
000
001
002
003
004

INTERVENTIONAL REGION ATTRIBUTION FOR STREET-VIEW PERCEPTION WITH HUMAN PAIRWISE JUDGMENTS

005
006
007
008
009
010
011
012

Anonymous authors

Paper under double-blind review

ABSTRACT

Street-view perception models can predict subjective attributes such as safety, wealth, or boringness, yet they rarely answer the question urban planning practitioners care about: for a particular street, what visual evidence drives the judgment, and what localized change would shift perception without altering unrelated cues? We propose an interventional approach to interpretability that treats counterfactual edits as *mechanism probes* for human perception. Given an image and a target attribute, we produce a testable explanation consisting of (i) a target object/region t , (ii) an evidence description e , and (iii) a generated image x' intended to differ from x only in the factor described by e . Our evaluation is intentionally *generation-model agnostic*: x' can be produced by prompt-only generators, inpainting/editing models, or proprietary systems. Rather than assuming edits are faithful, we formalize a human-judgment protocol that measures validity (same-place, locality, realism) and directional perception shift. We then estimate counterfactual effects using randomized pairwise human judgments (edited vs. original), producing per-street effect sizes with uncertainty and explicit failure diagnostics when faithful edits cannot be delivered. This reframes interpretability for urban perception from correlational saliency and narratives to testable evidence grounded in human feedback.

1 INTRODUCTION

Street-view imagery paired with human judgments has enabled models that predict perceived safety, beauty, wealth, or liveliness at scale (Salesse et al., 2013; Naik et al., 2014; Dubey et al., 2016). While these predictors are useful for mapping and correlational analysis, they do not answer a question central to planning, design, and policy: for this specific street, what visual evidence is responsible for the perception, and what targeted modification would causally change that perception? Existing explainability approaches—saliency maps, feature importance, or concept probes—often remain descriptive: they highlight pixels correlated with a score without validating that manipulating those pixels changes human judgment (Adebayo et al., 2018; Kim et al., 2018).

Recent progress in image editing and diffusion-based generation makes it tempting to “improve” streetscapes directly (Meng et al., 2022; Brooks et al., 2023; Zhang et al., 2023). However, using generative edits as evidence is scientifically risky. A generator is an imperfect intervention mechanism: it may fail to implement an intended change, and it may introduce correlated, non-target modifications (e.g., lighting, cleanliness cues, scene activity) that also affect perception. As a result, naive before/after comparisons can conflate the intended intervention with generator artifacts, and post-hoc explanations may not be faithful.

We argue that interpretability for urban perception should be reframed as interventional attribution: explanations should be hypotheses that can be tested by localized counterfactual edits, with explicit checks that the edit isolates the intended concept. Concretely, we target questions of the form:

For this street, what specific visual factor provides causal evidence for a perception attribute (e.g., safety), and can a minimal, localized edit to a single target region shift human judgments while leaving the rest of the scene unchanged?

A key practical complication is that counterfactual image edits are an imperfect intervention mechanism: generators may fail to implement the intended change or may introduce global, correlated shifts that also affect perception. We therefore focus on producing *single-target, minimal* edits and validating them with human judgments. Our main contributions:

- **Testable perception explanations via single-target counterfactuals.** For each street, we produce an evidence description paired with a counterfactual that is constrained to a single target object.
- **Critic-enforced confound control.** We use LLM-as-a-judge as critics to enforce target preservation, realism, and minimal/localized change without unintended global shifts.
- **Human-judgment evaluation protocol and metrics.** We introduce human experiments that directly measure same-place preservation, locality, realism, and directional perception shift success.
- **Human-grounded causal measurement with uncertainty.** Using randomized 2AFC comparisons, we estimate intervention effects while accounting for rater variability and generation stochasticity.

2 RELATED WORK

2.1 URBAN PERCEPTION FROM STREET-VIEW IMAGERY

Large-scale datasets of pairwise perceptual judgments over street-view images enabled learning-based models of perceived safety, wealth, beauty, liveliness, and related attributes [ref]. Early work introduced crowdsourced pairwise comparisons to quantify the “collaborative image” of cities and study spatial inequality in perception [ref]. Subsequent datasets expanded coverage across many cities and attributes, making it feasible to train vision models that predict perception at scale [ref]. This line of work established the now-standard paradigm of (i) collecting pairwise preferences, (ii) fitting ranking or regression models, and (iii) mapping predictions geographically. While effective for prediction and descriptive analysis, these models are not designed to answer *mechanistic* questions at the level of a specific scene: *what visual evidence drives a judgment, and what localized change would causally shift it?*

2.2 INTERPRETING PERCEPTION MODELS: FROM CORRELATIONAL FEATURES TO MECHANISMS

A common approach to interpretability in vision is to attribute predictions to pixels or features (e.g., gradient-based saliency), or to derive feature importance over semantic quantities (e.g., segmented object proportions) using post-hoc tools such as SHAP [ref]. For example, [ref] However, saliency-style explanations can be visually plausible yet unfaithful, and are known to fail basic sanity checks. Concept-based interpretability methods aim to quantify sensitivity to high-level concepts, but typically remain correlational unless paired with interventions that manipulate those concepts. In urban perception specifically, interpretability is often presented as associations between semantic elements (trees, roads, buildings) and predicted scores, rather than as *tested* causal hypotheses validated by human judgments under controlled edits. Our work positions interpretability as **interventional attribution**: explanations are hypotheses that must survive localized counterfactual tests.

2.3 COUNTERFACTUAL EXPLANATIONS AND CAUSAL INTERPRETABILITY IN VISION

Counterfactual explanations are widely used as actionable explanations in ML, typically framed as the smallest change needed to obtain a different outcome (Wachter et al., 2018). In computer vision, counterfactual visual explanations have been studied as edits that flip a *model’s* prediction by replacing informative regions or modifying inputs (Goyal et al., 2019). These approaches provide useful interpretability tools, but differ from our setting in two ways: (i) the outcome of interest is *human perceptual preference* (pairwise judgments) rather than model logits, and (ii) the intervention must be semantically plausible and localized in street scenes. Recently, the causal ML community has also emphasized principled evaluation of counterfactual generation under explicit causal constraints,

108 highlighting that counterfactual images are difficult to evaluate without observable ground truth and
109 that causal validity and spurious changes must be audited (Pawlowski et al., 2020; Melistas et al.,
110 2024). Our method is aligned with this perspective but targets *urban perception* and couples coun-
111 terfactual edits with randomized human experiments, treating edits as mechanism tests rather than
112 synthetic exemplars.

113

114 2.4 GENERATIVE IMAGE EDITING AND CONTROLLABLE INTERVENTIONS

115

116 Diffusion-based editing methods enable photorealistic edits, and controllability techniques improve
117 locality and structural preservation (Meng et al., 2022; Brooks et al., 2023; Hertz et al., 2022; Zhang
118 et al., 2023; Cao et al., 2023). However, even localized edits can introduce correlated non-target
119 changes (illumination, style, scene activity cues) that can dominate perceptual outcomes, which
120 is particularly problematic when using edits as evidence about perception mechanisms. Sepa-
121 rately, emerging urban-computing work uses diffusion models fused with perception data to gen-
122 erate streetscape improvements that increase predicted or measured perception scores (Zhao et al.,
123 2026). These efforts demonstrate feasibility of perception-guided generation, but typically optimize
124 *outcomes* rather than provide faithful *explanations* for a given street, and often do not treat generator
125 artifacts as a threat to validity. Our contribution complements this line by focusing on **evaluation**:
126 human-judgment protocols that quantify validity, directional shift, and coverage for counterfactual
127 edits.

128 **Positioning.** In summary, prior work provides (i) strong datasets and predictors for urban percep-
129 tion, (ii) post-hoc interpretability tools that are often correlational, (iii) counterfactual explanation
130 frameworks in ML and vision, and (iv) diffusion-based editing mechanisms and early attempts at
131 causal controllability. We unify these threads into a practical framework for **interventional expla-**
132 **nations** of street-view perception that are human-grounded and explicitly guarded against generator-
133 induced confounds.

134

135 3 METHODS

136

137

138 3.1 PROBLEM SETUP

139 Let $x \in \mathcal{X}$ be a street-view image and $a \in \mathcal{A}$ a subjective perceptual attribute (e.g., safety). Given
140 a target direction on A (increase/decrease), our objective is to produce a **testable interventional**
141 **explanation** for the specific scene x : a proposed target object/region t , a short evidence description
142 e , and a generated image x' intended to differ from x only in the visual factor described by e .

143 In our pipeline, the planning agent conditions on the *attribute* a (e.g., safety) to propose a plausible
144 evidence factor and target region; it does not require access to a separate estimate of the *directional*
145 *effect* of an edit. Directionality is specified as part of the edit intent in e (e.g., “make the street look
146 safer by improving lighting”), and is validated downstream via pairwise human judgments.

147 Crucially, we treat counterfactual edits as *mechanism probes* for human perception: the validity of
148 an explanation is evaluated by whether a localized, realistic change—perceived as the *same place*—
149 produces a consistent shift in human pairwise judgments.

150

151 3.2 MODEL-AGNOSTIC COUNTERFACTUAL GENERATION INTERFACE

152

153 We are intentionally agnostic to the image generation or editing mechanism. We define a generator
154 G as a black-box *editing operator* that maps an input image to a counterfactual image under an edit
155 specification:

$$x' = G(x; t, e, \phi), \quad (1)$$

156 where x' is the edited image, and G may be instantiated by any generative model (e.g., diffusion-
157 based editors, instruction-following image models, or inpainting systems). The inputs encode the
158 edit request: t specifies the intended locus of change (e.g., an object name, an approximate region,
159 or another handle), e describes the intended change in natural language, and ϕ denotes generator-
160 specific parameters such as random seed, guidance strength, number of steps, edit strength, or tem-
161 perature.

162 Importantly, G is not assumed to be a perfect intervention mechanism: it may fail to apply e at t ,
163 and it may introduce unintended collateral changes outside t . Our downstream llm-as-a-judge and
164 human-judgment protocol therefore evaluates whether x' preserves the same place, remains realistic,
165 and exhibits a sufficiently localized change before using it for causal tests.
166

167 **3.3 GENERATOR FAMILY AND PROMPT-ONLY BASELINES**
168

169 We evaluate multiple generators within the same human evaluation protocol. Our current generator
170 set includes:
171

- 172 • google/nano-banana-pro
- 173 • bytedance/seedream-4
- 174 • openai/gpt-image-1.5
- 175 • black-forest-labs/flux-kontext-max

176 This set is extensible; the protocol does not assume any generator-specific capability.
177

180 **Baselines (current stage).** In the current stage, our baselines use **prompt-only generation**: the
181 generator receives the original image and a text instruction specifying (i) the target object and (ii)
182 the edit intent, but does not receive any explicit mask or segmentation. We also include a *global*
183 *prompt* baseline that omits the target constraint (e.g., “make this street look safer”), to quantify how
184 much global drift can drive perception changes. NOTE TO STEPHEN: Should we constrain the
185 object we are editing?
186

187 **4 DATA CURATION (*placeholder*)**
188

189 **5 METHODOLOGY**
190

192 In this section, we describe our interventional framework for street-view perception interpretability.
193 As illustrated in Figure (x - placeholder), our method operationalizes counterfactual edits as *mech-*
194 *anism probes* for human perception. Given an input image x and a target attribute a (with a desired
195 direction), we orchestrate a lightweight team of agents to produce a testable explanation (t, e, x') :
196 a single target region/object t , an evidence description e , and an edited image x' intended to differ
197 from x *only* in the factor described by e at t . Because image generators can introduce confounds
198 (e.g., global lighting/style drift), we do not assume edits are faithful; instead, we explicitly audit
199 validity and report both **coverage** (how often valid edits can be produced) and **directional effects**
200 (how valid edits shift human judgments).
201

202 **Agentic workflow.** Our pipeline follows a Planner→Prompt→Generator→Critics workflow, that
203 improve visual artifact quality through planning and iterative critique, but specialized for *single-*
204 *target* interventions and *confound control* in street scenes.
205

206 **5.1 PLANNER AGENT**
207

208 The Planner Agent serves as the cognitive core of the system. Given the input image x , target
209 attribute a , and a desired attributes, the planner proposes a *single-target* intervention specification:
210 (i) a target handle t (the only allowed locus of change), and (ii) an evidence description e describing
211 the localized modification intended to affect a . We denote the planner output as:
212

$$(t, e) = \text{LLM}_{\text{plan}}(x, a). \quad (2)$$

213 In the current stage, t is a textual reference to a salient object/region (e.g., “streetlight”, “shopfront
214 shutter”, “graffiti on wall”); future versions can replace t with explicit segmentation or bounding
215 boxes.
216

5.2 PROMPT CONSTRUCTOR

The Prompt Constructor translates (t, e) into a structured edit instruction for the generator. The prompt explicitly encodes four constraints: **(1) same-place preservation** (retain identity, geometry, viewpoint), **(2) single-target locality** (only modify t), **(3) minimality** (avoid global style/lighting/cleanliness drift), and **(4) photorealism** (retain street-photo appearance). This yields a generator-ready specification:

$$\pi = \text{Prompt}(x; t, e, a), \quad (3)$$

which is paired with the input image x and passed to the generator.

5.3 IMAGE GENERATOR

We treat the image generator as a black-box editing operator G . Given the input image and prompt, the generator produces one or more candidate edits:

$$x' = G(x; \pi, \phi), \quad (4)$$

where ϕ denotes generator-specific sampling parameters (e.g., seed, guidance strength, edit strength). We are intentionally generator-agnostic: G can be instantiated by instruction-following image models, diffusion editors, inpainting systems, or proprietary tools. Crucially, G is not assumed to implement perfect interventions—it may under-edit, over-edit, or introduce collateral changes outside t .

5.4 CRITIC AGENTS

The Critic Agents form a closed-loop refinement mechanism with the generator by inspecting each candidate x' and issuing targeted feedback. For iteration k , given (x, x'_k) and the planned specification (t, e) , critics evaluate three validity criteria: **same-place preservation**, **locality** (minimal change outside t), and **realism**. They then output a pass/fail decision and a concise diagnostic message:

$$(d_k, \Delta_k) = \text{VLM}_{\text{critic}}(x, x'_k, t, e), \quad (5)$$

where $d_k \in \{0, 1\}$ indicates acceptance and Δ_k describes failure modes (e.g., “global lighting shift”, “scene identity changed”, “target not edited”).

Iterative refinement. If no candidate is accepted, the planner updates (t, e) using critic feedback and regenerates for a fixed number of rounds T :

$$(t_{k+1}, e_{k+1}) = \text{LLM}_{\text{revise}}(x, a, t_k, e_k, \Delta_k). \quad (6)$$

We cap T to control cost and prevent uncontrolled drift; the final output is the first accepted edit, or a recorded failure if none passes.

5.5 AUDIT-AND-FILTER: PRODUCING VALID INTERVENTIONS

Because generators can introduce non-target changes that confound interpretation, our evaluation distinguishes between *attempted edits* (all generated candidates) and *valid interventions* (candidates that satisfy validity criteria). We report **coverage**: the fraction of scenes for which the pipeline produces at least one valid intervention. Low coverage is treated as a meaningful outcome, providing explicit diagnostics on when faithful localized edits cannot be delivered.

Human grounding. Automated critics improve throughput but are not assumed perfect. In experiments, we compute validity metrics (SPR/LCC/RP?? TO BE DEFINED) and perception-shift outcomes from human judgments (Section ??). This separation prevents the pipeline from “grading its own homework” and makes failures observable.

Controls. Where feasible, we generate matched control edits x_{ctrl} that aim to match the *edit budget* (magnitude of change) while breaking alignment with the stated evidence (e.g., applying a comparable localized edit to a different target than t , or performing a neutral edit on t). These controls test whether observed shifts are specific to the hypothesized evidence rather than driven by “any edit” or generic generator drift.

270 6 BENCHMARK CONSTRUCTION
271

272 The lack of standardized benchmarks makes it difficult to rigorously compare interventional ex-
273 planations across generators and pipelines. We address this by constructing a benchmark of *single-*
274 *target interventional test cases* for street-view perception. Each benchmark item couples a real street
275 image with a constrained edit specification and supports paired evaluation of (i) intervention validity
276 and (ii) directional perception shift.

277
278 6.1 DATA CURATION
279

280 *Placeholder.* We will describe collection, filtering, and human verification in Section 4, includ-
281 ing dataset statistics (image quality, geographic coverage, and attribute distribution), and any pri-
282 vacy/bias considerations.

283
284 6.2 TEST CASE CONSTRUCTION
285

We represent each benchmark item as a tuple:

$$\mathcal{B}_i = (x_i, a_i, d_i, t_i, e_i, x_{\text{ctrl},i}), \quad (7)$$

where x_i is the original image, a_i is the target attribute, and $d_i \in \{+1, -1\}$ indicates the intended direction of change for that attribute (e.g., safer vs. less safe). The pair (t_i, e_i) is the planner-proposed single-target specification, and $x_{\text{ctrl},i}$ is an optional matched control plan/edit.

Design goals. We curate items to (i) stress-test **locality** (single-target edits), (ii) expose **generator confounds** (global drift), and (iii) support **paired human evaluation** at scale.

Controls (matched edit budget). When feasible, each item includes a control specification that matches the magnitude of change while breaking alignment with the evidence. This allows faithfulness testing beyond “edited vs. original”.

298 6.3 EVALUATION PROTOCOL
299

300 Our benchmark evaluation follows a referenced, paired comparison design. Unlike settings where a
301 human reference diagram is available, street-scene interventions lack a ground-truth counterfactual.
302 We therefore evaluate two complementary aspects: (i) **validity** (is x' a faithful localized intervention
303 on the same place?) and (ii) **attribute-shift effect** (does x' shift the target attribute in the intended
304 direction?).

305 **Validity dimensions.** We audit each candidate edit using three human-evaluatable criteria: same-
306 place preservation, locality (collateral change), and realism. Detailed rubrics and interface screen-
307 shots are provided in the appendix.

309 **Directional preference.** For accepted edits, we measure attribute-specific 2AFC preference (e.g.,
310 “Which looks safer?”) under randomized left/right presentation.

312 **Reporting principle.** We always report *attempted* vs. *accepted* outcomes separately, together with
313 coverage and failure modes, so that apparent perception shifts driven by confounded global drift are
314 not mistaken for evidence about localized mechanisms.

316 7 EXPERIMENTS
317

318 7.1 BASELINE METHODS AND MODELS
319

320 We compare our planner–generator–critic pipeline against baseline settings designed to isolate which
321 components contribute to valid interventions and directional shifts:

- 322 • **Vanilla (direct edit prompt).** Directly prompt the image generator with a generic instruc-
323 tion to modify the scene for the target attribute, without an explicit single-target plan.

-
- 324 • **Single-target (no critics).** Use the planner-generated (t, e) and structured prompt, but do
325 not apply critic screening or refinement.
326 • **Global prompt baseline.** Omit the single-target constraint (e.g., “make this street look
327 safer”), quantifying how much global drift can drive preference shifts.
328 • **Control edits (when available).** Matched edit-budget controls x_{ctrl} misaligned with e .
- 329

330 We evaluate multiple image generators through the same interface $x' = G(x; \pi, \phi)$ (see Appendix
331 for implementation details and prompts).
332

333 7.2 EVALUATION SETTINGS
334

335 We conduct two core human-subject experiments on the benchmark:
336

337 **(E1) Validity audit.** Participants compare (x, x') and judge (i) whether they depict the same place
338 with a small change, (ii) locality (collateral change), and (iii) realism.
339

340 **(E2) Directional preference.** Participants complete an attribute-specific 2AFC task (e.g., “Which
341 looks safer?”) comparing (x, x') .
342

343 We report coverage and validity metrics for all attempted edits, and preference-shift metrics for edits
344 that pass validity screening. Full participant criteria, quality controls, and rubrics are provided in
345 Section ?? and the appendix.
346

347 7.3 MAIN RESULTS
348

349 7.4 ABLATIONS
350

351 7.5 FAILURE ANALYSIS AND CASE STUDIES
352

353 8 CONCLUSION
354

355 We presented a framework that moves urban perception interpretability from passive observation
356 to active testing by formalizing *human evaluation* for minimal, localized, plausible counterfactual
357 edits. By comparing generators through a shared protocol—validity (same-place, locality, realism),
358 directional perception shift, and coverage—we provide testable evidence factors and explicit failure
359 diagnostics.
360

361 AUTHOR CONTRIBUTIONS
362

363 If you’d like to, you may include a section for author contributions as is done in many journals. This
364 is optional and at the discretion of the authors.
365

366 ACKNOWLEDGMENTS
367

368 Use unnumbered third level headings for the acknowledgments. All acknowledgments, including
369 those to funding agencies, go at the end of the paper.
370

371 REFERENCES
372

373 Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity
374 checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*,
375 2018. URL <https://arxiv.org/abs/1810.03292>.
376

377 Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow
378 image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision
379 and Pattern Recognition (CVPR)*, 2023. URL [https://openaccess.thecvf.com/
382 content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_
383 Image_Editing_Instructions_CVPR_2023_paper.html](https://openaccess.thecvf.com/
380 content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_
381 Image_Editing_Instructions_CVPR_2023_paper.html).
384

- 378 Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Ying Zheng.
379 Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and
380 editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*
381 (*ICCV*), 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cao_MasaCtrl_Tuning-Free_Mutual_Self-Attention_Control_for_Consistent_Image_Synthesis_and_ICCV_2023_paper.html.
- 384 Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. Deep learning
385 the city: Quantifying urban perception at a global scale. In *Computer Vision – ECCV 2016*,
386 volume 9905 of *Lecture Notes in Computer Science*, pp. 196–212. Springer, Cham, 2016. doi:
387 10.1007/978-3-319-46448-0_12. URL <https://arxiv.org/abs/1608.01769>.
- 389 Yash Goyal, Ziyuan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual
390 explanations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*,
391 volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384, 2019. URL <https://proceedings.mlr.press/v97/goyal19a.html>.
- 393 Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or.
394 Prompt-to-prompt image editing with cross attention control. *arXiv preprint*, 2022. URL
395 <https://arxiv.org/abs/2208.01626>.
- 397 Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas,
398 and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with con-
399 cept activation vectors (tcav). In *Proceedings of the 35th International Conference on Ma-*
400 *chine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, 2018. URL
401 <https://proceedings.mlr.press/v80/kim18d.html>.
- 402 Thomas Melistas, Nikos Spyrou, , et al. Benchmarking counterfactual image generation. In *Ad-*
403 *vances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*,
404 2024. URL <https://arxiv.org/abs/2403.20287>.
- 406 Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon.
407 Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International*
408 *Conference on Learning Representations (ICLR)*, 2022. URL <https://arxiv.org/abs/2108.01073>.
- 410 Nikhil Naik, Jonah Philipoom, Ramesh Raskar, and César A. Hidalgo. Streetscore – predicting
411 the perceived safety of one million streetscapes. In *2014 IEEE Conference on Computer Vision*
412 *and Pattern Recognition Workshops (CVPRW)*, 2014. doi: 10.1109/CVPRW.2014.121. URL
413 https://openaccess.thecvf.com/content_cvpr_workshops_2014/W20/papers/Naik_Streetscore_-_Predicting_2014_CVPR_paper.pdf.
- 415 Nick Pawłowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for
416 tractable counterfactual inference. In *Advances in Neural Information Processing Systems*
417 (*NeurIPS*), 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.
- 420 Philip Salesses, Katja Schechtner, and César A. Hidalgo. The collaborative image of the city:
421 Mapping the inequality of urban perception. *PLOS ONE*, 8(7):e68400, 2013. doi: 10.1371/journal.pone.0068400. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068400>.
- 424 Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening
425 the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):
426 841–887, 2018. URL <https://arxiv.org/abs/1711.00399>.
- 428 Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image
429 diffusion models. In *Proceedings of the IEEE/CVF International Conference on Com-*
430 *puter Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content_ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

432 Chenbo Zhao, Yoshiki Ogawa, Shenglong Chen, Takuya Oki, and Yoshihide Sekimoto. Street
433 space quality improvement: Fusion of subjective perception in street view image generation.
434 *Information Fusion*, 125:103467, 2026. doi: 10.1016/j.inffus.2025.103467. URL <https://www.sciencedirect.com/science/article/pii/S1566253525005408>.

436
437 **A APPENDIX**

438 You may include other additional sections here.

439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485