How Many Visual Levers Drive Urban Perception? Interventional Counterfactuals via Multiple Localised Edits

Jason Tang Stephen Law
University College London (UCL)
jasoncpits@outlook.com, stephen.law@ucl.ac.uk

Abstract

Street-view perception models predict subjective attributes such as safety at scale, but remain correlational: they do not identify which localized visual changes would plausibly shift human judgement for a specific scene. We propose a lever-based interventional counterfactual framework that recasts scene-level explainability as a bounded search over structured counterfactual edits. Each lever specifies a semantic concept, spatial support, intervention direction, and constrained edit template. Candidate edits are generated through prompt-conditioned image editing and retained only if they satisfy validity checks for same-place preservation, locality, realism, and plausibility. In a pilot across 50 scenes from five cities, the framework reveals preliminary proxy-based directional patterns and a practical failure taxonomy under prompt-only editing, with Mobility Infrastructure and Physical Maintenance showing the largest auxiliary safety shifts. Human pairwise judgements remain the ground-truth endpoint for future validation.

1. Introduction

Street-view perception models predict subjective attributes such as safety, wealth, and liveliness at scale [4, 17, 21], but they remain limited as explanatory tools: they do not show which local visual changes would plausibly shift judgement for a specific scene. Existing explainability methods, including saliency, SHAP, and concept-based probes, are largely correlational, identifying features associated with a score without testing whether manipulating them changes perceived safety [1, 10]. Urban-design research links cues such as greenery, maintenance, and visibility to perceived safety[1] [9, 12, 19], but does not establish which scene-specific local changes validly shift perception.

[1] *Perceived safety* here refers to the subjective feeling of safety from crime and disorder, as elicited in Place Pulse [21] and SPECS [20], rather than traffic safety. Some lever families, especially Mobility Infrastructure, may also evoke traffic-safety cues; whether these generalise to crime-safety perception remains open.

Following urban planning theory [8, 18], we study a restricted question: whether a scene admits multiple distinct *single-lever* interventions that, evaluated independently against the same original image, plausibly shift perception. We propose an *interventional counterfactual* protocol that makes this question testable through structured lever interventions defined by a semantic concept, spatial support, intervention direction, and constrained edit template. Candidate edits are generated with prompt-conditioned diffusion editing and retained only if a vision–language critic judges them to preserve the same place, remain localised, and appear realistic and plausible. Because generative editors may introduce non-target changes, each edit is treated as an auditable hypothesis rather than faithful evidence by default.

This paper introduces an auditable pilot framework rather than a validated causal estimate of human perception. We ask: (i) which lever families show positive auxiliary shifts after validity auditing, and (ii) how many distinct single-lever edits remain promising per scene after screening and thresholding?

Our contributions are:

1. a **lever-based interventional counterfactual** formulation for street-view perception, with structured lever interventions $l = (c, s, d, \tau)$ as the explanatory unit;
2. a **scene-specific generation-and-audit pipeline** that instantiates, realises, and validity-checks prompt-only edits; and
3. a **50-scene pilot benchmark** across five cities, reporting preliminary proxy-based directional patterns and feasibility diagnostics under prompt-only editing.

2. Related Work

Urban perception and design cues. Crowdsourced pairwise judgement datasets have enabled vision models for perceived safety, wealth, and liveliness at scale [4, 17, 20, 21], but these models are not designed to answer mechanistic scene-level questions. Urban-design theory links perceived safety to surveillance, active fontages, upkeep,

and legibility [8, 13, 18], and empirical street-view studies confirm roles for greenery, street-facing windows, maintenance, and lighting [3, 9, 12, 19]. However, none provides scene-level tests of localised, validity-audited single-lever interventions.

Interpretability and counterfactual explanations. Saliency and SHAP-style methods are known to be visually plausible yet unfaithful [1]; concept-based methods such as TCAV [10] remain correlational unless paired with explicit interventions. Counterfactual visual explanations [5] and causal evaluation benchmarks [14] provide stronger tools but typically target model logits, focusing on the effect of a single feature on the model output rather than the holistic perceptual experience of the scene which can be affected by multiple single visual changes.

Urban counterfactuals and diffusion-based editing. Law et al. [11] demonstrate plausible counterfactuals for urban image regressors, establishing that generative edits can serve as explanations in urban analytics but does not explicitly identify which edits are plausible for a given scene. UrbanPhysicalDisorder-4K [16] brings counterfactual reasoning into urban safety via annotated disorder features, but operates in feature space rather than through localised image edits with explicit validity auditing. Zhao et al. [22] propose perception-guided street-view generation, but target scene optimisation rather than scene-level interventional counterfactual analysis. Diffusion-based editors such as SDEdit [15], InstructPix2Pix [2], and Prompt-to-Prompt [6] make localised edits increasingly feasible, but can introduce non-target drift that demands explicit validity auditing.

Positioning. The diffusion-editing literature above optimises for editing fidelity, how faithfully a model executes an instruction, whereas our work sits at this intersection: rather than attributing a scene-level judgement to pixels or concepts alone, we search over structured, editor-feasible single-lever edits and retain only those that pass explicit validity auditing.

3. Method

3.1. Problem Setup

Let $x \in \mathcal{X}$ denote a street-view image and $a \in \mathcal{A}$ a perceptual attribute (e.g. safety). We take the explanatory atom to be a *lever intervention*

$$l = (c, s, d, \tau), \tag{1}$$

where c is a lever concept (e.g. greenery, graffiti), s is the grounded scene support, d the intervention direction (add, remove, repair), and τ a constrained edit template.

For image x, let $L(x)$ denote the scene-specific candidate set instantiated from an editor-feasible intervention vocabulary. Our method identifies the subset that can be realised as valid edits:

$$V(x) = \{l_i \in L(x) \mid \exists\, \tilde{x}_i \text{ passing the validity audit}\}. \tag{2}$$

Each lever is tested independently against the unedited original; edits are never composed sequentially.

3.2. Phase 1: Lever Candidate Construction

We define an *editor-feasible intervention vocabulary*: a bounded set of lever concepts chosen to satisfy three criteria: **(i)** *theoretical grounding* in the urban perception literature [8, 9, 12, 18, 19], **(ii)** *semantic distinctness* so concepts remain interpretable, and **(iii)** *prompt-only editability* without requiring segmentation masks or major structural changes. The full concept list is given in Appendix A1.

For each image, a VLM planner grounds applicable vocabulary concepts to visible scene regions, producing scene-specific candidates such as *remove graffiti from this wall* or *add modest greenery near this entrance*. These grounded candidates form $L(x)$.

3.3. Phase 2: Bounded Stochastic Intervention

For each candidate l_i, a prompt-only image generator G produces

$$x'_{i,j} = G(x;\, c_i, s_i, d_i, \tau_i, \varphi_j), \tag{3}$$

where φ_j indexes stochastic draws under a bounded retry budget T, with G being a prompt-only image editing model.

Each candidate is subjected to an automated validity audit performed by a vision–language model, evaluating: **(1)** same-place preservation, **(2)** locality to the intended support, **(3)** realism, and **(4)** plausibility of the intended lever. If at least one draw passes within T attempts, the first accepted edit $\tilde{x}_i$ is retained and l_i enters $V(x)$. Only edits passing all four criteria enter the retained set, so auxiliary shifts are measured only on edits with confirmed semantic and spatial integrity. We report valid-edit coverage $\text{Coverage}(x) = |V(x)|/|L(x)|$ as a property of edit realisability under bounded prompt control rather than an urban perception signal.

3.4. Phase 3: Model-based Evaluation

For each accepted edit $\tilde{x}_i$, we compute

$$\Delta_i = f_a(\tilde{x}_i) - f_a(x), \tag{4}$$

where f_a is a ViT-B/16 perception model pretrained on the MIT Place Pulse 2.0 pairwise safety dataset [7], which outputs a continuous score on a 0–10 scale (higher = safer). This is strictly a model-based evaluation, not the main estimand; f_a has not been tuned on edited images, so Δ_i should be treated as a ranking signal. The proxy-shortlisted set is

$$E_{\text{aux}}(x, a) = \{l_i \in V(x) \mid \Delta_i > \theta_{\text{aux}}\}, \tag{5}$$

where θ_{aux} is an exploratory threshold. Human-grounded pairwise evaluation remains future work.[2]

4. Preliminary Results

Dataset and setup. We use SPECS (Street Perception Evaluation Considering Socioeconomics) [20], a demographically balanced survey in which 1,000 participants from five countries rated street-view scenes on ten perceptual indicators. We sample $N{=}50$ scenes from five SPECS cities (Amsterdam, Abuja, San Francisco, Santiago, Singapore; 10 per city), stratified by baseline safety and visual complexity to cover a broad range of starting conditions. Each scene allows up to $K{=}5$ lever candidates with a stochastic budget of $T{=}5$ generation attempts per candidate. We instantiate the framework [3] with Qwen-Image-Edit as the prompt-only editor and GPT-5.4 as the LLM-as-judge critic, yielding a reproducible baseline. Validity auditing acts as an eligibility gate: all reported effect summaries are computed only on edits retained in $V(x)$.

Proxy safety-score shifts. For each valid edit $\tilde{x}_i$, we compute the model-based score delta Δ_i using the ViT-B/16 safety proxy defined in §3; positive values indicate a higher predicted safety score relative to the unedited original. Among the 177 valid edits, the proxy produces a mean shift of $+0.366$ (95% CI $[+0.199, +0.537]$, median $+0.184$, range $[-3.624, +3.842]$). Using threshold $\theta_{\text{aux}}{=}0.1$, 95 of 177 valid edits fall in $E_{\text{aux}}(x, a)$, corresponding to a mean of 1.90 proxy-shortlisted levers per scene (95% CI $[1.476, 2.324]$); 40 of the 50 scenes retain at least one proxy-shortlisted lever and 24 retain multiple. The largest positive proxy shifts are observed for lane-marking repainting, crosswalk repainting, and localized greenery addition.

Mobility Infrastructure yields the largest family-level mean auxiliary shift ($+0.579$), led by crosswalk and lane-marking repainting. These interventions are conventionally associated with traffic safety rather than crime safety; whether the proxy shift reflects a genuine crime-safety signal or a scorer artefact remains an open question (see safety-definition footnote in §1). Physical Maintenance is the broadest consistently positive family ($+0.344$), driven by surface cleaning, litter removal, facade repair, and graffiti removal. Environmental Amenity is more heterogeneous: localized greenery addition is strongly positive ($+0.650$), whereas lighting repair and tree canopy management are weak or negative.[4] Visual Legibility is small-sample and mixed in sign; the negative mean for signage decluttering (-0.631) may reflect the proxy associating commercial signage density with activity in Place Pulse training data. Per-lever proposal and valid-rate diagnostics are in Appendix Tables A2 and A3.

Table 1. Lever-type results grouped by intervention family. Mean Δ_{aux} and Δ_{aux} [95% CI] are computed over valid edits only where $\theta_{\text{aux}}{>=}0.1$; confidence intervals are bootstrap percentile intervals over the retained set.

Family / Lever concept	Valid	Mean Δ_{aux}	Δ_{aux} [95% CI]
Physical Maintenance			
Graffiti removal	4	$+0.396$	$[+0.121, +0.670]$
Litter removal	25	$+0.296$	$[-0.128, +0.743]$
Facade repair	4	$+0.519$	$[+0.105, +0.949]$
Surface cleaning	38	$+0.352$	$[+0.026, +0.684]$
Family total	71	$+0.344$	$[+0.107, +0.584]$
Environmental Amenity			
Localized greenery add.	32	$+0.650$	$[+0.241, +1.070]$
Lighting repair	8	-0.232	$[-0.598, +0.173]$
Tree canopy management	14	-0.245	$[-0.952, +0.446]$
Family total	54	$+0.287$	$[-0.041, +0.610]$
Visual Legibility			
Signage decluttering	5	-0.631	$[-1.294, -0.151]$
Storefront transparency	2	$+0.958$	$[+0.184, +1.732]$
Family total	7	-0.177	$[-0.927, +0.599]$
Mobility Infrastructure			
Crosswalk repainting	15	$+0.767$	$[+0.255, +1.261]$
Lane marking repainting	30	$+0.485$	$[+0.045, +0.934]$
Family total	45	$+0.579$	$[+0.244, +0.923]$
Overall	177	$+0.366$	$[+0.199, +0.537]$

Thresholded proxy-shortlisted lever counts. Appendix Figure A6 counts levers that are both valid and directionally notable at increasingly strict auxiliary thresholds. At $\theta_{\text{aux}}{=}0.1$, Physical Maintenance averages 0.76 proxy-shortlisted levers per scene, Mobility Infrastructure 0.56, and Environmental Amenity 0.54; Visual Legibility averages only 0.04. By city, San Francisco and Santiago dominate the positive tail with 2.5 and 2.6 proxy-shortlisted levers per scene respectively, while Abuja averages 1.1. As the cutoff rises to 1.0, the family ranking compresses but Environmental Amenity and Mobility Infrastructure retain the largest positive tails.

Santiago and San Francisco show the largest positive mean shifts; Abuja and Singapore are near zero or slightly negative. Valid-rate diagnostics by city are in Appendix Table A3.

Baseline score and gate-qualified editability. Appendix Figure A5 plots each scene's baseline safety score $f_a(x)$ against its valid lever count $|V(x)|$. At $N{=}50$, no meaningful monotonic relationship is observed (Spearman $\rho{=}0.10$, $p{=}0.511$): scenes with lower baseline safety do not admit more valid levers than higher-baseline scenes. We revisit this null pattern in §5.

Validity gate. Of 250 proposed candidates, 177 pass the

[2] Our method is constrained interventional search, not pixel-level attribution or causal identification. "Multiple levers" means distinct interventions that each pass validity checks independently; it does not imply additive effects. Locality is audited, not pixel-guaranteed.

[3] The framework is model-agnostic: any component (e.g., planner, editor, auditor) can be swapped, finetuned, or extended as needed.

[4] Tree canopy management refers to pruning or trimming overgrown canopy to improve sightlines and pedestrian visibility, the opposite direction from greenery addition. The ViT-B/16 safety model, trained on Place Pulse images where visible greenery correlates with higher safety, may penalise any net reduction in canopy cover regardless of the urbanistic intent.

Table 2. City-level summary over the full $N=50$ run. Mean Δ_{aux} and Δ_{aux} [95% CI] are computed over valid edits only. Coverage diagnostics are reported in Appendix Table A3.

City	Valid	Mean Δ_{aux}	Δ_{aux} [95% CI]
Amsterdam	40	$+0.400$	$[+0.073, +0.764]$
Abuja	32	-0.136	$[-0.441, +0.141]$
San Francisco	36	$+0.704$	$[+0.285, +1.105]$
Santiago	38	$+0.742$	$[+0.398, +1.089]$
Singapore	31	-0.012	$[-0.383, +0.352]$
Overall	177	$+0.366$	$[+0.199, +0.537]$

validity audit (rate 0.708); all 50 scenes retain at least one valid lever. Detailed coverage and failure analysis are in Appendix A.3; representative rejected counterfactuals are shown in Appendix Figure A3.

Qualitative examples. Three positive auxiliary-shift examples – lane-marking repainting in Amsterdam, facade repair in Singapore, and surface cleaning in Santiago – are shown in Appendix Figure A2. All three are spatially constrained, visually legible, and plausible as single-lever urban changes.

5. Discussion and Future Work

This pilot establishes two outputs: a feasibility profile for the generation-and-audit pipeline, and preliminary proxy-based directional patterns to prioritise for human evaluation. On feasibility, roadway-marking and maintenance interventions realise reliably across cities, while lighting repair remains the clearest bottleneck. No meaningful relationship is observed between baseline safety and valid lever count, suggesting that multi-lever richness depends more on scene geometry than starting safety level. On directionality, Mobility Infrastructure yields the largest retained auxiliary shifts, though it remains unclear whether this reflects human perceptual sensitivity, scorer priors, or edit realisability.

Environmental Amenity is more heterogeneous. Localised greenery addition is strongly positive, whereas lighting repair and canopy management are negative on average despite being theory-grounded cues. This contrast suggests that the auxiliary scorer may encode a simple "more green = safer" prior rather than a more nuanced sensitivity to sightlines or surveillance, and that family-level aggregation can obscure important lever-specific differences.

On the model-based evaluation proxy. The scorer f_a has not been validated on edited images, so Δ_i should not be read as a calibrated effect size. However, for monotone interventions (road markings appear, graffiti is removed, greenery is added), the directional signal is grounded in large-scale crowdsourced pairwise preference over this class of scene variation, making the sign and relative magnitude of Δ_{aux} a useful prioritisation signal for selecting which levers and scenes to send to human evaluation.

XAI contribution. Beyond urban perception, the methodological contribution is a stronger standard for visual explanation: the proposed change must be semantically coherent and generatively feasible before it enters the attribution result, providing a reusable scaffold for interventional XAI in other perceptual domains.

Future directions and human endpoint. The present pipeline is intentionally minimal, prompt-only generation, a prompted VLM critic, and an auxiliary proxy scorer are used here to establish a reproducible baseline rather than a fully validated explanation system. Four directions follow naturally from the pipeline structure. *(i) Planner tuning:* learning to predict lever viability from scene features, using the validity and Δ_{aux} outcomes reported here as supervision, so that the K-candidate budget is spent on levers most likely to be both realisable and promising under the auxiliary proxy. *(ii) Generator alignment:* optimising for validity-gate criteria via reward-from-feedback or mask-constrained generation; the lighting-repair failure mode (valid rate 0.30) is the clearest target, with per-family valid rates as the benchmark metric. *(iii) Critic calibration:* aligning the VLM critic with human raters and ultimately learning a critic that predicts 2AFC preference directly, reducing the gap between the auxiliary proxy and the human endpoint. *(iv) End-to-end optimisation:* jointly training planner, generator, and critic once each component is independently validated. The ground-truth estimand throughout is pairwise human safety ratings on passing edits, collected via a randomised 2AFC study. Until then, validity should be read as an eligibility control on image generation and auxiliary shifts as a prioritisation signal for which lever families and cities to send to human evaluation first.

6. Conclusion

The 50-scene pilot shows that structured, validity-audited single-lever edits can be generated at scale, with Mobility Infrastructure and Physical Maintenance showing the most consistent positive shifts and therefore the clearest priorities for human evaluation. More broadly, the framework shifts explanation from post-hoc pixel attribution to semantically grounded interventions. Realising that potential will require better planning, editing control, critic calibration, and ultimately pairwise human judgement as the ground-truth endpoint.

References

[1] Julius Adebayo, Justin Gilmer, Michael Christoph Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 1, 2

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[3] Marco De Nadai, Radu-Laurentiu Vieriu, Gabriele Zen, Suzana Dragicevic, Nikhil Naik, Michele Caraviello, Cesar A. Hidalgo, Nicu Sebe, and Bruno Lepri. Are safer looking neighborhoods more lively? a multimodal investigation into urban life. In *Proceedings of the 24th ACM International Conference on Multimedia*, 2016. 2

[4] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and Cesar A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *Computer Vision – ECCV 2016*, pages 196–212. Springer, 2016. 1

[5] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2376–2384, 2019. 2

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. 2

[7] Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes – a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:216–238, 2024. 2

[8] Jane Jacobs. *The Death and Life of Great American Cities*. Random House, 1961. 1, 2, 6

[9] Bin Jiang, Cecilia Nga Sze Mak, Hua Zhong, Linda Larsen, and Christopher John Webster. From broken windows to perceived routine activities: Examining impacts of environmental interventions on perceived safety of urban alleys. *Frontiers in Psychology*, 9:2450, 2018. 1, 2, 6

[10] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677, 2018. 1, 2

[11] Stephen Law, Rikuo Hasegawa, Brooks Paige, Chris Russell, and Andrew Elliott. Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals. *International Journal of Geographical Information Science*, 37:2575–2596, 2023. 2

[12] Xiaojiang Li, Chuanrong Zhang, and Weidong Li. Does the visibility of greenery increase perceived safety in urban areas? evidence from the place pulse 1.0 dataset. *ISPRS International Journal of Geo-Information*, 4(3):1166–1183, 2015. 1, 2, 6

[13] Linda J. Loewen, G. Daniel Steel, and Peter Suedfeld. Perceived safety from crime in the urban environment. *Journal of Environmental Psychology*, 13(4):323–331, 1993. 2

[14] Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A. Tsaftaris. Benchmarking counterfactual image generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. 2

[15] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[16] Felipe A. Moreno, Andres De La Puente, and Jorge Poco. Urbanphysicaldisorder-4k: Understanding urban perception via counterfactuals and street view signs of physical disorder. In *IEEE International Conference on Big Data*, pages 5194–5200, 2025. 2

[17] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar A. Hidalgo. Streetscore – predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 779–785, 2014. 1

[18] Oscar Newman. *Defensible Space: Crime Prevention through Urban Design*. Macmillan, New York, 1972. 1, 2, 6

[19] Boris A. Portnov, Rasha Saad, Tal Trop, Doron Kliger, and Anna Svechkina. Linking nighttime outdoor lighting attributes to pedestrians' feeling of safety: An interactive survey approach. *PLOS ONE*, 15(11):e0242172, 2020. 1, 2, 6

[20] Matias Quintana, Youlong Gu, Xiucheng Liang, Yujun Hou, Koichi Ito, Yihan Zhu, Mahmoud Abdelrahman, and Filip Biljecki. Global urban visual perception varies across demographics and personalities. *Nature Cities*, 2(11):1092–1106, 2025. 1, 3, 6

[21] Philip Salesses, Katja Schechtner, and Cesar A. Hidalgo. The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*, 8(7):e68400, 2013. 1

[22] Chenbo Zhao, Yoshiki Ogawa, Shenglong Chen, Takuya Oki, and Yoshihide Sekimoto. Street space quality improvement: Fusion of subjective perception in street view image generation. *Information Fusion*, 125:103467, 2026. 2

A. Appendix

A.1. Pipeline and Intervention Vocabulary

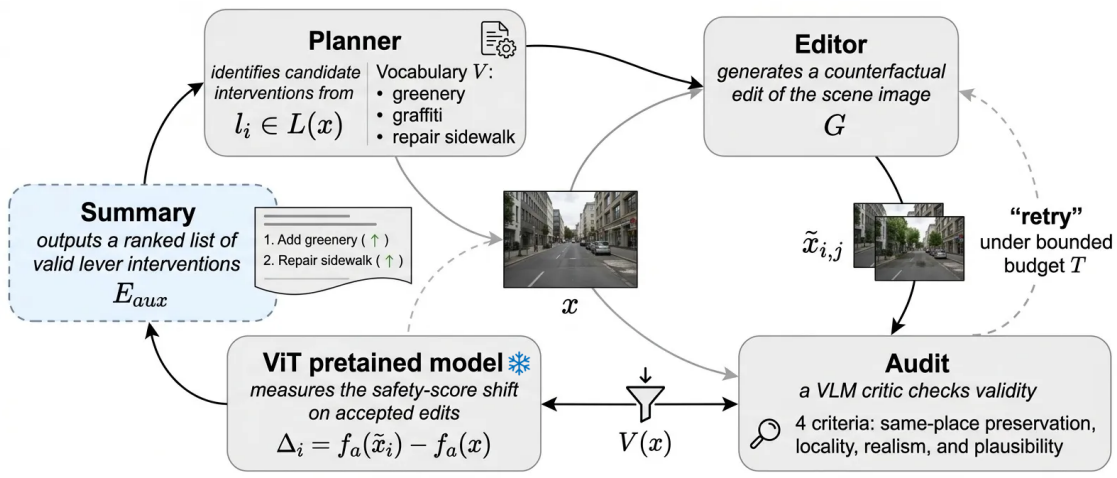

Figure A1. Overview of the lever-based interventional counterfactual pipeline. Each candidate is a structured lever intervention $l = (c, s, d, \tau)$. Phase 1 constructs scene-specific candidates from a curated ontology. Phase 2 performs bounded stochastic generation subject to a four-criterion validity audit. Phase 3 summarises accepted edits with auxiliary classifier scoring and, optionally, future human pairwise judgements.

Table A1. Intervention vocabulary. Each concept satisfies all three inclusion criteria defined in §3. Structural interventions and social-order cues are excluded as they are not realisable under prompt-only editing.

Family	Basis	Lever concepts
Physical Maintenance	Broken-windows / upkeep [9, 18]	Graffiti removal; litter removal; facade repair; surface cleaning; shutter repair
Environmental Amenity	Biophilic safety / visibility [12, 19]	Localised greenery addition; lighting repair; tree canopy management
Visual Legibility	Active-frontages / surveillance [8]	Signage decluttering; storefront transparency increase
Mobility Infrastructure	Walkability / road markings [20]*	Crosswalk repainting; lane marking repainting

*See safety-definition footnote in §1.

A.2. Qualitative Examples

Figure A2. Three qualitative examples from the retained set. From top: lane-marking repainting (Amsterdam), facade repair (Singapore), surface cleaning (Santiago). Red dashed bounding boxes indicate the edited region; surrounding areas are dimmed for contrast.

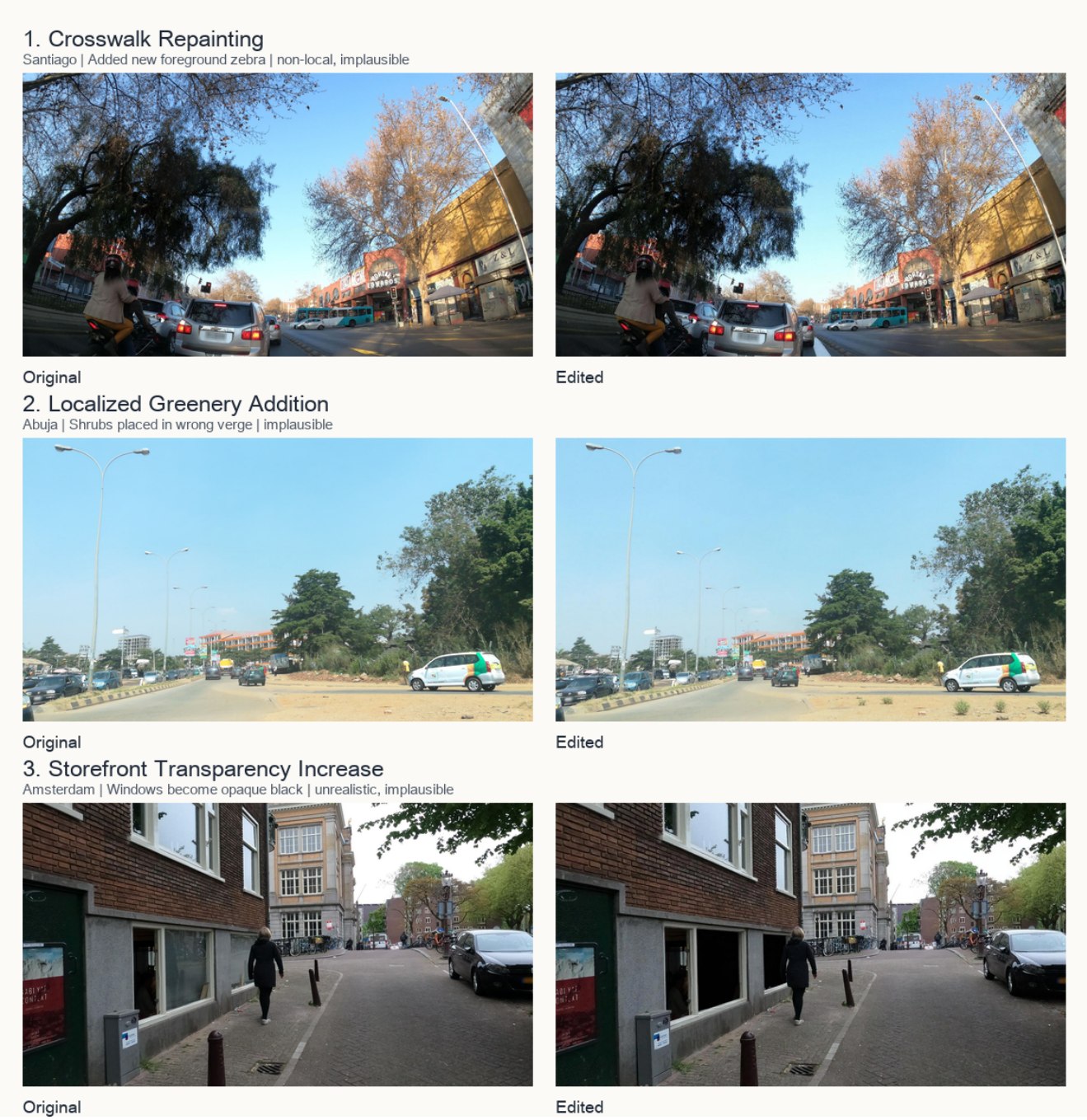

Figure A3. Three representative rejected counterfactuals from the audited-but-rejected set. From top: crosswalk repainting adds a new foreground zebra crossing rather than repairing the supported intersection marking (Santiago), localised greenery addition places shrubs in the wrong roadside verge (Abuja), and storefront transparency increase turns the target windows into opaque black openings rather than revealing a plausible interior (Amsterdam).

A.3. Validity and Coverage Details

Across 50 scenes the planner produces the full 250 candidate rows (5.0 per image on average), of which 177 pass the validity audit, yielding mean coverage $|V(x)|/|L(x)| = 0.708$. Figure A4 shows the distribution of valid counts per scene: 2 scenes have one valid lever, 3 have two, 16 have three, 24 have four, and 5 have all five. The dominant bottleneck is not candidate proposal but valid realisation. Among the 73 audited-but-rejected edits, plausibility dominates completely; 45 also show no

discernible target change, 25 exhibit non-local drift, 5 are unrealistic, and none fail same-place preservation. Representative rejected counterfactuals are shown in Figure A3. Generator failure before audit is zero in the cleaned $N{=}50$ run, so the remaining variance is concentrated in valid realisation rather than missing candidate production.

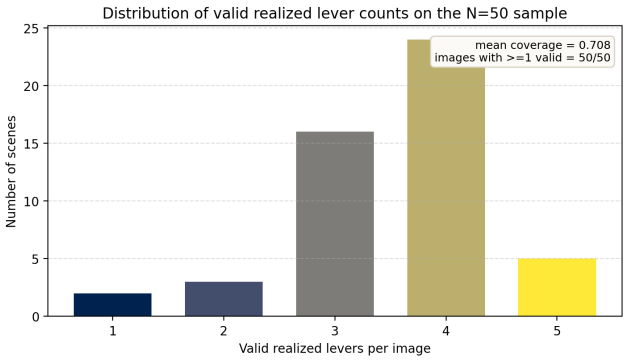

Figure A4. Distribution of valid realised lever counts per scene ($N{=}50$). Most scenes admit three or four independently valid single-lever interventions.

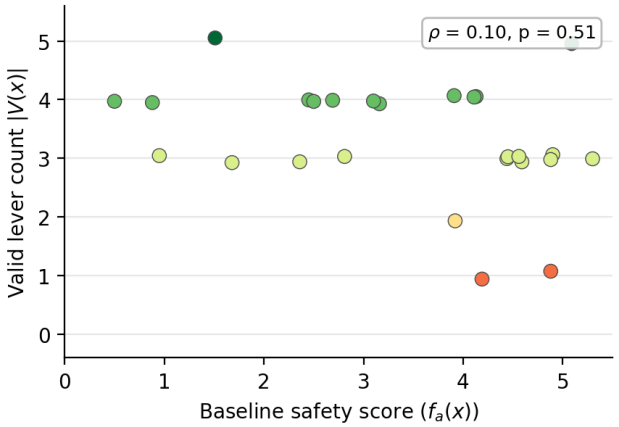

Figure A5. Baseline safety score vs. valid lever count per scene ($N{=}50$). No meaningful relationship is observed (Spearman $\rho{=}0.10$, $p{=}0.511$).

Table A2. Family-level summary over the full $N{=}50$ run. Valid-rate intervals are 95 % Wilson confidence intervals; mean Δ_{aux} and its 95 % CI are computed over valid edits only. Δ_{aux} values are on the 0–10 proxy scale.

Family	Valid / Prop.	Rate [95 % CI]	Mean Δ_{aux}	Δ_{aux} [95 % CI]
Physical Maintenance	71 / 92	0.77 [0.68, 0.85]	+0.344	[+0.107, +0.584]
Environmental Amenity	54 / 82	0.66 [0.55, 0.75]	+0.287	[−0.041, +0.610]
Visual Legibility	7 / 10	0.70 [0.40, 0.89]	−0.177	[−0.927, +0.599]
Mobility Infrastructure	45 / 66	0.68 [0.56, 0.78]	+0.579	[+0.244, +0.923]
Overall	177 / 250	0.71 [0.65, 0.76]	+0.366	[+0.199, +0.537]

Table A3. City-level coverage and directional summary over the full $N{=}50$ run. Each city contributes 10 scenes and therefore 50 proposed candidates. Valid-rate intervals are 95 % Wilson confidence intervals; mean Δ_{aux} and its 95 % CI are computed over valid edits only. Δ_{aux} values are on the 0–10 proxy scale.

City	Valid / Prop.	Rate [95 % CI]	Mean Δ_{aux}	Δ_{aux} [95 % CI]
Amsterdam	40 / 50	0.80 [0.67, 0.89]	+0.400	[+0.073, +0.764]
Abuja	32 / 50	0.64 [0.50, 0.76]	−0.136	[−0.441, +0.141]
San Francisco	36 / 50	0.72 [0.58, 0.83]	+0.704	[+0.285, +1.105]
Santiago	38 / 50	0.76 [0.63, 0.86]	+0.742	[+0.398, +1.089]
Singapore	31 / 50	0.62 [0.48, 0.74]	−0.012	[−0.383, +0.352]
Overall	177 / 250	0.71 [0.65, 0.76]	+0.366	[+0.199, +0.537]

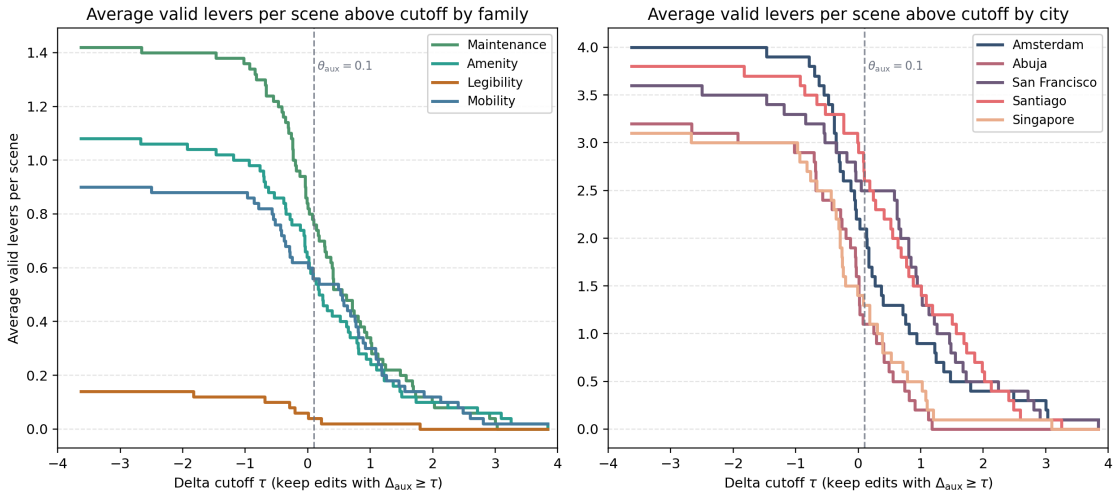

Figure A6. Average number of retained levers per scene satisfying $\Delta_{\text{aux}} \geq \tau$, split by lever family (left) and city (right). The vertical line marks the operating threshold θ_{aux}=0.1. Annotations report the overall mean Δ_{aux} with 95 % CI and the mean number of proxy-shortlisted levers per scene with 95 % CI.

A.4. Prompt Contracts

The planner, editor, and critic each operate under a structured prompt contract. Condensed versions are shown below; full prompts are in the released repository.

Planner prompt (condensed)

```
You are an urban perception planner. Given a street-view image and target percept, propose a constrained set of candidate lever interventions.

ONTOLOGY: Choose only from four families --- Physical Maintenance, Environmental Amenity, Visual Legibility, Mobility Infrastructure --- and prefer cross-family diversity when the scene supports it.

HARD CONSTRAINTS: (1) One lever per candidate. (2) Grounded in a visible scene element. (3) Local, plausible, prompt-only friendly. (4) No global relighting, weather, or camera changes. (5) Prefer the smallest plausible intervention. (6) Exclude theoretically relevant levers whose target element is not clearly visible/editable in the image.

DIVERSITY: Return distinct candidates using different lever concepts and avoid magnitude variants of the same intervention.

Return JSON with field ``candidates'', each containing: lever_concept, lever_family, scene_support, target_object, intervention_direction, edit_template, edit_plan.
```

Edit prompt (condensed)

```
Use the PROVIDED image as base. Preserve exact viewpoint, geometry, and layout.

ALLOWED: Only modify the target object as required by the plan. If repainting/retexturing, keep shape and placement identical.

FORBIDDEN: No global restyling, relighting, or recoloring. No adding/removing other objects. No readable text. No background, sky, road, or context changes.

Lever concept: {lever_concept}
Lever family: {lever_family}
Scene support: {scene_support}
Intervention direction: {intervention_direction}
Edit template: {edit_template}
Target object: {target_object}
Edit plan: {edit_plan}
```

Critic prompt (condensed)

Evaluate whether the edited image is a valid single-lever counterfactual relative to the original.

CONTEXT: Edit produced by prompt-only diffusion without masking. Minor incidental changes (tone drift, texture resampling) are expected artefacts and should not cause failure unless they materially alter scene meaning; plausibility and locality are still judged strictly against the requested lever and support.

CRITERIA:

(A) edit_attempted: generator made a visible change at the target (false only if output looks identical to original there).
(B) same_place_preserved: same underlying place.
(C) is_localised: primary meaningful change is in/near intended support; minor global tone shifts do not count as non-local.
(D) is_realistic: physically plausible and coherent.
(E) is_plausible: recognisably the requested lever at the stated support; fail if the edit type/support is wrong or too excessive.

CLEAR FAIL CONDITIONS: Viewpoint/geometry changed; large non-target objects added/removed; requested lever replaced by a different change; no discernible change at target.

Return JSON: {edit_attempted, same_place_preserved, is_localised, is_realistic, is_plausible, notes}, where notes = {failure_modes, diagnosis, repair_suggestion}.