Novelty Assessment Report

Paper: Joint Distribution–Informed Shapley Values for Sparse Counterfactual Explanations
PDF URL: https://openreview.net/pdf?id=3vIe5pNiUN
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Counterfactual explanations (CE) aim to reveal how small input changes flip a model's prediction, yet many methods modify more features than necessary, reducing clarity and actionability. We introduce COLA, a model- and generator-agnostic post-hoc framework that refines any given CE by computing a coupling via optimal transport (OT) between factual and counterfactual sets and using it to drive a Shapley-based attribution p-SHAP that selects a minimal set of edits while preserving the target effect. Theoretically, OT minimizes an upper bound on the W_1 divergence between factual and counterfactual outcomes and that, under mild conditions, refined counterfactuals are guaranteed not to move farther from the factuals than the originals. Empirically, across four datasets, twelve models, and five CE generators, COLA achieves the same target effects with only 26–45% of the original feature edits. On a small-scale benchmark, COLA shows near-optimality.

Core Task Landscape

This paper addresses: **Refining Counterfactual Explanations with Minimal Feature Modifications**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Optimization-Based Counterfactual Generation Methods**
- **Generative Model-Based Counterfactual Explanations**
- **Evolutionary and Search-Based Counterfactual Methods**
- **Case-Based and Retrieval-Augmented Counterfactuals**
- **Domain-Specific Counterfactual Explanation Applications**
- **Actionability, Feasibility, and User-Centric Counterfactuals**
- **Counterfactual Data Augmentation and Model Training**
- **Unified Frameworks and Theoretical Foundations**
- **Concept-Based and Semantic Counterfactual Explanations**
- **Large Language Model-Based Counterfactual Generation**

Complete Taxonomy Tree

- Refining Counterfactual Explanations with Minimal Feature Modifications Survey Taxonomy
- Optimization-Based Counterfactual Generation Methods
  - Gradient-Based and Perturbation Optimization (3 papers)
  - [8] Exploring Energy Landscapes for Minimal Counterfactual Explanations: Applications in Cybersecurity and Beyond (Veroni Eleni, 2025) View paper
  - [22] Enhancing counterfactual image generation using mahalanobis distance with distribution preferences in feature space (Yukai Zhang, 2024) View paper
  - [32] Counterfactual explanation based on gradual construction for deep networks (Jung, 2022) View paper
  - Shapley Value-Guided Optimization (2 papers)
  - [6] Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality (You Lei, 2024) View paper
  - [19] SVCE: Shapley Value Guided Counterfactual Explanation for Machine Learning-Based Autonomous Driving (Meng Li, 2024) View paper
  - Optimal Transport and Coupling-Based Refinement ★ (1 papers)
  - [0] Joint Distribution–Informed Shapley Values for Sparse Counterfactual Explanations (Anon et al., 2026) View paper
  - Multi-Objective and Constraint-Based Optimization (5 papers)
  - [2] Flexible and Robust Counterfactual Explanations with Minimal Satisfiable Perturbations (Wang YongJie, 2023) View paper
  - [18] Multi-SpaCE: Multi-Objective Subsequence-based Sparse Counterfactual Explanations for Multivariate Time Series Classification (Luengo, 2024) View paper
  - [27] TX-Gen: Multi-Objective Optimization for Sparse Counterfactual Explanations for Time-Series Classification (Qi Huang, 2024) View paper
  - [29] DiCE-Extended: A Robust Approach to Counterfactual Explanations in Machine Learning (Polat, 2025) View paper
  - [49] Feature-Driven Counterfactual Explanations: A SHAP-Based Approach to Dimensionality Reduction in XAI (Yu-Lun Chien, 2025) View paper
- Generative Model-Based Counterfactual Explanations
  - Diffusion Model-Based Counterfactuals (2 papers)
  - [9] Diffusion visual counterfactual explanations (Augustin, 2022) View paper

    - [46] CoLa-DCE - Concept-guided Latent Diffusion Counterfactual Explanations (Motzkus, 2024) View paper
    - GAN and Adversarial-Based Counterfactuals (2 papers)
    - [12] Adversarial counterfactual visual explanations (Jeanneret, 2023) View paper
    - [17] Flexible Counterfactual Explanations with Generative Models (Algaba, 2025) View paper
    - Latent Space and Autoencoder-Based Counterfactuals (2 papers)
    - [39] Explainable image classification with evidence counterfactual (Vermeire, 2022) View paper
    - [44] Looking in the mirror: A faithful counterfactual explanation method for interpreting deep image classification models (Chowdhury, 2025) View paper
- Evolutionary and Search-Based Counterfactual Methods (2 papers)
    - [31] UGCE: User-Guided Incremental Counterfactual Exploration (Pitoura, 2025) View paper
    - [45] AIM-CF: Fast and Precise Counterfactual Explanations via Approximate Inverse Models (Takafumi Nakanishi, 2025) View paper
- Case-Based and Retrieval-Augmented Counterfactuals (2 papers)
    - [15] DisCERN: Discovering Counterfactual Explanations using Relevance Features from Neighbourhoods (Nirmalie Wiratunga, 2021) View paper
    - [20] Measurable counterfactual local explanations for any classifier (White, 2019) View paper
- Domain-Specific Counterfactual Explanation Applications
    - Graph Neural Network Counterfactuals (2 papers)
    - [3] Cf-gnnexplainer: Counterfactual explanations for graph neural networks (Lucic, 2022) View paper
    - [7] COMBINEX: A Unified Counterfactual Explainer for Graph Neural Networks via Node Feature and Structural Perturbations (Giorgi, 2025) View paper
    - Time-Series Counterfactual Explanations (2 papers)
    - [28] MASCOTS: Model-Agnostic Symbolic COunterfactual explanations for Time Series (Spinnato, 2025) View paper
    - [43] PerCE: Hierarchical Perturbation-Based Counterfactual Explanations for Multivariate Time Series Classification (Betul Bayrak, 2025) View paper
    - Natural Language and Text Counterfactuals (3 papers)
    - [10] Counterfactual Explanations for Models of Code (Cito, 2022) View paper
    - [41] A survey on natural language counterfactual generation (Feng Yu-hong, 2024) View paper
    - [50] Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics (Liem, 2025) View paper
    - Visual and Image Counterfactual Explanations (1 papers)
    - [11] Counterfactual visual explanations (Yash Goyal, 2019) View paper
    - Recommender Systems and Molecular Counterfactuals (3 papers)
    - [5] Counterfactual explainable recommendation (Tan, 2021) View paper
    - [33] Counterfactual review-based recommendation (Kun Xiong, 2021) View paper
    - [35] Generation of Molecular Counterfactuals for Explainable Machine Learning Based on Coreâ□□Substituent Recombination (Alec Lamens, 2023) View paper
    - Specialized Application Domains (7 papers)
    - [13] Counterfactual explanations for deep learning-based traffic forecasting (Wang Ru-shan, 2024) View paper
    - [14] An investigation into creating counterfactual examples for non-linear Support Vector Machines (Luca Bergamin, 2025) View paper
    - [16] Enhanced Counterfactual Explanations for Optimizing Three-Dimensional Printing Parameters Using SHAP and Nearest-Neighbor Constraints With Physics-Based â□¦ (S Saleh, 2025) View paper
    - [24] Customer-Centric Decision-Making with XAI and Counterfactual Explanations for Churn Mitigation (S. Oprea, 2025) View paper
    - [25] Multimodal LLM for enhanced Alzheimerâ□□s Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data (Meiwei Zhang, 2025) View paper
    - [26] Counterfactual Explanation of a Classification Model for Detecting SQL Injection Attacks (BA Cumi-Guzman, 2024) View paper
    - [38] EEG-based motor imagery recognition via novel explainable ensemble learning architecture (A. L. Alfeo, 2025) View paper
- Actionability, Feasibility, and User-Centric Counterfactuals (3 papers)
    - [30] FACE: feasible and actionable counterfactual explanations (Poyiadzi, 2020) View paper
    - [36] Actionable and diverse counterfactual explanations incorporating domain knowledge and causal constraints (Szymon Bobek, 2025) View paper
    - [48] Introducing User Feedback-Based Counterfactual Explanations (UFCE) (Muhammad Suffian, 2024) View paper
- Counterfactual Data Augmentation and Model Training (2 papers)
    - [34] PairCFR: Enhancing Model Training on Paired Counterfactually Augmented Data through Contrastive Learning (Feng Yu-hong, 2024) View paper
    - [37] Cpl: Counterfactual prompt learning for vision and language models (He, 2022) View paper
- Unified Frameworks and Theoretical Foundations (3 papers)
    - [1] Counterfactual explanations and how to find them: literature review and benchmarking (Riccardo Guidotti, 2024) View paper
    - [23] Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End (Mothilal, 2021) View paper
    - [42] Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction (Goldberg, 2021) View paper
- Concept-Based and Semantic Counterfactual Explanations (3 papers)
    - [4] Counterfactual explanation generation with minimal feature boundary (Dianlong You, 2023) View paper
    - [40] Conceptual Edits as Counterfactual Explanations. (G Filandrianos, 2022) View paper
    - [47] Region-aware Minimal Counterfactual Rules for Model-agnostic Explainable Classification (Guido Gagliardi, 2025) View paper
- Large Language Model-Based Counterfactual Generation (1 papers)
    - [21] SenseCF: LLM-Prompted Counterfactuals for Intervention and Sensor Data Augmentation (Arefeen Asiful, 2025) View paper

Narrative

Core task: Refining counterfactual explanations with minimal feature modifications. The field of counterfactual explanations has evolved into a rich landscape organized around distinct methodological and application-oriented branches. Optimization-Based Counterfactual Generation Methods form a central pillar, encompassing gradient-driven approaches, constraint satisfaction techniques like Minimal

Satisfiable Perturbations[2], and specialized refinements using optimal transport or coupling strategies. Generative Model-Based Counterfactual Explanations leverage VAEs, GANs, and diffusion models such as Diffusion Visual Counterfactual[9] to produce realistic alternatives, while Evolutionary and Search-Based Counterfactual Methods employ genetic algorithms and heuristic search. Case-Based and Retrieval-Augmented Counterfactuals retrieve similar instances from data, and Domain-Specific Counterfactual Explanation Applications address tailored challenges in healthcare, finance, autonomous driving, and other areas. Actionability, Feasibility, and User-Centric Counterfactuals emphasize practical constraints and human interpretability, exemplified by works like Actionable Minimality[6]. Additional branches include Counterfactual Data Augmentation, Unified Frameworks, Concept-Based explanations, and emerging Large Language Model-Based generation approaches.

Within the optimization landscape, a key tension exists between achieving minimal perturbations and ensuring actionability or semantic coherence. Many studies focus on sparsity and proximity metrics, while others incorporate causal constraints or domain knowledge to enhance feasibility. Joint Distribution Shapley[0] situates itself within the Optimal Transport and Coupling-Based Refinement cluster, emphasizing distributional alignment to refine counterfactuals beyond simple distance minimization. This contrasts with gradient-based methods like CF-GNNExplainer[3] for graph data or boundary-focused approaches such as Minimal Feature Boundary[4], which prioritize decision surface proximity. The interplay between theoretical rigor—ensuring minimal yet meaningful changes—and practical deployment remains an active research question, with Joint Distribution Shapley[0] contributing a principled coupling perspective that complements existing optimization paradigms by addressing joint feature dependencies.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on optimal transport and coupling theory to refine counterfactuals by minimizing distributional divergence while maintaining sparsity. Sibling subtopics represent alternative optimization paradigms: gradient-based perturbation methods, multi-objective frameworks balancing competing criteria, and Shapley value-guided approaches for feature prioritization. All share the goal of minimal feature modifications but differ fundamentally in their mathematical frameworks and optimization strategies.

Similarities: - All subtopics aim to generate counterfactual explanations with minimal feature changes - Each approach addresses the trade-off between proximity to the original instance and validity of the counterfactual - Sparsity (minimizing the number of changed features) is a common objective across all methods - All methods require some form of optimization to balance competing desiderata

Differences: - Optimal transport uses distributional coupling and Wasserstein-like metrics, while gradient methods operate directly on feature or latent spaces - Multi-objective approaches explicitly balance multiple competing objectives (proximity, sparsity, diversity), whereas optimal transport embeds these within a single divergence measure - Shapley value methods leverage game-theoretic feature attribution to guide optimization, distinct from transport-theoretic or gradient-based strategies - Optimal transport provides theoretical guarantees on distribution matching, while gradient methods focus on local perturbation efficiency - Multi-objective frameworks may use Pareto optimization or constraint satisfaction, contrasting with the coupling-based refinement of optimal transport

Suggested Search Directions: - Hybrid methods combining optimal transport with Shapley-based feature selection for interpretable refinement - Comparative studies on computational efficiency: optimal transport vs. gradient descent for counterfactual generation - Integration of optimal transport into multi-objective frameworks to handle distributional constraints alongside proximity and sparsity

Sibling Subtopics

- **Gradient-Based and Perturbation Optimization** (leaves: 1, papers: 3)
- Scope: Approaches using gradient descent or direct perturbation optimization to find minimal changes in feature or latent space.
- Exclude: Methods incorporating Shapley values, optimal transport, or energy landscapes belong to specialized optimization subcategories.
- **Multi-Objective and Constraint-Based Optimization** (leaves: 1, papers: 5)
- Scope: Methods balancing multiple objectives such as proximity, sparsity, validity, and diversity through multi-objective optimization or constraint satisfaction.
- Exclude: Single-objective methods or those using genetic algorithms belong to other subcategories.
- **Shapley Value-Guided Optimization** (leaves: 1, papers: 2)
- Scope: Methods leveraging Shapley values or SHAP to identify and prioritize minimal feature changes for counterfactual generation.
- Exclude: Methods using Shapley values only for post-hoc refinement or dimensionality reduction belong to refinement frameworks.

Contributions Analysis

Overall novelty summary. The paper introduces COLA, a post-hoc refinement framework that uses optimal transport coupling and Shapley-based attribution to reduce feature edits in counterfactual explanations. Within the taxonomy, it resides in the 'Optimal Transport and Coupling-Based Refinement' leaf under 'Optimization-Based Counterfactual Generation Methods'. Notably, this leaf contains only the original paper itself—no sibling papers are present. This isolation suggests the specific combination of optimal transport coupling with Shapley-driven feature selection for counterfactual refinement represents a relatively sparse research direction within the broader optimization-based counterfactual landscape.

The taxonomy reveals that COLA's parent branch, 'Optimization-Based Counterfactual Generation Methods', contains several neighboring leaves: 'Gradient-Based and Perturbation Optimization' (3 papers), 'Shapley Value-Guided Optimization' (2 papers), and 'Multi-Objective and Constraint-Based Optimization' (5 papers). While Shapley values appear in adjacent work for feature prioritization, and optimal transport exists in theoretical frameworks, the taxonomy structure indicates that coupling-based refinement as a distinct methodological approach has not been extensively explored. The scope note clarifies this leaf excludes methods using optimal transport only for post-hoc refinement without coupling theory, suggesting a narrow definitional boundary.

Among 29 candidates examined across three contributions, the 'p-SHAP' component shows one refutable candidate from 10 examined, indicating some overlap with prior Shapley-based attribution methods. The 'COLA framework' contribution examined 10 candidates with zero refutations, suggesting the overall refinement architecture appears distinct within the limited search scope. The 'theoretical guarantees' contribution also shows no refutations across 9 candidates. These statistics reflect a top-K semantic search, not exhaustive coverage, meaning the apparent novelty of COLA's coupling-driven refinement may stem partly from the nascent state of this specific methodological intersection rather than comprehensive field saturation.

Given the limited search scope of 29 candidates and the taxonomy's structural sparsity in this leaf, COLA appears to occupy a relatively unexplored niche combining optimal transport coupling with Shapley attribution for counterfactual refinement. The single refutation for p-SHAP suggests incremental overlap in attribution mechanics, while the framework's overall architecture shows distinctiveness within the examined literature. However, the analysis cannot rule out relevant work outside the top-K semantic matches or in adjacent optimization paradigms not captured by the current taxonomy boundaries.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: COLA framework for sparse counterfactual explanations

Description: The authors propose COLA (COunterfactuals with Limited Actions), a general post-hoc framework that refines counterfactual explanations across different models and CE generators. It uses optimal transport to compute a coupling between factual and counterfactual sets, which then guides Shapley-based attribution to select minimal feature edits while preserving target effects.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality

URL: View paper

Brief Assessment

Actionable Minimality[6] addresses the same problem of minimizing feature changes in counterfactual explanations using optimal transport and Shapley values, but appears to be the same work or a closely related variant rather than prior work that refutes novelty.

---

2. The matrix reloaded: Towards counterfactual group fairness in machine learning

URL: View paper

Brief Assessment

Matrix Reloaded[55] focuses on counterfactual group fairness evaluation in ML models, not on post-hoc refinement of counterfactual explanations using optimal transport for sparse feature edits. The candidate addresses bias detection and mitigation through counterfactual generation for fairness assessment, whereas the original paper addresses minimizing feature modifications in counterfactual explanations to achieve target model outputs.

---

3. Post-event modeling via causal optimal transport for ctr prediction

URL: View paper

Brief Assessment

Causal Optimal Transport[53] focuses on CTR prediction in recommendation systems using optimal transport for post-event modeling, not on refining counterfactual explanations with Shapley-based attribution for sparse feature edits.

---

4. Distributional Counterfactual Explanations With Optimal Transport

URL: View paper

Brief Assessment

Distributional Optimal Transport[56] focuses on distributional counterfactual explanations where entire input-output distributions are treated as distributions, not on post-hoc refinement of individual counterfactual explanations for sparsity. The candidate addresses a fundamentally different problem of aligning statistical distributions rather than selecting minimal feature edits from existing counterfactuals.

---

5. DISCOUNT: Distributional Counterfactual Explanation With Optimal Transport

URL: View paper

Brief Assessment

DISCOUNT[51] focuses on distributional counterfactual explanations using optimal transport to align entire input-output distributions, not on post-hoc refinement of individual counterfactual explanations for sparsity as COLA does.

---

6. Conservative inference for counterfactuals

URL: View paper

Brief Assessment

Conservative Counterfactual Inference[54] focuses on identifying conservative joint laws of counterfactual random variables for continuous treatments in causal inference, not on post-hoc refinement of counterfactual explanations using optimal transport for feature selection in machine learning models.

---

7. A consistent extension of discrete optimal transport maps for machine learning applications

URL: View paper

Brief Assessment

Discrete Transport Extension[58] focuses on extending discrete optimal transport maps to new observations for general machine learning applications, not on post-hoc refinement of counterfactual explanations or Shapley-based feature selection for minimal edits.

---

8. When adversarial attacks become interpretable counterfactual explanations

URL: View paper

Brief Assessment

Adversarial Interpretable[59] focuses on saliency maps and gradient-based counterfactual explanations in 1-Lipschitz neural networks using optimal transport for training, not on post-hoc refinement of counterfactual explanations across different CE generators using Shapley-based attribution.

---

9. Relative Explanations for Contextual Problems with Endogenous Uncertainty: An Application to Competitive Facility Location

URL: View paper

Brief Assessment

Relative Explanations[57] focuses on contextual stochastic optimization with endogenous uncertainty in facility location, not on post-hoc refinement of counterfactual explanations using optimal transport for feature attribution across different models and CE generators.

---

10. Collective Counterfactual Explanations via Optimal Transport

URL: View paper

Brief Assessment

Collective Optimal Transport[52] focuses on collective counterfactual explanations that account for population dynamics and competition among individuals, rather than post-hoc refinement of individual counterfactual explanations using optimal transport for feature sparsity as proposed in COLA.

Contribution 2: Joint distribution-informed Shapley values (p-SHAP)

Description: The authors introduce p-SHAP, a Shapley value method that integrates an algorithm returning joint probability between factual and counterfactual instances. This method unifies other commonly used Shapley methods under appropriate couplings and provides a modular interface for attribution and edit selection.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality

URL: View paper

Brief Assessment

Actionable Minimality[6] describes computing joint distributions between observed and counterfactual data to inform Shapley values, which is the same technical approach rather than prior work challenging the novelty claim.

---

2. Explaining Reinforcement Learning: A Counterfactual Shapley Values Approach

URL: View paper

Brief Assessment

RL Counterfactual Shapley[67] focuses on explaining RL agent decisions by comparing optimal vs. non-optimal actions using counterfactual analysis, not on general counterfactual explanation refinement or joint distributions between factual and counterfactual instances for feature selection in supervised settings.

---

3. Calculating and Visualizing Counterfactual Feature Importance Values

URL: View paper

Brief Assessment

Visualizing Counterfactual Importance[63] focuses on calculating Shapley values between factual-counterfactual pairs for counterfactual explanations in classification tasks, not on general RL frameworks with optimal transport-based joint distributions for sparse counterfactual refinement.

---

4. Decomposition of inequality of opportunity in India: An application of data-driven machine learning approach

URL: View paper

Brief Assessment

Inequality Decomposition India[66] applies Shapley value decomposition to inequality analysis in India, focusing on counterfactual distributions for socioeconomic variables. This is a domain-specific application unrelated to the original paper's framework for sparse counterfactual explanations in machine learning models using optimal transport-based joint distributions.

---

5. The Counterfactual-Shapley Value: Attributing Change in System Metrics

URL: View paper

Brief Assessment

Counterfactual-Shapley Value[64] focuses on attributing change in system metrics using time-series models and structural causal models for real-world systems, not on feature selection via joint distributions between factual and counterfactual instances as in p-SHAP.

---

6. -test: Global Feature Selection and Inference for Shapley Additive Explanations

URL: View paper

Brief Assessment

Global Feature Selection[68] focuses on global feature selection with selective inference for post-selection p-values and confidence intervals, not on joint distribution-informed Shapley values for counterfactual explanations. The candidate uses Shapley values for feature importance ranking and screening, whereas the original paper introduces p-SHAP specifically for sparse counterfactual explanations via optimal transport coupling.

---

7. Counterfactual Shapley Values for Explaining Reinforcement Learning

URL: View paper

Brief Assessment

Counterfactual Shapley RL[60] focuses on reinforcement learning explanations using counterfactual analysis to compare optimal vs. non-optimal actions, not on counterfactual explanations for sparse feature selection in general ML models as in the original paper.

---

8. Counterfactual shapley additive explanations

URL: View paper

Prior Art Analysis

Counterfactual Shapley Additive[61] demonstrates that prior work exists on using counterfactual information within Shapley value frameworks. The candidate paper proposes CF-SHAP, which incorporates counterfactual information to produce a background dataset for use within the marginal Shapley value framework. This directly addresses the same core concept as p-SHAP: integrating counterfactual/joint probability information into Shapley value computation. Both methods aim to improve feature attribution by considering the relationship between factual and counterfactual instances, though they differ in implementation details (CF-SHAP uses counterfactual-generated background datasets, while p-SHAP uses optimal transport-based joint distributions).

Evidence

Evidence 1 - **Rationale**: The original paper explicitly references CF-SHAP as prior work in their preliminaries section, acknowledging it as an existing method that uses counterfactual distributions. This demonstrates that the concept of using counterfactual information in Shapley values predates the original paper's p-SHAP contribution. - **Original**: counterfactual shapley (cf-shap) further sets d to the counterfactual distribution conditioned on xi: v(i) cf (s) = er~d(xi)[f(xi,s; rf\s)] - er~d(xi)[f(r)], which assumes a probabilistic alignment and has shown advantages for contrastive attribution - **Candidate**: we propose a variant of shap, counterfactual shap (cf-shap), that incorporates counterfactual information to produce a background dataset for use within the marginal (a.k.a. interventional) shapley value framework.

---

9. Group Shapley Value and Counterfactual Simulations in a Structural Model

URL: View paper

Brief Assessment

Group Shapley[65] focuses on decomposing parameter changes in structural economic models using group-level Shapley values, not on joint distributions between factual and counterfactual instances for feature attribution in counterfactual explanations.

---

10. To Select or Not to Select? The Role of Meta-features Selection in Meta-learning Tasks with Tabular Data

URL: View paper

Brief Assessment

Meta-features Selection[62] focuses on meta-feature selection for meta-learning tasks with tabular data, not on counterfactual explanations or Shapley value methods for feature attribution in the context of sparse counterfactual explanations. The candidate's counterfactual-based selection uses counterfactual generation efficiency as a criterion for meta-feature selection, which is fundamentally different from the original's p-SHAP method that integrates joint probability between factual and counterfactual instances for attribution and edit selection in counterfactual explanations.

Contribution 3: Theoretical guarantees for OT-based counterfactual refinement

Description: The authors provide theoretical results showing that optimal transport minimizes an upper bound on the 1-Wasserstein divergence between factual and counterfactual outcomes. They also prove that under mild conditions, refined counterfactuals remain no farther from factuals than the original counterfactuals.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Distributional Counterfactual Explanations With Optimal Transport

URL: View paper

Brief Assessment

Distributional Optimal Transport[56] provides theoretical results on Wasserstein distance bounds in the context of distributional counterfactual explanations (Theorem 3.2, equations 6 and 8), not for refining individual counterfactuals with distance guarantees. The theoretical framework addresses chance-constrained optimization for distributional alignment, which is distinct from the original paper's refinement guarantees.

---

2. DISCOUNT: Distributional Counterfactual Explanation With Optimal Transport

URL: View paper

Brief Assessment

DISCOUNT[51] provides theoretical results for distributional alignment using Wasserstein distance with confidence intervals, but does not address refinement guarantees showing counterfactuals remain no farther from factuals than originals.

---

3. Trajectory-Based Representation Balancing with Decomposed Uncertainty for Dynamic Treatment Regimes

URL: View paper

Brief Assessment

Trajectory-Based Balancing[74] focuses on trajectory-based Wasserstein distance for dynamic treatment regimes, not on optimal transport minimizing divergence bounds for counterfactual refinement with distance guarantees as in the original paper.

---

4. Conservative inference for counterfactuals

URL: View paper

Brief Assessment

Conservative Counterfactual Inference[54] uses optimal transport to find conservative joint distributions in causal inference settings, not to minimize Wasserstein divergence bounds for refining counterfactual explanations in predictive models with distance guarantees.

---

5. Inter-and intra-similarity preserved counterfactual incentive effect estimation for recommendation systems

URL: View paper

Brief Assessment

Similarity Preserved Incentive[70] uses optimal transport for similarity preservation in uplift modeling for recommendation systems, not for counterfactual refinement with Wasserstein divergence bounds and distance guarantees as in the original paper's theoretical framework.

---

6. Proximity Matters: Local Proximity Preserved Balancing for Treatment Effect Estimation

URL: View paper

Brief Assessment

Local Proximity Balancing[71] focuses on treatment effect estimation in causal inference, not counterfactual explanations for model predictions. The theoretical guarantees address different problems: the candidate bounds PEHE for treatment effects, while the original bounds W1 divergence for counterfactual refinement in explainability.

---

7. Counterfactual fairness by combining factual and counterfactual predictions

URL: View paper

Brief Assessment

Counterfactual Fairness[69] focuses on fairness in ML predictions by combining factual and counterfactual predictions, not on counterfactual explanation refinement or sparsity optimization as in the original paper.

---

8. Causal optimal transport for treatment effect estimation

URL: View paper

Brief Assessment

Causal Treatment Transport[72] focuses on treatment effect estimation in causal inference, not counterfactual explanation refinement. The theoretical guarantees address different problems with different objectives and settings.

---

9. ModÃ¨les contrefactuels pour un apprentissage machine explicable et juste: une approche par transport de masse

URL: View paper

Brief Assessment

Contrefactual Transport Masse[73] focuses on counterfactual fairness and mass transportation theory for causal reasoning, not on counterfactual refinement with Wasserstein divergence bounds and distance guarantees as in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 25 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality

Detected in: Contribution: contribution_1, Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Joint Distribution–Informed Shapley Values for Sparse Counterfactual Explanations View paper
- [1] Counterfactual explanations and how to find them: literature review and benchmarking View paper
- [2] Flexible and Robust Counterfactual Explanations with Minimal Satisfiable Perturbations View paper
- [3] Cf-gnnexplainer: Counterfactual explanations for graph neural networks View paper
- [4] Counterfactual explanation generation with minimal feature boundary View paper
- [5] Counterfactual explainable recommendation View paper
- [6] Refining Counterfactual Explanations With Joint-Distribution-Informed Shapley Towards Actionable Minimality View paper
- [7] COMBINEX: A Unified Counterfactual Explainer for Graph Neural Networks via Node Feature and Structural Perturbations View paper
- [8] Exploring Energy Landscapes for Minimal Counterfactual Explanations: Applications in Cybersecurity and Beyond View paper
- [9] Diffusion visual counterfactual explanations View paper
- [10] Counterfactual Explanations for Models of Code View paper
- [11] Counterfactual visual explanations View paper
- [12] Adversarial counterfactual visual explanations View paper
- [13] Counterfactual explanations for deep learning-based traffic forecasting View paper
- [14] An investigation into creating counterfactual examples for non-linear Support Vector Machines View paper
- [15] DisCERN: Discovering Counterfactual Explanations using Relevance Features from Neighbourhoods View paper
- [16] Enhanced Counterfactual Explanations for Optimizing Three-Dimensional Printing Parameters Using SHAP and Nearest-Neighbor Constraints With Physics-Based â□¦ View paper
- [17] Flexible Counterfactual Explanations with Generative Models View paper
- [18] Multi-SpaCE: Multi-Objective Subsequence-based Sparse Counterfactual Explanations for Multivariate Time Series Classification View paper
- [19] SVCE: Shapley Value Guided Counterfactual Explanation for Machine Learning-Based Autonomous Driving View paper
- [20] Measurable counterfactual local explanations for any classifier View paper
- [21] SenseCF: LLM-Prompted Counterfactuals for Intervention and Sensor Data Augmentation View paper
- [22] Enhancing counterfactual image generation using mahalanobis distance with distribution preferences in feature space View paper
- [23] Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End View paper
- [24] Customer-Centric Decision-Making with XAI and Counterfactual Explanations for Churn Mitigation View paper
- [25] Multimodal LLM for enhanced Alzheimerâ□s Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data View paper
- [26] Counterfactual Explanation of a Classification Model for Detecting SQL Injection Attacks View paper
- [27] TX-Gen: Multi-Objective Optimization for Sparse Counterfactual Explanations for Time-Series Classification View paper
- [28] MASCOTS: Model-Agnostic Symbolic COunterfactual explanations for Time Series View paper
- [29] DiCE-Extended: A Robust Approach to Counterfactual Explanations in Machine Learning View paper
- [30] FACE: feasible and actionable counterfactual explanations View paper
- [31] UGCE: User-Guided Incremental Counterfactual Exploration View paper
- [32] Counterfactual explanation based on gradual construction for deep networks View paper
- [33] Counterfactual review-based recommendation View paper
- [34] PairCFR: Enhancing Model Training on Paired Counterfactually Augmented Data through Contrastive Learning View paper
- [35] Generation of Molecular Counterfactuals for Explainable Machine Learning Based on Coreâ□Substituent Recombination View paper
- [36] Actionable and diverse counterfactual explanations incorporating domain knowledge and causal constraints View paper
- [37] Cpl: Counterfactual prompt learning for vision and language models View paper
- [38] EEG-based motor imagery recognition via novel explainable ensemble learning architecture View paper
- [39] Explainable image classification with evidence counterfactual View paper
- [40] Conceptual Edits as Counterfactual Explanations. View paper
- [41] A survey on natural language counterfactual generation View paper
- [42] Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction View paper
- [43] PerCE: Hierarchical Perturbation-Based Counterfactual Explanations for Multivariate Time Series Classification View paper
- [44] Looking in the mirror: A faithful counterfactual explanation method for interpreting deep image classification models View paper
- [45] AIM-CF: Fast and Precise Counterfactual Explanations via Approximate Inverse Models View paper
- [46] CoLa-DCE - Concept-guided Latent Diffusion Counterfactual Explanations View paper
- [47] Region-aware Minimal Counterfactual Rules for Model-agnostic Explainable Classification View paper
- [48] Introducing User Feedback-Based Counterfactual Explanations (UFCE) View paper
- [49] Feature-Driven Counterfactual Explanations: A SHAP-Based Approach to Dimensionality Reduction in XAI View paper
- [50] Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics View paper
- [51] DISCOUNT: Distributional Counterfactual Explanation With Optimal Transport View paper
- [52] Collective Counterfactual Explanations via Optimal Transport View paper
- [53] Post-event modeling via causal optimal transport for ctr prediction View paper

- [54] Conservative inference for counterfactuals View paper
- [55] The matrix reloaded: Towards counterfactual group fairness in machine learning View paper
- [56] Distributional Counterfactual Explanations With Optimal Transport View paper
- [57] Relative Explanations for Contextual Problems with Endogenous Uncertainty: An Application to Competitive Facility Location View paper
- [58] A consistent extension of discrete optimal transport maps for machine learning applications View paper
- [59] When adversarial attacks become interpretable counterfactual explanations View paper
- [60] Counterfactual Shapley Values for Explaining Reinforcement Learning View paper
- [61] Counterfactual shapley additive explanations View paper
- [62] To Select or Not to Select? The Role of Meta-features Selection in Meta-learning Tasks with Tabular Data View paper
- [63] Calculating and Visualizing Counterfactual Feature Importance Values View paper
- [64] The Counterfactual-Shapley Value: Attributing Change in System Metrics View paper
- [65] Group Shapley Value and Counterfactual Simulations in a Structural Model View paper
- [66] Decomposition of inequality of opportunity in India: An application of data-driven machine learning approach View paper
- [67] Explaining Reinforcement Learning: A Counterfactual Shapley Values Approach View paper
- [68] -test: Global Feature Selection and Inference for Shapley Additive Explanations View paper
- [69] Counterfactual fairness by combining factual and counterfactual predictions View paper
- [70] Inter-and intra-similarity preserved counterfactual incentive effect estimation for recommendation systems View paper
- [71] Proximity Matters: Local Proximity Preserved Balancing for Treatment Effect Estimation View paper
- [72] Causal optimal transport for treatment effect estimation View paper
- [73] ModÃ¨les contrefactuels pour un apprentissage machine explicable et juste: une approche par transport de masse View paper
- [74] Trajectory-Based Representation Balancing with Decomposed Uncertainty for Dynamic Treatment Regimes View paper