Beyond Correlation: Incorporating Counterfactual Guidance to Better Support Exploratory Visual Analysis

Arran Zeyu Wang, David Borland, and David Gotz

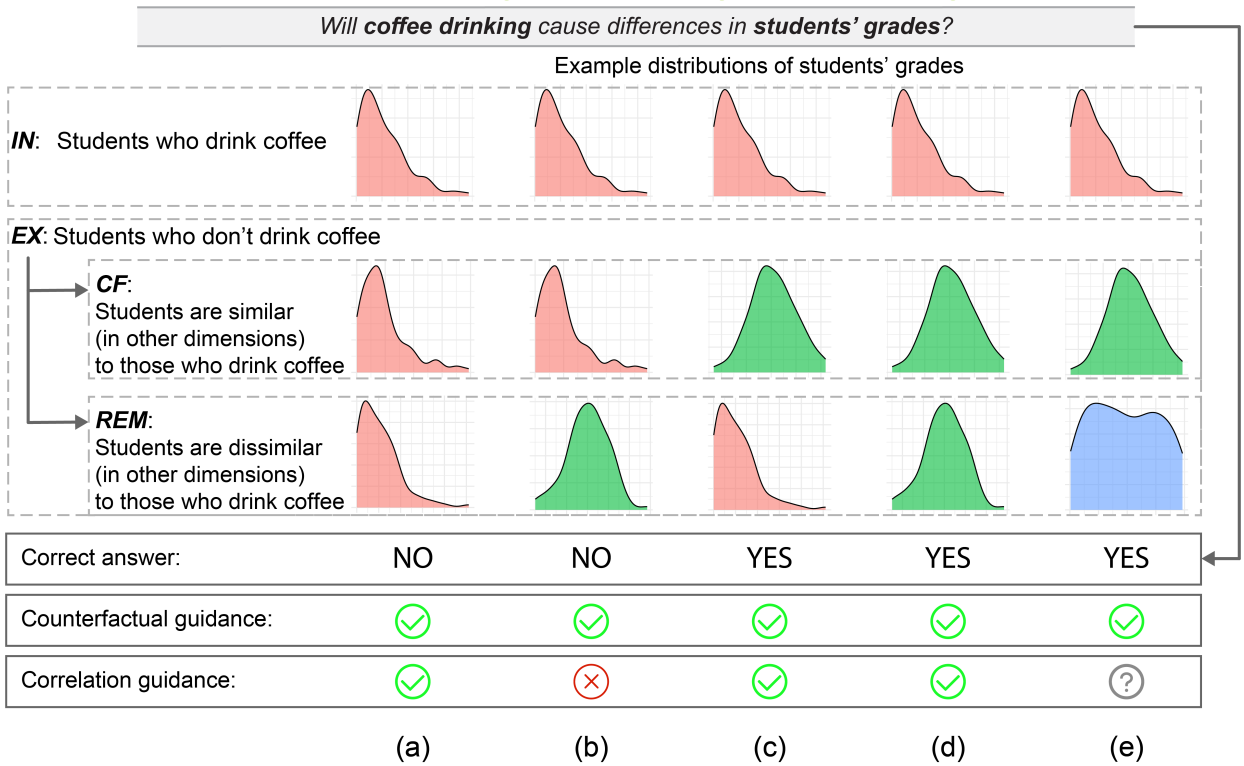

Fig. 1: The proposed counterfactual guidance technique can be compared with traditional correlation-based guidance by considering five archetypal scenarios. Consider the example question "*Will coffee drinking cause differences in students' grades?*" An analyst might compare data based on whether students drink coffee or not, and attempt to answer the question based on differences in the distribution of grades for the resulting subsets. The leftmost column lists the subsets created in this process (see Section 3.1 for details), and the various charts illustrate five potential combinations (a-e) of distributions across the different subsets which suggest different possible answers to the analytical question (see Section 3.2 for details). Across the bottom of the figure, the symbols indicate which methods more accurately reflect the correct interpretation of the data. As the example illustrates, counterfactual-based approaches have advantages in two of the five scenarios while they perform equally on the other three.

Abstract— Providing effective guidance for users has long been an important and challenging task for efficient exploratory visual analytics, especially when selecting variables for visualization in high-dimensional datasets. Correlation is the most widely applied metric for guidance in statistical and analytical tools, however a reliance on correlation may lead users towards false positives when interpreting causal relations in the data. In this work, inspired by prior insights on the benefits of counterfactual visualization in supporting visual causal inference, we propose a novel, simple, and efficient counterfactual guidance method to enhance causal inference performance in guided exploratory analytics based on insights and concerns gathered from expert interviews. Our technique aims to capitalize on the benefits of counterfactual approaches while reducing their complexity for users. We integrated counterfactual guidance into an exploratory visual analytics system, and using a synthetically generated ground-truth causal dataset, conducted a comparative user study and evaluated to what extent counterfactual guidance can help lead users to more precise visual causal inferences. The results suggest that counterfactual guidance improved visual causal inference performance, and also led to different exploratory behaviors compared to correlation-based guidance. Based on these findings, we offer future directions and challenges for incorporating counterfactual guidance to better support exploratory visual analytics.

Index Terms—Counterfactual, Guidance, Exploratory visual analysis, Visual causal inference, Correlation

1 Introduction

Supporting efficient discoveries of insights within complex datasets is a primary goal for exploratory data analysis. Visual analytics tools often employ guided approaches to lead users to find meaningful inferences from high-dimensional data [11, 25, 53]. The most typical and widely applied guidance metric is correlation [2,60], however correlation-based guidance may mislead users by suggesting false causal relationships [6, 41, 53].

- *Arran Zeyu Wang and David Gotz are with the University of North Carolina at Chapel Hill. E-mail: zeyuwang@cs.unc.edu, gotz@unc.edu*
- *David Borland is with RENCI at the University of North Carolina at Chapel Hill. E-mail: borland@renci.org*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Filtering is a common step in visual analytics workflows [7, 36, 46], whereby users can create data subsets of interest based on specified constraints to help answer analytical questions. However, ad-hoc filtering operations can also lead to mistaken assumptions regarding the strength and causal nature of relationships between variables. Recent advances have employed *counterfactuals* in visualization—visualizing additional data subsets designed to provide improved context—to provide benefits to various visualization and visual analytics tasks, such as better interpretations of machine learning models [14, 20, 59] and improved visual causal inference [35, 54, 55]. However, counterfactual-based methods require more complicated and nuanced interpretation of visualizations, potentially leading to more time-consuming and complex analyses [6].

This paper aims to capitalize on the benefits of counterfactual approaches, while also reducing complexity for the user, by using counterfactuals to improve guidance in visual analytics systems with respect to causal interpretations of data. Inspired by existing insights and expert interviews, we introduce a novel counterfactual-based guidance technique designed to capture differences between subsets created by counterfactual visualization techniques [55]. Similar to correlation-based methods, our approach outputs a numeric value that can be used to guide users' exploration, thus simplifying the complexity typically associated with explaining counterfactuals to users. Our technique therefore combines the benefits of counterfactuals to better support guided visual exploration while mitigating its limitations. Unlike previous counterfactual visualization work, our study incorporates counterfactuals into a guidance technique that enables effective exploration of datasets while significantly reducing visualization complexity. In addition, we provide a more thorough analysis of users' exploratory patterns.

We illustrate the benefits of counterfactual guidance compared to correlation-based guidance via a theoretical scenario using a simple students' coffee drinking example (Figure 1). Through this use case, we show how counterfactual guidance can avoid incorrect inferences and more effectively lead to correct inferences compared to correlation-based guidance.

Furthermore, through a comparative user study with a prototype exploratory visual analysis system using a synthetic dataset with ground-truth causal relationships, we demonstrate that counterfactual guidance leads to improved performance in visual causal inference tasks compared to correlation-based guidance. Based on the findings, we propose design implications for better integrating counterfactual guidance into exploratory visual analytics systems, aiming to facilitate more efficient and insightful data exploration.

Specifically, the contributions of this paper include:

- **A counterfactual-based guidance technique with an open-sourced library to support exploratory visual analysis.** We propose a new counterfactual-based method to compute guidance for visual analytics systems. Furthermore, we provide an open-source Python library to compute counterfactual guidance.
- **Theoretical and empirical evidence demonstrating the benefits of counterfactual guidance for visual analysis.** We demonstrate a theoretical use case and present results from an empirical user study to illustrate the benefits of counterfactual guidance versus traditional correlation-based guidance.
- **Reflecting on prior work and discussing future research directions.** We discuss how our study can reflect and confirm prior insights and indicate future research directions to better incorporate counterfactual guidance with exploratory analytics.

2 Background and Related Work

The methods presented in this paper build on prior work in two broad areas of related research. First, our approach is informed by prior work exploring counterfactuals and their applications in support of visual causal inference. Second, our contributions are designed to extend previous approaches to guided exploratory visual analysis.

2.1 Counterfactuals in Visual Causal Inference

Causal inference techniques are designed to help characterize the causal relationships between various factors within a dataset, showing how one factor may lead to changes in another. Pearl [41] established *counterfactual reasoning* as the most advanced level of his proposed statistical causal inference hierarchy, involving the exploration of hypothetical alternatives to observed events. The statistical and machine learning communities have proposed many techniques to support causal inference. For example, instrumental variables have been employed to explore causal structures among datasets [1]. In other work, machine learning approaches were utilized to perform causal inference from complex data [33]. Alternatively, score matching methods can be used to extract impact factors for target outcomes to help create causal models [31, 51].

Due to their utility in exploring outcome relations [39], counterfactuals have been increasingly applied in visualization research to enhance the understanding of datasets [6]. For example, Kaul et al. introduced the first general-purpose counterfactual-based exploratory visual analytics system, *CoFact* [35]. Through a user study, they found that *CoFact* could assist users in inferring feature-to-outcome relations from datasets. However, *CoFact* utilized complex counterfactual visualizations without well-designed guidance techniques, making effective data exploration challenging.

Further, preliminary research has explored the potential of counterfactual visualization—visualizing data subsets that do not match filter inclusion criteria, but are similar to the included subset in other ways—to benefit users' causal inferences in exploratory tasks. Wang et al. proposed a causality comprehension model and found that counterfactual visualizations benefit users' causality comprehension for juxtaposed visualizations [54]. However, their studies focused on static statistical charts and did not address the use of counterfactuals in an exploratory context.

However, challenges remain in effectively conveying these complex causal relationships through visual means [6]. The most effective ways to present information in order to improve causal inference is still an area of active research. Prior studies also found that although visualizing counterfactuals can provide improved performance in causal inference, due to their increased visual and conceptual complexity users typically took longer to explore and interpret them [54]. These studies are also limited by a lack of ground truth causal relationships to validate any advantages of counterfactuals.

These approaches have shown promise in helping users form more accurate interpretations of data, although the field is still exploring the most effective ways to integrate counterfactuals into visual data communication. Built upon existing insights, we aim to maintain the benefits of visualizing counterfactuals while mitigating their increased complexity in interpreting visualizations. In this paper, we present a simple yet effective counterfactual measure to guide exploratory visual analysis.

2.2 Guided Exploratory Visual Analysis

Guided exploratory analysis in visual analytics refers to the process of leading users through a structured exploration of data to uncover insights [11]. Correlation between data variables is the most widely applied metric to guide users towards potentially interesting insights for exploratory visual analysis, and has been applied in various domains such as statistical software [60], visual analysis tools [38], and biostatistics methodologies [12].

Ceneda et al. explored a taxonomy of guidance in the context of visual analytics [9]. They emphasized that the major goal of such guidance is to mitigate the effects of the knowledge gap across different guidance degrees [10]. They further developed theoretical frameworks to better characterize guidance in visual analytics through analyzing designers' requirements [8], descriptively connecting visualization onboarding and guidance [49], and specifying practical guidance strategies [48].

Other approaches have explored practical methods for guided visual analysis. The progressive visual analytics workflow [50] enables user exploration of partial results to quickly lead to the next exploration step by inferring early and meaningful clues. Feedback-driven visual analytics [5] can provide benefits by providing relevant feedback in guiding users during the analysis of large multidimensional

datasets. *SOMFlow* [44] enables guided exploration for cluster analysis using time-series self-organizing maps. *EVM* [34] incorporates model checks [30] into visual analytics systems to guide users in better examining the efficiency of data exploration and interpretation based on statistical models. Indexing [27] and faceted [28] guidance approaches were reported to improve users' exploration efficiency in exploratory search tasks. AI-supported guidance [29] can also benefit users' trust and exploration during visual analysis, especially in more difficult tasks.

Recommendation techniques have also been shown to be beneficial for visual analytics guidance [66]. For example, modeling user behavior [23] and analytical focus during visual analysis [67] can lead to improved user exploration in various usage scenarios such as mass text document analysis [24], web search [15], data pre-fetching [3], and combating bias [22]. Task-driven approaches for recommendations have also been employed to guide advanced mixed-initiative visual analytics of users [16]. Other recommendation systems [40, 57, 61] utilize design principle-driven recommendations to efficiently guide user exploration in exploratory visual analytics.

Although this breadth of research offers significant insights on how to effectively guide user exploration for various specific scenarios, correlation-based guidance is still the most widely applied in analytical and statistical software such as SYSTAT [60] and Tableau [2]. To our knowledge, there have been no prior studies examining the use of counterfactual guidance for visual analytics systems. In this work, we compare counterfactual guidance to correlation-based guidance with respect to their performance supporting causal inference.

3 Archetypal Usage Scenarios

In this section, we describe five archetypal scenarios for counterfactual and correlation-based guidance in data analysis. Through these scenarios, we aim to demonstrate how and when counterfactual guidance can offer benefits over correlation-based methods.

3.1 Data Subsets

First we briefly introduce the definitions of data subsets related to computing counterfactuals to enable counterfactual guidance. For more detailed definitions see [54].

IN: The included (IN) subset comprises the data samples that match user-chosen inclusion criteria when filtering a dataset.

EX: The excluded (EX) subset comprises the data samples that do not match the inclusion criteria for IN.

CF: The counterfactual (CF) subset comprises data samples from EX that are the most similar to those from IN based on variables in the data other than the inclusion criteria. Following prior studies [35, 54], we employ the Euclidean distance as the default similarity measure.

REM: The remainder (REM) subset comprises the data samples from EX that are not included in CF.

Figure 1 illustrates instances for these subsets. In this case, the corresponding data subsets refer to: **IN:** Students who drink coffee. **EX:** Students who don't drink coffee. **CF:** Students who don't drink coffee but are similar to IN across other variables. **REM:** Students who neither drink coffee nor are similar to IN across other variables.

In the following sections, we employ $D_{IN,CF}$ and $D_{IN,REM}$ as terms to represent the differences between IN and CF, and between IN and REM, respectively. We also refer to a low guidance value as one that reflects a low evidence for a causal effect of the filter variable on the outcome, and a high value as one that reflects greater evidence for a causal effect.

3.2 Archetypes Overview

The five archetypes presented in this section are determined based on the degree of similarity between the key subsets defined earlier in this paper: IN, CF, and REM. More specifically, if we take a simplified binary view of similarity between two subsets, we can specify five different scenarios relating these three sets: (1) IN is similar to both CF and REM; (2) IN and CF are similar, but REM is different; (3) IN and REM are similar, but CF is different; (4) CF and REM are similar, with both different from IN; and (5) all subsets are different from each other.

We illustrate these five archetypal scenarios with examples from a guided exploratory analysis of data describing how coffee drinking relates to students' grades as depicted in Figure 1. In all of these examples, the filter *students who drink coffee* is used to define IN while the distributions represent the outcome variable *students' grades*.

3.3 Case 1: All Subsets have Similar Distributions

When all subsets exhibit similar data distributions (Figure 1 (a)), the conclusion is relatively clear: there is no evidence that the filter variable (coffee consumption) meaningfully influences the outcome (students' grades). This can be seen by comparing the grades of those who consume coffee with those who don't (both similar and non-similar students). In all cases, the grades are similarly distributed.

Mathematically, this scenario will lead to a very low $D_{IN,CF}$ value as well as a very low $D_{IN,REM}$ value, and should result in a low counterfactual guidance value. Similarly, the correlation between coffee drinking and grades (as evidenced by the small difference in outcomes between IN and EX) is low, resulting in a correspondingly low correlation guidance value.

3.4 Case 2: REM is Different

In this case, IN and CF exhibit similar outcomes (similar grades), while the REM subset shows a different outcome distribution (Figure 1 (b)). This case suggests that students who drink coffee earn grades similar to those of non-coffee drinkers who are "just like them" except for their coffee drinking. In contrast, students in the REM subset do not drink coffee but are also dissimilar from coffee-drinking students in other ways beyond their coffee intake. The REM students, in this case, earn different grades from the others and these can be attributed to factors other than coffee given the similar grades within IN and CF.

For these reasons, a user would ideally avoid focusing on these types of variables, and guidance during exploratory analysis would not push users toward such a pattern. When considering counterfactual guidance approaches, even though $D_{IN,REM}$ can be relatively high, guidance should still be low because of the low $D_{IN,CF}$ value. In sharp contrast, correlation-based guidance might very well lead users directly to this less interesting pattern because the correlation between coffee drinking and grades (as evidenced by the difference between IN and EX) is substantial given that *REM* is part of *EX* along with the large $D_{IN,REM}$.

3.5 Case 3: CF is Different

A third possible pattern shows IN and REM having similar outcomes while CF is different as shown (Figure 1 (c)). This case is less common and reflects a more complex circumstance. Using the coffee and grades example, the difference in students' grades between IN and CF suggests that coffee drinking is an important factor given that IN and CF contain similar students except for their coffee consumption. However, the similar grades between IN and REM, which also differ in coffee consumption, suggest that other factors that distinguish between REM and CF may also influence the students' grades. Moreover, the other factors may influence grades in a way that is similar to coffee consumption.

This pattern could be reflected in both counterfactual and correlation-based guidance approaches. Counterfactual guidance would lead users to explore these cases because of the large $D_{IN,CF}$ value. Correlation guidance, meanwhile, would capture the difference between IN and EX, though the strength of the signal may be weaker due to it reflecting a combination of the large $D_{IN,CF}$ and the low $D_{IN,REM}$.

3.6 Case 4: IN is Different

A fourth common pattern is when the CF and REM subsets have similar outcomes that are both different from the IN subset (Figure 1 (d)). In our example, this pattern would reflect coffee drinkers earning grades that are different from those who don't drink coffee and that this difference was seen for all non-coffee drinkers regardless of how similar or different the students were to their coffee-drinking counterparts.

In this case, the difference in students' grades for both REM and CF to IN will result in large values for both $D_{IN,REM}$ and $D_{IN,CF}$. Therefore, this pattern would be identified by both correlation-based guidance and counterfactual guidance.

3.7 Case 5: All Subsets are Different

In the fifth and final case, the outcome distributions are different across all three subsets IN, CF, and REM (Figure 1 (e)). This situation would reflect that all three groups have different grade distributions: coffee-drinking students, non-coffee-drinking students who are like their coffee-drinking peers, and non-coffee-drinking students who are dissimilar from their coffee-drinking peers.

This scenario combines aspects from both cases 2 and 3. As in case 3, counterfactual guidance would highlight this pattern for exploration given the large difference between IN and CF. However, as in case 2, CF and REM are also different. The correlation-based guidance would therefore be more difficult to predict because it depends on the combined distributions of those two subsets.

3.8 Summary

As outlined in the description of these five archetypes and their depiction in Figure 1, correlation-based and counterfactual guidance can both be used effectively under multiple conditions. More specifically, they can both be used to help productively guide users towards cases 3 and 4 and away from case 1. In contrast, cases 2 and 5 are more problematic for correlation-based guidance and could result in incorrect or misleading guidance, whereas counterfactual guidance should be able to provide correct results for both cases.

These archetypes help demonstrate the theoretical rationale for, and benefits of, counterfactual guidance as summarized in the table at the bottom of Figure 1. Motivated by these observations, the study presented in Section 5 provides empirical evidence about the benefits of counterfactual guidance during exploratory analysis when compared to a correlation-based approach.

4 Counterfactual Guidance

In this section, we first share results from formative interviews with visualization experts which aimed to distill key design requirements. Then, informed by the findings from those interviews we present the details of our counterfactual guidance methodology.

4.1 Insights and Concerns from Expert Interviews

Building upon prior findings on counterfactuals in visualization, we gathered suggestions from qualitative expert interviews to identify rationales and practical key insights to incorporate counterfactuals into visual analysis systems. The interviews were conducted with 6 experts, comprising three visualization researchers and three visualization engineers. The interviews lasted 30 minutes on average.

During each interview, experts were introduced to and shown existing counterfactual visualizations and visual analytic tools. They then discussed their concerns and proposed suggestions for building efficient and easy-to-use counterfactual-based analytics systems. We summarize the main insights and concerns raised during these interviews within the following themes.

Complex counterfactual concepts. One common theme was that counterfactual visualizations were seen as helpful tools, but that they are not typically employed in their analytical workflows and may be a complex concept for new users to understand. Four out of six experts reported they had no prior knowledge of counterfactuals, and that understanding the definitions involved in counterfactual visualizations was challenging for them even though examples were provided. All experts suggested that if we want to effectively incorporate counterfactuals into the visual analytics workflow for general users, it would be best if it was done in a way that avoids introducing new complex concept definitions.

Difficulty in understanding the impact of the REM subset. Another challenge raised by the experts focused on the REM subset. When exploring datasets, users are primarily focused on the impact of chosen filters, i.e., the IN subset. Moreover, the experts all agreed that visualizing the CF subset was useful as a way to help them understand the differential impact of their chosen filters. However, five experts mentioned that they didn't easily understand the usefulness of looking at data from the REM subset. This aligns with the more complex conceptual basis required to meaningfully account for the REM subset when interpreting a chart. It involves a three-way subset comparison which requires a deeper understanding of counterfactuals (e.g., 'if IN and CF are different, but REM and CF are alike...'). Given this difficulty, five experts suggested that visualizing REM be de-emphasized. This reflects a difficulty in implementation, however, as the REM subset is critical in counterfactual interpretation.

Lack of simple explanatory use cases. Previous studies provided use cases to explain how counterfactuals in visualizations could be interpreted. However, five experts expressed concern that these use cases may be too complex for users to understand. Furthermore, they reported that the examples didn't effectively illustrate at a glance how counterfactuals can be useful for data exploration rather than the interpretation of an individual chart. This concern helped guide the development of the archetypes in Section 3.

Complexity in visualizations. Many current counterfactual visualization systems typically combine multiple charts to render the different subsets. However, four experts suggested that this kind of presentation significantly increases the complexity of interpreting visualizations, especially for general users. Users have to interpret the individual charts and then mentally combine them to draw the correct insight based on their understanding of counterfactual reasoning. Therefore, our experts suggested that we find ways to simplify the representations of counterfactual visualizations.

4.2 Computing Counterfactual Guidance

Given these insights, we aimed to use counterfactual information in a manner that hides some of the complexity from users. By calculating a value based on user-selected filters in the computed data subsets, counterfactual information can be used to provide improved guidance for exploratory visual analytics. Examples illustrating the counterfactual guidance metric are provided in Section 3.

Note that all the guidance computations described here incorporate only user-selected variables (filters and outcome) and do not include other dataset variables that are used by the computation of data subsets (see Section 3.1).

4.2.1 Counterfactual Dissimilarity

The counterfactual guidance is based on the similarity under user-selected variables for filters and the outcome of interest (e.g., *coffee drinking* and *students' grades* in Figure 1) between the previously introduced data subsets. To compute the distance between data subsets, we first define the *Similarity* between two individual data points as:

$$Similarity(i,j) = exp^{-distance(i,j)}, \tag{1}$$

where i and j are two data points and the exponential function maps the *distance* to a range of [0, 1] with a lower *distance* leading to a higher *Similarity*. The *distance* can be any kind of distance measurement between data points. In this study we employed Euclidean distance for simplicity, familiarity, and continuity with prior work [35],

$$distance(i,j) = \sqrt{(|x_i - x_j|)^2 + (|y_i - y_j|)^2}. \tag{2}$$

Note that the distance measure should be calculated based on all dimensions in the datasets for Equation 2.

Although the results in this paper use the Euclidean distance, other distance measures may be more appropriate for certain domain-specific applications, datasets with unequal importance between variables, or datasets containing large amounts of noise. For example, methods such

as the Mahalanobis distance [18] or propensity score matching [13] could be applied for treatment analyses and other healthcare applications. We therefore provide several commonly used distance measures in the counterfactual library used to implement our guidance system [55]. Developers can also implement or use their own preferred distance measures to replace the implemented ones if they have specific needs.

For typical analyses, users are seeking filter variables that indicate differences in outcomes between subsets. Therefore, the more similar CF is to IN, the less likely it is that the filter constraints is one of interest in the analysis, and vice versa. We therefore define $D_{IN,CF}$, as introduced in Section 3.1, as the normalized dissimilarity between IN and CF:

$$D_{IN,CF} = \frac{1}{|S_{IN}||S_{CF}|} \sum_{i \in S_{IN}}^{|S_{IN}|} \sum_{j \in S_{CF}}^{|S_{CF}|} (1 - Similarity(i,j)), \quad (3)$$

where S_{IN} and S_{CF} are the IN and CF subsets, and $1 - Similarity(i-j)$ is the dissimilarity between points i and j. This formulation of $D_{IN,CF}$ results in a range of [0, 1].

4.2.2 Remainder Dissimilarity

The similarity between IN and REM also impacts data interpretation, as outcome differences between IN and REM could suggest the importance of non-filter variables in the dataset. We therefore define the dissimilarity between IN and REM subsets $D_{IN,REM}$ in a similar way to Equation 3 However, we replace the CF subset S_{CF} with the REM subset S_{REM} throughout the equation:

$$D_{IN,REM} = \frac{1}{|S_{IN}||S_{REM}|} \sum_{i \in S_{IN}}^{|S_{IN}|} \sum_{j \in S_{REM}}^{|S_{REM}|} (1 - Similarity(i,j)), \quad (4)$$

4.2.3 Guidance Score

To compute an overall guidance score, we incorporate $D_{IN,CF}$ and $D_{IN,REM}$ to represent how IN and CF are dissimilar and how IN and REM are dissimilar, respectively. For each, larger values indicate that the selected filters may be of more importance to explore during guided analysis.

To reflect this design, we incorporate both dissimilarities together, weighting $D_{IN,CF}$ more heavily than $D_{IN,REM}$ to capture the focus on the inclusion criteria, as:

$$Guidance_{CF} = \frac{1}{2}(D_{IN,CF} + \sqrt{D_{IN,CF} * D_{IN,REM}}). \quad (5)$$

To reduce the weight of REM we calculate the geometric mean with the CF subset (i.e., the square root item). This ensures that REM is impactful only when CF is impactful. This guides away from archetype case 2. With this guidance equation, we combine the impacts of both subsets, while emphasizing the impact of the CF subset.

As a baseline for our control group in our evaluation study, we also developed a correlation-based guidance measure. Correlation guidance ($Guidance_{corr}$) follows a somewhat similar computation process in our prototype, but replaces the calculation of counterfactual subset distances with correlations between IN and EX.

4.2.4 Subset Distribution Score

A threshold-based method is used to create data subsets. The size n of IN is directly determined by user-selected filters. To create CF we select the n closest point to IN from EX, resulting in IN and CF having the same size, with the remaining points belonging to REM, following previous work [54]. However, if IN contains more than $\frac{1}{3}$ of all data points, we split the data points in EX evenly between CF and REM.

We note that the effectiveness of these subsets may be impacted by the relative sizes of their data samples. For example, when IN includes almost all data samples from the target dataset, no matter how large the $Guidance_{CF}$ or $Guidance_{corr}$ value may be, we cannot conclude that the data subsets would have a high impact. Similarly, if the size of IN were very small, the dissimilarity between these subsets would not be very informative.

Further, we define subset distribution scores $Distribution_{S1,S2}$ to measure the difference between the sizes of different subsets, based on the $SizeDifference_{S1,S2}$:

$$SizeDifference_{S1,S2} = \frac{|S_{S2}|}{|S_{S2}| + |S_{S1}|}, \quad (6)$$

$$Distribution_{S1,S2} = 1 - 2 * |SizeDifference_{S1,S2} - \frac{1}{2}| \quad (7)$$

$$= 1 - 2 * |\frac{|S_{S2}| - |S_{S1}|}{|S_{S2}| + |S_{S1}|}|, \quad (8)$$

which results in a normalized value between 0 and 1.

For example, $Guidance_{CF}$ relies primarily on the differences between IN and CF, so we use $SizeDifference_{IN,CF}$. In ideal cases, the number of samples in IN and CF would be similar, such that $SizeDifference_{IN,CF}$ would be close to 0.5, and the overall $Distribution_{IN,CF}$ would be close to 1. Whenever IN has an extremely large or small number of samples, $SizeDifference_{IN,CF}$ will approach 0 or 1, and $Distribution_{IN,CF}$ will approach 0. Therefore, this measure can be employed as an empirical validation of $Guidance_{IN,CF}$, where lower values of $Distribution_{IN,CF}$ imply a smaller impact of the CF subset. Empirically, when $Distribution_{IN,CF}$ is smaller than 0.1, we find that the subset distribution cannot reliably support insights from $Guidance_{IN,CF}$.

For correlation-based guidance, since $Guidance_{corr}$ measures the differences between IN and EX, we calculated its subset distribution score as $Distribution_{IN,EX}$.

4.3 Implementation

A Python implementation of the proposed counterfactual guidance technique can be found in the *cf_guidance* file accompanied by the *Co-op* library [55] as two functions, *get_cf_guidance_score* and *get_distribution_score*. This library also contains basic computation mechanisms for creating counterfactual subsets, built on efficient scientific computing packages including NumPy, SciPy, and Pandas. The integrated open-source library is available at GitHub.

5 User Study

We conducted a comparative user study using a prototype visual interface to evaluate the performance of counterfactual guidance and compare it to correlation-based guidance. The user study was approved by the UNC-Chapel Hill Institutional Review Board. This section provides detailed descriptions of the study design, analysis process, and results.

5.1 Visual Interface

We designed the functionalities and interactions of the prototype system based on insights from a prior counterfactual-based exploration system [35] and expert interviews. The system enables guided exploration by providing a feature guidance view to help users pick interesting variables and an analytical summary view to help them explore selected filters in detail. Here we describe the two primary views, feature guidance and analytical details, and supported atomic interactions. We compare it with the *CoFact* [35] interface to illustrate how our interface can provide a simplified counterfactual-guided exploration experience.

Note that following prior studies [4, 64], the task, dataset, and outcome variable were fixed in our study design (see Section 5.2 and Section 5.5), therefore we disabled the configuration page for selecting datasets and outcome variables during the study. The same outcome variable is therefore always displayed, and the user is able to filter based on other variables in the dataset to examine changes in the distribution of this outcome variable and interpret relationships between these variables and the outcome variable.

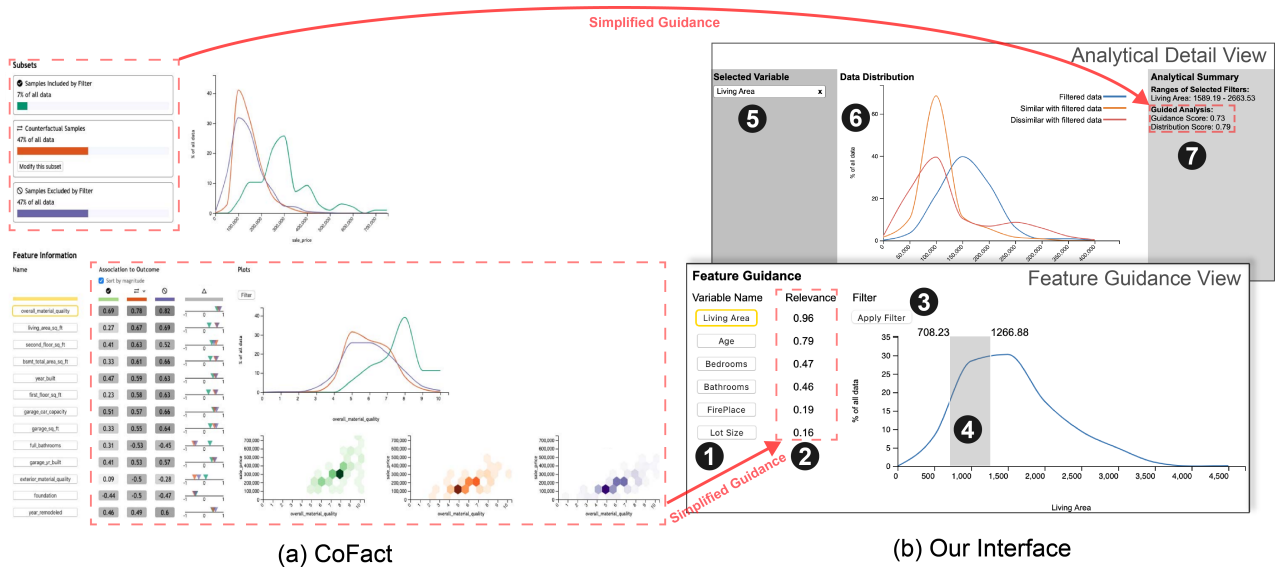

Fig. 2: A comparison between *CoFact* [35] (a) and our interface (b). The red indicates the visualized counterfactual information shown to guide user analysis in each interface, demonstrating how our technique simplifies the counterfactual information shown to users. The labels (b.1) to (b.7) refer to different information and functionality shown in the interface, see Section 5.1 for details.

Feature Guidance. The bottom view in Figure 2 (b) shows the feature guidance view for users to select variables for filtering. Selectable variables (i.e., those other than the outcome and already selected variables) are shown under *Variable Name*, see Figure 2 (b.1). The variables are ordered by their guidance values, shown in *Relevance*, see Figure 2 (b.2). These values may reflect counterfactual guidance or correlation-based guidance, based on different user groups in the study. The term *Relevance* is used for both guidance types. Users can select a variable name displayed in Figure 2 (b.1), and then a visualization of the distribution of the selected variable will be shown in *Filter*, see Figure 2 (b.3). In this *Filter* view, users can click the distribution chart to control the filter ranges they want to apply, as shown in Figure 2 (b.4), and click the *Apply Filter* button to apply the selected filter range. Once the button is clicked, the feature guidance view will be updated based on guidance values calculated using user-applied filters.

Analytical Detail. After applying filters in the feature guidance view, users can switch to a more detailed view, as shown at the top of Figure 2 (b). Selected variables are shown in the *Selected Variable* panel, see Figure 2 (b.5), in the order they were added. Users can also remove a selected variable by clicking the × button next to each variable name in Figure 2 (b.5). Data distributions of the outcome variable for IN, EX, and CF subsets are shown in the *Data Distribution* panel, see Figure 2 (b.6). Note that to ease understanding of the subset distributions, we did not show the subset names explicitly, instead explaining the IN, CF, and REM subsets' patterns as *filtered data*, *those similar with filtered data*, and *those dissimilar with filtered data*, as shown in the data legend in Figure 2 (b.6). Similarly, when using correlation-based guidance, the distributions of IN and EX are shown, where the legend of EX shows *those not in filtered data*. Detailed analytical and guidance information, including all filter ranges and guidance values (guidance score and subset distribution score), are shown in the *Analytical Summary* panel Figure 2 (b.7).

Atomic Interaction Types. Two atomic interactions are available in the system. **Changing filter variables** refers to user interaction to add or remove different variables from the filter inclusion criteria to explore their impact on the outcome. **Changing filter ranges** refers to user interaction to adjust the range of values used for a filter variable.

Comparison with *CoFact* [35]. Figure 2 compares our interface with *CoFact*. The red dashed boxes emphasize the panels and views that show counterfactual and guidance information to help users explore data in each interface. Our interface simplifies much of the information related to guidance, whereas *CoFact* (see Figure 2 (a)) shows more complex subset information and visualizations for achieving the same goal.

5.2 Synthetic Data

One key limitation of prior empirical studies on counterfactuals in visualization is the lack of ground truth causal relations in the studied datasets. To address this issue we generated a dataset with a defined ground truth causality between variables based on a causal graph by selecting example variable names from typical healthcare data [21], such as *blood pressure* and *cholesterol*.

Our data generation is based on prior work on graphical causal inference mechanisms. Since constructing causal models from a dataset is usually complex and may not always guarantee ground-truth causality, we instead first defined a directed acyclic causal graph using our selected variables as nodes and assigned causal relationship strengths between these variables as links using *DoWhy* [45]. We then generated the synthetic data based on the constructed causal graph [32] (see *CausalSynth* [58] for a web application). Due to the difficulties in controlling the causal strength of categorical variables in a causal graph, all of the variables are continuous.

Figure 3 shows the defined causal graph and corresponding values of the ground truth causal relationships shown near each causal link. In this causal graph, we define *mortality risk* as the target outcome variable with different causal relationship strengths to other variables. The top five causal links are shown in red. The causal relationship strengths are assigned with a 0.05 interval (i.e., one's strength to outcome is 0.05 higher than its nearest-lower factor) to avoid potential effects caused by unbalanced differences in causal strength. Several causal links between factors were randomly added to increase data complexity (e.g., *cholesterol* → *blood pressure*).

To facilitate exploratory analysis, we predefined default filter ranges for each variable, determined by two expert analysts to highlight potential interesting filter ranges to serve as inclusion criteria for IN.

5.3 Participants

We recruited 20 users (14 male, 6 female; 19–30 years old) to participate in the study via mailing lists and contacts within professional networks. All participants were at least 18 years old, had or were pursuing a university degree, and had experience using visualization and data analysis such as taking information visualization or data science courses. We employed a between-subjects design by randomly assigning 10 participants to use counterfactual guidance (the CFACT group) while the other 10 participants used correlation guidance (the CORR group)

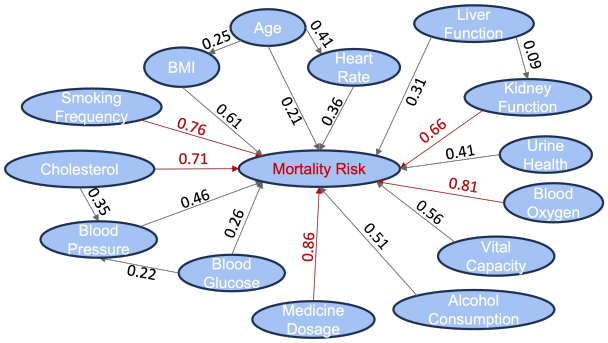

Fig. 3: The defined casual graph in the synthetic data. The middle node is the target outcome, *mortality risk*, shown in red. The values near each link are the causal relationship strengths. The top 5 causal links are shown in red. Causal strengths range from 0.21 to 0.86, with a 0.05 interval between each causal strength in the graph.

in the same prototype system. On average, the study took around 20-30 minutes for users in each group.

5.4 Hypotheses

We aimed to explore the effectiveness of counterfactual guidance for finding causal relationships in a dataset using exploratory visual analysis. Based on this goal, we hypothesized that:

- **H1:** Counterfactual guidance would lead to higher accuracy in finding causal relationships compared to correlation guidance.
- **H2:** Counterfactual guidance would lead to higher confidence in users' findings.
- **H3:** Counterfactual guidance would lead to fewer wrong attempts.
- **H4:** Users using counterfactual guidance may spend more time in analysis.

Through these hypotheses, our study aims to better characterize the impact of counterfactual guidance and compare its effectiveness with correlation-based guidance.

5.5 Procedure and Task

After accepting the informed consent form including the study's purpose and participants' rights, we gave the users a tour of the visual interface and introduced the available interactions and functionalities. With *mortality risk* chosen as the target outcome variable for the visual interface, participants were asked to complete two tasks:

- **T1:** Identify which variables may be most likely to cause higher *mortality risk*, choosing up to 5 variables.
- **T2:** Rank those variables from most likely to least likely.

Participants recorded their confidence on a 5-point Likert scale following each task. The ground truth results of these tasks are five variables with the highest causal relationship strengths to *mortality risk*, as shown by the red links in Figure 3.

5.6 Analysis

We measured our results as accuracy, confidence, interactions, and time spent. The only independent factor was the two user groups, so we calculated individual t-tests based on these measures.

5.7 Results

Table 1 provides the results of the main measures from our analysis. Of note is that all users reported 5 variables, despite the freedom to choose fewer.

Table 1: The t-tests results for main measures in our study. Significant effects are indicated by bold text and the corresponding rows are highlighted in green.

	t-value	p-value	r
T1 Accuracy	-4.40	**< .0001**	.84
T2 Offset	3.43	**.003**	.76
T1 Confidence	-0.56	.58	.02
T2 Confidence	-0.18	.86	.002
Total Interactions	-3.56	**.002**	.51
Changing Filter Variables	1.25	.23	.08
Changing Filter Ranges	-4.17	**.0005**	.79
Wrong Attempts	7.85	**< .0001**	.77
Time	0.23	.82	.003

5.7.1 Accuracy

We assess the accuracy of **T1** by computing the relative ratio between the number of correct answers and 5, i.e., the number of provided answers that were in the top 5 of the ground truth causal relationships, independent of order.

To measure the accuracy for **T2** we use an offset distance, similar to the edit distance [43]:

$$Offset_{T2} = \sum_{i=1}^{5} |i - Rank_{GT}(Answer_i)|, \tag{9}$$

where $Answer_i$ is the variable at the i-th position ($i \in [1,5]$) in the user's ranking, and $Rank_{GT}(Answer_i)$ is the ground truth (GT) ranking of this variable. This metric measures the accuracy between two rankings by summing how much each variable's position in the user ranking deviates from its correct position in the ground truth ranking.

We found significant differences in accuracy between the CFACT and CORR groups for both **T1** ($t = -4.40, p < .0001$) and **T2** ($t = 3.43, p = .003$). The mean differences between CFACT and CORR are 0.28 for **T1** and 2.80 for **T2**, indicating that counterfactual guidance performed better. The results on accuracy therefore support **H1**: we found that counterfactual guidance can lead to higher accuracy in finding causal relationships.

5.7.2 Confidence

We did not find any significant differences for confidence between the CFACT and CORR groups. This indicates that the results do not support **H2**: we found users' confidence was not significantly impacted by guidance type.

5.7.3 Atomic Interactions

We also examined differences with respect to the overall atomic interactions (changing filter variables and changing filter ranges) employed by users in each group. Users in CFACT employed significantly more atomic interactions ($t = -3.56, p = .002$). We also found that the CFACT group performed significantly more changes to filter ranges ($t = -4.17, p = .0005$), but found no significant difference in changing filter variables.

Furthermore, we identified wrong attempts as the total number of atomic interactions involving a variable that does not exist in the top 5 ground truth causal variables. We found a significant difference in wrong attempts ($t = 7.85, p =< .0001$), with a mean difference of -8.7, indicating that users in CFACT had fewer wrong attempts. Our results on interactions therefore support **H3**: we found that counterfactual guidance can help users explore fewer incorrect variables.

5.7.4 Time

We found no significant difference between groups in the overall time spent on analysis, indicating that the results do not support **H4**: counterfactual guidance did not lead to a significant difference in time spent.

Table 2: Significance results for different interaction behaviors in our study. Significant effects are indicated by bold text and the corresponding rows are highlighted in green.

	t-value	p-value	r
Go-back	0.16	.88	.001
Go-next	0.95	.36	.04
Go-back w/ Changing Range	-3.17	**.005**	.49
Go-back w/o Changing Range	-2.98	**.008**	.36
Go-next w/ Changing Range	2.71	**.01**	.39
Go-next w/o Changing Range	4.56	**.0002**	.54
Tree Depth	-4.34	**.0003**	.51
Tree Width	0.61	.55	.02
Filter-Range Width	-3.75	**.001**	.44
Filter-Variable Width	3.86	.07	.18

5.7.5 Exploratory Analysis for Interaction Behaviors

Interaction strategies and behaviors can also impact users' exploration [4, 64]. Given the significant differences in accuracy and number of interactions, we performed a more fine-grained analysis of detailed interaction behaviors to identify any differences between the CFACT and CORR groups. This section follows the previous definition of atomic interaction types from Section 5.1 and extends them based on user behaviors. See Table 2 for a summary of the results for each interaction behavior.

Figure 4 illustrates two sets of key interaction behaviors—go-back and go-next—which we identified from users' exploration behaviors. Each node represents an atomic interaction (changing filter variables or filter ranges). The gray nodes are starting points, which could be any interaction. The green nodes represent a filter variable interaction and the red nodes represent a filter range interaction. Go-back behaviors are cases in which users first add a variable, and then remove it after exploring insights with this variable. Go-next behaviors are cases in which users add a second variable during the analysis of the current variable. Go-next may therefore indicate that users are trying to explore the combined impact of multiple variables. In addition, we subdivide go-back and go-next interaction behaviors into two subtypes according to users' filter range interactions. Go-back interactions include go-back after changing filter ranges (Figure 4 (a)) and go-back without changing filter ranges (Figure 4 (b)). The same subtypes are also included for go-next (Figure 4 (c, d)).

No significant differences were found between the overall go-back and go-next behaviors. However, significant differences were found between each sub-type. The CFACT group exhibited more interaction behaviors of go-back ($t = -3.17, p = .005$) and go-next ($t = 2.71, p = .01$) after changing filter ranges, with mean differences of 1.8 and 1.2 respectively, and exhibited fewer interaction behaviors of go-back ($t = -2.98, p = .008$) and go-next ($t = 4.56, p = .0002$) after changing filter variables, with mean differences of -1.8 and -2.2 respectively.

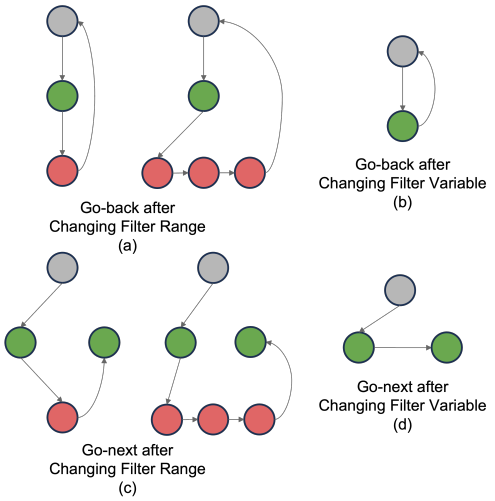

Fig. 4: Four types of identified interaction behaviors in our exploratory analysis consist of atomic interactions. (a-b) are go-back behaviors and (c-d) are go-next behaviors.

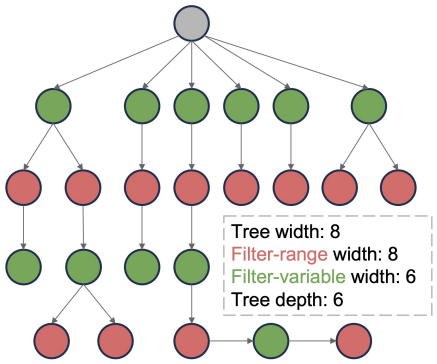

Fig. 5: A search tree of a user in the CFACT group, with a width of 8, a filter-range layer of 8, a filter-variable layer width of 6, and a depth of 6.

Additionally, we examined the impact of counterfactual guidance on the overall interaction structures of users. Following prior work [4], we identified each variable change and range change as a node and removed the backward edges to construct search trees for each user and assessed the data patterns. The tree structure captures the exploration path and helps us better evaluate users' analysis behaviors [4]. Figure 5 illustrates a search tree created based on a user in the CFACT group of our study, with red and green nodes denoting changing filter variables and changing filter ranges respectively. We denote the layer with red nodes as the filter-range layer and the layer with green nodes as the filter-variable layer.

We employed tree width and height as measurements [4] to perform significance tests between the CFACT and CORR groups, using the largest width and the largest height, as well as the largest width on the filter-range layer (filter-range width) and the filter-variable layer (filter-variable width), for each tree (see Figure 5). We found a significant difference for tree depth ($t = -4.34, p = .0003$), with a 2.5 mean difference, indicating that users in CFACT had deeper search trees. For tree width we found a significant difference for the filter-range layer ($t = -3.75, p = .001$), with a 2.2 mean difference, indicating that users in CFACT had wider filter-range layer widths.

Taken together, these exploratory patterns indicate that counterfactual guidance could lead to more filter range explorations within each filter variable, and visualizations of more filter variable combinations.

6 Discussion

The analysis results indicate that counterfactual guidance was effective for improving the accuracy of finding causal relationships in exploratory analysis, and that counterfactual guidance led to different interaction behaviors compared to correlation-based guidance. Here we discuss how our results relate to prior work, identify limitations of this work, and point out future research directions.

6.1 Reflection on Prior Work

Our results indicate the effectiveness of using counterfactuals as guidance in exploratory visual analysis workflows. Here we discuss how our results reflect on insights from prior work.

6.1.1 Counterfactual Visualization

The main insights of general-purpose counterfactual visualizations were reported by Kaul et al. [35] and Wang et al. [54], which both demonstrate the potential utility of counterfactuals in fostering a deeper understanding of causal relationships within datasets. Our method, even though largely simplifying the counterfactual information in the interface, also found that using counterfactual information as guidance

can benefit data exploration. One of the reported limitations of these studies was increased analysis time and complexity. Our counterfactual guidance method was able to address these limitations while retaining the benefits of counterfactuals for analyzing causal relationships.

However, our results are inconsistent with their findings in terms of confidence. Kaul et al. [35] reported counterfactual visualizations would reduce or confirm users' confidence in different filter choices in analytics systems and Wang et al. [54] also found counterfactuals would impact confidence in static charts. We anticipate that these contradictions may be a result of our simplification of counterfactual information, as shown in Figure 2. But further work should better examine this assumption.

6.1.2 Exploratory Behaviors

Our results indicate that when exploring data with counterfactual guidance, users' interactions are more likely to construct deeper search trees. This insight aligns with Battle and Heer [4], where they found exploratory sessions are more likely to be depth-oriented (using Tableau).

Moreover, we also found correlation-based guidance leads to a relatively low overall causal inference accuracy. This result aligns with Zgraggen et al. [65], however, Battle and Heer [4] found that users performed very well in their tasks. They explain these differences as being due to recruiting more experienced users for their datasets (i.e., from professional Tableau User Groups). In addition, we propose that these differences may also be due to differences in tasks and datasets, for instance, our task is more open-ended, similar to [65].

6.2 Limitations and Future Directions

This research, while providing valuable insights into the use of counterfactual guidance in visual analytics, acknowledges several limitations that pave the way for future exploration:

6.2.1 Empirical Elements in Counterfactual Guidance Computation

When computing counterfactual guidance, we mitigated the REM subset's impact using the geometric mean ($\sqrt{D_{IN,CF} * D_{IN,REM}}$), and added it to the impact of the CF subset ($D_{IN,CF}$) in our counterfactual guidance technique. This formulation was determined empirically. Future work should aim to provide analytical evidence to support these weight choices or explore adaptive weighting schemes to better emphasize the significance of the impacts of the two subsets. In addition, for the subset size and subset distribution score, we provided empirical suggestions to users, but did not perform any statistical analysis to measure at which scale the subset size and its distribution score may imply a significant impact for the effectiveness of the guidance. More statistical analysis and usage suggestions for empirical thresholds related to the data subsets should be thoroughly examined in future work.

6.2.2 Automated Computing Counterfactual Filtering

The counterfactual guidance technique does not incorporate automated methods for determining filter ranges for computing counterfactual subsets, therefore we employed expert-designed filters as the default for both counterfactual and correlation guidance. This limits the potential usage of guidance as an overview of the dataset prior to any user interactions. Future improvements should integrate automated approaches to suggest default filter ranges, such as machine learning models, which may enhance the guidance's ability to create meaningful counterfactual subsets automatically, leading to more efficient data exploration.

6.2.3 Synthetic Data

The synthetic data used cannot accurately reflect the complex causal relationships as well as noise existing in real-world datasets [19]. The noise in these datasets may make counterfactual analysis more challenging. Future studies should investigate the applicability of our guidance technique to more real-world and complex datasets. Further, other distance measures such as Mahalanobis distance [17] should be explored in the future to mitigate the impact of noise. Further, real-world data is often influenced by numerous factors beyond immediate causal relations between two variables. These factors, such as confounders and colliders, can introduce significant biases and distortions into users' analyses, potentially leading to incorrect conclusions [26]. However, the synthetic data did not account for confounders and colliders, limiting its ability to simulate real-world cases. Addressing these elements is essential for a more comprehensive evaluation of the counterfactual guidance method. In the future, we plan to incorporate more causal structure techniques to better identify confounders and colliders, such as causal diagrams, statistical testing, and domain knowledge-assisted identification [47]. Furthermore, our study did not test unbalanced distributions or out-of-distribution cases between the employed subsets. Investigating these scenarios is important for understanding the robustness of counterfactual guidance.

6.2.4 Task Design

Our study tested only open-ended tasks even though it required users to explore many variables. Providing more tasks and designing them with a more focused scope (e.g., "*What relationships do you observe involving weather conditions and strike frequency?* [4]") may reveal different exploratory insights. At the same time, experimenting with questions that are even more open-ended, such as inferential tasks, may better examine users' abilities to reason from data [52]. Future research should explore a wider range of exploration tasks designed to be both more focused and more open-ended.

6.2.5 Impact of Prior Beliefs

Our study did not consider the impact of users' pre-existing beliefs when selecting variables in the synthetic data, which may impact users' data interpretation. For example, studies have shown that users' prior beliefs can impact their estimations of correlations [63] and judgments of causalities [56] from visualizations. Meanwhile, techniques such as Bayesian-guided inference can be effectively used to assist users' belief updating [37]. Future research should examine how beliefs influence users' exploration process when using counterfactual-based guidance, and investigate potential technologies to assist in evaluating and estimating priors.

6.2.6 Visual Interface

The employed exploratory visual interface did not support sophisticated visualizations or more complex exploratory interactions, such as users' random exploration strategies [4], varying data aggregation levels [62], and different visual encodings [42]. However, these factors may impact users' data understanding and exploratory behaviors. Enhancing the interface to accommodate a wider range of interactions, user customizations, and visualizations could improve user experience and analytical outcomes and may lead to more interesting interaction behavior patterns, which should be explored in future work.

7 Conclusion

This paper presented a novel counterfactual-based guidance technique to support exploratory data analytics. By transforming the subsets generated for counterfactual visualizations into guidance values, our approach addresses the limitations of complexity and time-consuming analysis associated with counterfactual visualizations. Employing counterfactual guidance, we retain the advantages of counterfactual reasoning while mitigating the cognitive load on users. Through the usage scenario analysis and empirical study, our investigation has demonstrated the efficacy of counterfactual guidance for exploratory visual analysis. The empirical evidence shows that counterfactual guidance significantly enhances the accuracy of interpreting causal relationships in datasets, suggesting that counterfactual guidance can serve as a powerful tool in guided visual analytics without overwhelming users. In summary, our research advocates for the integration of counterfactual reasoning into visual analytics systems more widely using counterfactual guidance. This technique promises to facilitate more precise and insightful data exploration, ultimately empowering users to make informed decisions based on robust exploratory analysis and causal inferences.

Acknowledgments

We thank the reviewers for their insightful comments. This material is based upon work supported by the National Science Foundation under Grant No. 2211845.

References

[1] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902 2

[2] S. Batt, O. R. Harmon, and P. Tomolonis. Learning tableau: A data visualization tool. *The Journal of Economic Education*, 51(3-4):317–328, 2020. doi: 10.2139/ssrn.3438993 1, 3

[3] L. Battle, R. Chang, and M. Stonebraker. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 International Conference on Management of Data*, pp. 1363–1375. ACM, New York, 2016. doi: 10.1145/2882903.2882919 3

[4] L. Battle and J. Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Computer Graphics Forum*, 38(3):145–159, 2019. doi: 10.1111/cgf.13678 5, 8, 9

[5] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 43–52. IEEE, Washington DC, 2014. doi: 10.1109/VAST.2014.7042480 2

[6] D. Borland, A. Z. Wang, and D. Gotz. Using counterfactuals to improve causal inferences from visualizations. *IEEE Computer Graphics and Applications*, 44(1):95–104, 2024. doi: 10.1109/MCG.2023.3338788 1, 2

[7] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, Cambridge, 1999. doi: 10.5555/300679.300826 2

[8] D. Ceneda, N. Andrienko, G. Andrienko, T. Gschwandtner, S. Miksch, N. Piccolotto, T. Schreck, M. Streit, J. Suschnigg, and C. Tominski. Guide me in analysis: A framework for guidance designers. *Computer Graphics Forum*, 39(6):269–288, 2020. doi: 10.1111/cgf.14017 2

[9] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2016. doi: 10.1109/TVCG.2016.2598468 2

[10] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, M. Streit, and C. Tominski. Guidance or no guidance? a decision tree can help. In *EuroVA@ EuroVis*, pp. 19–23. Eurographics, Eindhoven, 2018. doi: 10.2312/eurova.20181107 2

[11] D. Ceneda, T. Gschwandtner, and S. Miksch. A review of guidance approaches in visual data analysis: A multifocal perspective. *Computer Graphics Forum*, 38(3):861–879, 2019. doi: 10.1111/cgf.13730 1, 2

[12] Y. Chan. Biostatistics 104: correlational analysis. *Singapore Med J*, 44(12):614–619, 2003. 2

[13] J. W. Chen, D. R. Maldonado, B. L. Kowalski, K. B. Miecznikowski, C. Kyin, J. A. Gornbein, and B. G. Domb. Best practice guidelines for propensity score methods in medical research: consideration on theory, implementation, and reporting. a review. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 38(2):632–642, 2022. doi: 10.1016/j.arthro.2021.06.037 5

[14] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020. doi: 10.1109/TVCG.2020.3030342 2

[15] W.-H. Cheng and D. Gotz. Context-based page unit recommendation for web-based sensemaking tasks. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pp. 107–116, 2009. doi: 10.1145/1367497.1367662 3

[16] K. Cook, N. Cramer, D. Israel, M. Wolverton, J. Bruce, R. Burtner, and A. Endert. Mixed-initiative visual analytics using task-driven recommendations. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 9–16. IEEE, Washington DC, 2015. doi: 10.1109/VAST.2015.7347625 3

[17] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000. doi: 10.1016/S0169-7439(99)00047-7 9

[18] J. L. Flores-Guerrero, M. A. Grzegorczyk, M. A. Connelly, E. Garcia, G. Navis, R. P. Dullaart, and S. J. Bakker. Mahalanobis distance, a novel statistical proxy of homeostasis loss is longitudinally associated with risk of type 2 diabetes. *EBioMedicine*, 71, 2021. doi: 10.1016/j.ebiom.2021.103550 5

[19] A. Gentzel, D. Garant, and D. Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32, 2019. doi: 10.5555/3454287.3455338 9

[20] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 531–535. ACM, New York, 2020. doi: 10.1145/3377325.3377536 2

[21] D. Gotz and D. Borland. Data-driven healthcare: challenges and opportunities for interactive visualization. *IEEE Computer Graphics and Applications*, 36(3):90–96, 2016. doi: 10.1109/MCG.2016.59 6

[22] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 85–95. ACM, New York, 2016. doi: 10.1145/2856767.2856779 3

[23] D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pp. 315–324. ACM, New York, 2009. doi: 10.1145/1502650.1502695 3

[24] D. Gotz, Z. When, J. Lu, P. Kissa, N. Cao, W. H. Qian, S. X. Liu, and M. X. Zhou. Harvest: an intelligent visual analytic tool for the masses. In *Proceedings of the first International Workshop on Intelligent Visual Interfaces for Text Analysis*, pp. 1–4. ACM, New York, 2010. doi: 10.1145/2002353.2002355 3

[25] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009. doi: 10.1057/ivs.2008.31 1

[26] S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003. doi: 10.1097/01.EDE.0000042804.12056.6C 9

[27] M. Guo, D. Gotz, and Y. Wang. How does imperfect automatic indexing affect semantic search performance? In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pp. 588–596. IEEE, Washington DC, 2023. doi: 10.1109/ICHI57859.2023.00105 3

[28] M. Guo, Z. Zhou, D. Gotz, and Y. Wang. Grafs: Graphical faceted search system to support conceptual understanding in exploratory search. *ACM Transactions on Interactive Intelligent Systems*, 13(2):1–36, 2023. doi: 10.1145/3588319 3

[29] S. Ha, S. Monadjemi, and A. Ottley. Guided by ai: Navigating trust, bias, and data exploration in ai-guided visual analytics. *Computer Graphics Forum*, 43(3):e15108, 2024. doi: 10.1111/cgf.15108 3

[30] J. Hullman and A. Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3):10–1162, 2021. doi: 10.1162/99608F92.3AB8A587 3

[31] S. M. Iacus, G. King, and G. Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012. doi: 10.1093/pan/mpr013 2

[32] R. Jarry, S. Kobayashi, and K. Fukuda. A quantitative causal analysis for network log data. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1437–1442. IEEE, Washington DC, 2021. doi: 10.1109/COMPSAC51774.2021.00213 6

[33] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029. JMLR.org, New York, 2016. doi: 10.5555/3045390.3045708 2

[34] A. Kale, Z. Guo, X. L. Qiao, J. Heer, and J. Hullman. Evm: Incorporating model checking into exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3326516 3

[35] S. Kaul, D. Borland, N. Cao, and D. Gotz. Improving visualization interpretation using counterfactuals. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):998–1008, 2021. doi: 10.1109/TVCG.2021.3114779 2, 3, 4, 5, 6, 8, 9

[36] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, New York, 2008. doi: 10.1007/978-3-540-70956-5_7 2

[37] Y.-S. Kim, P. Kayongo, M. Grunde-McLaughlin, and J. Hullman. Bayesian-assisted inference from visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):989–999, 2020. doi: 10.1109/TVCG.2020.3028984 9

[38] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, and W. Huang. A correlative analysis process in a visual analytics environment. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Washington DC, 2012. doi: 10.1109/VAST.2012.6400491 2

[39] S. L. Morgan and C. Winship. *Counterfactuals and causal inference*. Cambridge University Press, Cambridge, 2015. doi: 10.1017/CBO9781107587991 2

[40] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 2018. doi: 10.1109/TVCG.2018.2865240 3

[41] J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2009. doi: 10.1017/CBO9780511803161 1, 2

[42] G. J. Quadri, A. Z. Wang, Z. Wang, J. Adorno, P. Rosen, and D. A. Szafir. Do you see what i see? a qualitative study eliciting high-level visualization comprehension. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–26. ACM, New York, 2024. doi: 10.1145/3613904.3642813 9

[43] E. S. Ristad and P. N. Yianilos. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998. doi: 10.1109/34.682181 7

[44] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim. Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE transactions on visualization and computer graphics*, 24(1):120–130, 2017. doi: 10.1109/TVCG.2017.2744805 3

[45] A. Sharma and E. Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020. doi: 10.48550/arXiv.2011.04216 6

[46] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343. IEEE, Washington DC, 1996. doi: 10.1109/VL.1996.545307 2

[47] A. Sjölander and J. Zetterqvist. Confounders, mediators, or colliders: what types of shared covariates does a sibling comparison design control for? *Epidemiology*, 28(4):540–547, 2017. doi: 10.1097/EDE.0000000000000649 9

[48] F. Sperrle, D. Ceneda, and M. El-Assady. Lotse: A practical framework for guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1124–1134, 2022. doi: 10.1109/TVCG.2022.3209393 2

[49] C. Stoiber, D. Ceneda, M. Wagner, V. Schetinger, T. Gschwandtner, M. Streit, S. Miksch, and W. Aigner. Perspectives of visualization onboarding and guidance in va. *Visual Informatics*, 6(1):68–83, 2022. doi: 10.1016/j.visinf.2022.02.005 2

[50] C. D. Stolper, A. Perer, and D. Gotz. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, 2014. doi: 10.1109/TVCG.2014.2346574 2

[51] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010. doi: 10.1214/09-STS313 2

[52] A. Suh, A. Mosca, S. Robinson, Q. Pham, D. Cashman, A. Ottley, and R. Chang. Inferential Tasks as an Evaluation Technique for Visualization. In *EuroVis 2022 - Short Papers*. Eurographics, Eindhoven, 2022. doi: 10.2312/evs.20221086 9

[53] J. W. Tukey et al. *Exploratory Data Analysis*, vol. 2. Addison-Wesley, Reading, 1977. doi: 10.1007/978-3-031-20719-8_2 1

[54] A. Z. Wang, D. Borland, and D. Gotz. An empirical study of counterfactual visualization to support visual causal inference. *Information Visualization*, 23(2):197–214, 2024. doi: 10.1177/14738716241229437 2, 3, 5, 8, 9

[55] A. Z. Wang, D. Borland, and D. Gotz. A framework to improve causal inferences from visualizations using counterfactual operators. *Information Visualization*, 2024. doi: 10.1177/14738716241265120 2, 5

[56] A. Z. Wang, D. Borland, T. Peck, W. Wang, and D. Gotz. Causal priors and their influence on judgements of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE VIS 2024)*, 2025. 9

[57] J. Wang, X. Li, C. Li, D. Peng, A. Z. Wang, Y. Gu, X. Lai, H. Zhang, X. Xu, J. Zhou, X. Liu, and C. Wei. AVA: An automated and ai-driven intelligent visual analytics framework. *Visual Informatics*, 2024. doi: 10.1016/j.visinf.2024.06.002 3

[58] Z. Wang, A. Z. Wang, D. Borland, and D. Gotz. Causalsynth: An interactive web application for synthetic dataset generation and visualization with user-defined causal relationships. In *IEEE VIS Posters*. 2024. 6

[59] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2019. doi: 10.1109/TVCG.2019.2934619 2

[60] L. Wilkinson. Systat. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):256–257, 2010. doi: 10.1002/wics.66 1, 2, 3

[61] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, (1):1–1, 2016. doi: 10.1109/TVCG.2015.2467191 3

[62] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2019. doi: 10.1109/TVCG.2019.2934399 9

[63] C. Xiong, C. Stokes, Y.-S. Kim, and S. Franconeri. Seeing what you believe or believing what you see? belief biases correlation estimation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):493–503, 2022. doi: 10.1109/TVCG.2022.3209405 9

[64] C.-H. E. Yen, A. Parameswaran, and W.-T. Fu. An exploratory user study of visual causality analysis. *Computer Graphics Forum*, 38(3):173–184, 2019. doi: 10.1111/cgf.13680 5, 8

[65] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 2018. doi: 10.1145/3173574.3174053 9

[66] Z. Zhou, W. Wang, M. Guo, Y. Wang, and D. Gotz. A design space for surfacing content recommendations in visual analytic platforms. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):84–94, 2022. doi: 10.1109/TVCG.2022.3209445 3

[67] Z. Zhou, X. Wen, Y. Wang, and D. Gotz. Modeling and leveraging analytic focus during exploratory visual analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, New York, 2021. doi: 10.1145/3411764.3445674 3