

CauTab: Counterfactual Tabular Data Synthesis with Functional and Causal-Dependency Constraints

Xingyu Yang^{1,2*}, Mingchuan Ma^{1,2*}, Dexuan Xu^{1,3}, Xiangyang Long^{1,2}, Yu Huang^{1,2,3†}

¹ National Engineering Research Center for Software Engineering, Peking University

² School of Software & Microelectronics, Peking University

³ School of Computer Science, Peking University

Abstract

High-quality tabular data is essential for data-driven tasks, yet often limited by scarcity and privacy concerns. While synthetic data generated by recent Deep Generative Models and LLMs improves fidelity, existing methods struggle to preserve rigorous attribute dependencies. As a result, the generated data may appear statistically realistic but lack logical consistency. To address this challenge, we propose CauTab, a novel counterfactual framework for tabular data synthesis that preserves functional dependencies and causal-dependency structures. CauTab operates in a three-stage pipeline: (1) it employs an LLM-guided mining strategy to efficiently identify functional dependencies; (2) it constructs a causal-dependency graph by integrating data-driven optimization with LLM-based semantic reasoning; and (3) it executes a dependency-constrained generation process via the Abduction-Intervention-Prediction paradigm. By enforcing deterministic constraints and using the learned graph for structured propagation, CauTab improves statistical fidelity while preserving logical consistency. Experiments demonstrate that CauTab outperforms baselines in data utility and fidelity, while maintaining competitive empirical privacy and strong logical consistency.

1 Introduction

Tabular data serves as a cornerstone of structured information and is widely used in many domains. The application of machine learning techniques to tabular data has enabled numerous downstream tasks, ranging from disease classification (Abdullah et al., 2024) and petroleum exploration (Khalifa et al., 2023) to stock market prediction (Mintarya et al., 2023). However, the efficacy of these data-driven approaches critically depends on the availability of large-scale, high-quality datasets. In practice, data collection is often constrained by high

* These authors contributed equally to this work.

† Corresponding author: Yu Huang, hy@pku.edu.cn.

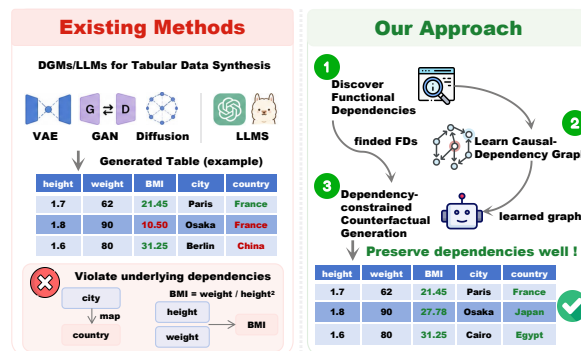


Figure 1: A comparison between our approach and existing methods. Existing methods often fail to preserve underlying functional dependencies in the generated data.

costs and privacy concerns. Consequently, tabular data synthesis has emerged as a promising research direction. Synthetic data can not only mitigate privacy and ethical concerns but also augment limited datasets, thereby improving model robustness and generalization.

Previous studies on tabular data synthesis primarily rely on Deep Generative Models (DGMs), such as Variational Autoencoders (VAEs) (Apelániz et al., 2024), Generative Adversarial Networks (GANs) (Xu et al., 2019), and diffusion models (Kotelnikov et al., 2023). Despite their success, these methods suffer from two major limitations. First, these methods primarily model statistical correlations and often fail to capture the semantic relationships inherent in the data. Second, most DGMs are relatively data-hungry, especially for diffusion-based approaches, which typically require thousands of training samples to achieve stable performance. Such requirements are often impractical in moderate- or low-data regimes.

More recently, Large Language Models (LLMs) have demonstrated strong potential for few-shot and zero-shot tabular data synthesis by leveraging extensive prior knowledge (Liu et al., 2025; Seedat

et al., 2024; Ling et al., 2024). Nevertheless, both DGM-based and LLM-based methods still face a fundamental challenge: *how to preserve the underlying dependencies among tabular attributes in low-data regimes*. In real-world datasets, attributes are rarely independent. For instance, in healthcare data, certain clinical factors are strongly correlated with specific physiological indicators. More critically, some attributes are governed by deterministic logical rules. As illustrated in Figure 1, a city uniquely determines its country, and body mass index (BMI) is a deterministic function of height and weight.

To formally characterize such dependencies, we draw inspiration from causal theory and model attribute relationships using Structural Causal Models (SCMs). Under this framework, each attribute is modeled as a function of its direct causes (parents) and an exogenous noise term. While SCMs typically capture probabilistic dependencies, they naturally accommodate deterministic constraints as a special case where the exogenous noise vanishes. From the perspective of database theory, these deterministic patterns correspond precisely to Functional Dependencies (FDs) (Liu et al., 2010), where the value of one attribute (or a set of attributes) uniquely determines another.

In this paper, we propose **CauTab**, a three-stage counterfactual framework for tabular data synthesis that preserves functional dependencies and causal-dependency structures. First, to identify invariant functional dependencies, CauTab integrates an LLM-guided mining strategy with the equivalence-partition validation algorithm. This design avoids exhaustive dependency search while filtering out accidental rules caused by sparse or low-sample observations. Second, we construct a plausible causal-dependency graph by synergizing data-driven optimization with LLM-based semantic reasoning. We utilize the DAGMA algorithm to infer a preliminary statistical skeleton, which is subsequently refined by the LLM to correct causal orientations and prune spurious correlations based on domain knowledge. Finally, CauTab executes dependency-constrained counterfactual generation. Following a structural causal model, the framework approximates exogenous noise, intervenes on target attributes, and propagates the effects through the learned dependency graph. During propagation, a specialized strategy is applied to enforce constraints for FD-governed attributes.

To sum up, our main contributions can be summarized as follows:

- We propose CauTab, a novel counterfactual framework for tabular data synthesis. By reformulating the generation task through the Abduction-Intervention-Prediction paradigm, our framework not only enables precise control over attribute synthesis but also preserves the underlying dependencies.
- We explicitly model deterministic functional dependencies and probabilistic causal dependencies as two distinct constraints within CauTab. And we propose to synergize statistical learning with LLM-based semantic reasoning to estimate these dependencies.
- Extensive experiments demonstrate the superiority of CauTab. Our method outperforms state-of-the-art baselines in utility and statistical fidelity, while effectively preserving the data privacy and logical consistency.

2 Background

In this section, we first formalize the problem of constrained tabular data synthesis. Subsequently, we introduce the theoretical foundations of functional dependencies and structural causal models.

2.1 Problem Formulation

Consider a tabular training dataset, denoted as $\mathcal{D} = (\mathbf{X}, \mathcal{V})$. Here, $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ represents a set of N instances (or rows), and $\mathcal{V} = \{V_1, \dots, V_d\}$ defines the schema of the data through its d attributes (or columns). We assume the samples are independent and identically distributed (i.i.d.), drawn from an underlying joint probability distribution $P(\mathcal{D})$. In real-world scenarios, tabular data often obeys strict logical rules. We adopt functional dependency from database theory to formally characterize such rules. Specifically, a functional dependency is denoted as $V_L \rightarrow V_R$, where the subset $V_L \subset \mathcal{V}$ is referred to as the Left-Hand Side (LHS) and the dependent attribute $V_R \in \mathcal{V}$ as the Right-Hand Side (RHS). It asserts that for any two tuples $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \mathcal{D}$, if $\mathbf{x}^{(i)}[V_L] = \mathbf{x}^{(j)}[V_L]$, then $\mathbf{x}^{(i)}[V_R] = \mathbf{x}^{(j)}[V_R]$. Let Σ be the set of all valid functional dependencies holding in $\mathcal{D}$.

Our goal is to learn a generative model that produces a synthetic dataset $\mathcal{D}_{syn} = (\mathbf{X}_{syn}, \mathcal{V})$. The generation process must satisfy two key objectives:

- **Statistical Fidelity:** The empirical distribution of the synthetic data $\hat{P}_{syn}$ should be close to the real distribution $\hat{P}_{train}$ (i.e., $\hat{P}_{syn} \approx \hat{P}_{train}$).
- **Logical Consistency:** The synthetic data must respect the deterministic constraints found in the real data. Formally, for every dependency $(V_L \rightarrow V_R) \in \Sigma$, no tuple in $\mathbf{X}_{syn}$ should violate this mapping.

2.2 Functional Dependencies

Functional dependencies (FDs) provide a formal way to describe deterministic relationships among attributes in tabular data. An FD $V_L \rightarrow V_R$ states that the values of determinant attributes V_L uniquely determine the value of a dependent attribute V_R . A standard way to validate an FD is based on equivalence partitions. Let $\pi(V_L)$ denote the partition of tuples induced by identical values on V_L . A candidate dependency $V_L \rightarrow V_R$ holds if adding V_R does not further refine the partition:

$$\pi(V_L) = \pi(V_L \cup \{V_R\}). \quad (1)$$

Classical data-driven FD discovery methods like TANE algorithm (Huhtala et al., 1999) can be used to enumerate such dependencies. However, in low-data regimes, such methods may identify accidental dependencies caused by small sample size, sparse categories, or quasi-identifiers.

2.3 Causal Theory Foundation

While functional dependencies capture deterministic constraints, they fail to model probabilistic relationships. To enable controllable data synthesis, we ground our framework in causal theory.

Structural Causal Model An SCM represents each endogenous variable as a deterministic function of its direct parent variables and an exogenous noise term. Formally, it is defined by $(\mathcal{V}, \mathcal{F}, \epsilon)$, where $\mathcal{V} = \{V_1, \dots, V_d\}$ denotes the observed attributes, $\mathcal{F} = \{f_1, \dots, f_d\}$ is the set of structural functions, and $\epsilon = \{\epsilon_1, \dots, \epsilon_d\}$ captures unobserved factors. Each variable V_i is determined by

$$V_i = f_i(\text{Pa}(V_i), \epsilon_i), \quad (2)$$

where $\text{Pa}(V_i) \subseteq \mathcal{V}$ denotes the direct parent variables of V_i .

Counterfactual Generation. The core idea of counterfactual is presented as a query, such that “Given an observed sample $\mathbf{x}$, what would V_j be if V_i had been set to a ?”. This is computed through a standard three-stage process: (i) *Abduction*: Infer the posterior distribution of exogenous noise ϵ given the observed instances $\mathcal{D}$. (ii) *Intervention*: Modify the structural equations by fixing the value of the target variable (e.g., $V_i \leftarrow a$). (iii) *Prediction*: Propagate the changes through the structural equations f_i using the inferred noise ϵ to compute the counterfactual outcome $\mathbf{x}^*$.

3 Methodology

In this section, we present the proposed CauTab framework. The whole process is shown in Figure 2. Specifically, the workflow unfolds in three sequential phases. First, we identify deterministic relations through *Functional Dependency Discovery*. Next, we construct a plausible causal-dependency graph via *Causal Dependency Discovery*. Finally, based on the learned DAG, we perform *Dependency Constrained Counterfactual Generation* to produce diverse dependency-constrained samples.

3.1 Functional Dependency Discovery

The first step in CauTab is to uncover the functional dependencies present in the training tabular dataset. In the low-data setting, classical algorithms such as TANE (Huhtala et al., 1999) are prone to discovering many accidental dependencies due to data imbalance or sparsity. Instead, we propose an LLM-Guided Two-Stage FD Discovery method, where the LLM first identifies a set of semantically plausible candidates and then verify these candidates on sampled data.

Given attribute metadata and dataset description, denoted as $\mathcal{D}_{des}$, along with few-shot examples $\mathcal{D}_{few}$ drawn from the training data, the prompt $\mathcal{P}_{FD}$ asks the LLM to output candidate rules of the form $V_L \rightarrow V_R$, where $V_L \subseteq \mathcal{V}$ denotes the determinant attributes and $V_R \in \mathcal{V}$ is the dependent attribute. The prompt asks for deterministic mappings such as category-to-code relations or simple formulas, rather than ordinary correlations. Formally, the candidate set is obtained via

$$\mathcal{C}_{FD} = \text{LLM}(\mathcal{P}_{FD}(\mathcal{V}, \mathcal{D}_{des}, \mathcal{D}_{few})). \quad (3)$$

Each candidate $c = (V_L, V_R) \in \mathcal{C}_{FD}$ is then validated using the equivalence-partition criterion in

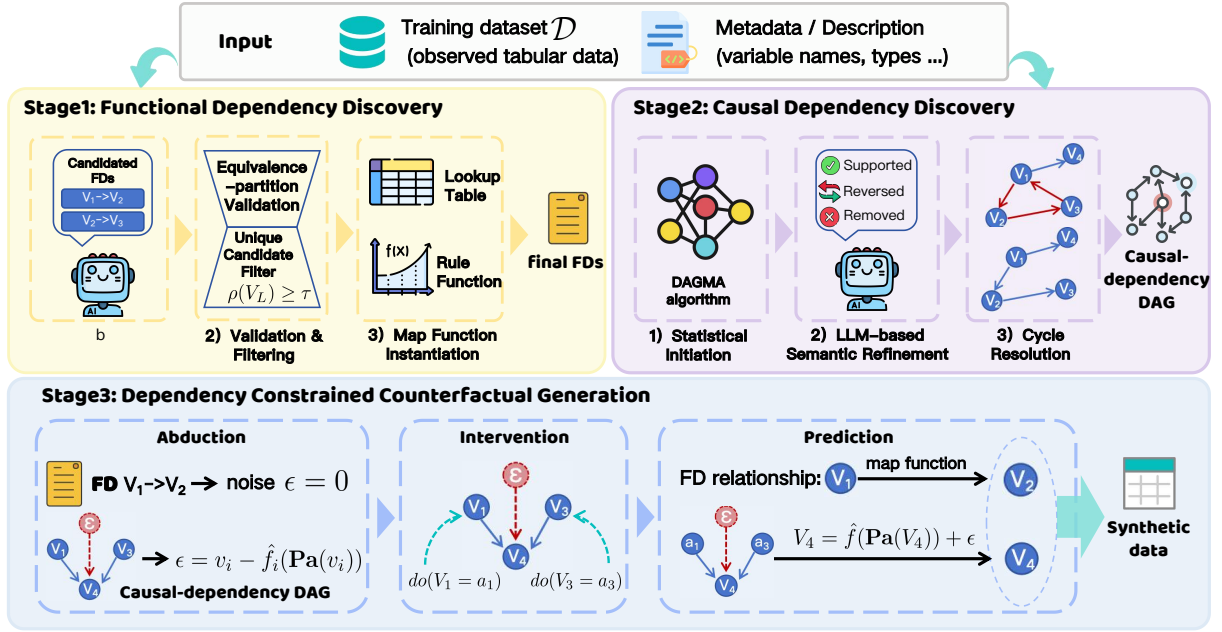


Figure 2: Overview of the proposed CauTab framework.

Eq. 1. Consequently, candidates that satisfy this condition are added to the FD set Σ .

We also discard candidates whose LHS is nearly unique in the sampled table, since such attributes can accidentally determine many other columns without representing a meaningful rule. Specifically, we compute

$$\rho(V_L) = \frac{|\pi(V_L)|}{|\mathcal{D}_{few}|}. \quad (4)$$

If $\rho(V_L) \geq \tau$, the candidate is discarded unless it is supported by an explicit deterministic formula. The final FD set Σ is then provided to the counterfactual generator. For each validated dependency $V_L \rightarrow V_R$, we instantiate a deterministic map $g_R : \text{Dom}(V_L) \rightarrow \text{Dom}(V_R)$ from the training tuples. Specifically, g_R is either extracted as a lookup table for discrete categorical mappings, or formalized as an explicit rule function for arithmetic and logical transformations.

3.2 Causal Dependency Discovery

In this section, we aim to estimate a causal-dependency structure among attributes $\mathcal{V}$. We represent attribute dependencies as a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each edge $V_i \rightarrow V_j$ denotes a plausible directed causal dependency used for propagation. Since reliable causal discovery from observational tabular data is challenging in low-data regimes, inspired by (Du et al., 2025), we combine data-driven structure learning with semantic refinement to learn

the causal-dependency graph. The discovery procedure consists of two stages, namely statistical structure initialization and LLM-assisted semantic refinement with cycle removal.

Statistical structure initialization We first learn a preliminary statistical dependency graph from observational data using the DAGMA algorithm (Bello et al., 2022). Given the training matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, DAGMA estimates a weighted adjacency matrix $W \in \mathbb{R}^{d \times d}$ by minimizing a reconstruction loss with sparsity regularization under a differentiable acyclicity constraint:

$$\min_W \frac{1}{2N} \|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda_1 \|W\|_1 \quad (5)$$

s.t. $h_s(W) = 0$,

where $h_s(W) = -\log \det(sI - W \odot W) + d \log s$ denotes the DAGMA acyclicity characterization. After thresholding the learned weights, we obtain a preliminary adjacency matrix A_{stat} . This graph is not directly used as the final structure; instead, it provides data-grounded statistical hints for the following semantic refinement stage.

Semantic refinement with cycle removal. The preliminary graph may contain spurious edges or implausible directions, especially when the training data is limited. Therefore, we use the LLM to refine the causal ordering suggested by A_{stat} . For each target variable V_i , the prompt $\mathcal{P}_{\text{CD}}$ provides

variable descriptions together with the DAGMA-discovered parents and children of V_i . The LLM is asked to examine these statistically suggested relations and classify them as semantically supported, reversed and removed, such that

$$\mathcal{R}_i = \text{LLM}(\mathcal{P}_{\text{CD}}(V_i, \mathcal{D}_{\text{des}}, A_{\text{stat}})), \quad (6)$$

where $\mathcal{R}_i$ denotes the refinement decisions for the local neighborhood of V_i . Supported relations are retained, reversed relations are re-oriented, removed relations are pruned. This design anchors the final graph in the data-driven DAGMA result while using semantic knowledge to correct locally implausible relations.

Since refinements are performed independently for different variables, aggregating them may introduce directed cycles. To resolve such conflicts, we infer a topological ordering. Specifically, we assign each variable a score

$$S(V_i) = \sum_{j \neq i} \mathbb{I}(d_{ij} \Rightarrow) - \sum_{k \neq i} \mathbb{I}(d_{ki} \Rightarrow), \quad (7)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This scoring function encourages variables with more outgoing than incoming causal relations to appear earlier in the ordering. Edges that contradict the topological ordering induced by $S(\cdot)$ are pruned. The remaining valid relations form the final edge set $\mathcal{E}$.

3.3 Dependency Constrained Counterfactual Generation

Given the learned DAG $\mathcal{G}$ from Section 3.2, CauTab synthesizes new records through the Pearlian Abduction-Intervention-Prediction procedure (Galles and Pearl, 1998). For each factual record, we first infer record-specific exogenous noise, then intervene on selected root variables, and finally propagate the changes along the topological order under functional-dependency constraints.

Abduction. For a factual record, CauTab abduces the exogenous variables that explain its observed attribute values under the learned structural mechanisms. We first separate deterministic FD-governed variables from stochastic variables. If V_i is determined by a functional dependency, its value is fully specified by its determinant attributes, and no stochastic noise is introduced:

$$\forall \mathbf{Pa}(V_i) \rightarrow V_i \in \Sigma : V_i = f_i(\mathbf{Pa}(V_i)), \epsilon_i = 0. \quad (8)$$

For the remaining stochastic variables, we consider numerical and categorical attributes separately. Similar to previous work (Hoyer et al., 2008), we assume an additive noise model for numerical attributes, such that

$$V_i = f_i(\mathbf{Pa}(V_i)) + \epsilon_i, \quad (9)$$

where ϵ_i is independent of its parent variables $\mathbf{Pa}(V_i)$. We approximate f_i with a predictive model $\hat{f}_i$ and recover the record-specific residual as

$$\epsilon_i = v_i - \hat{f}_i(\mathbf{Pa}(v_i)), \quad (10)$$

where v_i and $\mathbf{Pa}(v_i)$ denote the factual values.

For categorical attributes, an additive scalar residual is not well-defined. We therefore model the categorical structural equation with the Gumbel-Max mechanism (Maddison et al., 2014). Let $\mathcal{C}_i$ denote the category set of V_i . The structural assignment is defined as

$$V_i = \arg \max_{c \in \mathcal{C}_i} [\log P(V_i = c \mid \mathbf{Pa}(V_i)) + \epsilon_{i,c}], \quad (11)$$

where $\epsilon_i = \{\epsilon_{i,c}\}_{c \in \mathcal{C}_i}$ is a record-specific vector of standard Gumbel noise. During abduction, given the factual category x_i and factual parent values $\mathbf{x}_{\mathbf{Pa}(V_i)}$, CauTab retains a Gumbel noise vector consistent with the observed class.

Intervention. In the intervention phase, we assign a fixed value a_i to each selected variable V_i , where the value is sampled from the empirical distribution within the training data. To maintain consistency with the DAG, we restrict the intervention candidate set to exogenous root variables in the learned graph. Moreover, we exclude all variables which belong to the RHS of any functional dependency from the intervention candidate set. Since the values of these variables are logically determined by their parent nodes, we treat them as dependent consequences rather than intervention targets.

Prediction. Finally, in the prediction stage, we propagate the intervention effects through the causal DAG $\mathcal{G}$ to compute the counterfactual outcome. Consider a specific observed instance $\mathbf{x} = [x_1, \dots, x_d]$. Our goal is to generate its counterfactual counterpart $\mathbf{x}^* = [x_1^*, \dots, x_d^*]$. To respect the learned DAG, we generate values for each attribute V_i sequentially, following the topological order of $\mathcal{G}$. The value x_i^* is determined as follows.

- If V_i is an intervention variable, we enforce the fixed value $x_i^* = a_i$.

- If V_i is determined by a functional dependency (i.e., $\exists(V_L \rightarrow V_i) \in \Sigma$), its value is directly derived from the already updated determinant values as $x_i^* = g_i(\mathbf{x}_{V_L}^*)$, where g_i is the deterministic map function obtained in Section 3.1.
- For all other variables, we synthesize the counterfactual value using the structural equation model derived from the abduction phase. Concretely, for numerical variables, given the updated parent values $\mathbf{x}_{\text{Pa}(V_i)}^*$, we compute the expected outcome via the learned predictive function $\hat{f}_i$ and add the recovered noise ϵ_i as:

$$x_i^* = \hat{f}_i(\mathbf{x}_{\text{Pa}(V_i)}^*) + \epsilon_i. \quad (12)$$

For categorical variables, we keep the abducted Gumbel noise vector ϵ_i fixed and combine it with the updated interventional probabilities. Let $\mathcal{C}_i$ denote the category set of V_i . The counterfactual class is determined by

$$x_i^* = \arg \max_{c \in \mathcal{C}_i} \left[\log P(V_i = c \mid \mathbf{x}_{\text{Pa}(V_i)}^*) + \epsilon_i \right]. \quad (13)$$

Thus, a categorical value changes only when the intervention shifts the structural probabilities enough to overcome the factual record-specific noise.

4 Experiment

4.1 Experiment Setup

Datasets: We evaluate CauTab on four diverse, real-world tabular datasets, namely Adult¹, Abalone², Student³ and Apple⁴. These datasets cover sociology, biology, education, and agriculture. Notably, Student and Apple are recent 2024 datasets, helping assess limited pre-training exposure. Detailed specifications for each dataset are provided in Appendix A.

Baselines: We compare our method against seven baselines from three categories, namely (1) Conventional methods, including ARF (Watson et al., 2023), DECAF (Van Breugel et al., 2021); (2) DGM-based methods, including TVAE (Xu et al., 2019), CTGAN (Xu et al., 2019), TabDDPM (Kotelnikov et al., 2023); and (3) LLM-based

methods, including GReaT (Borisov et al., 2022) and CLLM (Seedat et al., 2024). The implementation details are provided in Appendix B.2.

Evaluation Metrics: Similar to previous works (Liu et al., 2025; Ling et al., 2024; Jiang et al., 2025; Yang et al., 2025; Umesh et al., 2025a), we evaluate synthetic data by utility, fidelity, privacy, and logical consistency. **Utility** is measured by the TSTR strategy (Jiang et al., 2025; Ling et al., 2024), where downstream models are trained on synthetic data and tested on real test data. We report weighted F1 for classification tasks and R-squared (R^2) for regression tasks. **Fidelity** is assessed by the Distance to Closest Record (DCR), $\Delta(x, D) = \min_{x' \in D} \|x - x'\|_1$. We compare DCR distributions computed against both training and test sets, where similar distributions indicate that the generator captures general data manifold rather than only matching the training records. **Privacy** is evaluated by a nearest-neighbor membership inference attack (MIA), which tests whether real training records can be distinguished from held-out records based on their proximity to the synthetic data. Attack performance close to random guessing (0.5) indicates weaker membership leakage. **Logical consistency** is measured by the violation rate of dataset-specified deterministic rules.

Experiment Settings: We employ *gpt-4o-mini* at a temperature of 0.9 for all LLM-based components. To evaluate performance under moderate- and low-data regimes, we generate 1,000 synthetic samples from training sets of varying sizes ($N \in \{50, 100, 200\}$). The downstream utility is evaluated using a Random Forest model. To ensure robustness, all experiments are repeated across 10 random seeds, and results are reported as mean $\pm$ standard deviation. Due to space constraints, we report the main results in the following sections and provide complete results across additional sample sizes and evaluation settings in Appendix D.

4.2 Main Results

Utility. Table 1 reports downstream TSTR performance across four datasets and three training sizes ($N \in \{50, 100, 200\}$). CauTab obtains the best average rank of 1.08 and ranks first in 11 of 12 settings, with the only exception being Apple at $N = 200$, where TabDDPM is slightly better. The gains are especially clear in low-data settings and on the 2024 Student and Apple datasets, suggesting that dependency-constrained counterfactual gener-

¹<https://archive.ics.uci.edu/ml/datasets/adult>

²<https://archive.ics.uci.edu/dataset/1/abalone>

³<https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset>

⁴<https://www.kaggle.com/datasets/nelgiriwethana/apple-quality>

Table 1: Main utility results on four benchmark datasets. We report weighted F1(%) for classification tasks and R^2 for regression tasks. The background colors distinguish the extent of LLM pre-training exposure, where purple rows indicate seen datasets, while pink rows denote unseen datasets. We vary the sample sizes ($N \in \{50, 100, 200\}$). The best result in each row is shown in **bold**, and the second-best is underlined.

Datasets	Conventional		Deep Generative Models			LLM-based		Ours
	ARF	DECAF	TVAE	CTGAN	TabDDPM	GReaT	CLLM	CauTab
Adult ($n = 50$)	72.17 $\pm$ 2.43	50.32 $\pm$ 21.00	69.69 $\pm$ 6.35	69.69 $\pm$ 5.94	70.64 $\pm$ 3.20	59.84 $\pm$ 5.54	<u>76.29$\pm$0.48</u>	78.17$\pm$1.05
Abalone ($n = 50$)	0.28 $\pm$ 0.03	-0.27 $\pm$ 0.38	0.23 $\pm$ 0.11	0.21 $\pm$ 0.16	-2.05 $\pm$ 0.61	<u>0.31$\pm$0.04</u>	0.17 $\pm$ 0.04	0.34$\pm$0.02
Apple ($n = 50$)	60.13 $\pm$ 3.18	37.00 $\pm$ 6.95	57.74 $\pm$ 3.76	55.48 $\pm$ 5.58	42.77 $\pm$ 6.79	52.76 $\pm$ 5.88	48.31 $\pm$ 3.67	68.64$\pm$0.54
Student ($n = 50$)	55.09 $\pm$ 5.02	27.68 $\pm$ 11.23	50.62 $\pm$ 4.13	46.20 $\pm$ 10.91	37.09 $\pm$ 2.28	76.05 $\pm$ 4.04	<u>79.71$\pm$2.50</u>	87.40$\pm$0.95
Adult ($n = 100$)	72.80 $\pm$ 1.28	49.09 $\pm$ 19.91	71.34 $\pm$ 4.11	70.96 $\pm$ 4.37	63.42 $\pm$ 11.94	59.61 $\pm$ 8.01	<u>75.34$\pm$1.32</u>	78.50$\pm$0.63
Abalone ($n = 100$)	0.27 $\pm$ 0.04	-0.45 $\pm$ 0.62	0.26 $\pm$ 0.11	0.17 $\pm$ 0.15	-3.00 $\pm$ 1.57	<u>0.33$\pm$0.04</u>	0.24 $\pm$ 0.00	0.39$\pm$0.04
Apple ($n = 100$)	62.17 $\pm$ 2.51	40.39 $\pm$ 7.79	62.83 $\pm$ 3.75	57.07 $\pm$ 8.43	63.14 $\pm$ 9.10	54.46 $\pm$ 3.58	47.49 $\pm$ 4.71	69.66$\pm$2.43
Student ($n = 100$)	62.12 $\pm$ 3.26	32.43 $\pm$ 2.41	53.50 $\pm$ 6.41	54.12 $\pm$ 5.31	36.79 $\pm$ 7.82	79.32 $\pm$ 2.49	<u>80.94$\pm$2.83</u>	88.78$\pm$0.69
Adult ($n = 200$)	74.60 $\pm$ 1.83	53.17 $\pm$ 8.09	73.90 $\pm$ 3.07	75.12 $\pm$ 1.43	65.98 $\pm$ 7.36	67.55 $\pm$ 4.60	<u>76.54$\pm$0.77</u>	78.73$\pm$0.53
Abalone ($n = 200$)	0.35 $\pm$ 0.03	-0.39 $\pm$ 0.50	0.26 $\pm$ 0.08	0.06 $\pm$ 0.47	-4.62 $\pm$ 3.90	<u>0.36$\pm$0.04</u>	0.31 $\pm$ 0.02	0.44$\pm$0.00
Apple ($n = 200$)	66.72 $\pm$ 2.29	44.93 $\pm$ 9.59	66.57 $\pm$ 2.03	59.45 $\pm$ 6.27	72.13$\pm$1.56	63.83 $\pm$ 2.29	53.69 $\pm$ 3.47	71.05 $\pm$ 0.38
Student ($n = 200$)	67.15 $\pm$ 1.92	33.32 $\pm$ 9.15	58.76 $\pm$ 6.41	60.59 $\pm$ 3.02	56.57 $\pm$ 13.91	85.68 $\pm$ 1.80	82.04 $\pm$ 0.78	90.08$\pm$0.84
Avg Rank	<u>3.33</u>	7.75	4.54	5.12	6.00	4.17	4.00	1.08

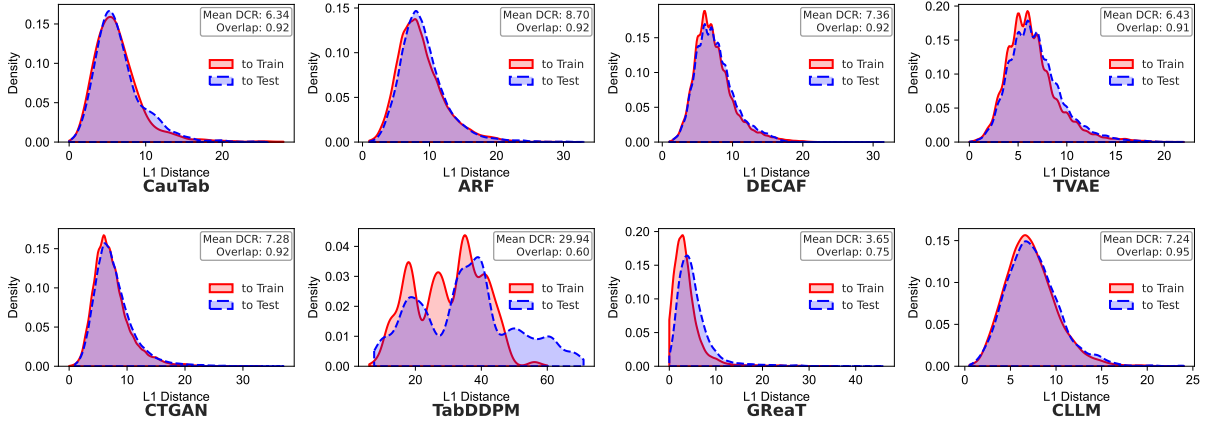


Figure 3: Comparison of DCR distributions on the Adult dataset. Red-shaded areas and blue-dashed lines denote distances to the training and test sets, respectively. High fidelity is indicated by the high overlap between the two distributions, showing that the model captures the general data manifold. High privacy is reflected by the distance values to the training set; as a distribution shifted from zero suggests that the model avoids memorizing samples.

ation improves utility even when LLM pre-training exposure is limited.

Fidelity. Figure 3 compares DCR distributions on Adult at $N = 200$. CauTab achieves a moderate train DCR of 6.34, a small train–test gap of 0.18, and a high overlap of 0.92, indicating that its samples remain close to the real data manifold without being disproportionately closer to training records. By contrast, GReaT has a much lower train DCR and lower overlap, suggesting a stronger memorization tendency, while TabDDPM shows anomalously large DCR values and a gap of 5.32, indicating off-manifold drift.

Privacy. Table 2 reports a nearest-neighbor membership inference attack at $N = 200$. CauTab

obtains the lowest attack AUC on Adult and Apple, and remains within 0.021 of random guessing on Abalone. On Student, CauTab is less private than DECAF and CLLM, likely because enforcing the functional dependency makes some records easier to distinguish. Overall, CauTab shows low empirical leakage on most datasets.

Logic Consistency. We evaluate whether the generated records satisfy explicit schema-level rules. There exist two rules explicitly encoded in the original datasets. Namely, the Adult dataset contains a one-to-one code mapping between Education and Education_Num, and the Student dataset defines GradeClass by GPA ranges. The evaluation results are presented in Table 3. We

Table 2: Nearest-neighbor membership inference attack results at $N = 200$. We report effective attack AUC, defined as $\max(\text{AUC}, 1 - \text{AUC})$, where 0.5 indicates random guessing (weakest membership leakage). The best result in each row is shown in **bold**, and the second-best is underlined.

Dataset	Conventional		Deep Generative Models			LLM-based		Ours
	ARF	DECAF	TVAE	CTGAN	TabDDPM	GReaT	CLLM	CauTab
Adult	0.564±0.035	0.542±0.029	0.552±0.027	0.541±0.029	0.535±0.028	0.648±0.043	0.519±0.018	0.512±0.009
Abalone	0.550±0.019	<u>0.515±0.011</u>	0.520±0.017	0.526±0.020	0.515±0.008	0.651±0.025	0.531±0.017	0.521±0.010
Apple	<u>0.520±0.018</u>	0.521±0.021	0.526±0.018	0.522±0.019	0.521±0.015	0.569±0.025	0.525±0.008	0.516±0.009
Student	0.621±0.013	0.533±0.021	0.586±0.027	0.563±0.026	0.571±0.042	0.665±0.016	<u>0.536±0.016</u>	0.580±0.008

Table 3: Violation rates (%) of explicit deterministic rules.

Method	Adult	Student
ARF	× (82.3)	× (47.9)
DECAF	× (76.7)	× (62.1)
TVAE	× (75.1)	× (46.8)
CTGAN	× (75.0)	× (48.5)
TabDDPM	× (97.1)	× (55.1)
GReaT	× (4.4)	× (12.3)
CLLM	× (1.7)	× (13.0)
CauTab	✓ (0.0)	✓ (0.0)

Table 4: FD ablation study in Adult and Student datasets with $N = 200$. We report RF-F1 and FD violation rate (%).

Dataset	CauTab	RF-F1 ↑		CauTab	FD Viol. ↓	
		w/o FD	Data-only FD		w/o FD	Data-only FD
Adult	77.49	77.21	78.93	0.00	2.33	0.00
Student	89.62	85.98	85.08	0.00	11.33	5.93

observe that conventional and DGM baselines frequently violate these constraints, often with violation rates above 75% on Adult. LLM-based methods reduce violations, showing that semantic priors help, but they still cannot guarantee exact rule satisfaction. CauTab achieves 0.0% violation on both rules, showing good performance on preserving inherent functional dependency.

4.3 Ablation Study

To evaluate the role of functional dependencies, we compare CauTab with two variants, where **w/o FD** removes deterministic FD constraints during generation, and **Data-only FD** replaces LLM-guided FD discovery with purely data-driven rules. As shown in Table 4, removing FD constraints introduces clear rule violations on both Adult and Student, whereas CauTab maintains 0.00% violation. Data-only FD works on Adult but fails to fully preserve the Student rule, showing that semantic filtering is important for avoiding incomplete or accidental dependencies. In terms of utility, the differences are smaller on Adult but more visible on Student,

Table 5: Causal graph ablation at $N = 200$. We report RF-F1 (%) for classification datasets and R^2 for Abalone.

Dataset	CauTab	PC	DAGMA-only	LLM-only
Adult	78.73±0.53	78.01±0.51	72.42±1.51	77.85±1.46
Apple	71.05±0.38	67.40±5.24	71.80±3.24	65.39±1.76
Student	90.08±0.84	75.93±4.22	74.96±5.06	87.91±1.05
Abalone	0.44±0.00	0.33±0.02	0.27±0.06	0.39±0.06

where the target attribute is tightly governed by explicit rules.

For the causal graph ablation, we construct three alternative graphs using **PC**, **DAGMA-only**, and **LLM-only discovery**, respectively. Each graph is then used in the same dependency-constrained counterfactual generation pipeline, with the FD constraints kept unchanged. As shown in Table 5, the full CauTab graph achieves the best utility on Adult, Student, and Abalone, and remains competitive on Apple, where DAGMA-only is slightly higher. This indicates that statistical structure learning and semantic refinement provide complementary signals for constructing a useful causal-dependency graph.

5 Related Work

5.1 Causal Discovery

Causal discovery, the process of inferring causal structures from observational data, has traditionally relied on statistical independence tests and score optimization. Standard methodologies predominantly fall into constraint-based algorithms, such as PC and FCI (Spirites et al., 1995), which utilize conditional independence queries, and score-based approaches like GES (Chickering, 2002), which search for structures that maximize goodness-of-fit metrics (e.g., BIC). These methods provide principled statistical foundations, but they can be brittle in low-data settings and do not use semantic information about variables. Recent work has therefore explored using large language models (LLMs) for

causal analysis, either to identify pairwise causal relations (Kıcıman et al., 2023; Yang et al., 2025) or to provide priors and constraints for data-driven discovery (Long et al., 2023; Abdulaal et al., 2023; Ban et al., 2025). However, exclusive reliance on LLMs is contested; studies suggest they may merely retrieve memorized associations rather than perform genuine causal reasoning (Willig et al., 2022), and their performance degrades significantly with increased graph complexity (Jin et al., 2024). Consequently, current state-of-the-art research advocates for a synergistic integration, where LLM-derived semantic knowledge supplements, rather than replaces, the rigorous mathematical framework of traditional causal discovery algorithms.

5.2 Tabular Data Synthesis

General-purpose synthesis. Tabular data synthesis has evolved from interpolation and statistical dependency modeling, such as SMOTE (Chawla et al., 2002) and Copulas (Xue-Kun Song, 2000), to deep generative models. Representative approaches include VAEs and GANs, such as GT-Coder (Zhang et al., 2024) and CTGAN (Xu et al., 2019), CTAB-GAN+ (Zhao et al., 2024), as well as diffusion- and transformer-based models such as TabDDPM (Kotelnikov et al., 2023), TabSyn (Zhang et al., 2023a), and GraDe (Zhang et al., 2025). These methods can capture complex statistical distributions, but they often require sufficient training data and provide limited control over deterministic relations among attributes. LLM-based methods further exploit pretrained semantic knowledge. Fine-tuning-based approaches, such as GReaT (Borisov et al., 2022), TAPTAP (Zhang et al., 2023b), and HARMONIC (Wang et al., 2024), cast tabular synthesis as language modeling, while prompt-based methods such as CLLM (Seedat et al., 2024), EPIC (Kim et al., 2024), TABGEN-ICL (Fang et al., 2025), and ReFine (Jiang et al., 2025) use in-context generation. Despite their flexibility, these models may still violate deterministic logical or functional dependencies.

Constraint-aware synthesis. A growing line of work studies how to enforce or preserve constraints in synthetic tabular data. Stoian et al. (Stoian et al., 2024; Stoian and Giunchiglia, 2025) incorporate background-knowledge constraints into deep generative models, including quantifier-free real linear constraints. CuTS (Vero et al., 2023) sup-

ports customizable generation with declarative statistical and logical specifications. Recent studies further examine the preservation of logical and functional dependencies (Umesh et al., 2025b), hierarchical dependency reconstruction (Umesh et al., 2026), and dependency-graph-guided generation (Dhooghe et al., 2026). Nevertheless, these studies typically address causal or functional dependencies in isolation, and often compromise data utility by rigidly enforcing fixed rules.

6 Conclusion

In this paper, we propose CauTab, a novel counterfactual framework for tabular data generation that addresses the critical challenge of preserving both functional and learned causal dependencies. Through the Abduction-Intervention-Prediction paradigm, CauTab effectively enforces deterministic rules and propagates changes through a learned dependency graph, ensuring that the generated samples are not only statistically high-fidelity but also logically consistent. Extensive experiments across multiple real-world datasets demonstrate that CauTab outperforms state-of-the-art baselines in utility, fidelity, privacy and logical consistency, especially in low-data regimes.

Limitations

CauTab is evaluated on four tabular benchmarks under moderate- and low-data regimes, and broader validation on larger, noisier, and domain-specific datasets remains necessary. The framework also depends on LLM-based semantic judgments and these components may introduce model-specific biases. Besides, the causal graph in CauTab is inferred from observational data and may not represent the unique ground-truth causal structure.

Ethics Statement

This work studies synthetic tabular data generation as a way to reduce data scarcity and privacy risks. Synthetic data can still reflect biases present in the source data or in LLM priors, and generated records should not be treated as inherently unbiased or privacy-preserving without task-specific auditing. We recommend checking downstream fairness, sensitive-attribute behavior, and memorization risk before deployment in high-stakes domains such as healthcare, finance, or education.

Acknowledgment

This paper was supported in part by the National Key R&D Program of China (Oil & Gas Major Project, No. 2025ZD1401505), the Sanya Science and Technology Special Fund (No. 2024KFJX04), the Beijing Natural Science Foundation (No. L257018), and the Beijing Nova Program.

References

- Ahmed Abdulaal, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C Castro, Daniel C Alexander, et al. 2023. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*.
- Ahmed Aizaldeen Abdullah, Ahmed Aldhahab, and Hanaa M Al Abboodi. 2024. Review of eye diseases detection and classification using deep learning techniques. In *BIO Web of Conferences*, volume 97, page 00012. EDP Sciences.
- Patricia A Apellániz, Juan Parras, and Santiago Zazo. 2024. An improved tabular data generator with vae-gmm integration. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1886–1890. IEEE.
- Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen. 2025. Llm-driven causal discovery via harmonized prior. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1943–1960.
- Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. 2022. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239.
- Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.
- Mary Grace Dhoooghe, Murat Kantarcioglu, and Bhavani Thuraisingham. 2026. Moe-t: Dependency graph-gated mixture of experts for tabular generation with functional dependency. In *Proceedings of the Sixteenth ACM Conference on Data and Application Security and Privacy*, pages 58–70.
- Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 543–554.
- Liancheng Fang, Aiwei Liu, Hengrui Zhang, Henry Peng Zou, Weizhi Zhang, and Philip S Yu. 2025. Tabgen-icl: Residual-aware in-context example selection for tabular data generation. *arXiv preprint arXiv:2502.16414*.
- David Galles and Judea Pearl. 1998. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182.
- Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.
- Yka Huhtala, Juha Kärkkäinen, Pasi Porkka, and Hannu Toivonen. 1999. Tane: An efficient algorithm for discovering functional and approximate dependencies. *The computer journal*, 42(2):100–111.
- Mingxuan Jiang, Yongxin Wang, Ziyue Dai, Yicun Liu, Hongyi Nie, Sen Liu, and Hongfeng Chai. 2025. Limited reference, reliable generation: A two-component framework for tabular data generation in low-data regimes. *arXiv preprint arXiv:2509.09960*.
- Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.
- Houdaifa Khalifa, Olusegun Stanley Tomomewo, Uchenna Frank Ndulue, and Badr Eddine Berrehal. 2023. Machine learning-based real-time prediction of formation lithology and tops using drilling parameters with a web app integration. *Eng*, 4(3):2443–2467.
- Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.
- Jinhee Kim, Taesung Kim, and Jaegul Choo. 2024. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models. *Advances in Neural Information Processing Systems*, 37:31504–31542.
- Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pages 17564–17579. PMLR.

- Yaobin Ling, Xiaoqian Jiang, and Yejin Kim. 2024. Mallm-gan: Multi-agent large language model as generative adversarial network for synthesizing tabular data. *arXiv preprint arXiv:2406.10521*.
- Jixue Liu, Jiuyong Li, Chengfei Liu, and Yongfeng Chen. 2010. Discover dependencies from data—a review. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):251–264.
- Siyi Liu, Yujia Zheng, and Yongqi Zhang. 2025. Struct-synth: Leveraging llms for structure-aware tabular data synthesis in low-data regimes. *arXiv preprint arXiv:2508.02601*.
- Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*.
- Chris J Maddison, Daniel Tarlow, and Tom Minka. 2014. A* sampling. *Advances in neural information processing systems*, 27.
- Latrisha N Mintarya, Jeta NM Halim, Callista Angie, Said Achmad, and Aditya Kurniawan. 2023. Machine learning approaches in stock market prediction: A systematic literature review. *Procedia Computer Science*, 216:96–102.
- Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. 2020. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869.
- Nabeel Seedat, Nicolas Huynh, Boris Van Breugel, and Mihaela Van Der Schaar. 2024. Curated llm: synergy of llms and data curation for tabular augmentation in low-data regimes. In *Proceedings of the 41st International Conference on Machine Learning*, pages 44060–44092.
- Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, page 499–506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Mihaela Stoian and Eleonora Giunchiglia. 2025. Beyond the convexity assumption: Realistic tabular data generation under quantifier-free real linear constraints. In *International Conference on Learning Representations*, volume 2025, pages 28812–28841.
- Mihaela CĂ Stoian, Salijona Dyrnishi, Maxime Cordy, Thomas Lukasiewicz, and Eleonora Giunchiglia. 2024. How realistic is your synthetic data? constraining deep generative models for tabular data. *arXiv preprint arXiv:2402.04823*.
- Chaithra Umesh, Kristian Schultz, Manjunath Mahendra, Saptarshi Bej, and Olaf Wolkenhauer. 2025a. Dependency-aware synthetic tabular data generation. *arXiv preprint arXiv:2507.19211*.
- Chaithra Umesh, Kristian Schultz, Manjunath Mahendra, Saptarshi Bej, and Olaf Wolkenhauer. 2025b. Preserving logical and functional dependencies in synthetic tabular data. *Pattern Recognition*, 163:111459.
- Chaithra Umesh, Kristian Seegel-Schultz, Manjunath Mahendra, Saptarshi Bej, and Olaf Wolkenhauer. 2026. Dependency-aware synthetic tabular data generation. *Pattern Recognition*, page 113819.
- Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. 2021. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233.
- Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Cuts: Customizable tabular synthetic data generation. *arXiv preprint arXiv:2307.03577*.
- Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. 2024. Harmonic: Harnessing llms for tabular data synthesis and privacy protection. *Advances in Neural Information Processing Systems*, 37:100196–100212.
- David S Watson, Kristin Blesch, Jan Kapar, and Marvin N Wright. 2023. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 5357–5375. PMLR.
- Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. 2022. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*.
- Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.
- Peter Xue-Kun Song. 2000. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320.
- Shuo Yang, Zheyu Zhang, Bardh Prenkaj, and Gjergji Kasneci. 2025. Doubling your data in minutes: Ultra-fast tabular data generation via llm-induced dependency graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10369.
- Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2023a. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*.
- Rongchao Zhang, Yiwei Lou, Dexuan Xu, Yongzhi Cao, Hanpin Wang, and Yu Huang. 2024. A learnable discrete-prior fusion autoencoder with contrastive learning for tabular data synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16803–16811.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Li Jian, and Qian Liu. 2023b. Generative table pre-training empowers models for tabular prediction. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14836–14854.

Zheyu Zhang, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2025. Not all features deserve attention: Graph-guided dependency learning for tabular data generation with language models. *arXiv preprint arXiv:2507.18504*.

Zilong Zhao, Aditya Kunar, Robert Birke, Hiek Van der Scheer, and Lydia Y Chen. 2024. Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data*, 6:1296508.

Appendix

A Dataset Descriptions

We evaluate our framework on four real-world tabular datasets spanning social science, biology, agriculture, and education. These datasets vary in sample size, feature composition (numerical vs. categorical), and learning objectives (regression, binary, and multi-class classification), providing a diverse benchmark for evaluation. Dataset statistics are summarized in Table 6.

- **Adult.**⁵ A widely used benchmark derived from the 1994 U.S. Census, where the task is to predict whether an individual’s annual income exceeds \$50K based on 14 demographic and employment attributes.
- **Abalone.**⁶ A biological dataset for age estimation, aiming to predict the number of shell rings from physical measurements such as length, diameter, and weight, with one categorical attribute (Sex).
- **Student.**⁷ An educational dataset containing profiles of 2,392 high school students. The task is to classify academic performance into five GPA-based categories, enabling evaluation of multi-class classification.
- **Apple.**⁸ An agricultural dataset with seven continuous attributes describing fruit characteristics (e.g., size, sweetness, and acidity), where the goal is to predict fruit quality.

B Experimental Settings

B.1 Hardware and Software Environment

All experiments were conducted on a Linux-based server. For reproducibility, we fixed the random seed to 42 for model initialization and used seeds ranging from 0 to 9 for training data sampling. The detailed hardware and software configurations are summarized as follows:

- **System:** Ubuntu 20.04 LTS; Intel Xeon Silver 4210 CPU (2.20 GHz); 128 GB RAM.

- **GPU:** $4 \times$ NVIDIA GeForce RTX 4090D (24 GB VRAM each).
- **Software:** Python 3.11, PyTorch 2.1, CUDA 12.1.

B.2 Baseline Implementation Details

We compare CauTab against seven state-of-the-art baselines categorized into Conventional Methods, Deep Generative Models (DGMs), and LLM-based Methods. To ensure a fair comparison, we utilized standardized implementations from widely adopted open-source libraries where available. The specific configurations are given as follows.

- **ARF** (Adversarial Random Forests) (Watson et al., 2023): A method inspired by GANs that implements a recursive procedure where unsupervised random forests gradually learn the data structure through alternating rounds of generation and discrimination. We utilized the default configuration provided by synthcity⁹, specifically setting `num_trees=30`, `min_node_size=5`, `max_iters=10`, and enabling `early_stop=True`.
- **DECAF** (Van Breugel et al., 2021): A GAN-based framework that embeds a Structural Causal Model (SCM) explicitly into the generator’s architecture, allowing variables to be reconstructed based on their causal parents to ensure structural integrity. We adopted the default settings from synthcity, which include training for `n_iter=100` iterations (with `n_iter_baseline=1000` for initialization), a batch size of 200, and a learning rate of 1×10^{-3} . Both the generator and discriminator were configured with 2 hidden layers of 500 units each.
- **TVAE** (Xu et al., 2019): A VAE-based architecture specifically adapted for mixed-type tabular data, utilizing mode-specific normalization to handle multi-modal numerical distributions. We used the standard synthcity implementation with the following default hyperparameters: an encoder/decoder structure with 3 hidden layers of 500 units each, a learning rate of 1×10^{-3} and a batch size of 200.

⁵<https://archive.ics.uci.edu/ml/datasets/adult>

⁶<https://archive.ics.uci.edu/dataset/1/abalone>

⁷<https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset>

⁸<https://www.kaggle.com/datasets/nelgiriyeewithana/apple-quality>

⁹<https://github.com/vanderschaarlab/synthcity>

Table 6: Summary of Datasets Used in Experiments. (# Num: Number of Numerical Features; # Cat: Number of Categorical Features).

Dataset	Domain	Samples	# Num.	# Cat.	Task	Target	Source/Year
Adult	Social Science	32,561	6	8	Binary Class.	Income ($> / \leq 50K$)	UCI (1996)
Abalone	Biology	4,177	7	1	Regression	Rings (Age)	UCI (1995)
Student	Education	2,392	3	9	Multi-class	GradeClass (0-4)	Kaggle(2024)
Apple	Agriculture	4,000	7	0	Binary Class.	Quality (Good/Bad)	Kaggle(2024)

Table 7: Complete list of raw functional dependencies discovered by exhaustive TANE on the Adult dataset. This list is used only as a data-driven reference for comparison.

ID	LHS	RHS	ID	LHS	RHS
1	educational-num	education	38	age, education, fnlwtg, income, marital-status	gender
2	education	educational-num	39	age, education, fnlwtg, marital-status, native-country	gender
3	age, fnlwtg, hours-per-week	race	40	age, education, fnlwtg, marital-status, race	gender
4	age, fnlwtg, relationship	race	41	age, educational-num, fnlwtg, income, marital-status	gender
5	age, education, fnlwtg, hours-per-week	gender	42	age, educational-num, fnlwtg, marital-status, native-country	gender
6	age, education, fnlwtg, occupation	gender	43	age, educational-num, fnlwtg, marital-status, race	gender
7	age, education, fnlwtg, gender	race	44	age, fnlwtg, hours-per-week, income, native-country	gender
8	age, education, fnlwtg, relationship	gender	45	age, fnlwtg, hours-per-week, native-country, workclass	gender
9	age, education, fnlwtg, income	race	46	capital-loss, education, fnlwtg, hours-per-week, occupation	gender
10	age, education, fnlwtg, native-country	race	47	capital-loss, educational-num, fnlwtg, hours-per-week, occupation	gender
11	age, education, fnlwtg, occupation	race	48	education, fnlwtg, hours-per-week, income, occupation	gender
12	age, educational-num, fnlwtg, hours-per-week	gender	49	education, fnlwtg, hours-per-week, occupation, workclass	gender
13	age, educational-num, fnlwtg, occupation	gender	50	education, fnlwtg, marital-status, race, workclass	gender
14	age, educational-num, fnlwtg, gender	race	51	education, fnlwtg, marital-status, relationship, workclass	gender
15	age, educational-num, fnlwtg, relationship	gender	52	education, fnlwtg, occupation, relationship, workclass	gender
16	age, educational-num, fnlwtg, income	race	53	educational-num, fnlwtg, hours-per-week, income, occupation	gender
17	age, educational-num, fnlwtg, native-country	race	54	educational-num, fnlwtg, hours-per-week, occupation, workclass	gender
18	age, educational-num, fnlwtg, occupation	race	55	educational-num, fnlwtg, marital-status, race, workclass	gender
19	age, fnlwtg, hours-per-week, marital-status	gender	56	educational-num, fnlwtg, marital-status, relationship, workclass	gender
20	age, fnlwtg, hours-per-week, occupation	gender	57	educational-num, fnlwtg, occupation, relationship, workclass	gender
21	age, fnlwtg, hours-per-week, relationship	gender	58	fnlwtg, hours-per-week, occupation, relationship, workclass	race
22	age, fnlwtg, gender, marital-status	race	59	age, capital-gain, education, fnlwtg, hours-per-week, occupation	workclass
23	age, fnlwtg, income, marital-status	race	60	age, capital-gain, educational-num, fnlwtg, hours-per-week, occupation	workclass
24	age, fnlwtg, marital-status, native-country	race	61	age, education, fnlwtg, hours-per-week, marital-status, occupation	capital-loss
25	age, fnlwtg, marital-status, occupation	race	62	age, education, fnlwtg, hours-per-week, occupation, relationship	capital-loss
26	age, fnlwtg, native-country, occupation	race	63	age, educational-num, fnlwtg, hours-per-week, marital-status, occupation	capital-loss
27	age, fnlwtg, occupation, workclass	race	64	age, educational-num, fnlwtg, hours-per-week, occupation, relationship	capital-loss
28	education, fnlwtg, hours-per-week, marital-status	gender	65	age, education, fnlwtg, hours-per-week, income, occupation	workclass
29	education, fnlwtg, hours-per-week, relationship	gender	66	age, educational-num, fnlwtg, hours-per-week, income, occupation	workclass
30	education, fnlwtg, hours-per-week, relationship	race	67	education, fnlwtg, income, occupation, race, workclass	gender
31	education, fnlwtg, occupation, relationship	race	68	educational-num, fnlwtg, income, occupation, race, workclass	gender
32	educational-num, fnlwtg, hours-per-week, marital-status	gender	69	age, education, fnlwtg, income, marital-status, occupation, workclass	capital-loss
33	educational-num, fnlwtg, hours-per-week, relationship	gender	70	age, education, fnlwtg, income, occupation, relationship, workclass	capital-loss
34	educational-num, fnlwtg, hours-per-week, relationship	race	71	age, educational-num, fnlwtg, income, marital-status, occupation, workclass	capital-loss
35	educational-num, fnlwtg, occupation, relationship	race	72	age, educational-num, fnlwtg, income, occupation, relationship, workclass	capital-loss
36	fnlwtg, hours-per-week, occupation, relationship	gender	73	capital-loss, fnlwtg, income, marital-status, occupation, relationship, workclass	race
37	age, capital-gain, fnlwtg, hours-per-week, native-country	gender	74	fnlwtg, income, marital-status, native-country, occupation, relationship, workclass	race

- **CTGAN** (Xu et al., 2019): A GAN-based model specifically designed for mixed-type tabular data. It employs mode-specific normalization to capture complex multi-modal continuous distributions and a conditional generator to address class imbalance in discrete columns. We utilized the default configuration from `synthcity`, which consists of a generator and discriminator each with 2 hidden layers of 500 units. The model was trained for 2,000 iterations with a batch size of 200, a learning rate of 1×10^{-3} , and a dropout rate of 0.1 for both networks.
- **TabDDPM** (Kotelnikov et al., 2023): A state-of-the-art diffusion-based model specifically designed to handle the inherent heterogeneity of tabular data (i.e., mixed continuous and discrete features). It adapts the denoising diffusion paradigm to generate high-fidelity samples by reversing a forward noise corruption

process. We utilized the default configuration from `synthcity`, employing an MLP backbone with 3 hidden layers of 256 units. The diffusion process was configured with 1,000 timesteps and a cosine noise scheduler. The model was trained for 1,000 epochs with a batch size of 1,024 and a learning rate of 2×10^{-3} .

- **GReaT** (Borisov et al., 2022): An LLM-based approach that leverages the generative capabilities of auto-regressive transformers to produce realistic tabular data. It excels at modeling heterogeneous feature types and supports arbitrary conditional generation to capture complex data distributions. In our implementation, we followed the official library¹⁰ and utilized `Qwen3-0.3B-distil` as the backbone model. We fine-tuned the model for 100 epochs with a batch size of 8 and a sampling

¹⁰https://github.com/kathrinse/be_great

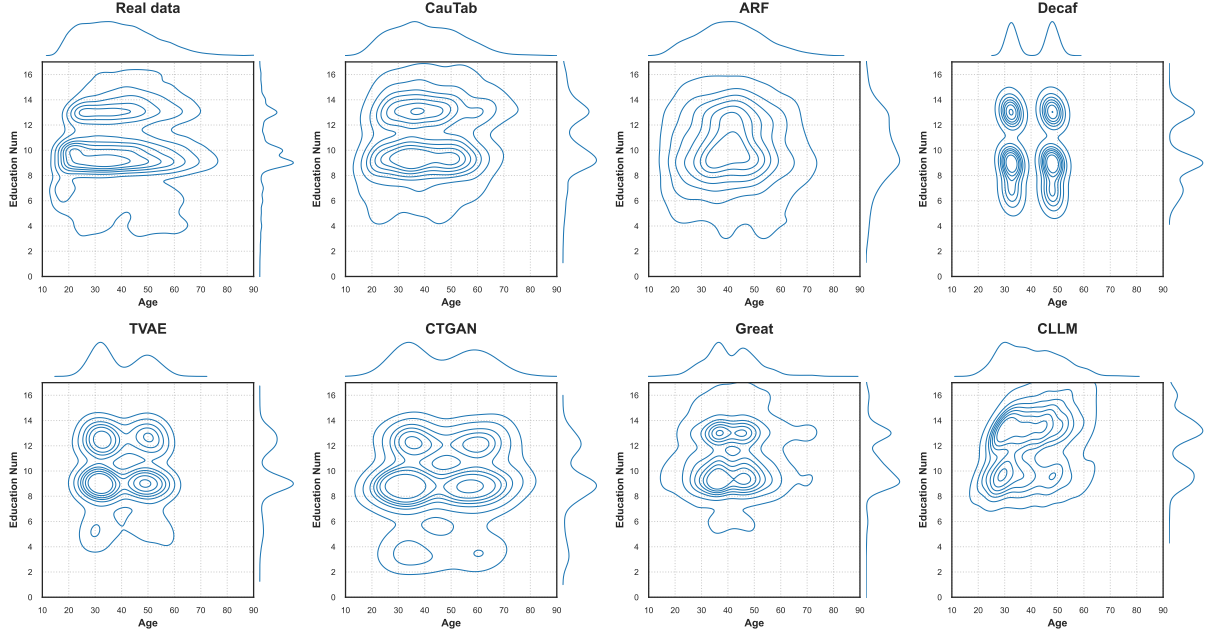


Figure 4: Bivariate density estimation of Age vs. Education Num. CauTab faithfully recovers the multi-modal distribution and marginals of the real data, whereas baselines like Decaf exhibit severe mode collapse.

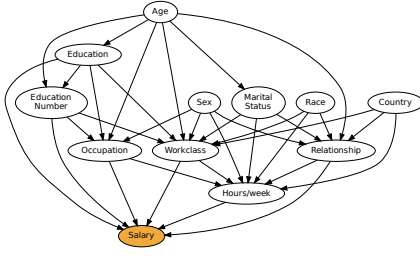


Figure 5: Learned DAG from the real Adult dataset.

temperature of 0.9.

- **CLLM** (Curated LLM) (Seedat et al., 2024): An LLM-based data augmentation method designed for low-data scenarios, which utilizes a specific curation mechanism involving confidence and uncertainty metrics to filter generated samples and enhance the quality of the augmented dataset. The implementation followed the official repository¹¹. Specifically, we incorporated the dataset description into the langchain_templates function, mirroring the schema of existing datasets. Subsequently, we executed the run_llm_generator module with the temperature set to 0.9 and the maximum token limit set to 1000.

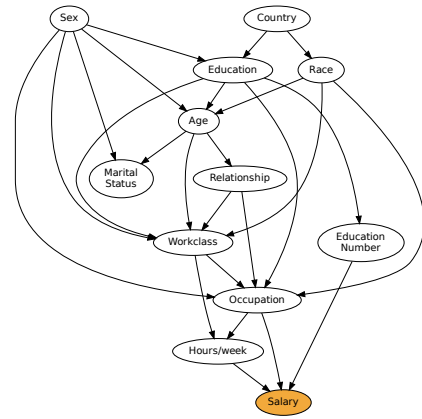


Figure 6: Learned DAG from the synthetic data generated by CauTab.

C Case Study

To provide a comprehensive qualitative analysis of our framework, we conduct a detailed case study using the Adult dataset.

C.1 Discovery of Functional Dependency

CauTab discovers functional dependencies in two steps. It first asks the LLM to nominate semantically plausible deterministic rules, such as code-label mappings or explicit formulas, and then validates these candidates using the equivalence-partition criterion. On the Adult dataset, this

¹¹<https://github.com/seedatnabeel/CLLM>

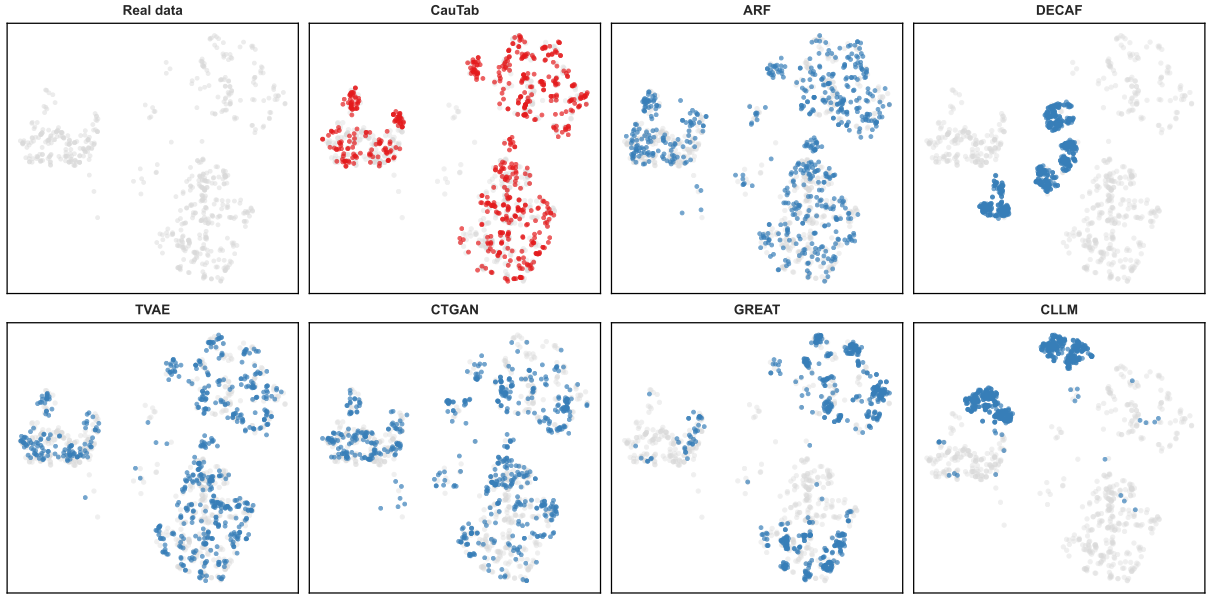


Figure 7: t-SNE visualization of the global data manifold. Gray points represent real data (background), and colored points represent synthetic data. CauTab (red) demonstrates superior coverage and alignment with the real data distribution compared to other baselines.

procedure identifies two meaningful mappings: “education-num $\rightarrow$ education” and “education $\rightarrow$ education-num”.

As a data-only comparison, we also run classical TANE algorithm on the same dataset and report the raw output in Table 7. TANE discovers 74 dependencies, but 72 of them involve `fnlwgt`. Since `fnlwgt` is a high-precision continuous attribute that acts as a quasi-identifier, these dependencies mainly reflect accidental uniqueness rather than meaningful deterministic rules.

We also compare the runtime of the two strategies. TANE algorithm takes approximately 53 seconds to enumerate the full search space. In contrast, the LLM-guided pipeline takes about 10 seconds in total, including 8 seconds for LLM candidate screening and 2 seconds for partition-based validation. Thus, the LLM-first design improves both efficiency and rule quality by narrowing the search space before exact validation.

C.2 Discovery of Causal Dependency

To qualitatively validate our framework’s ability to capture and preserve plausible directed dependency patterns, we applied our causal-dependency discovery module to both the original Adult dataset and the synthetic data generated by CauTab. The resulting DAGs are visualized in Figure 5 (Real) and Figure 6 (Synthetic), respectively.

The real-world structure aligns well with so-

cioeconomic logic, where demographic attributes (e.g., Age, Sex) correctly appear as upstream nodes driving life-stage variables (e.g., Age $\rightarrow$ Marital Status), with Salary positioned as the downstream sink node determined by work-related features. Crucially, by comparing Figure 6 with Figure 5, we find that that CauTab faithfully preserves this hierarchical backbone. Despite minor variations, the synthetic graph successfully reconstructs key dependencies—such as the influence of Education on Occupation. This alignment serves as evidence that CauTab effectively preserves key directed dependency patterns observed in the learned graph.

C.3 Qualitative Analysis Results

To qualitatively evaluate the fidelity of the generated data, we visualize both pairwise feature correlations and the global data manifold. Figure 4 presents the Kernel Density Estimation (KDE) of “Age” versus “Education-Num”, a pair typically exhibiting complex correlations in the Adult dataset. As observed, CauTab closely reproduces the multimodal structure of the real data, faithfully preserving both joint density modes and marginal distributions. In contrast, baseline methods such as DECAF suffer from severe mode collapse, and CLLM fails to capture fine-grained density variations. We further examine the global data geometry using t-SNE projections in Figure 7, where real sam-

Table 8: Constraint-aware generation results aggregated over $N \in \{50, 100, 200\}$. Adult and HFGF/CauTab Case 4 results use seeds 0–2; CuTS Case 4 uses one run per sample size. Values are mean $\pm$ standard deviation across the corresponding runs. Best results within each benchmark are shown in **bold**.

Benchmark	Method	Constraint access	FD Viol. (%) $\downarrow$	Coverage (%) $\uparrow$	Weighted F1 (%) $\uparrow$
Adult	HFGF (TVAE)	Discovered FDs	0.00 $\pm$ 0.00	21.53 $\pm$ 3.30	67.74 $\pm$ 3.11
Adult	CuTS	Supplied constraints	1.20 $\pm$ 0.43	19.44 $\pm$ 2.08	77.40 $\pm$ 2.12
Adult	CauTab	Discovered FDs + graph	0.00 $\pm$ 0.00	59.03 $\pm$ 9.94	75.37 $\pm$ 3.38
HFGF Case 4	HFGF (TVAE)	Predefined hierarchy	0.00 $\pm$ 0.00	77.04 $\pm$ 21.71	42.24 $\pm$ 15.17
HFGF Case 4	CuTS	Supplied constraints	0.34 $\pm$ 0.07	94.26 $\pm$ 3.39	36.75 $\pm$ 5.61
HFGF Case 4	CauTab	Discovered FDs + graph	0.00 $\pm$ 0.00	100.00 $\pm$ 0.00	48.64 $\pm$ 2.47

Table 9: Reference-DAG recovery. Each entry reports SHD / directed-edge F1. The best result for each benchmark is shown in **bold**.

Dataset	LLM-only	DAGMA-only	PC	CauTab
ASIA	9 / 0.640	8 / 0.000	13 / 0.286	0 / 1.000
CANCER	5 / 0.615	4 / 0.000	2 / 0.500	1 / 0.889
ALARM	132 / 0.359	47 / 0.075	N/A	46 / 0.576

Table 10: Agreement between CauTab counterfactuals and oracle counterfactuals from the synthetic SCM at $N = 50$.

Intervention	CF-NRMSE $\downarrow$	Categorical CF Acc. $\uparrow$
do(Age = min(Age + 10, 80))	0.6219	73.60%
do(Height = 175)	0.2898	98.67%
do(Region = East)	0.5014	94.80%

ples are shown in gray as a reference background. CauTab demonstrates the most comprehensive coverage of the real data manifold, indicating strong global fidelity. Conversely, methods such as DE-CAF, TVAE and CLLM exhibit pronounced mode dropping, as shown by large exposed gray areas. These qualitative results confirm that CauTab consistently outperforms state-of-the-art baselines in preserving structural integrity of tabular data.

D Additional Experimental Results

This section reports additional experiments that extend the main evaluation in three directions: reference-graph recovery on ASIA, CANCER, and ALARM; agreement with oracle counterfactuals from a known SCM; and constraint preservation on Adult and HFGF Case 4 with external constraint-aware baselines. We also provide supplementary fidelity, privacy, ablation, and downstream-model robustness analyses that could not be fully reported in the main text due to space constraints.

D.1 Reference-Structure and Counterfactual Validation

Because observational data do not by themselves identify a unique causal DAG, we use the learned graph as a directed dependency structure for generation. The generated records are counterfactuals under the estimated structural model, not verified real-world counterfactuals. We evaluate graph recovery and counterfactual agreement separately in controlled settings with known references.

Reference-DAG recovery. We evaluate graph recovery on ASIA, CANCER, and ALARM, which are three widely used causal-discovery and Bayesian-network benchmarks with published reference DAGs (Ban et al., 2025). They contain 8/8, 5/4, and 37/46 nodes/edges, respectively. All methods receive the same 1,000 observational samples, and the reference DAGs are used only for evaluation. We compare CauTab with LLM-only, DAGMA-only, and PC under identical data and evaluation settings. We report Structural Hamming Distance (SHD; lower is better) and directed-edge F1 (higher is better).

As shown in Table 9, CauTab exactly recovers the ASIA reference DAG and achieves the strongest recovery on CANCER. On ALARM, it reduces SHD by 65.2% relative to LLM-only and improves directed-edge F1 from 0.359 to 0.576. DAGMA-only obtains a similar SHD on ALARM but predicts only seven edges and recovers two directed reference edges, yielding an F1 of 0.075. PC did not converge on ALARM under our implementation. These results show that combining observational evidence with semantic refinement improves recovery under this controlled protocol; they do not establish causal identification on real-world observational data.

Table 11: DCR summary of CauTab on the four benchmark datasets. We report the mean DCR to the real training set, the mean DCR to a size-matched real test reference set, their absolute gap, and the distributional overlap between the two DCR distributions.

Dataset	Train DCR	Test DCR	Gap ↓	Overlap
Adult ($n = 50$)	9.51	9.17	0.34	0.79
Adult ($n = 100$)	8.91	8.49	0.42	0.86
Adult ($n = 200$)	6.34	6.52	0.18	0.92
Abalone ($n = 50$)	0.68	0.78	0.09	0.81
Abalone ($n = 100$)	0.52	0.54	0.02	0.89
Abalone ($n = 200$)	0.30	0.32	0.02	0.88
Apple ($n = 50$)	5.86	5.82	0.04	0.91
Apple ($n = 100$)	5.31	5.16	0.15	0.90
Apple ($n = 200$)	4.55	4.56	0.01	0.92
Student ($n = 50$)	8.85	9.17	0.33	0.91
Student ($n = 100$)	7.69	8.02	0.33	0.91
Student ($n = 200$)	6.64	6.96	0.32	0.90

Oracle counterfactual agreement. Synthetic SCMs are commonly used for graph-recovery and oracle-counterfactual evaluation because their reference structures, mechanisms, and exogenous variables are known (Pawlowski et al., 2020). We construct a 12-variable mixed-type healthcare SCM with a known DAG and structural equations. CauTab is trained on only $N = 50$ observational records without access to the true DAG, structural equations, or exogenous noise. For evaluation, we apply the same intervention to the true SCM while retaining each factual record’s exogenous noise, and evaluate only true descendants of the intervention target. For numerical descendants, CF-NRMSE is the per-variable RMSE normalized by the variable’s factual standard deviation and then averaged across variables. For categorical descendants, we report cell-level counterfactual accuracy.

Table 10 shows that approximation error varies across interventions. The Height and Region interventions yield high categorical agreement, whereas the Age intervention is more difficult because Age is an upstream variable with many descendants, allowing estimation errors to accumulate along longer propagation paths. These results provide controlled evidence of model-based counterfactual approximation under the specified SCM and low-data setting, but do not imply assumption-free counterfactual identification on observational real-world data.

D.2 Constraint Preservation

The main evaluation considers two explicit deterministic rules. We broaden this analysis by com-

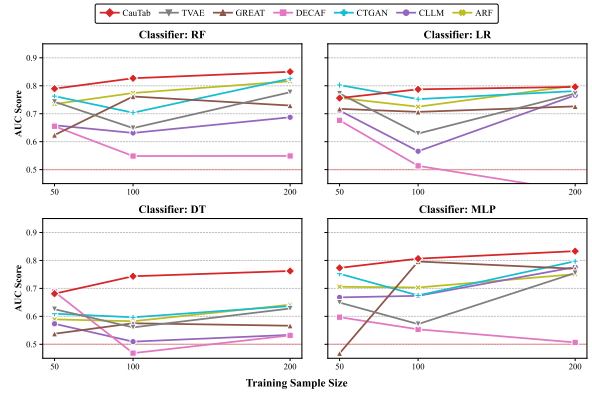


Figure 8: Comparison of utility evaluation using varying downstream classifiers.

paring CauTab with the constraint-aware methods CuTS (Vero et al., 2023) and HFGF (Umesh et al., 2026) on Adult and HFGF Case 4 dataset. Case 4 includes single- and multi-attribute determinants, many-to-one mappings, and multi-step FD chains. CuTS receives declarative constraints supplied by the user, while HFGF receives the predefined hierarchy on Case 4. CauTab instead discovers FDs and learns its directed dependency graph solely from each low-data training split.

We evaluate three aspects. FD Violation is the percentage of evaluated row–constraint pairs that violate the reference deterministic constraints. Coverage is the percentage of reference determinant-value patterns represented in the generated data. Weighted F1 measures downstream utility under the Train-on-Synthetic, Test-on-Real (TSTR) protocol.

Table 8 shows that CauTab achieves zero FD violations and the highest coverage on both benchmarks. On Adult, it reaches 59.03% coverage, compared with 21.53% for HFGF and 19.44% for CuTS, while remaining close to CuTS in utility (75.37% versus 77.40%). On Case 4, CauTab reaches 100% coverage and the highest downstream utility, with a weighted F1 of 48.64%, compared with 42.24% for HFGF and 36.75% for CuTS. CauTab obtains these results without access to the predefined Case 4 hierarchy. This broader evaluation shows that its deterministic enforcement extends beyond the two explicit rules considered in the main experiments and remains effective for more complex dependency structures in low-data settings.

Table 12: Nearest-neighbor membership inference attack results at $N = 50$. We report effective attack AUC, defined as $\max(\text{AUC}, 1 - \text{AUC})$, where 0.5 indicates random guessing and lower values indicate weaker membership leakage. The best result in each row is shown in **bold**, and the second-best is underlined.

Dataset	Conventional		Deep Generative Models			LLM-based		Ours
	ARF	DECAF	TVAE	CTGAN	TabDDPM	GReaT	CLLM	CauTab
Adult	0.592±0.054	0.548±0.027	0.592±0.046	0.582±0.057	<u>0.547±0.060</u>	0.668±0.064	0.538±0.022	0.593±0.050
Abalone	0.588±0.037	0.530±0.024	0.565±0.046	0.567±0.044	<u>0.534±0.029</u>	0.673±0.020	0.613±0.045	0.564±0.012
Apple	0.541±0.030	0.539±0.027	0.546±0.029	0.545±0.031	<u>0.531±0.026</u>	0.631±0.053	0.567±0.075	0.524±0.031
Student	0.713±0.046	<u>0.559±0.034</u>	0.722±0.049	0.650±0.052	0.537±0.026	0.717±0.056	0.585±0.034	0.560±0.028

Table 13: Nearest-neighbor membership inference attack results at $N = 100$. We report effective attack AUC, defined as $\max(\text{AUC}, 1 - \text{AUC})$, where 0.5 indicates random guessing and lower values indicate weaker membership leakage. The best result in each row is shown in **bold**, and the second-best is underlined.

Dataset	Conventional		Deep Generative Models			LLM-based		Ours
	ARF	DECAF	TVAE	CTGAN	TabDDPM	GReaT	CLLM	CauTab
Adult	0.573±0.039	0.545±0.032	0.555±0.045	0.549±0.037	0.533±0.016	0.641±0.033	0.537±0.030	<u>0.537±0.020</u>
Abalone	0.556±0.036	0.529±0.016	0.561±0.042	0.544±0.038	<u>0.530±0.016</u>	0.648±0.026	0.564±0.043	0.546±0.014
Apple	0.529±0.019	<u>0.528±0.021</u>	0.534±0.025	0.535±0.021	0.534±0.022	0.561±0.030	0.540±0.033	0.525±0.007
Student	0.669±0.039	<u>0.563±0.025</u>	0.654±0.025	0.597±0.047	0.540±0.043	0.642±0.046	0.569±0.015	0.579±0.010

D.3 Fidelity

We provide the complete DCR statistics of CauTab across all benchmark datasets and sample sizes.

Table 11 shows that CauTab maintains small train-test DCR gaps on Adult, Abalone, Apple, and Student. Since DCR depends on feature scaling, dimensionality, and reference-set size, we do not compare its absolute values across datasets or treat larger values as automatically better. Instead, we focus on whether the synthetic samples are similarly close to training and unseen test records under a matched-reference protocol. Together with the Adult visualization in Figure 3, these results provide complementary evidence that CauTab avoids both near-duplicate memorization and severe off-manifold drift.

D.4 Privacy

Tables 12–13 report the supplementary nearest-neighbor membership inference attack results at $N = 50, 100$.

Table 12 shows the nearest-neighbor MIA results at $N = 50$. CauTab is close to random guessing on Apple and Student, but is less competitive on Adult and Abalone under this smallest training budget. This indicates that membership distinguishability is more variable when only very limited real samples are available.

Table 13 shows the results at $N = 100$. CauTab improves on Adult, Abalone, and Apple compared with the $N = 50$ setting, achieving the best result

on Apple and remaining close to the strongest baselines on Adult. Student remains more challenging, likely because preserving the explicit rule makes some local neighborhoods more distinguishable to nearest-neighbor attacks.

D.5 Ablation

FD ablation. Table 14 reports the FD ablation results at $N = 50$ and $N = 100$, complementing the $N = 200$ results in Table 4. Removing FD constraints leads to non-negligible violation rates, especially on Adult at $N = 50$ and Student across both settings. Data-only FD discovery can preserve the simple Adult mapping, but remains less reliable on Student, where the rule is range-based. These results further support the need for semantic FD screening before exact validation.

Causal graph ablation. To further examine the causal graph construction module under different training budgets, we repeat the graph ablation at $N = 50$ and $N = 100$. Specifically, we replace the full CauTab graph with graphs obtained from PC, DAGMA-only, and LLM-only discovery, while keeping the FD constraints and counterfactual generator unchanged. Tables 15 and 16 report the resulting utility. At $N = 50$, CauTab achieves the best score on all four datasets. At $N = 100$, CauTab remains best on Adult, Student, and Abalone, and is competitive on Apple, where DAGMA-only is slightly higher. These results suggest that combining data-driven structure learning

Table 14: FD ablation at smaller training sizes. We report RF-F1 and FD violation rate (%).

Dataset	RF-F1 $\uparrow$			FD Viol. $\downarrow$		
	CauTab	w/o FD	Data-only FD	CauTab	w/o FD	Data-only FD
Adult ($N = 50$)	78.17	78.47	78.87	0.00	15.97	0.00
Adult ($N = 100$)	77.64	77.96	78.88	0.00	3.47	0.00
Student ($N = 50$)	87.40	83.30	73.61	0.00	11.43	7.23
Student ($N = 100$)	88.78	83.23	77.81	0.00	12.83	6.50

Table 15: Causal graph ablation at $N = 50$. We report RF-F1 (%) for classification datasets and R^2 for Abalone.

Dataset	CauTab	PC	DAGMA-only	LLM-only
Adult	78.17$\pm$1.05	72.38 $\pm$ 2.09	71.15 $\pm$ 1.33	72.16 $\pm$ 3.74
Apple	68.64$\pm$0.54	62.87 $\pm$ 2.31	63.34 $\pm$ 1.28	58.67 $\pm$ 6.35
Student	87.40$\pm$0.95	86.95 $\pm$ 2.51	85.00 $\pm$ 1.58	86.79 $\pm$ 2.19
Abalone	0.34$\pm$0.02	0.21 $\pm$ 0.07	0.05 $\pm$ 0.12	0.24 $\pm$ 0.08

with semantic refinement yields more useful propagation paths than relying on a single graph source in low-data settings.

D.6 Robustness Across Downstream Models

We further investigate the robustness of synthetic data utility by varying the downstream classification algorithms, namely Random Forest (RF), Logistic Regression (LR), Decision Tree (DT) and Multilayer Perceptron (MLP). Figure 8 reveals that the efficacy of baseline models often fluctuates with the choice of classifier—specifically showing poor generalization in linear models like LR. But CauTab consistently achieves state-of-the-art performance across all tested scenarios. Regardless of whether the downstream task relies on linear decision boundaries (LR) or complex non-linear mappings (MLP), CauTab preserves the necessary feature dependencies, demonstrating that our method generates high-fidelity data suitable for a wide range of analytical applications.

E Prompt Designs

In this section, we provide the two prompt templates used in CauTab. Placeholders such as `{{Variable}}` represent dynamic inputs derived from the dataset.

E.1 FD Candidate Proposal Prompt ($\mathcal{P}_{\text{FD}}$)

This prompt screens semantically implausible determinant-dependent pairs before validation. It asks the LLM to identify attribute pairs (V_i, V_j) that are logically impossible to form a functional dependency.

Table 16: Causal graph ablation at $N = 100$. We report RF-F1 (%) for classification datasets and R^2 for Abalone.

Dataset	CauTab	PC	DAGMA-only	LLM-only
Adult	78.50$\pm$0.63	64.87 $\pm$ 0.00	72.65 $\pm$ 1.80	76.12 $\pm$ 4.16
Apple	69.66 $\pm$ 2.43	63.89 $\pm$ 4.24	70.67$\pm$2.68	65.92 $\pm$ 1.33
Student	88.78$\pm$0.69	75.23 $\pm$ 6.97	70.10 $\pm$ 3.54	88.17 $\pm$ 1.58
Abalone	0.39$\pm$0.04	0.19 $\pm$ 0.02	0.19 $\pm$ 0.03	0.33 $\pm$ 0.03

E.2 Causal Discovery Prompt ($\mathcal{P}_{\text{CD}}$)

This prompt is used to refine the statistically learned structure. We query the LLM to assess the causal direction between a target variable V_i and all other variables in the dataset, while injecting DAGMA suggestions as statistical hints.

Prompt 1: FD Candidate Proposal $\mathcal{P}_{\text{FD}}$

Role: Database Domain Expert

Task Description: Your goal is to identify candidate functional dependencies from the given tabular schema. A candidate dependency $X \rightarrow Y$ should describe a deterministic mapping, where the value of attribute X or an attribute set X uniquely determines the value of attribute Y .

Input Context:

- **Attributes List:** $\{\{\text{attribute_list}\}\}$
- **Dataset Description:** $\{\{\text{dataset_description}\}\}$
- **Data Samples (Few-shot):** $\{\{\text{representative_samples}\}\}$

Reasoning Instructions:

1. Analyze the semantic meaning of each attribute based on its name and sample values.
2. Propose only deterministic rules, such as code–label mappings, category-to-index mappings, or explicit arithmetic/logical formulas.
3. Do not report ordinary correlations, predictive associations, or rules that only hold statistically.
4. Avoid quasi-identifier rules caused by nearly unique attributes unless the dependency is supported by an explicit deterministic formula.

Output Format: Return the result strictly in JSON format containing the candidate dependencies:

```
{
  "functional_dependencies": [
    { "lhs": ["Attribute_X"], "rhs": "Attribute_Y",
      "type": "lookup/formula", "reason": "brief deterministic rationale" },
    ...
  ]
}
```

Figure 9: Prompt template for LLM-guided FD candidate proposal.

Prompt 2: Causal Discovery ($\mathcal{P}_{CD}$)

Role: Domain Expert in $\{\{\text{domain_field}\}\}$ Data Analysis

Task Description: You are investigating the cause-and-effect relationships between variables in the $\{\{\text{dataset_name}\}\}$ dataset. Your goal is to determine the true causal direction relative to a specific target variable based on domain knowledge and provided statistical cues.

Input Context:

- **Variable Definitions:** $\{\{\text{variable_list_with_descriptions}\}\}$
- **Target Variable:** $\{\{\text{target_variable}\}\}$

Statistical Hints ($\mathcal{H}_{stat}$): The following relationships were suggested by a traditional causal discovery algorithm (DAGMA) based on the dataset:

- Indicators that statistically change **BEFORE** $\{\{\text{target_variable}\}\}$: $\{\{\text{stat_before_list}\}\}$
- Indicators that statistically change **AFTER** $\{\{\text{target_variable}\}\}$: $\{\{\text{stat_after_list}\}\}$

Reasoning Instructions:

1. Combine your expert domain knowledge with the reference information provided above.
2. **Critical Thinking:** Statistical algorithms may mistake correlation for causation or miss non-linear relationships. Use your semantic understanding to correct any logical errors in the reference information (e.g., "Age" cannot be caused by "Marital Status").
3. **Categorization:**
 - Which indicators **!undoubtedly!** lead to changes in $\{\{\text{target_variable}\}\}$? (Causes)
 - Which indicators **!undoubtedly!** change after $\{\{\text{target_variable}\}\}$ changes? (Effects)
 - Which indicators cannot determine the order of change?

Output Format: Please answer strictly in the following format. No explanation is required.

before:
variables that cause the target
after:
variables caused by the target
not confirm:
variables with uncertain causal order

Figure 10: Prompt template for LLM-assisted causal graph refinement.