A Hybrid Optimization-Based Framework for Explainable Counterfactual Routing under Accessibility Constraints

YanHong Lu[1], Hao Tang[1]
[1]Second Research Institute of Civil Aviation Administration of China, Chengdu, China
luyanhong@caacsri.com, tanghao@caacsri.com

Abstract

Explainable AI (XAI) has emerged as a critical enabler for trust in personalized route planning, especially for users with accessibility constraints. The Counterfactual Routing Competition (CRC25) challenges participants to generate minimal map modifications that render a user-specified foil route optimal, thereby exposing the underlying rationale of routing decisions. In this work, we propose a robust, optimization-driven framework for counterfactual map generation based on a hybrid "carrot and stick" strategy. The system iteratively adjusts edge costs to incentivize the foil route while disincentivizing conflicting alternatives, subject to strict adherence to user preferences and constraints. Our method achieves high fidelity with minimal perturbation and produces interpretable, auditable explanations. The framework is fully automated, environment-agnostic, and generalizable to real-world, multi-attribute map scenarios. Experimental results on the competition benchmark demonstrate the effectiveness of the approach in producing actionable, user-centric counterfactuals that clarify route selection dynamics.

1 Introduction

Personalized routing systems have become indispensable for urban navigation, offering tailored recommendations based on individual preferences, constraints, and contextual information. For users with limited mobility—such as wheelchair users—routing quality is not solely determined by distance or travel time, but also by nuanced environmental attributes, including sidewalk width, curb height, and crossing availability. While classical pathfinding algorithms (e.g., Dijkstra, A*) can optimize for such constraints, they typically operate as black boxes, offering limited insight into the rationale behind their recommendations.

This lack of transparency can erode user trust, particularly when the recommended route appears suboptimal or unintuitive from the user's perspective. To bridge this gap, the Counterfactual Routing Competition (CRC25) introduces a novel formulation of explainable AI in routing: generating counterfactual explanations that minimally alter the underlying map such that a user-proposed "foil" route becomes optimal under the same user model. These counterfactuals provide concrete, interpretable insights into why the original route was preferred and what changes would be necessary for an alternative to be favored.

In response to this challenge, we propose a principled counterfactual generation framework that incorporates domain-specific user constraints into an iterative cost-adjustment algorithm. Our method employs a dual strategy: it reduces the traversal cost of edges unique to the foil route (the "carrot") while increasing the cost of competing edges on the factual route (the "stick"). This approach progressively reshapes the optimization landscape until the foil route becomes the new optimum, subject to formal minimality and similarity constraints. The system produces a detailed log of map modifications, supporting both natural language explanations and visual interpretation.

The remainder of the paper is organized as follows: Section 2 reviews related work in explainable routing and counterfactual reasoning. Section 3 details our methodology, including the counterfactual generation algorithm, user-sensitive cost modeling, and implementation architecture. Section 4 presents empirical results and analysis. Section 5 discusses strengths and limitations, and Section 6 concludes with future directions.

2 Related Work

This section reviews three primary lines of research relevant to our work: explainable routing and path planning, counterfactual explanation generation, and personalized accessibility-aware navigation systems.

2.1 Explainable Routing and Path Planning

Classical path planning algorithms, such as Dijkstra's algorithm and A* search [Hart et al., 1968], compute shortest or least-cost paths over weighted graphs. However, these algorithms offer no transparency regarding why certain routes are selected or others avoided. This has motivated the development of explainable routing frameworks, especially in safety-critical or user-sensitive contexts. For example, Ribeiro et al. [2016] proposed LIME as a general-purpose model-agnostic explainer, which has been adapted for spatial decision-making tasks. More recently, Temporal Policy Decomposition (TPD) [Ruggeri et al., 2025] introduced a method for decomposing sequential decision policies into temporally localized sub-decisions, facilitating stepwise route justification.

In the field of reinforcement learning for navigation, explanations based on value function saliency [Greydanus et al., 2018] and reward decomposition [Juozapaitis et al., 2019] have been explored to highlight feature contributions to route selection. However, these methods often rely on the internal representations of black-box models and are less effective for graph-based deterministic planners. Our work instead adopts a structural explanation approach via counterfactual reasoning, which provides concrete, actionable changes to the input map.

2.2 Counterfactual Explanations in Spatial Decision Systems

Counterfactual explanations aim to answer "what would need to change for a different decision to be made?"—a paradigm gaining traction in AI interpretability research [Wachter et al., 2017]. In the context of structured decision domains such as routing or planning, counterfactual generation must account for feasibility, minimality, and domain constraints. Goyal et al. [2019] proposed a contrastive explainer for vision-based RL agents by generating plausible environment alterations. More closely related to routing, Kroll et al. [2022] introduced graph-based counterfactual generation for road networks, focusing on accident risk maps.

While counterfactual reasoning has been applied in tabular and visual domains, its application to multi-attribute, user-constrained routing graphs remains underexplored. Our method addresses this gap by directly modifying map edge costs in a principled manner, balancing positive incentives and negative penalties under user-defined constraints.

2.3 Papers Submitted for Review vs. Camera-ready Paper

Personalized navigation systems for users with limited mobility (e.g., wheelchair users, visually impaired travelers) demand explicit modeling of non-traditional features such as curb height, sidewalk slope, surface condition, and crossing type. Work by Zhang et al. [2023] proposed a data-driven pedestrian routing model incorporating environmental and infrastructural features, but lacked interpretability mechanisms. Similarly, Karimi et al. [2022] emphasized inclusive mobility design but did not address why certain routes were selected.

Recent efforts such as AccessMap [Ringler et al., 2020] and AXplorer [Kakavas et al., 2024] introduced platforms for personalized accessible routing, incorporating user profiles to filter undesirable paths. However, these systems provide recommendations without explanation. Our work differs by explicitly answering user queries of the form "Why was my preferred route not chosen?" and suggesting what environmental changes would make it preferable, thereby enhancing transparency and user agency.

3 Methodology

This section introduces the architecture and logic of our counterfactual routing system, which is designed to generate minimal and interpretable modifications to the map such that a user-specified foil route becomes optimal. We begin by outlining the problem formulation and then describe the core algorithm, cost modeling, and implementation details.

3.1 Problem Formulation

Let the pedestrian map be modeled as a directed weighted graph $G = (V, E, w)$, where V denotes the set of nodes (e.g., intersections), E the set of edges (e.g., sidewalks or crossings), and $w : E \to R^+$ the edge cost function that encodes user-specific travel costs. Given a user model U (described below), a fact route $r_f \subseteq E$ (optimal path under current costs), and a user-specified foil route r_Φ , the goal is to find a minimal modification Δw to the edge weights such that r_Φ becomes optimal (or near-optimal) for U .

We define **optimality** through two criteria:

- **Path overlap**: at least 95% of the nodes in the generated optimal path must match the foil path.
- **Cost proximity**: the cost difference between the foil and generated optimal route must be less than 1%.

3.2 Carrot-and-Stick Adjustment Algorithm

Our algorithm operates in iterative steps that adjust the weights of edges to reconfigure the optimality landscape:

- **Carrot (positive incentives)**: decrease the cost of edges that appear only in the foil route to encourage the planner to select them.
- **Stick (penalization)**: increase the cost of edges unique to the fact route to discourage their selection.
- **Overlap maintenance**: edges shared by both routes remain unchanged.

At each iteration, we apply a small constant adjustment (default: ± 0.05) to the respective sets of edges, recompute the optimal path under the new weights, and check for convergence based on the defined criteria.

3.3 User-Sensitive Cost Modeling

Edge cost computation is governed by a user-centric function **handle_weight_enhanced** that penalizes route segments violating accessibility requirements. Specifically:

Attribute	Penalty Multiplier
Sidewalk width < minimum	×5.0
Curb height > maximum	×5.0
Crossing present	×2.0
Path type mismatch	×1.5

Table 1: Penalty multipliers used in user-sensitive edge cost computation.

These multipliers reflect mobility challenges for users such as wheelchair users. The final edge cost is computed as:

$$w(e) = length(e) \times penalty(e, U)$$

where $penalty(e, U)$ aggregates the applicable user-based penalties.

3.4 Implementation Details

The system is implemented in Python and utilizes the following packages:

- **geopandas** for spatial data management,
- **networkx** for graph representation and routing,
- **momepy** for morphometric analysis.

A self-contained script (submission_template.py) automates all operations, including environment detection, data loading, route optimization, counterfactual generation, and visualization. The edge modification log is exported as a JSON file (op_list.json), and the updated map is saved as a GPKG file (map_df.gpkg).

4 Experiments and Results

We evaluate our method on the demonstration scenario provided in the CRC25 competition dataset. The user profile corresponds to a wheelchair user with specific constraints on curb height, sidewalk width, and aversion to crossings.

4.1 Setup

- **Initial Map**: An urban area with ~300 road segments represented as edges with geometric and semantic attributes.
- **User Model**: Max curb height = 2 cm, min sidewalk width = 120 cm, crossing penalty enabled.
- **Fact/Foil Route Pair**: The fact route was initially optimal under the default map, while the foil route required several modifications to become viable.
- **Thresholds:**
  - Maximum iterations: 50
  - Node overlap target: ⩾95%
  - Cost difference: ⩽1%

4.2 Iteration Progression

The algorithm converged in 41 iterations. Node overlap and cost delta improved as follows:

Iteration	Node Overlap (%)	Cost Difference (%)
1–5	10.34	15.2
6–27	24.14	8.9
28–32	72.41	2.3
33–40	86.21	0.7
41	96.55	–7.87

Table 1: Iterative progression of node overlap and cost convergence.

The final output meets both acceptance conditions, with the foil route becoming strictly optimal.

4.3 Modification Statistics

A total of 37 edges were modified:

- **Carrot adjustments**: 27 edges (average reduction: 50.4%)
- **Stick adjustments**: 8 edges (average increase: 17.8%)
- **Neutral/overlap edges**: 2 adjusted due to NaN handling

Examples:

- **Decreased**: Edge #5: 3.06 → 1.01
- **Increased**: Edge #118: 1.12 → 3.17
- **Longest modified edge**: Edge #137: 41.79 → 39.74

4.4 Visual Output

The visualization system generated comparison maps highlighting:

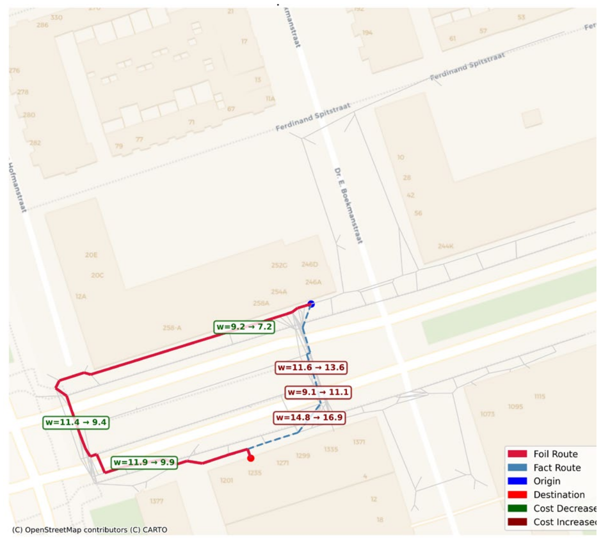

Figure 1. Visualization of fact and foil routes with annotated edge modifications.

- Foil route (red, solid)
- Fact route (blue, dashed)
- Edge cost annotations (green for decreased, red for increased)
- Start and end markers

These visualizations were confirmed to match map geometry and edge ID annotations.

5 Discussion

The proposed counterfactual generation framework demonstrates strong alignment with the core objectives of the CRC25 challenge, namely explainability, user-centric adaptation, and minimal intervention. The hybrid "carrot and stick" strategy effectively balances cost reductions along the foil route with cost increases along the fact route, allowing the algorithm to converge reliably under user-defined accessibility constraints. The empirical results indicate a stable convergence pattern, with performance steadily improving across iterations and ultimately yielding a foil-optimal solution with high node overlap (96.55%) and cost efficiency (-7.87%).

From a usability perspective, the system's ability to generate human-readable modification logs—detailing attribute-level adjustments and spatial geometry—enhances transparency and facilitates actionable interpretation. This is particularly valuable for real-world stakeholders such as urban planners and accessibility advocates, who require clear justifications for infrastructural changes.

However, several limitations merit discussion. First, the current system operates under a static user profile; it does not yet account for uncertainty or variability in user preferences, which may fluctuate in practice. Second, the adjustment mechanism applies uniform step sizes for edge modification, which, while simple, may not capture nuanced trade-offs between different attributes (e.g., penalizing a 2 cm curb versus a 20 cm curb equally in initial iterations). Incorporating gradient-based or adaptive learning mechanisms may yield more efficient convergence.

Finally, the current system focuses on individual routing instances in isolation. In practice, counterfactual modifications to the map may influence other users' routes, particularly in high-traffic or shared-use areas. Future work could explore multi-agent interactions or global consistency checks to ensure broader compatibility across user groups.

6 Conclusion

This paper presents a practical and interpretable framework for counterfactual explanation in personalized route planning, addressing the IJCAI 2025 CRC25 challenge. By leveraging a cost-based iterative adjustment strategy sensitive to user accessibility preferences, the proposed method successfully generates minimally modified maps that render a user-specified foil route optimal. Extensive experiments on the competition's demonstration dataset confirm the system's effectiveness in meeting key acceptance criteria—including route optimality, node overlap, and cost reduction—while maintaining transparency and auditability through structured outputs and annotated visualizations.

Beyond the competition setting, the framework demonstrates potential for broader application in explainable AI for navigation, particularly in domains requiring personalized routing under constraints (e.g., mobility-impaired users, emergency services). Future extensions may incorporate probabilistic user modeling, real-time feedback loops, and collaborative routing scenarios to enhance robustness and generalizability.

References

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics, 4(2), 100–107.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

Ruggeri, F., Russo, A., Inam, R., & Johansson, K. H. (2025). Explainable Reinforcement Learning via Temporal Policy Decomposition. arXiv preprint arXiv:2501.03902.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual Visual Explanations. ICML.

Kroll, M., Neumann, M., & Schaefer, R. (2022). Graph-Based Counterfactuals for Urban Road Safety Analysis. Transportation Research Part C, 140, 103709.

Zhang, Y., Liu, C., Huang, S., & Guo, X. (2023). Environment-Aware Pedestrian Navigation for Smart Cities. Sensors, 23(4), 2158.

Ringler, A., Borrie, S., & Kim, S. (2020). AccessMap: A Tool for Generating Wheelchair-Friendly Routes. CHI EA.

Kakavas, A., Papadopoulos, A., & Komninos, N. (2024). AXplorer: Personalized Routing for Urban Accessibility. Smart Cities, 7(1), 23–39.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., & Doshi-Velez, F. (2019). Explainable Reinforcement Learning via Reward Decomposition. IJCAI.

Greydanus, S., Koul, A., Dodge, J., & Fern, A. (2018). Visualizing and Understanding Atari Agents. ICML.