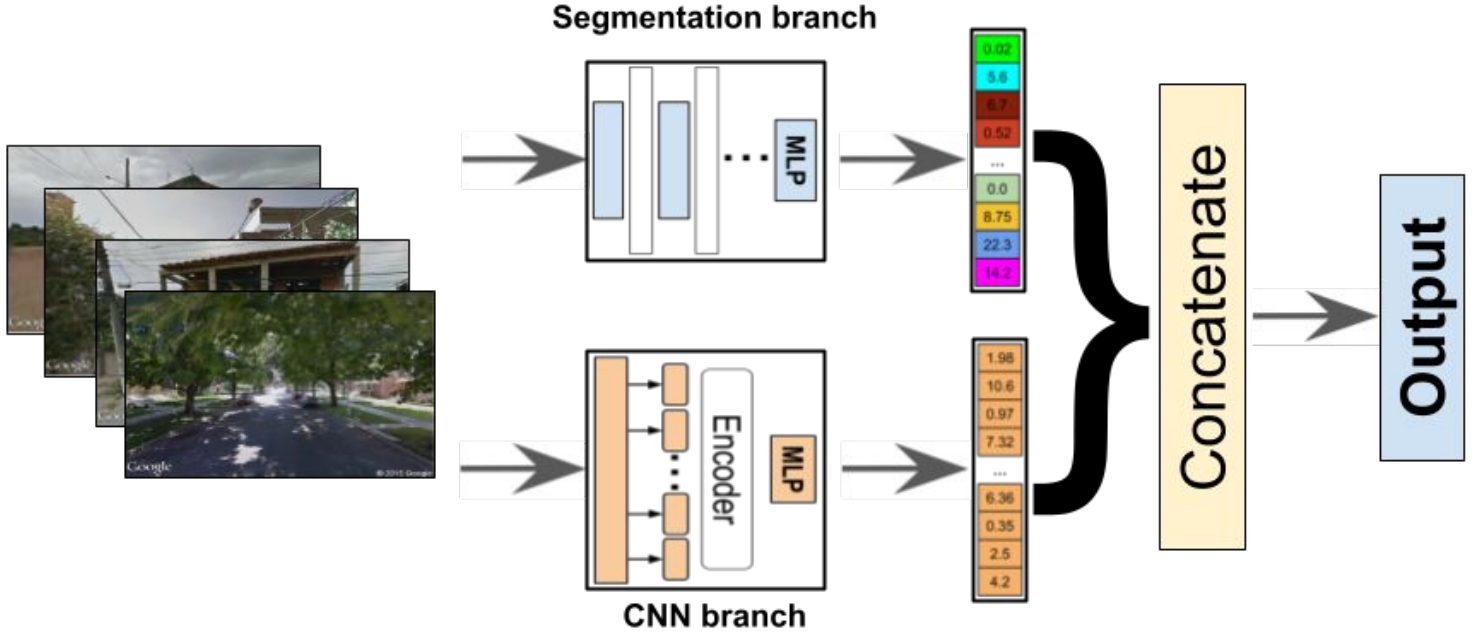
- **CNN-based models**
  — Learn rich visual representations from street-view images.
  — Explainability (e.g., Grad-CAM) highlights **regions**, but cannot directly attribute importance to **semantic objects**.

- **Segmentation-based models**
  — Use semantic pixel-ratio features (e.g., road, trees, buildings).
  — Explanations are obtained from feature importance in linear models.
  — Spatial layout and visual context are discarded.

- **Limitation**
  — Current methods provide either visual localization or semantic interpretability—but not **both simultaneously**.

Proposed direction: UrbanFormer

- Combine visual features (CNN) with semantic features (segmentation).
  — Rich visual representation.
  — Explicit semantic information.



UrbanFormer: results

- Model comparisons: **two experiments**
  — Binary classification.
  — 10-label classification.

Model	Binary (safe/unsafe)	10-label
Best prior segmentation-based	70.63%	–
Best prior CNN-based	71.16%	78.10%
UrbanFormer (ours)	**75.22% — strongest reported**	**78.51% — competitive**

Which visual elements shape perceived safety?

- Idea: **Visual Element Impact** (Grad-CAM ∩ masks, IoU).
  — Object-level explainability by comparing Grad-CAM with segmentation masks

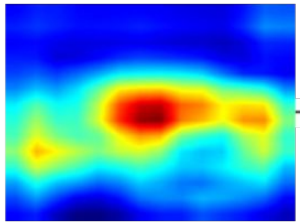
Which visual elements shape perceived safety?

- Example: **Visual Element Impact** for tree element

Which visual elements shape perceived safety?

- Example: **Visual Element Impact** for tree element

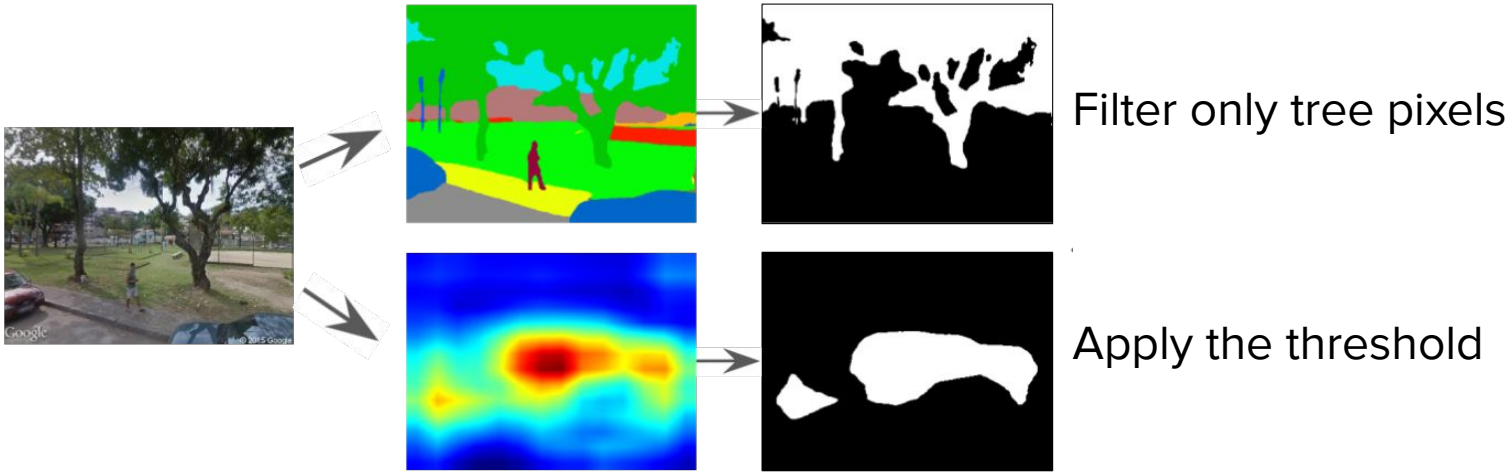
Obtain segmentation mask



Obtain the grad-cam

Which visual elements shape perceived safety?

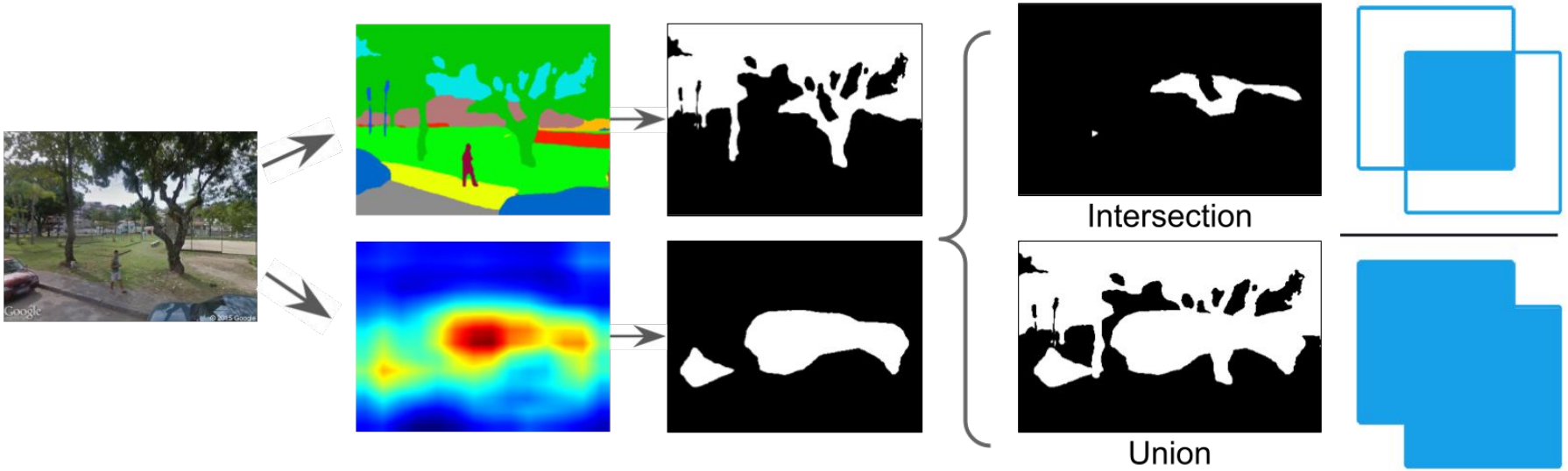
- Example: **Visual Element Impact** for tree element



We use a **threshold of 0.3** for Grad-CAM values (Adebayo, 2018; Pinciroli, 2021; Yuan, 2025).

Which visual elements shape perceived safety?

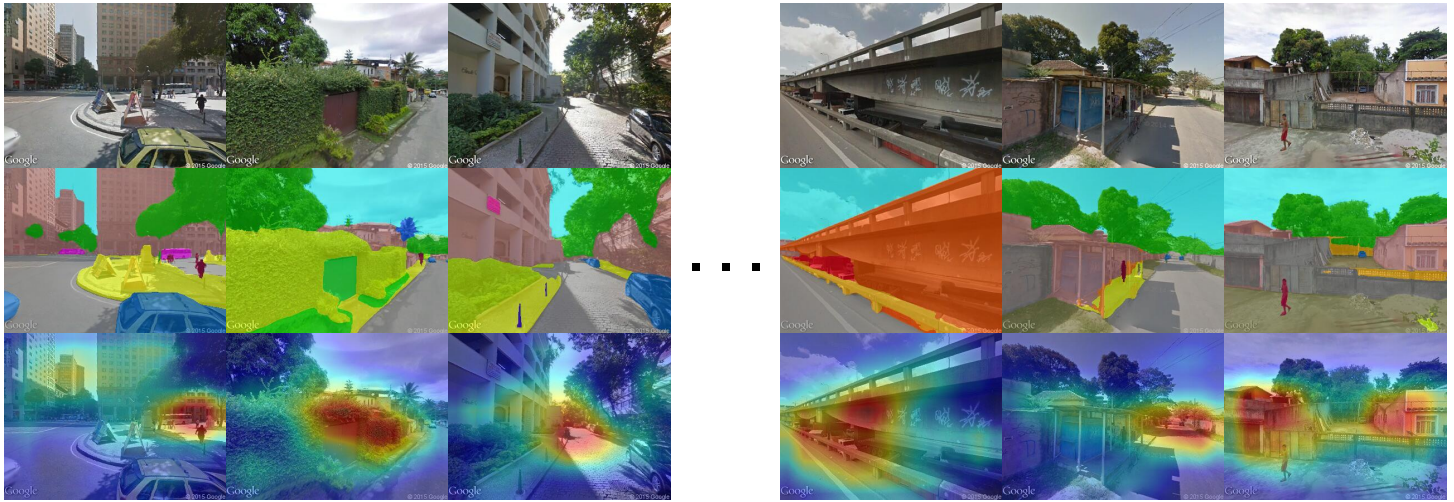
- Example: **Visual Element Impact** for tree element



The **Visual Element Impact** for tree element in this sample is 0.35

Visual Element Impact

safe ⟷ unsafe



Associated with SAFE
trees · cars · grass · plants

Neutral elements
buildings · sky · pole · sidewalks

Associated with UNSAFE
walls · fences · earth · trash

Takeaway: Visual elements shape perceived safety

✓ What we have (GAP 1)

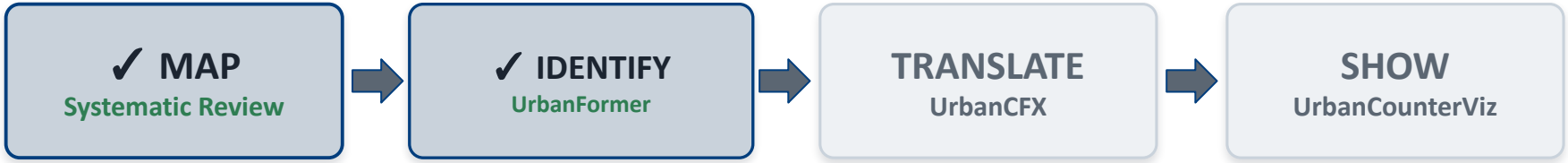
Certain urban elements are consistently associated with **higher** or **lower** perceived safety.

✗ What remains (GAP 2)

Knowing **which** elements matter **does not tell** us how they **should change** to improve perceived safety.

Identifying influential elements is not the same as knowing how to improve them.

Two to go



This work was published in *IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 2024*

Felipe Moreno-Vera and Jorge Poco

Translating to human-readable language

Question: What would make this street feel safer?

1. Limited insight into visual factors
   → IDENTIFY
2. **Static and non-readable explanations**
   → **TRANSLATE**
3. Unrealistic generative edits
   → SHOW
4. No interactive, human-in-the-loop tools
   → SHOW

Previous approaches

- **Existing explanations describe streets using standard object labels.**
  — Segmentation models explain predictions using semantic objects.
  — Typical labels include tree, road, wall, sidewalk, building...

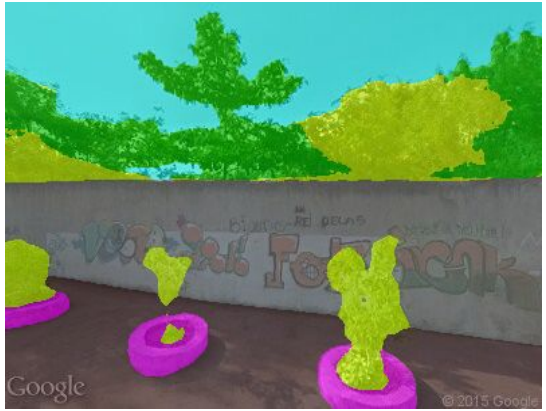
- **A model can only explain what it has words for.**
  — Existing vocabularies cannot describe **disorder cues** (e.g., graffiti, damaged sidewalks).
  — Counterfactual explanations are presented as **feature changes**, not as **human-readable** recommendations.

- **Limitation**
  — Current approaches provide **neither** a vocabulary that captures **urban disorder** nor understandable **explanations** of **what should change** to improve perceived safety**.**

Proposed direction: UrbanPD4k

- Extend both the vocabulary and the explanation.
  — **Group semantically related objects** into more interpretable concepts.
  — Introduce **disorder cues** that capture the condition of urban elements.

Input image



Segmentation

Groups+Disorder

Proposed direction: UrbanCFX

- Extend both the vocabulary and the explanation.
  — **Translate** counterfactual feature changes into **human-readable recommendations**.

Input image

Counterfactual

(add) tree +0.14
(add) plants +0.11
(remove) graffiti −0.09
(remove) broken sidewalk −0.08

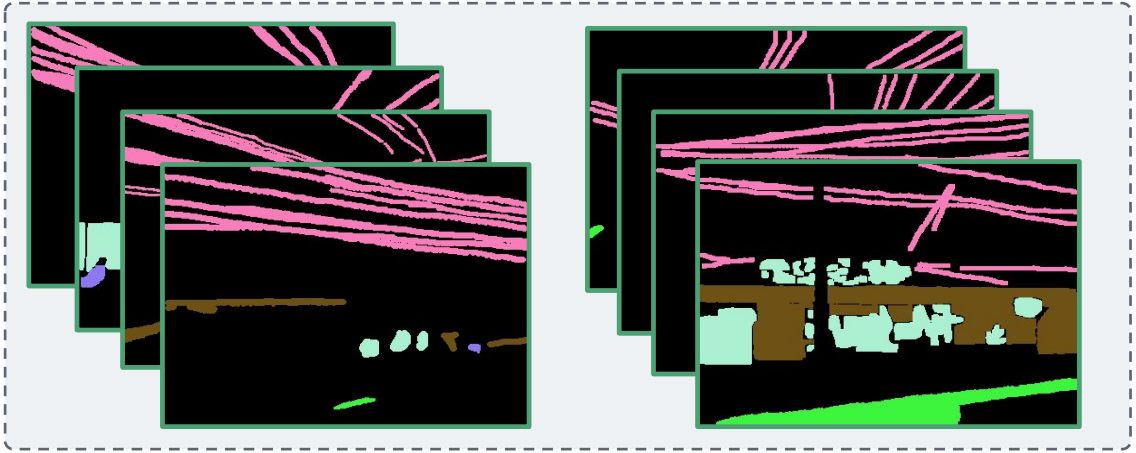
Generated recommendation

“Increase street trees moderately; remove graffiti from the boundary wall; repair the sidewalk surface, add some plants…”

Condition Shapes Perception

- Standard segmentation recognizes objects – e.g., they see "a building, sidewalks, roads".
- But perceived safety also **depends** on their **condition** — e.g., they see “a **building**” but not "a **neglected building** with **graffiti** and **broken windows**".

UrbanPD4k



3,659 SVI

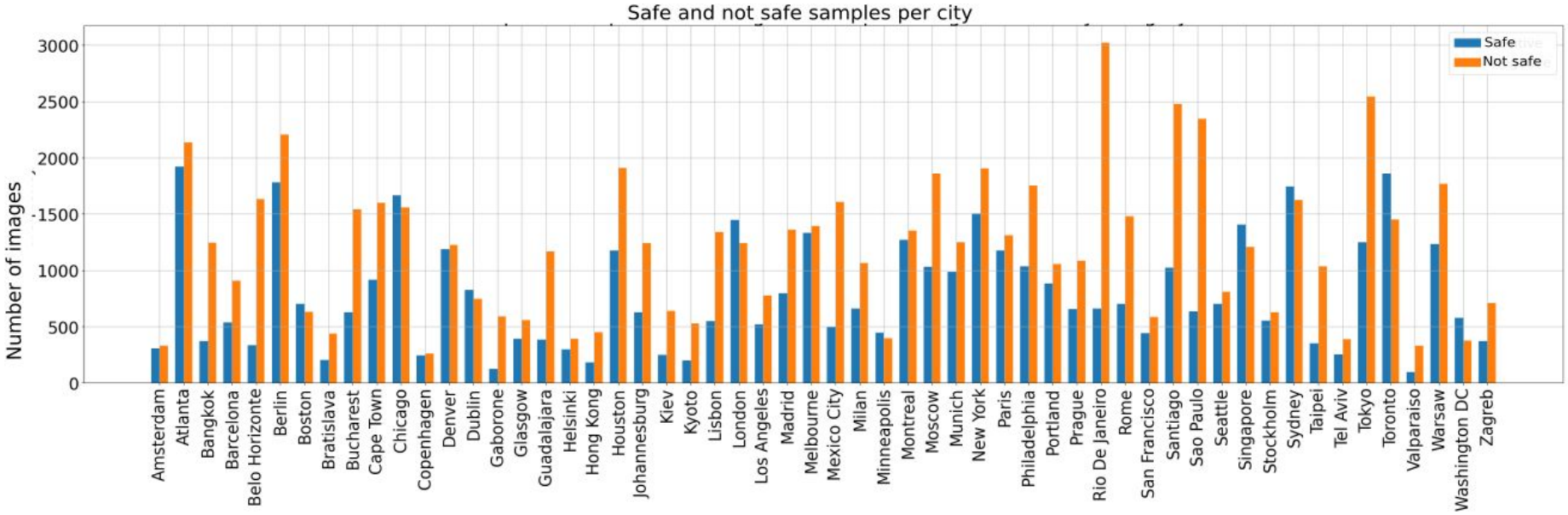
from Rio de Janeiro

13 disorder cues

pixel-level annotations

trashcan broken_pavement car_police graffiti garbage_bag
broken_window homeless overhead_cable damaged_traffic_sign
kiosk car_taxi broken_damaged_bricks_wall garbage_box

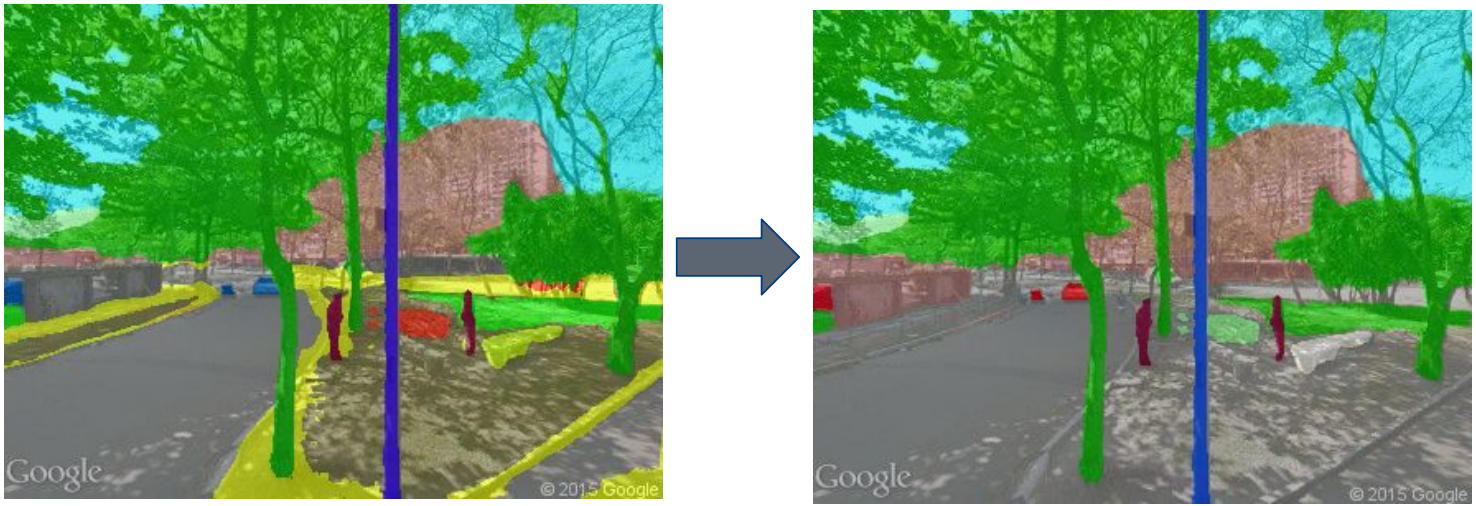
Why Rio de Janeiro?



Why Rio: Lowest mean perceived safety in Place Pulse 2.0 ➜ the most informative context for studying disorder.

The vocabulary is too fine-grained for human reasoning

- People reason about concepts—not hundreds of individual object classes
- Human-readable recommendations require human-readable concepts.
  - **“vegetation” = “plant, tree, palm, etc.”.**
  - **“wall” ≠ “graffiti-covered wall”.**



From fine-grained to human-readable concepts

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

Other categories have pixels occupy **less than 1% of image pixels** (e.g., door, basket, etc.)

Disorder and group elements carry real signal

Setting	Categories	Perception	Metrics		
			Precision	Recall	F-1
Pixel ratios	ADE20K (150 classes)	not safe	0.67	0.73	0.69
		safe	0.71	0.63	0.67
	ADE20K+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**

Disorder and group elements carry real signal

Setting	Categories	Perception	Metrics		
			Precision	Recall	F-1
	ADE20K (150 classes)	not safe	0.67	0.73	0.69
		safe	0.71	0.63	0.67
	ADE20K+Disorder (163 classes)	not safe	0.70	0.76	0.73
		safe	0.73	0.68	0.70
Pixel ratios	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**

+1–2%

unsafe-detection F1 when **group elements** are used

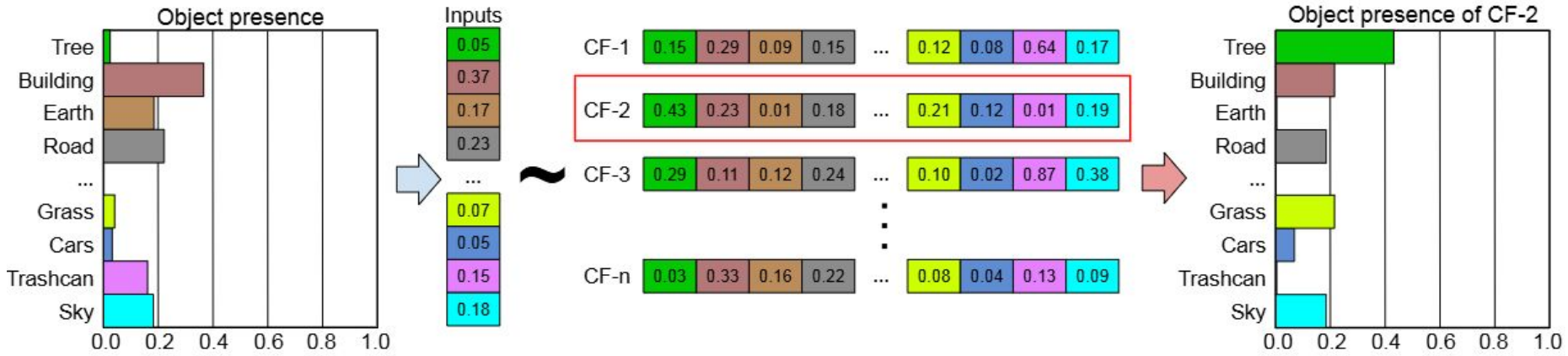
+3–5%

unsafe-detection F1 when **disorder elements** are added

+6%

unsafe-detection F1 when **group+disorder elements** are added

UrbanCFX: the minimal change that flips the prediction



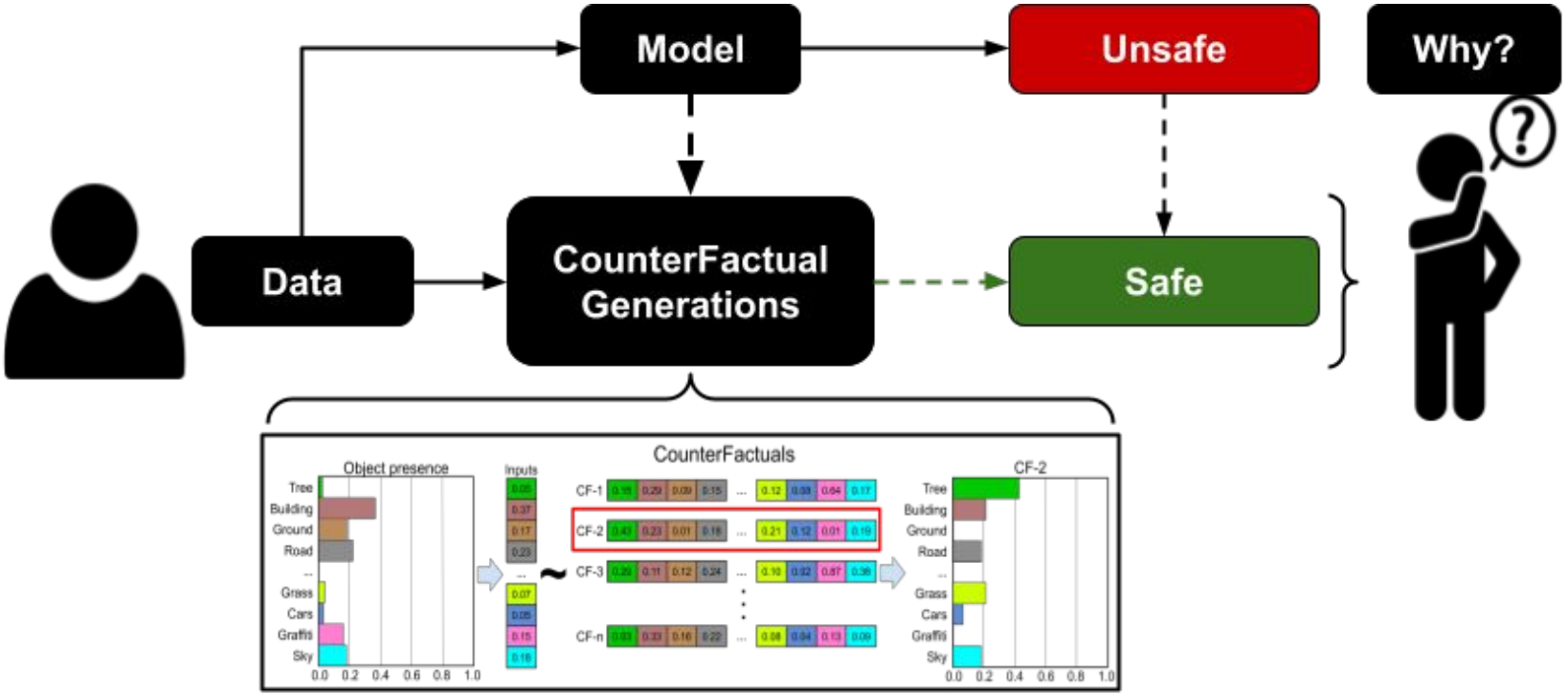
Counterfactual

the minimal feature change that flips the prediction: unsafe → safe

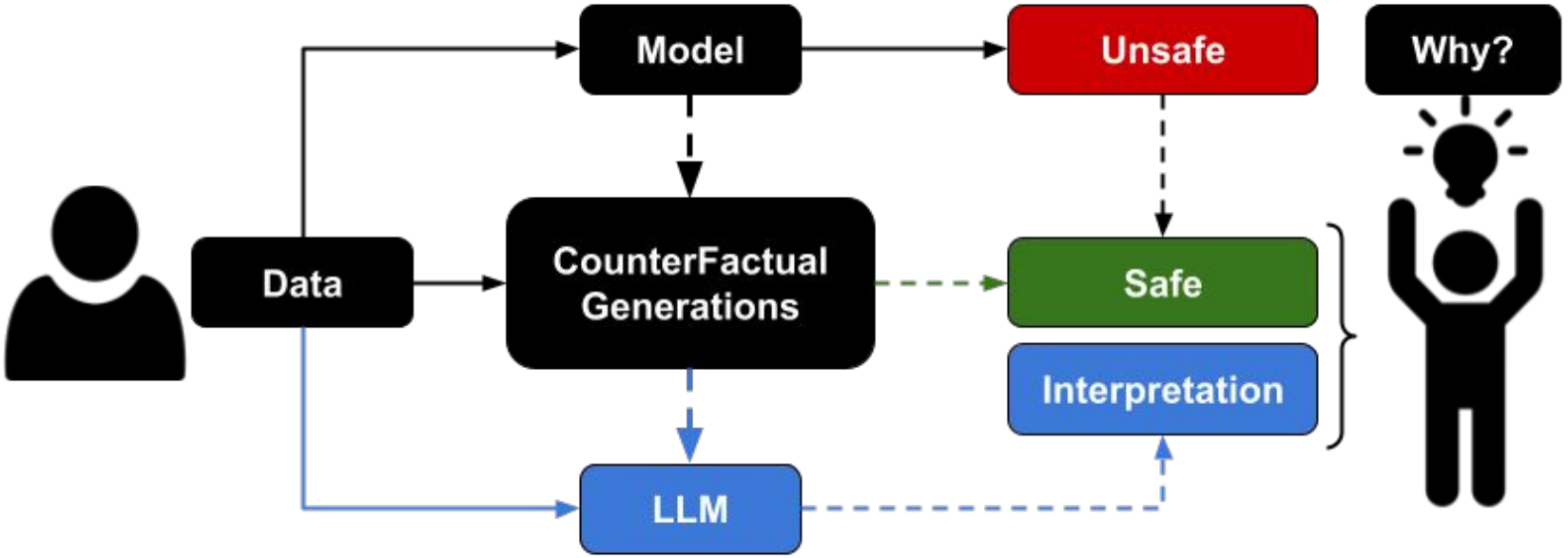
Output

a signed difference vector — which elements should increase or decrease, and by how much

From vector to sentence — with the LLM on a leash



From vector to sentence — with the LLM on a leash



The LLM does NOT decide what to change — it verbalizes a pre-computed, model-derived vector.

Outputs are planning-oriented hypotheses — not ready-to-implement interventions (yet).

What changes would make this street feel safer?

Counterfactual

(remove) overhead cables -0.08
(remove) damaged walls -0.12
(add) tree +0.15
(add) plants +0.06

Generated recommendation

"The overhead cables should be removed, the brick wall should be finished, and adding some tree or plants will help"

Counterfactual

(remove) garbage bags -0.04
(add) grass +0.04

Generated recommendation

"This street is already in a good condition, but taking away the garbage bags will make this street cleaner."

Takeaway: Explanations become human-readable

✓ What we have (GAP 2)

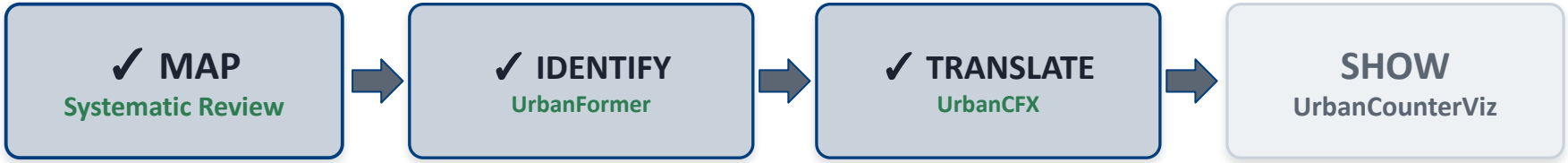
We can **translate** model **explanations** into specific and understandable **recommendations**.

✗ What remains (GAP 3 & 4)

Recommendations are still **text** —they cannot be directly **visualized**, **compared**, or **interactively explored**.

Reading what should change is different from seeing what those changes would look like.

One to go



This work was published in *IEEE International Conference on Big Data 2025*

*Felipe Moreno-Vera, **Andres De-la-Puente**, and Jorge Poco*

Showing changes clearly and understandably

Question: What would those changes look like?

1 Limited insight into visual factors
→ IDENTIFY

2 Static and non-readable explanations
→ TRANSLATE

3 **Unrealistic generative edits**
→ SHOW

4 **No interactive, human-in-the-loop tools**
→ SHOW

Previous approaches

- **Existing methods generate edited street images.**
  – **Prompt-based** image editing produces visual counterfactuals.
  – Generated edits are often **unrealistic** or **exaggerated**.

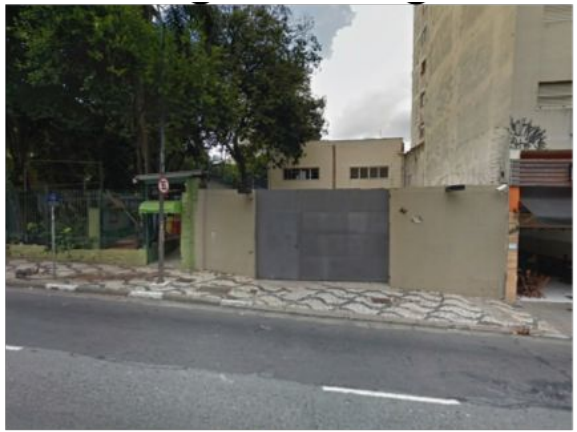
- **Existing explainability methods produce static explanations.**
  – Explain **what** should change.
  – Do **not show** what those **changes** would actually look like.

- **Limitation**
  – Current approaches either generate **ungrounded** images or provide **text-only** explanations—neither offers **realistic**, **traceable visual** evidence.

Previous approaches: prompt-based generations

make it unsafe

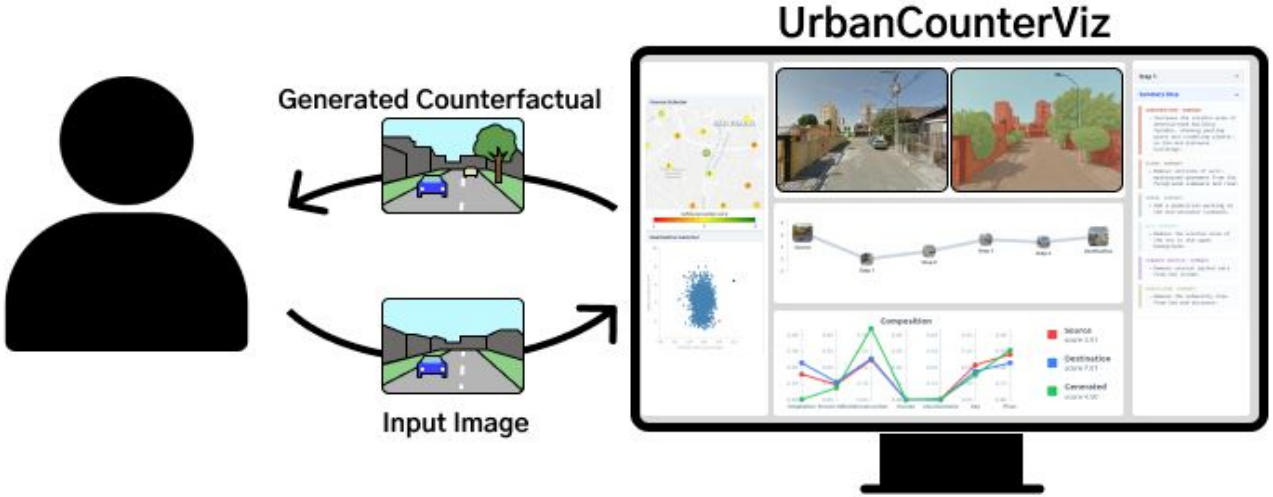
Original image



make it safe

Proposed direction: UrbanCounterViz

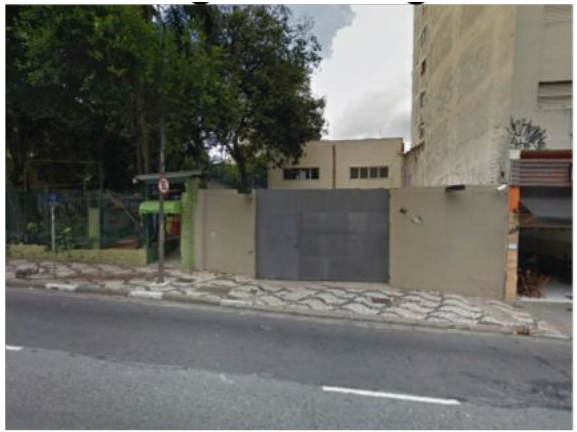
- Ground image edits on real visual evidence.
  – Derive changes from differences between **real streets**.
  – Preserve realism through **semantic edits**.
  – Keep every modification **traceable** and **inspectable**.



UrbanCounterViz

make it unsafe

Original image

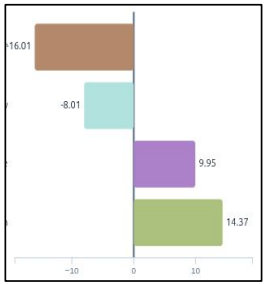


make it safe

Realistic changes rarely happen in one step

- Large visual edits often produce unrealistic transformations.
  – **small semantic changes** and **incremental changes**.

Original image



Semantic difference

Generated actions
1- Increase plants
2- Add a car
3- Repair sidewalk
4- ...

Edited image

Increase plants → Add a car → Repair sidewalk → ...

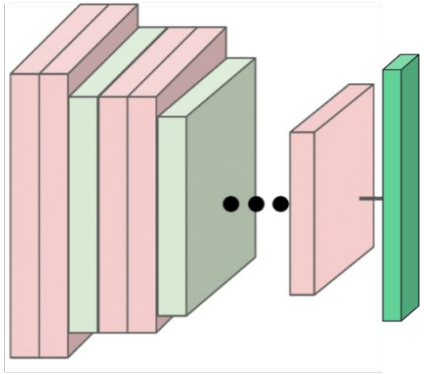
How do we find these small edits?

- Large visual changes are difficult to generate reliably
  – For each image, we search a **discordant pair**
  – Two images **considered** similar images with a large perception difference

Building a graph of realistic transitions

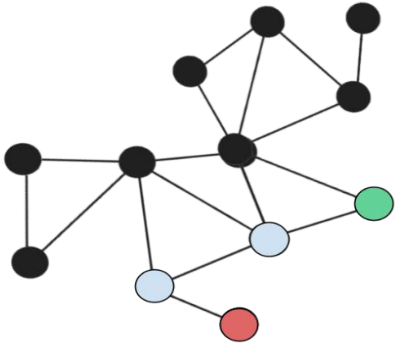
Nodes

Feature extraction



ResNet101 model

Similarity edges

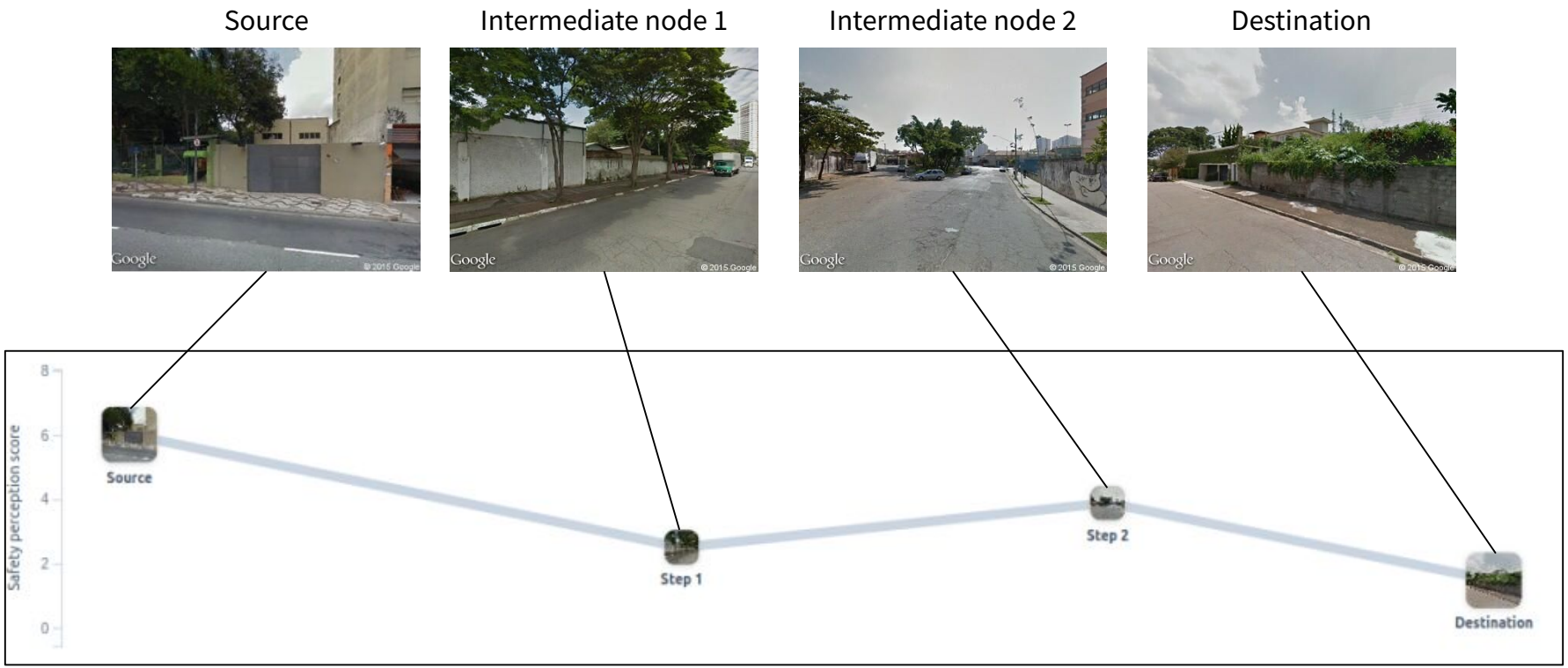


Cosine similarity **higher** than 0.8
up to 5 neighbors

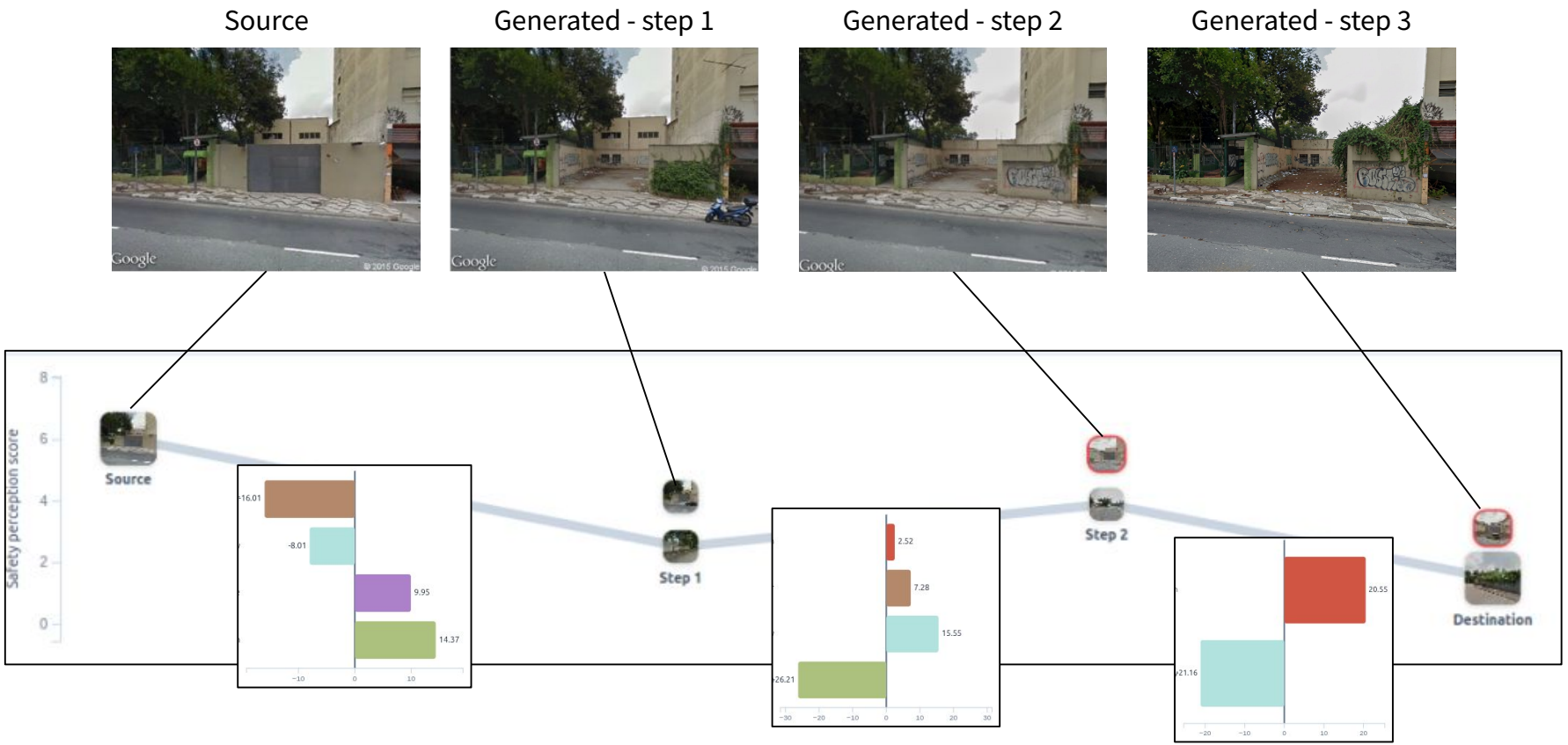
Building a graph of realistic transitions

- Connect visually similar streets.
- Each edge represents a small **semantic transition.**
- Longer transformations are obtained by traversing the graph.

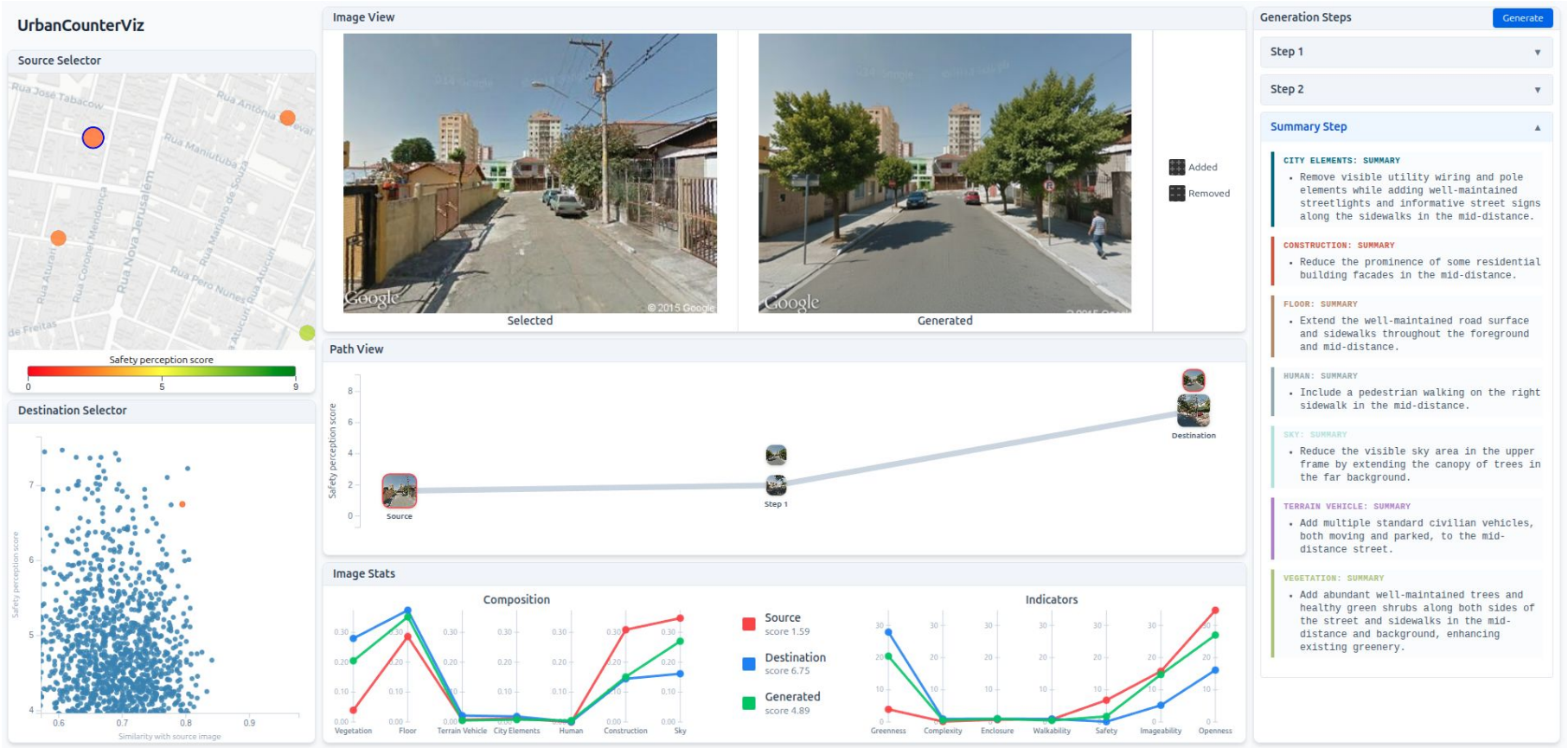
Finding realistic transformations (shortest path)



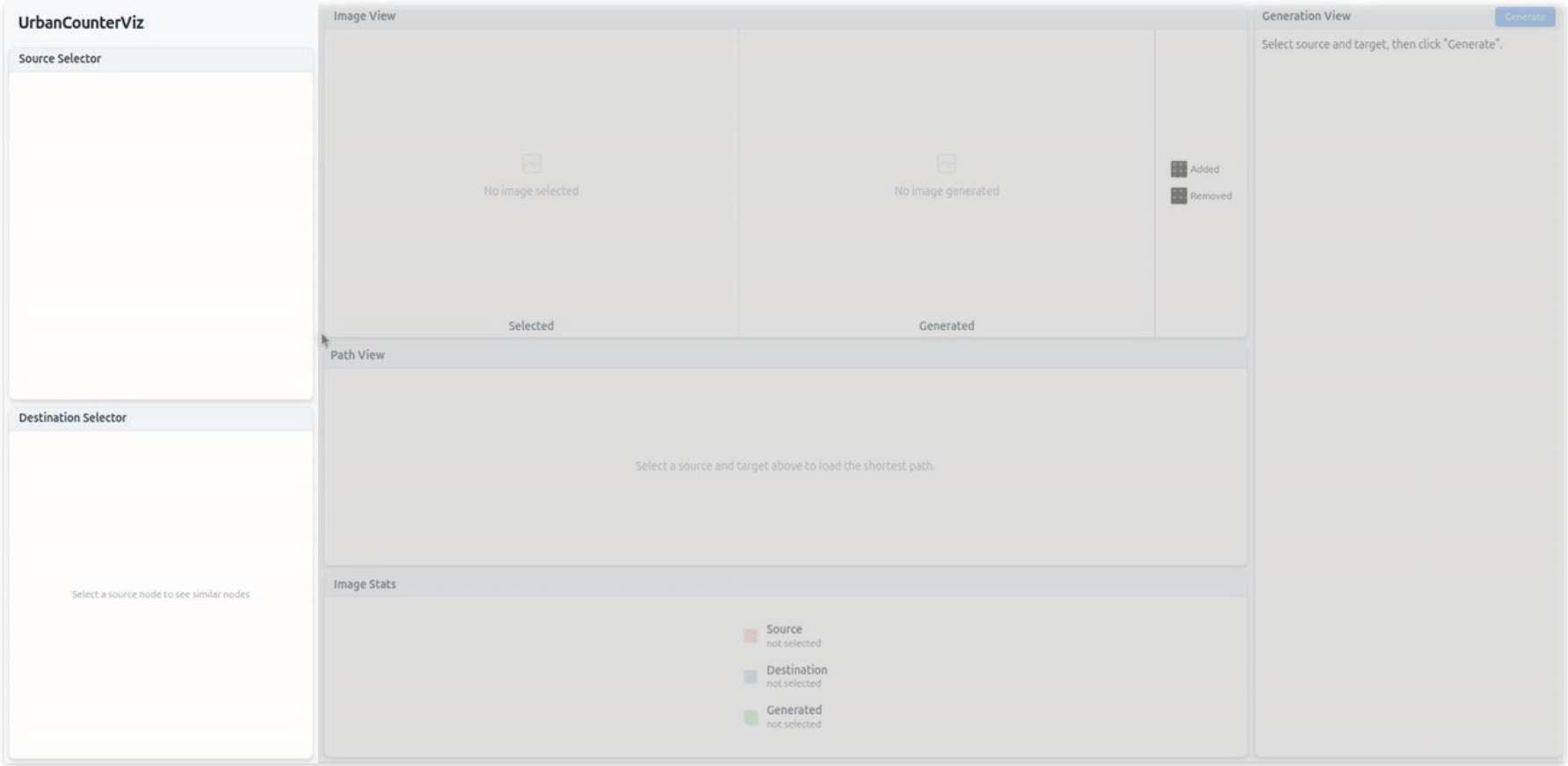
From semantic changes to visual changes



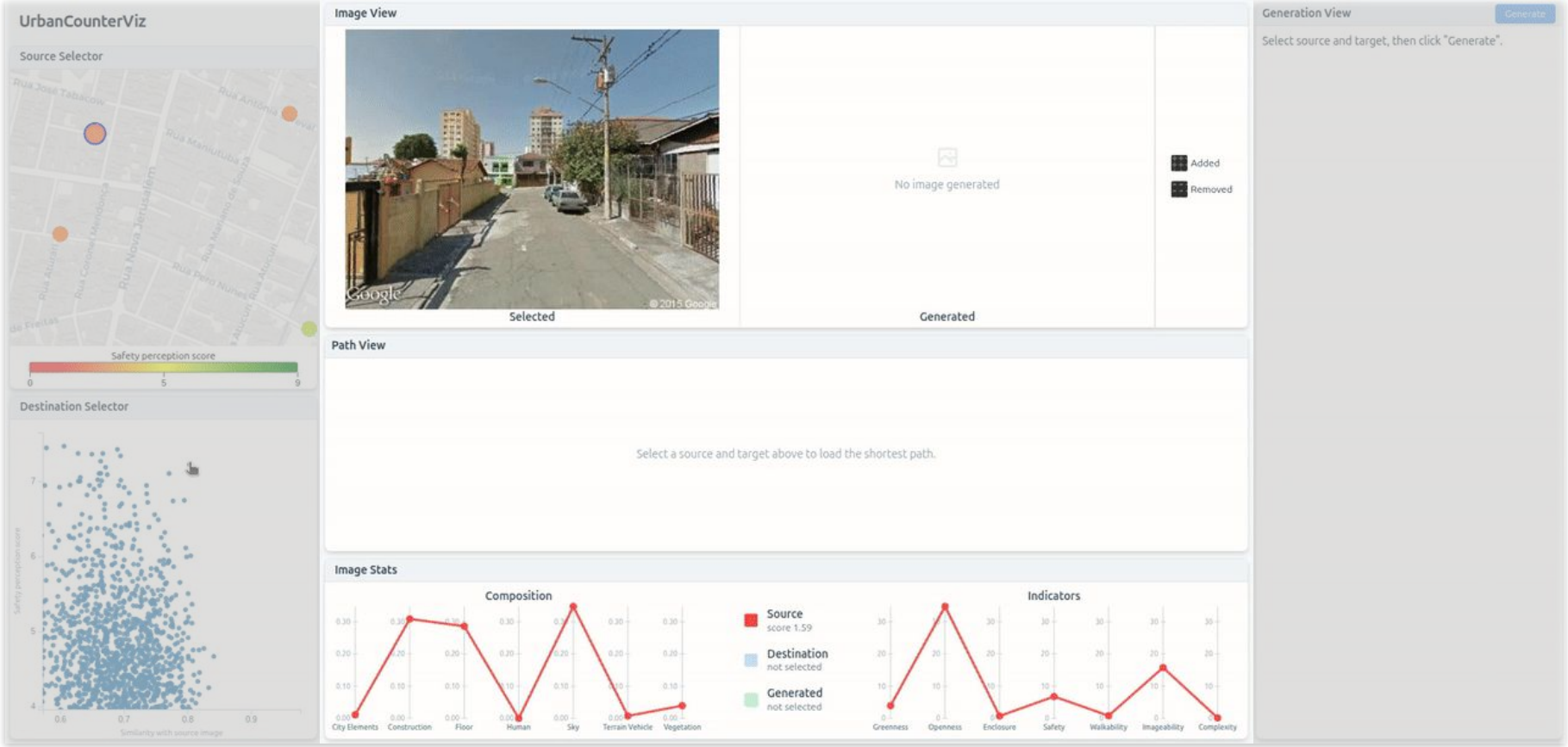
UrbanCounterViz



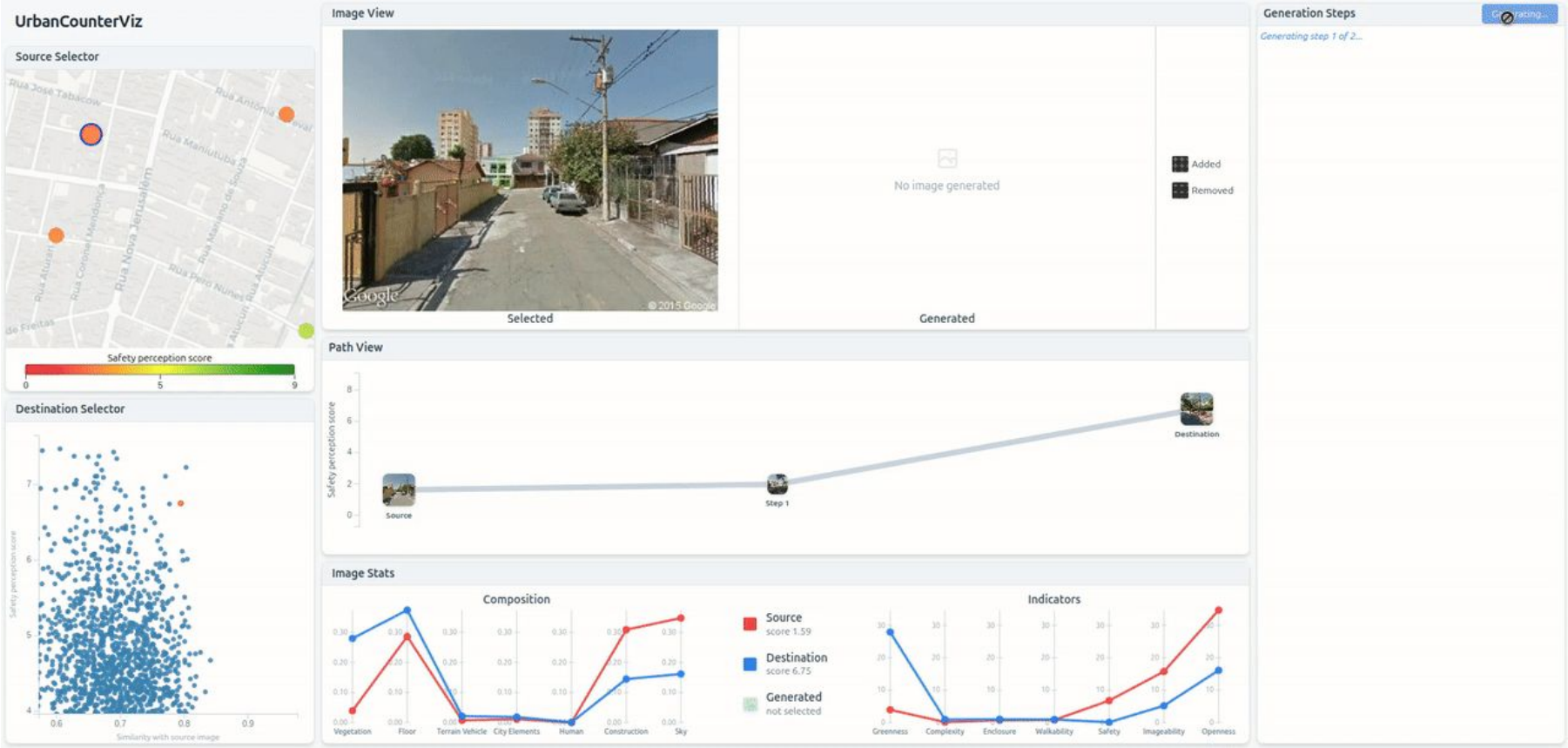
UrbanCounterViz



UrbanCounterViz



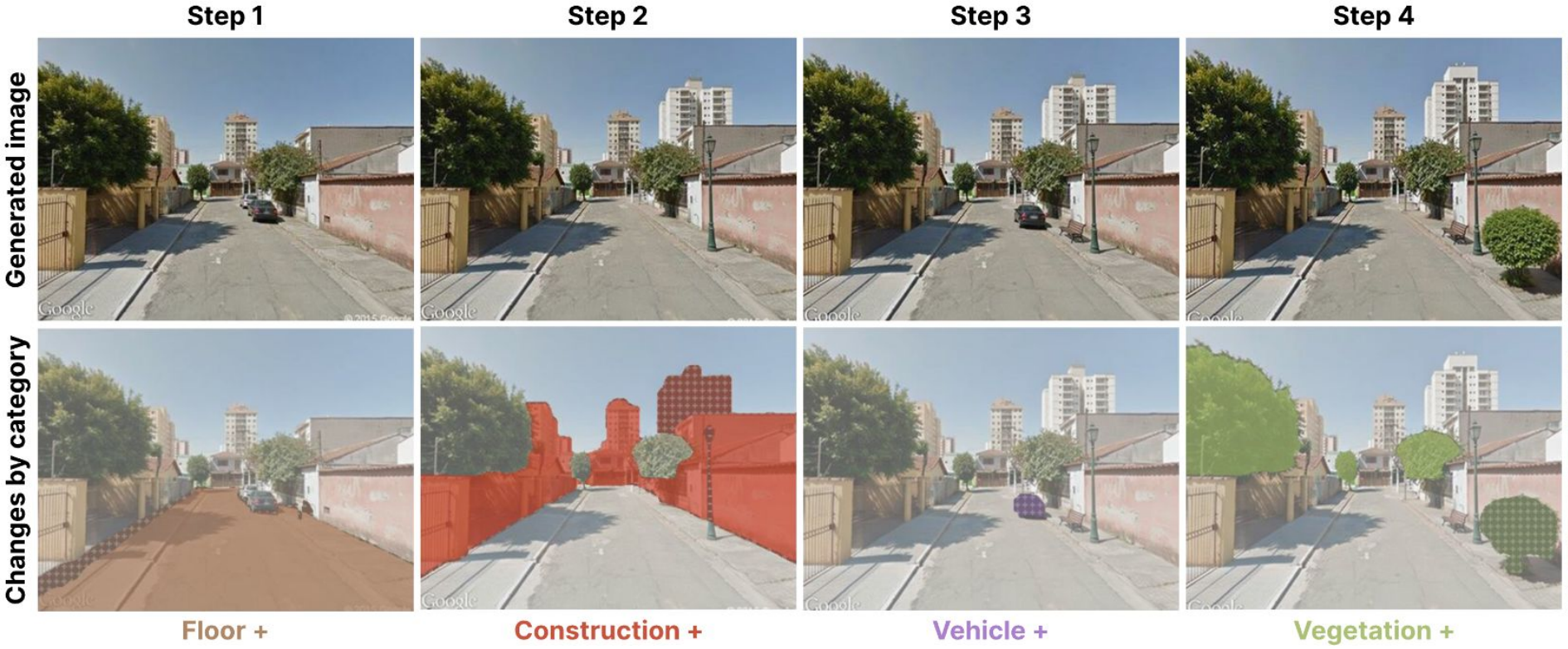
UrbanCounterViz



UrbanCounterViz

UrbanCounterViz

Counterfactual analysis: step-by-step



Expert interviews

- Three experts no co-authors.
  – Architect, Urbanism, and Urban development

- Experts explored UrbanCounterViz and 3 scenarios in a 45 minute session.

- Usage scenarios:
  – **Enhancing** Urban Safety Perception Through Discordant Pairs
  – Progressive **Decrease** in Perceived Safety
  – Multi-Step Trajectory Analysis

Scenario 1

Source score: 1.56

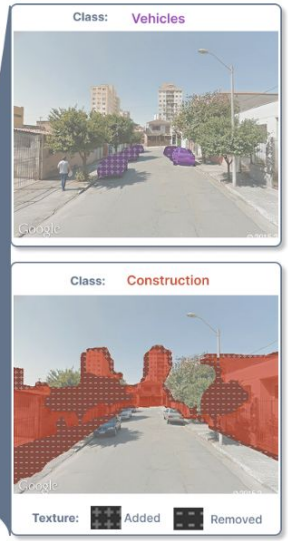
Remove the wall on the left.

Add trees on both sides of the road

Add light pole on the right

Add a vehicle below the tree

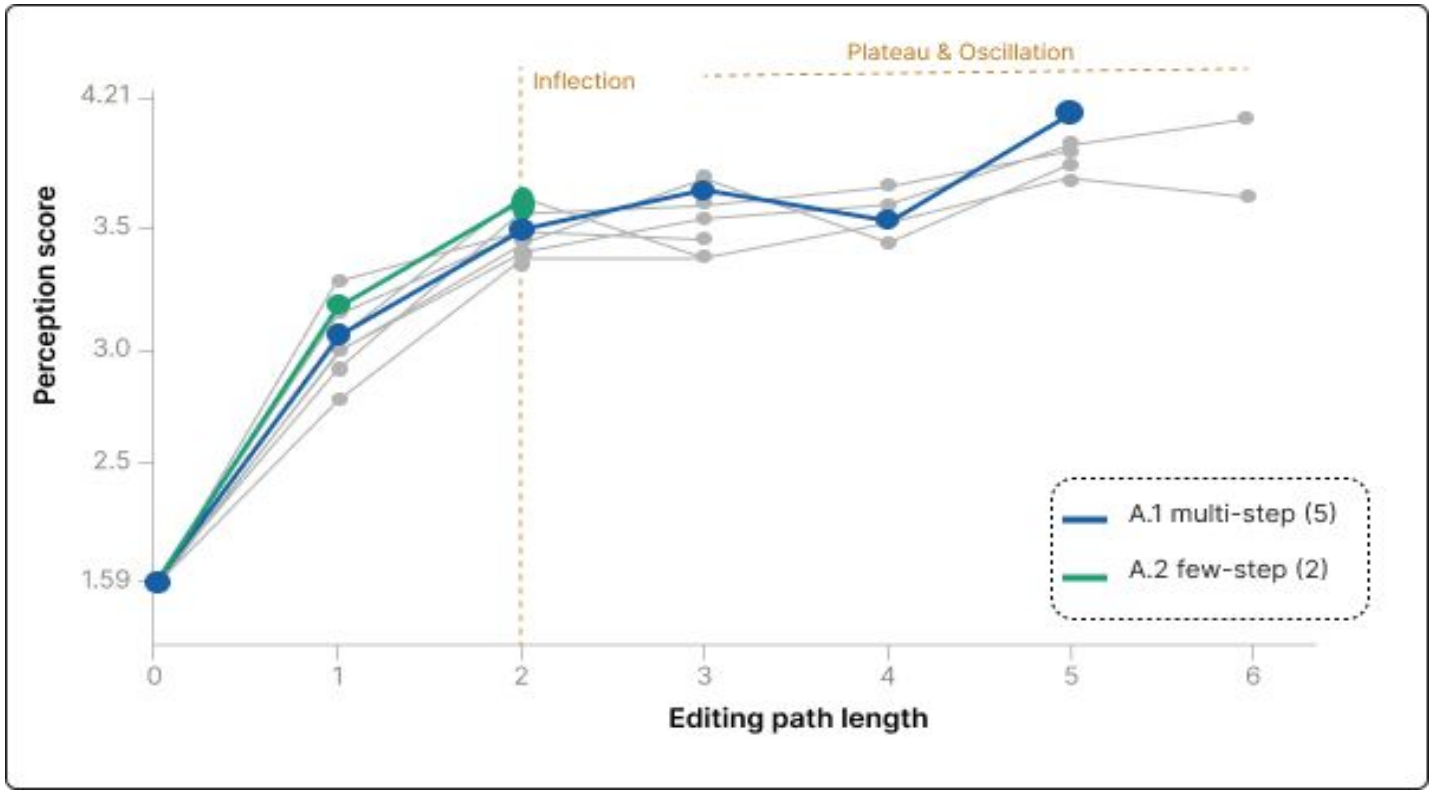
Final generated score: 6.08



Less Safe

Safer

Scenario 2


Plateau & Oscillation
Inflection
4.21
3.5
3.0
2.5
1.59
Perception score
0
1
2
3
4
5
6
Editing path length
A.1 multi-step (5)
A.2 few-step (2)

Scenario 2

A.1

construction –
vehicle –
vegetation ++

vegetation ++
human –

Scenario 2

A.2

vegetation ++
construction -

construction +
city elements +

sky -
vehicle +
vegetation ++
city elements -

construction ++
vehicle -

construction +
vehicle +

Scenario 3

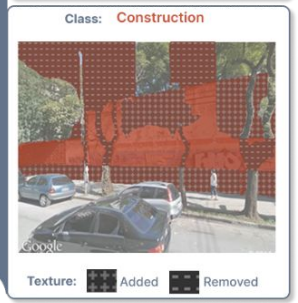
Source score: 5.48

Add a facade with peeling surfaces and graffiti.

Remove the hedge and palm trees.

Remove the background building facades.

Add overgrown street trees and vines.



Final generated score: 2.10

Safer

Less Safe

Expert feedback: Strengths

Usability

- Experts quickly understood the interface and transformation pipeline.

- The **Path View** was particularly valued for explaining how a street changes.

- The interface was considered intuitive, with **semantic change masks** helping interpret **subtle modifications**.

Usefulness

- Grounding edits on **real street differences** increased realism and credibility.

- **Small**, **localized** edits were considered the most useful and plausible.

- Recommendations involving **vegetation**, **maintenance**, and **graffiti removal** were viewed as practically meaningful.

Expert suggestions

Suggestions

- Provide a short **onboarding/tutorial** explaining perceived safety scores.

- **Clarify** that scores represent **perceived safety**, not objective crime risk.

- Improve consistency between **textual recommendations** and generated images.

Ethical considerations

- **Promote transparent interpretation** through appropriate contextual information.

- **Keep humans in the decision loop** for urban planning decisions.

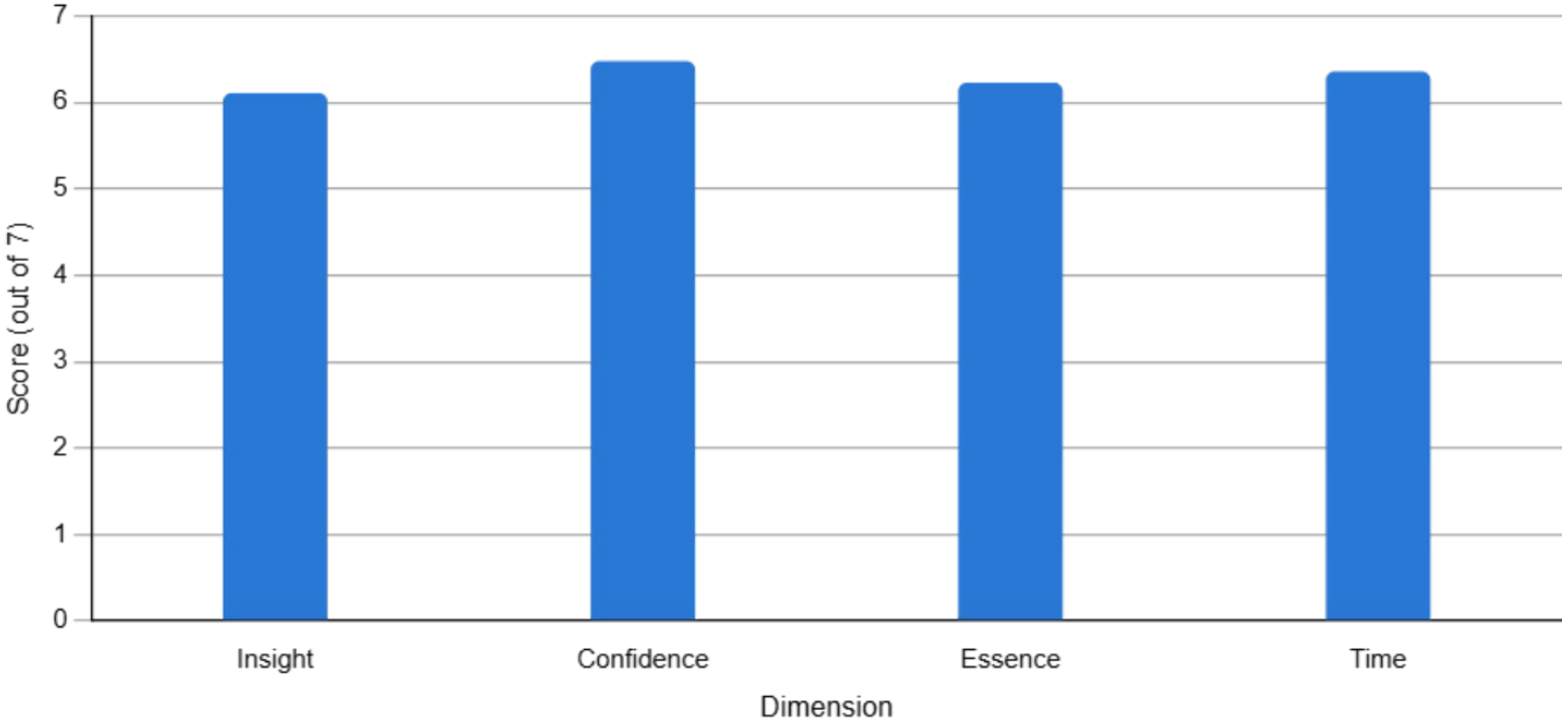
- **Urban Indicators** are an interesting metrics to measure **conditions, performance, and changes** in cities.

User study

- **22 participants** from diverse backgrounds
- No prior expertise in urban computing
- Evaluated using an **ICE-T framework subset**, focused in:
  - Insight
  - Confidence
  - Essence
  - Time

User study

All evaluations scored above 6.0/7, indicating positive usability and analytical effectiveness.



Main findings

Strengths

- **Image View** and **Generation View** made visual changes easy to understand.
- **Guided** transformations helped explain **how** perception changes.
- Interactive **tooltips reduced visual clutter** while preserving contextual information.

Improvements

- Increase prominence of the **map** during scene selection.
- Make **urban indicators** easier to interpret.
- Allow users to propose and test **their own edits**.

Takeaway: Recommendations become visual evidence

✓ What we have (GAP 3 & 4)

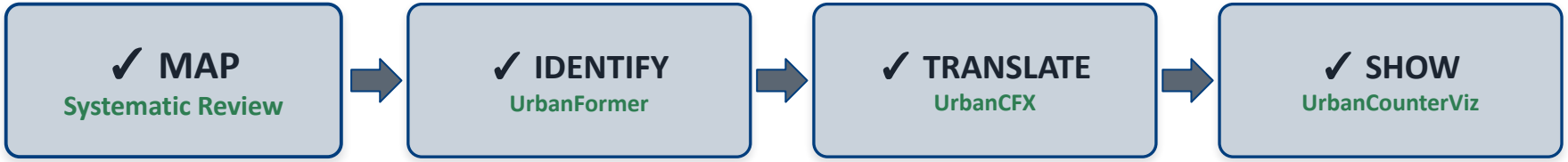
Counterfactual recommendations can be **transformed** into **realistic** and **traceable visual edits**, allowing inspection and comparisons.

✗ What remains

Generated edits are still **hypotheses**, not validated in **real** urban interventions.

Visual counterfactuals should support human judgment, not replace it.

All components presented



This work will be submitted to *EuroVis 2027*

Felipe Moreno-Vera, ***Juanpablo Heredia***, ***Mauro Diaz***, *and Jorge Poco*

Conclusions

Limitations

- **Data**
  – Relies primarily on **Place Pulse 2.0**, limiting geographical diversity.
  – Perceived safety reflects **anonymous subjective perception**.

- **Methodology**
  – Counterfactual edits represent **plausible hypotheses**, not real interventions.
  – Benchmarking and plausible metrics for **evaluating VLMs** results.

- **Evaluation**
  – Expert and user studies demonstrated the system's usefulness, but **larger** and more **diverse** participant groups are needed.
  – Larger edition steps can generate **oscillatory regime**.

Future directions

- **Interactive human-in-the-loop analysis**
  – Allow users to edit and refine recommendations before image generation.
  – Support free-form exploration of alternative urban interventions.

- **Local and cross-city adaptation**
  – Extend the framework to multiple cities, **adding disorder cues**.
  – Conduct larger studies with planners, architects, policymakers, and residents.

Conclusions

This thesis advances grounded visual analytics for urban safety perception by answering four fundamental questions.

Contribution	Question Answered	Main Outcome
MAP	*What do we know?*	A systematic review of SVI-based urban perception, identifying four research gaps.
IDENTIFY *(UrbanFormer)*	*What matters?*	Identified influential visual elements.
TRANSLATE *(UrbanCFX)*	*What should change?*	Generated human-readable recommendations using an enriched urban vocabulary.
SHOW *(UrbanCounterViz)*	*What would those changes look like?*	Produced grounded, traceable visual counterfactuals for urban analysis.

THANKS!