# AFT-Tab: Adversarial Fine-Tuning for Tabular Data Synthesis with Long Text Columns

**Yuhao Zhang[*1,3], Liang Yan[*1,4], Shaoming Duan[†3], Xinyu Zha[1], Jinhang Su[1], Peiyi Han[1,2], Chuanyi Liu[†1,2]**

[1]Harbin Institute of Technology, Shenzhen,
[2]Pengcheng Laboratory,
[3]Mindflow.ai,
[4]Inspur Cloud Information Technology Co., Ltd

**Correspondence:** shaomingduan@gmail.com, liuchuanyi@hit.edu.cn

## Abstract

Traditional tabular data synthesis methods often overlook the cross-modal heterogeneity of real-world tables, where structured continuous and discrete attributes coexist with unstructured long-text columns. Existing synthesis approaches struggle to simultaneously achieve accurate statistical fidelity for non-textual attributes and consistent semantic constraints between textual and non-textual attributes. In this work, we establish the first benchmark for long-text tabular data synthesis and introduce a novel metric, termed Textual Column Correlation Fidelity (TCCF), to quantify cross-modal semantic alignment. We propose AFT-Tab, an adversarial fine-tuning framework that synergistically trains an LLM-based text generator and a deep-learning-based non-textual generator. Through a dual-feedback mechanism guided by an LLM discriminator, AFT-Tab ensures both precise statistical distributions and rigorous semantic constraints. Experimental results show that AFT-Tab significantly outperforms state-of-the-art baselines in statistical fidelity, TCCF, diversity, and downstream task utility.

## 1 Introduction

In recent years, tabular data synthesis has attracted growing attention due to its ability to support data sharing(Hernandez et al., 2022), data augmentation(Cui et al., 2024), and data cleaning(Reis et al., 2024). However, existing studies are limited to traditional tabular data composed solely of continuous and discrete columns(Fang et al., 2024). In contrast, real-world tabular data is increasingly characterized by cross-modal heterogeneity(Hulsebos et al., 2023), often containing both structured numerical or categorical attributes and unstructured long-text attributes, such as user reviews on e-commerce platforms or project descriptions on crowdfunding platforms (Jensen and Özkil, 2018; Hulsebos et al., 2023). Synthesizing such mixed-type tabular data requires not only modeling the joint distribution of continuous and discrete columns (Wu et al., 2025), but also capturing the semantic relationships between textual and non-textual columns(Chen et al., 2020). For example, when a text review expresses strong positive sentiment, the corresponding rating column should be more likely to exhibit a high value. To the best of our knowledge, there has been no prior work specifically addressing the synthesis of tabular data with long-text attributes.

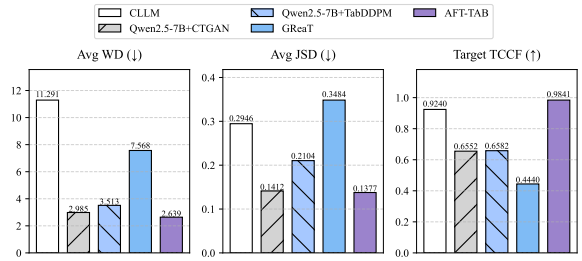


Figure 1: Comparison results on the Wine Dataset. CLLM(Seedat et al., 2024), GReaT(Borisov et al., 2023a), and AFT-Tab are all based on the Qwen2.5-7B model. The rightmost plot illustrates the TCCF between the text column and the target column.

*This problem motivates us to investigate whether existing synthesis methods can be effectively applied to tabular data containing long-text attributes.* Due to the lack of benchmark datasets for long-text tabular data and evaluation metrics that capture the semantic constraints between textual and non-textual columns(Zhang et al., 2024), we construct a tabular dataset containing long-text attributes, all of which are collected from real-world production systems. In addition, we propose a new evaluation metric, termed Textual Column Correlation Fidelity (TCCF), to quantify the discrepancy in semantic associations between textual and non-textual columns in synthetic and real data.

Based on the constructed dataset and evaluation

[*]Equal contribution.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4581–4594

metric, we conduct a systematic evaluation of existing synthesis methods. Feasible approaches for synthesizing long-text tabular data can be broadly categorized into two classes. The first class directly leverages large language models (LLMs) to generate entire tables(Borisov et al., 2023b; Solatorio and Dupriez, 2023; Seedat et al., 2024). In this approach, each table row is serialized into a text sequence, and carefully designed prompts or fine-tuning strategies are employed to improve the quality of the generated serialized data(Hegselmann et al., 2023; Sui et al., 2024). The model outputs are then decoded back into tabular form, followed by data filtering strategies to remove noise from the synthetic data. As shown in Figure 1, LLM-based methods are able to better preserve semantic correlations in textual columns, but perform poorly in terms of the statistical distributions of non-textual columns. This is because serializing non-textual attributes together with lengthy textual columns often introduces semantic noise, causing the model to prioritize linguistic fluency over statistical accuracy. As a result, structural distortions arise, where the synthesized numerical columns deviate from the original distributions.

The second feasible solution consists of hybrid approaches. These methods first employ conventional generative adversarial networks (GANs)(Xu et al., 2019; Armanious et al., 2020; Alshantti et al., 2024) or diffusion models(Kotelnikov et al., 2023; Shi et al., 2024; Lee et al., 2023) to generate continuous and discrete columns, and then use an LLM to generate textual columns conditioned on the non-textual attributes. Experimental results in Figure 1 indicate that hybrid methods outperform single LLM-based approaches in modeling the statistical distributions of non-textual columns, but exhibit inferior performance in terms of textual similarity fidelity. This limitation arises because, when generating textual columns, LLMs are constrained by their context window and can only rely on a limited number of examples, which prevents them from effectively modeling the global distribution of the textual data and, consequently, weakens the association constraints between textual and non-textual columns.

Motivated by the above experimental findings, we integrate the strengths of GANs and LLMs to improve the synthesis of long-text tabular data. To this end, we propose AFT-Tab, an adversarial fine-tuning–based framework for long-text tabular data synthesis. The core idea of AFT-Tab is to combine adversarial generative training with reinforcement-based incentives, enabling the joint and coordinated training of LLMs and GANs in a unified framework. Through adversarial rewards, AFT-Tab not only preserves the statistical distributions of non-textual columns in the original data, but also ensures that the constraint relationships between textual and non-textual attributes in the synthetic data are consistent with those in the real data. Specifically, in AFT-Tab, the LLM serves as the generator for textual columns, producing textual attributes conditioned on the non-textual columns of the table, while a deep generative model acts as the generator for non-textual columns and is responsible for synthesizing non-textual attributes. Meanwhile, a separate LLM–based discriminator is employed to distinguish between real and synthetic data. To ensure coordinated optimization between the textual and non-textual generators, we introduce a dual-feedback optimization mechanism. The first feedback loop applies a reinforcement learning–based optimization strategy that uses discriminator feedback to strengthen the LLM, enabling it to capture global dependencies in the textual attributes. The second feedback loop updates the non-textual generator via an experience replay mechanism, allowing it to better model the statistical properties of non-textual attributes.

The main contributions of this paper are as follows:

- We establish the first standardized benchmark for long-text tabular data synthesis and propose TCCF, a metric quantifying semantic discrepancies between textual and structured attributes. Utilizing this framework, we conduct a systematic assessment of existing synthesis methodologies.
- We propose AFT-Tab, an adversarial framework designed to preserve the statistical fidelity of structured columns while ensuring that the cross-modal dependencies between textual and non-textual attributes remain consistent with the real distribution.
- We design a dual-feedback mechanism through which discriminator signals are leveraged to jointly optimize the LLM and the GAN within the adversarial framework.
- Extensive experimental results demonstrate that AFT-Tab significantly outperforms state-

of-the-art baseline models in terms of non-textual distribution fidelity, textual correlation fidelity, synthetic data diversity, and downstream task utility.

## 2 Related Works

### 2.1 GAN-based Tabular Data Synthesis

Generative Adversarial Networks (GANs) have been extensively adapted for tabular data to model complex feature dependencies. Early works like medGAN (Armanious et al., 2020) and FCT-GAN (Zhao et al., 2022) introduced autoencoders and Fourier transforms to handle tabular structures. To improve control and fairness, conditional architectures were developed: TabFairGAN (Rajabi and Garibay, 2022) addresses bias, while CasTGAN (Alshantti et al., 2024) employs a cascaded generation strategy. Notably, CTGAN (Xu et al., 2019) and CTAB-GAN (Zhao et al., 2021) utilize conditional sampling and mixed-type encoding to effectively handle class imbalance and skewed distributions. RC-TGAN (Gueye et al., 2023) further extends this to relational data. Despite their success in numerical modeling, GANs suffer from training instability and mode collapse. Crucially, they lack the inherent capability to generate long-text sequences, necessitating hybrid approaches that loosely couple GANs with text generators for mixed-modal synthesis.

### 2.2 Diffusion-based Tabular Data Synthesis

Diffusion Models (DMs) (Yang et al., 2023; Cao et al., 2024) offer a stable alternative to GANs via iterative denoising. TabDDPM (Kotelnikov et al., 2023) pioneered this by mapping tabular rows to continuous vectors, while TABDIFF (Shi et al., 2024) and CoDi (Lee et al., 2023) introduced specialized mechanisms for heterogeneous and mixed-type features. Recent works like CTSyn (Lin et al., 2024) and RelDDPM (Liu et al., 2024) scale DMs to cross-table and relational settings. While DMs surpass GANs in distribution coverage and stability, they face significant computational overhead and struggle with discrete feature modeling. similar to GANs, standard tabular DMs cannot natively synthesize semantic-rich long text, often requiring separate pipelines that fail to capture the fine-grained correlation between structured attributes and textual content.

### 2.3 LLM-based Tabular Data Synthesis

Recent approaches leverage Large Language Models (LLMs) via prompting or fine-tuning. Prompt-based methods like CLLM (Seedat et al., 2024) and TABGEN-ICL (Fang et al., 2025) utilize in-context learning for low-resource scenarios, with variants like EPIC (Kim et al., 2024) and LITO (Yang et al., 2024a) optimizing for class imbalance. Fine-tuning methods, represented by GReaT (Borisov et al., 2023a), serialize tabular data into text sequences to train autoregressive models. Other works explore token compression (Tabula (Zhao et al., 2025)), masked modeling (TabMT (Gulati and Roysdon, 2023)), and privacy-preserving tuning (HARMONIC (Wang et al., 2024)). Although P-TA (Yang et al., 2024b) attempts to combine PPO with GANs, most LLM-based methods treat tabular synthesis as a pure text generation task. This serialization often discards the topological structure of tables, leading to "hallucinations" where the generated long text is linguistically fluent but logically inconsistent with the accompanying structured columns.

## 3 Benchmark for Long-Text Tabular Data Synthesis

Given the absence of standardized benchmarks for long-text tabular data synthesis, we establish the first comprehensive benchmark comprising diverse datasets and a multi-dimensional evaluation protocol tailored for cross-modal fidelity.

### 3.1 Benchmark Datasets

Table 1: Statistics of Benchmark Datasets. "Text Cols" denotes the number of unstructured text columns.

| Dataset | Train | Test | Text Cols | Task | Metric |
|---|---|---|---|---|---|
| Cloth | 18,788 | 4,698 | 3 | Regression | $R^2$ |
| Kickstarter | 86,502 | 21,626 | 3 | Binary Class. | AUC |
| Wine | 84,123 | 21,031 | 3 | Multi-class | Acc. |

We employ three public datasets characterized by long-text columns that exhibit strong semantic correlations with structured target attributes. Adhering to the benchmark protocols established by (Shi et al., 2021), our evaluation dataset comprises: **Kickstarter**(Jensen and Özkil, 2018), a crowdfunding platform dataset utilized for the binary classification of project success; **Cloth**(Agarap, 2018), an e-commerce dataset that maps customer reviews to numerical ratings as a regression task; and **Wine**(Zynicide, 2017), a

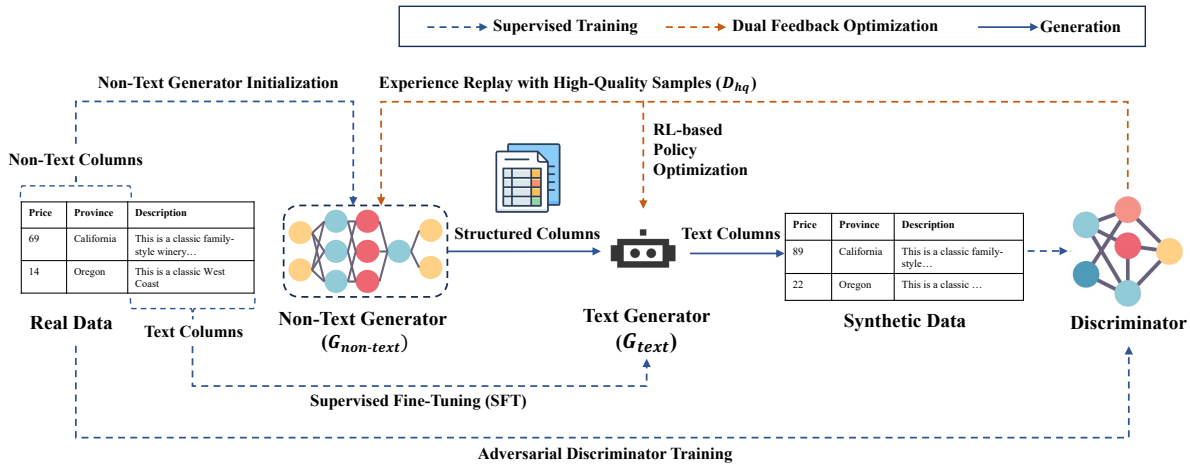


Figure 2: Overview of the AFT-Tab framework. It orchestrates a non-text generator ($G_{\text{non-text}}$) and a text generator ($G_{\text{text}}$) within a dual-feedback loop supervised by a discriminator ($D$).

collection of reviews used for variety identification within a multi-class classification framework. These datasets provide a comprehensive testbed for assessing the joint distribution modeling capabilities of generative models across diverse domains. Statistical summaries are presented in Table 1, while detailed descriptions are available in Appendix A.1.

### 3.2 Textual Column Correlation Fidelity

Conventional marginal metrics often overlook the intrinsic high-order dependencies between unstructured text and structured attributes. To quantify this cross-modal alignment, we propose *Textual Column Correlation Fidelity (TCCF)*. Unlike task-specific utility metrics, TCCF assesses whether the synthetic text retains the consistent predictive power regarding structured attributes as observed in the real distribution. To eliminate hyperparameter sensitivity and ensure determinism, we standardly employ TF-IDF feature extraction combined with basic linear models (Logistic Regression for classification and Ridge Regression for regression) to evaluate the predictive strength of the text.

Formally, let $\mathcal{D} = \{T, \mathbf{S}\}$ denote a dataset comprising textual attributes $T$ and structured columns $\mathbf{S}$. To evaluate the correlation for each $S_i \in \mathbf{S}$, we employ a protocol $\Phi$ where the text column $T_{\text{max}}$ with the maximum sequence length serves as the predictor. Using disjoint training and evaluation subsets, we train a model on the encoded $T_{\text{max}}$ to predict each target $S_i$. The correlation scores for real ($\mathcal{D}_{\text{real}}$) and synthetic ($\mathcal{D}_{\text{syn}}$) data are defined as:

$$C^{(i)}_{\text{real}} = \Phi(\mathcal{D}_{\text{real}}, S_i), \quad C^{(i)}_{\text{syn}} = \Phi(\mathcal{D}_{\text{syn}}, S_i) \quad (1)$$

The TCCF score for column $S_i$ is derived as the complement of the absolute performance discrepancy:

$$\text{TCCF}_i = \max\left(0, 1 - \left|C^{(i)}_{\text{syn}} - C^{(i)}_{\text{real}}\right|\right) \quad (2)$$

where $\max(0, \cdot)$ ensures non-negativity. We report both the Target TCCF for the designated downstream column and the Average TCCF across all structured attributes. A higher TCCF score signifies superior fidelity, indicating that the synthetic text preserves the underlying predictive relationships characteristic of the real data.

## 4 Tabular Data Synthesis with Long Text via Adversarial Fine-Tuning

### 4.1 Overall Framework

Figure 2 illustrates AFT-Tab, a framework for synthesizing tabular data with long text via adversarial fine-tuning. Built upon the GAN paradigm, our architecture employs two distinct generators: a model for non-text attributes ($G_{\text{non-text}}$) and a LLM ($G_{\text{text}}$) for textual content. A discriminator ($D$) is concurrently employed to evaluate the fidelity of the synthesized samples.

In this adversarial setting, $G_{\text{non-text}}$ and $G_{\text{text}}$ collaborate to generate composite samples, comprising structured features $\tilde{s}$ and textual sequences $\tilde{t}$, aiming to deceive $D$. Conversely, $D$ learns to distinguish between the real distribution $p_{\text{data}}$ and the synthetic distribution. This interaction is formalized as a minimax game with the objective function

$V(G_{\text{non-text}}, G_{\text{text}}, D)$:

$$\min_{G_{\text{non-text}}, G_{\text{text}}} \max_{D} V(D, G_{\text{non-text}}, G_{\text{text}}) = \mathbb{E}_{(s,t)\sim p_{\text{data}}}[\log D(s,t)] + \mathbb{E}_{\substack{\tilde{s}\sim G_{\text{non-text}} \\ \tilde{t}\sim G_{\text{text}}(\cdot|\tilde{s})}}[\log(1 - D(\tilde{s}, \tilde{t}))] \tag{3}$$

Distinguished from standard GANs, AFT-Tab implements a dual-feedback mechanism where signals from $D$ independently guide both generators. This approach shifts the paradigm from isolated generation to co-evolution, iteratively enhancing the joint fidelity of structured and textual modalities.

### 4.2 Model for Non-Text Attributes Synthesis

We employ the CTGAN architecture (Xu et al., 2019) as the backbone for $G_{\text{non-text}}$. This choice is motivated by its proven efficacy in modeling complex joint distributions $p(s)$ characterized by multi-modal continuous variables and imbalanced categorical features. While initialized on real structured attributes, our approach diverges from standard independent training by incorporating continuous feedback from the global discriminator $D$. This integration dynamically optimizes the generation strategy, ensuring that the synthesized structured data serves as a statistically faithful and semantically coherent condition for the subsequent text generator $G_{\text{text}}$.

### 4.3 Reinforcement Learning based Long Text Columns Synthesis

We utilize a LLM as the foundation for text synthesis, employing a two-stage optimization process to ensure semantic coherence between the generated text and structured attributes.

**Stage 1: Supervised Fine-Tuning.** Initially, we adapt the LLM (policy network $\pi_\theta$) to the specific task domain using real data. Non-text columns $s_{non-text}$ are serialized into text prompts, with the corresponding long text columns $s_{text}$ serving as target outputs. This initialization enables the model to capture fundamental conditional dependencies. The training objective is defined as:

$$L_{SFT}(\theta) = -\sum \log \pi_\theta(s_{text}|s_{non-text}) \tag{4}$$

**Stage 2: Reinforcement Learning Optimization.** Subsequently, we integrate discriminator feedback via an RL loop to refine the generation strategy. We formulate conditional text generation as a policy optimization problem: the state $s_t$ corresponds to the synthesized structured data $s_{non-text}$ from $G_{non-text}$, and the action $a_t$ is the generated text sequence $s_{text}$. The reward signal $R$ is directly derived from the discriminator $D$, such that $R(s_{non-text}, s_{text}) = D(s_{non-text}, s_{text})$, thereby propagating adversarial gradients to the text generator.

To mitigate the instability of sparse rewards, we introduce a value network $V_\phi(s_t)$ to estimate the expected return. The advantage function $\hat{A}_t$, quantifying the value of action $a_t$ relative to the baseline at state $s_t$, is computed using the immediate reward $R_t$:

$$\hat{A}_t = R_t - V_\phi(s_t) \tag{5}$$

Adopting Proximal Policy Optimization (PPO), we employ a clipped surrogate objective to constrain policy updates. The single-step objective $L_t(\theta)$ applies asymmetric clipping based on the sign of the advantage function:

$$L_t(\theta) = \begin{cases} \min(r_t(\theta), 1+\epsilon) \cdot \hat{A}_t & \text{if } \hat{A}_t \geq 0 \\ \max(r_t(\theta), 1-\epsilon) \cdot \hat{A}_t & \text{if } \hat{A}_t < 0 \end{cases} \tag{6}$$

The global objective maximizes $L^{PPO}(\theta) = \hat{\mathbb{E}}_t[L_t(\theta)]$. Here, $r_t(\theta)$ denotes the probability ratio between the current and old policies, while $\epsilon$ defines the trust region. This mechanism encourages actions with positive advantages while limiting excessive policy shifts to ensure stability.

Our application of RL diverges from standard static environments in two key aspects. First, the reward function is non-stationary, dynamically provided by a co-evolving discriminator $D$, necessitating continuous adaptation by the policy. Second, this RL loop is integral to the dual-feedback architecture; high-quality samples reciprocally enhance the non-text generator $G_{non-text}$. Consequently, the LLM actively explores generation strategies to maximize global consistency rather than merely mimicking surface-level patterns.

### 4.4 Dual Feedback Optimization

The dual feedback optimization mechanism fosters the co-evolution of the non-text and text generators, enabling mutual enhancement through interaction rather than isolated improvement to maximize overall data fidelity.

The first feedback loop implements **RL-based Policy Optimization** to target the text generator. Here, the reward signal provided by the discriminator $D$ continuously refines the LLM policy $\pi_\theta$ via the PPO algorithm. By maximizing this reward, the

LLM is incentivized to achieve rigorous alignment between the textual output and the structured conditions, ensuring semantic and logical consistency.

The second feedback loop utilizes **Experience Replay** with High-Quality Samples ($\mathcal{D}_{hq}$) to refine the non-text generator $G_{\text{non-text}}$. Post-generation, the discriminator $D$ assesses the authenticity of the synthesized samples, yielding a binary classification:

$$D(s', t') = \begin{cases} 1 & \text{Sample classified as real} \\ 0 & \text{Sample classified as fake} \end{cases} \quad (7)$$

We isolate the high-quality subset $\mathcal{D}_{hq}$ from the synthesized batch. Specifically, for generated pairs $(s', t')$ sampled via $s' \sim G_{\text{non-text}}$ and $t' \sim \pi_\theta(\cdot|s')$, we retain those satisfying the authenticity criterion:

$$\mathcal{D}_{hq} = \{(s', t') \mid D(s', t') = 1\} \quad (8)$$

The structured components $\{s'|(s', t') \in \mathcal{D}_{hq}\}$ are subsequently extracted and added to an experience replay buffer. Combined with the original real data, they constitute an augmented training set for the subsequent iteration of $G_{\text{non-text}}$. This process serves as a dynamic data augmentation mechanism, guiding $G_{\text{non-text}}$ to produce structured attributes that are amenable to coherent textual description.

These two feedback loops establish a synergistic cycle. An improved text generator creates more convincing composite samples, thereby enriching the high-quality training data for the non-text model. Conversely, a refined non-text generator produces more realistic structured conditions, facilitating the task of conditional text synthesis. Through this iterative adversarial process, the AFT-Tab framework drives both modules to jointly approximate the true data distribution, ensuring high fidelity in both low-level statistical features and high-level semantic relations.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets and Metrics.** We use the benchmark datasets and evaluation metrics introduced in Section 3. To quantify distributional discrepancies, we employ the Wasserstein distance for continuous attributes and the Jensen-Shannon divergence for categorical features (Arjovsky et al., 2017; Xu and Veeramachaneni, 2018). Textual fidelity is assessed via Semantic Similarity, calculated as the cosine similarity between real and synthetic embeddings. Detailed definitions and formulations of WD, JSD, and Semantic Similarity are provided in Appendix A.2. For practical utility, we adopt the Train-on-Synthetic, Test-on-Real (TSTR) paradigm, reporting AUC, $R^2$, and Accuracy appropriate to each task. We benchmark performance using three modeling strategies: **All-text** (Raffel et al., 2020), which fine-tunes a BERT predictor on generated text; **Fuse-early** (Hu and Singh, 2021), which concatenates text embeddings with structured features for MLP processing; and **Fuse-late** (Audebert et al., 2019), which aggregates prediction logits from independent BERT and XGBoost models.

**Baselines and Implementation.** To evaluate the efficacy of AFT-Tab, we benchmark it against four state-of-the-art methods spanning diverse paradigms: (1) CLLM (Seedat et al., 2024), a prompt-based approach leveraging LLM priors for low-resource synthesis; (2) CTGAN+LLM (Xu et al., 2019), a hybrid pipeline utilizing CTGAN for structured attributes and a fine-tuned LLM for conditional text generation; (3) TabDDPM+LLM (Kotelnikov et al., 2023), which pairs a diffusion model for tabular generation with an LLM; and (4) GReaT (Borisov et al., 2023b), a purely autoregressive LLM-based approach that synthesizes serialized rows. Regarding our implementation, we initialize the text generator $G_{\text{text}}$ with `Qwen2.5-7B-Instruct`. Comprehensive hyperparameters and engineering details are provided in Appendix A.3.

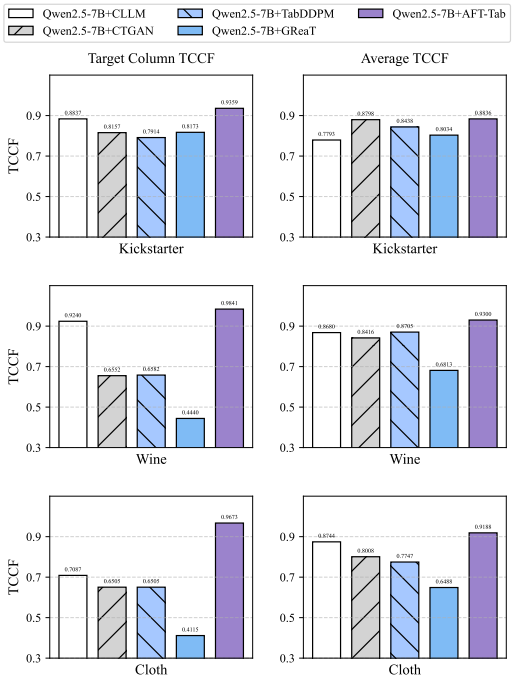

Figure 3: Textual Column Correlation Fidelity (TCCF) comparison.

Table 2: Combined results of Statistical Fidelity and Semantic Similarity. **Bold** indicates best performance. .

| Dataset | Method | Statistical Fidelity | | Semantic Similarity | |
|---|---|---|---|---|---|
| | | Avg WD ↓ | Avg JSD ↓ | Text-Col Sim. ↑ | Table-wise Sim. ↑ |
| Kickstarter | Qwen2.5-7B+CLLM | 6230.8 | 0.1894 | 0.5503 | 0.8524 |
| | Qwen2.5-7B+CTGAN | 6113.45 | 0.1577 | 0.5521 | 0.8620 |
| | Qwen2.5-7B+TabDDPM | 140988.4 | **0.0136** | **0.6059** | 0.8439 |
| | Qwen2.5-7B+GReaT | 460446.9 | 0.1138 | 0.5343 | 0.7874 |
| | **Qwen2.5-7B+AFT-Tab (Ours)** | **4930.8** | 0.1626 | 0.5492 | **0.8717** |
| Cloth | Qwen2.5-7B+CLLM | 248.77 | 0.1733 | **0.6018** | 0.8626 |
| | Qwen2.5-7B+CTGAN | 6.836 | 0.1445 | 0.5047 | 0.8535 |
| | Qwen2.5-7B+TabDDPM | 27.012 | 0.2615 | 0.5073 | 0.8641 |
| | Qwen2.5-7B+GReaT | 27.41 | 0.2353 | 0.4602 | 0.8478 |
| | **Qwen2.5-7B+AFT-Tab (Ours)** | **4.112** | **0.1339** | 0.5316 | **0.8950** |
| Wine | Qwen2.5-7B+CLLM | 11.291 | 0.2946 | 0.7698 | 0.9200 |
| | Qwen2.5-7B+CTGAN | 2.985 | 0.1412 | 0.7557 | 0.9316 |
| | Qwen2.5-7B+TabDDPM | 3.513 | 0.2104 | 0.7310 | 0.9193 |
| | Qwen2.5-7B+GReaT | 7.568 | 0.3484 | 0.5227 | 0.8507 |
| | **Qwen2.5-7B+AFT-Tab (Ours)** | **2.639** | **0.1377** | **0.7779** | **0.9387** |

## 5.2 Evaluation of TCCF

Figure 3 presents the evaluation of cross-modal dependency preservation between structured attributes and text, measured by both Target Column TCCF and Average TCCF. Our findings reveal that Qwen2.5-7B+AFT-Tab consistently outperforms all baseline methods across diverse datasets, particularly those characterized by complex semantic associations. On the Cloth benchmark, AFT-Tab achieves a remarkable Target TCCF of 0.9673, surpassing the leading competitor CLLM by a substantial margin. Similarly, on the Wine dataset, our model maintains near-perfect consistency with a score of 0.9841, whereas the purely generative baseline GReaT exhibits a marked performance decline to 0.4440. This empirical advantage validates the efficacy of our RL-driven consistency reward, which explicitly mitigates semantic misalignment during training. In contrast, hybrid pipelines such as CTGAN+LLM are often constrained by their decoupled generation processes, while serialization-based approaches face challenges in capturing long-range dependencies between feature tokens and generated text, leading to the observed degradation in fidelity. Detailed column-wise results are provided in Appendix A.4.

## 5.3 Statistical and Semantic Fidelity

Table 2 presents a comprehensive quantitative analysis spanning both statistical fidelity and semantic alignment. Regarding statistical metrics, AFT-Tab achieves the most balanced trade-off between continuous (WD) and categorical (JSD) fidelity. Specifically, on the Cloth dataset, our method significantly outperforms the second-best approach, CTGAN+LLM, in terms of WD. Notably, while baselines such as TabDDPM+LLM may attain low JSD scores on Kickstarter, they suffer from disproportionately high WD. This discrepancy suggests that such baselines effectively match discrete frequencies but fail to capture the underlying continuous manifold, highlighting the necessity of preserving the joint distribution—a capability that AFT-Tab enhances through its dual-feedback mechanism.

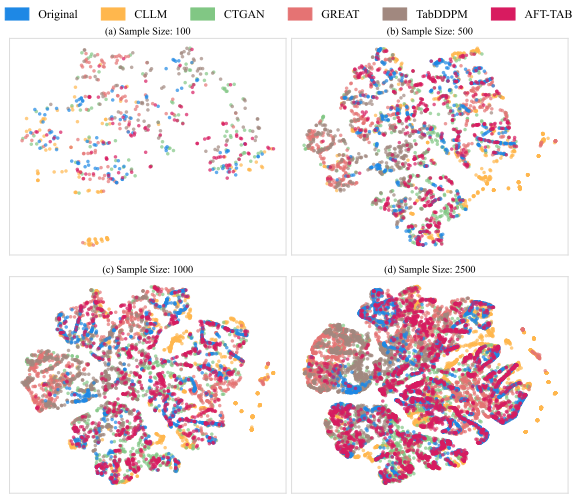


Figure 4: t-SNE visualization of real (Original) and synthetic data distributions on Cloth.

Beyond marginal distributions, the right-hand section of Table 2 details the **semantic similarity** results, revealing a critical distinction between linguistic plausibility and actual data utility. On the Wine dataset, for instance, CLLM produces text that closely mirrors real reviews, achieving high Text-Column Similarity; however, it exhibits weak logical correlations with the associated structured attributes. In contrast, AFT-Tab demonstrates robust performance across both metrics, maintaining

Table 3: Comparative analysis of downstream machine learning utility using $R^2$, AUC, and Accuracy. AFT-Tab demonstrates consistent superiority over baseline methods.

| Method | Kickstarter (AUC) | | | Cloth ($R^2$) | | | Wine (Acc) | | |
|---|---|---|---|---|---|---|---|---|---|
| | All-text | Fuse-Late | Fuse-Early | All-text | Fuse-Late | Fuse-Early | All-text | Fuse-Late | Fuse-Early |
| *Real Data* | *80.2* | *77.1* | *77.7* | *74.8* | *74.3* | *74.6* | *82.4* | *82.1* | *82.2* |
| Qwen2.5-7B+CLLM | 57.0 | 58.5 | 56.9 | 54.4 | 55.3 | 59.9 | 40.9 | 42.3 | 42.3 |
| Qwen2.5-7B+CTGAN | 57.5 | 56.6 | 56.6 | 12.6 | 24.2 | -13.9 | 40.3 | 37.9 | 40.4 |
| Qwen2.5-7B+TabDDPM | 59.1 | 56.9 | 56.7 | 18.3 | 16.0 | 5.6 | 24.1 | 26.4 | 24.0 |
| Qwen2.5-7B+GReaT | 55.5 | 53.1 | 56.5 | -16.1 | -16.6 | -14.9 | 18.2 | 17.0 | 16.2 |
| **Qwen2.5-7B+AFT-Tab (Ours)** | **66.3** | **64.8** | **66.5** | **61.0** | **58.9** | **62.8** | **67.4** | **68.6** | **68.0** |

comparable Text-Column Similarity while securing superior Table-wise Similarity. These findings imply that while prompt-based baselines may achieve surface-level fluency, AFT-Tab succeeds in generating naturalistic text that simultaneously maintains rigorous table-level consistency, driven by the alignment constraints imposed during the RL optimization phase.

### 5.4 Visualization of Data Distributions

To qualitatively assess the preservation of global data structure, Figure 4 visualizes the alignment between real and synthetic data distributions via t-SNE projections on the Cloth dataset. Observations indicate that samples generated by AFT-Tab exhibit substantial overlap with the real data, accurately recovering multiple density modes. Conversely, CLLM displays severe mode collapse, whereas GReaT generates diffuse noise with poor structural definition. These distinct patterns confirm that the experience replay mechanism effectively stabilizes training, enabling the tabular generator to learn and reproduce robust structural representations that mirror the complexity of the real distribution.

### 5.5 Machine Learning Utility

Table 3 validates the practical utility via the Train-on-Synthetic, Test-on-Real paradigm. Our results indicate that AFT-Tab consistently achieves state-of-the-art performance across all evaluation scenarios, effectively bridging the utility gap relative to real data. In the context of the Cloth regression task, baselines such as CTGAN+LLM and GReaT yield negative $R^2$ scores under the early fusion setting. This failure suggests that the textual modality generated by these baselines introduces detrimental noise, which compromises the predictive signal derived from structured features. Conversely, AFT-Tab secures robust positive results, demonstrating that its synthesized text maintains strong alignment with the target variables. Furthermore, AFT-Tab exhibits superior stability across fusion strategies compared to competing methods. While baselines like TabDDPM suffer marked performance declines when transitioning from late to early fusion, AFT-Tab maintains consistently high predictive capability irrespective of the fusion strategy. This stability underscores the efficacy of the cross-modal coupling within our framework, ensuring that textual and structured attributes provide complementary rather than conflicting information.

Table 4: Ablation study of component contributions on the Cloth dataset.

| Configuration | WD ↓ | TCCF ↑ | Utility ($R^2$) ↑ |
|---|---|---|---|
| **AFT-Tab (Full Model)** | **4.112** | **0.9116** | **61.0** |
| w/o RL Optimization ($G_{\text{text}}$) | 4.575 | 0.8932 | 59.7 |
| w/o Experience Replay ($G_{\text{non-text}}$) | 6.109 | 0.8915 | 59.1 |

### 5.6 Ablation Study

To isolate the individual contributions of the components within our dual-feedback mechanism, we conducted an ablation study on the Cloth dataset, with results detailed in Table 4. First, we assess the role of Reinforcement Learning in the text generator; the removal of the PPO phase, relying exclusively on SFT, results in a notable decline in the TCCF score. This reduction suggests that supervised learning predominantly captures surface-level regularities but lacks the incentive to enforce the rigorous instance-level consistency provided by the discriminator's reward signal, leading to generic text generation. Furthermore, we evaluate the impact of feedback on the non-text generator ($G_{\text{non-text}}$). Disabling the experience replay mechanism leads to a significant deterioration in the WD. This finding underscores that replaying text-compatible structured samples is critical for guiding the generator away from attribute combinations that are statistically plausible yet semantically incoherent, thereby mitigating the propagation of errors to the final output.

## 6 Conclusion

In this paper, we introduce AFT-Tab, an adversarial framework designed to synthesize tabular data with long-text attributes. By leveraging a novel dual-feedback mechanism, our approach coordinates a Large Language Model and a non-text generator to ensure strict semantic consistency between textual and numerical fields. We also propose a standardized benchmark and the Textual Column Correlation Fidelity (TCCF) metric for rigorous evaluation. Extensive experiments demonstrate that AFT-Tab significantly outperforms state-of-the-art baselines in preserving both statistical distribution and cross-modal dependencies.

## Limitations

Despite its superior performance, AFT-Tab has specific limitations. First, the integration of an LLM within a reinforcement learning loop incurs higher computational costs compared to traditional GAN-based methods. However, given the substantial leap in cross-modal consistency and the core TCCF metric, we believe this one-time training investment is justified. A detailed empirical comparison of computational complexity is provided in Appendix A.5. Second, the length of synthesized text is constrained by the context window of the underlying language model, which may limit the generation of extremely long documents. Finally, our current framework does not explicitly model differential privacy, a critical requirement for certain sensitive applications. While high recall indicates the model does not merely memorize training data, privacy-preserving tabular synthesis remains a distinct and crucial research field. As a foundational generative framework, AFT-Tab can theoretically be integrated with existing privacy techniques like DP-SGD, which we leave for future work.

## Acknowledgments

This study is supported by the National Key Research and Development Program of China under Grant 2023YFB3106504, the China Postdoctoral Science Foundation under Grant Number 2024M751555, the Major Key Project of PCL under Grant PCL2024A04 and PCL2025A16, Shenzhen Science and Technology Program under Grant ZDSYS20210623091809029.

## References

Abien Fred Agarap. 2018. Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network (rnn). *arXiv preprint arXiv:1805.03687*.

Abdallah Alshantti, Damiano Varagnolo, Adil Rasheed, Aria Rahmati, and Frank Westad. 2024. Castgan: Cascaded generative adversarial network for realistic tabular data synthesis. *IEEE Access*, 12:13213–13232.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, Sydney, Australia. PMLR.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. 2020. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684.

Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. 2019. Multimodal deep networks for text and image-based document classification. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 427–443. Springer.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023a. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023b. Language models are realistic tabular data generators. In *ICLR*, Kigali, Rwanda.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2024. A survey on generative diffusion models. *IEEE transactions on knowledge and data engineering*, 36(7):2814–2830.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942.

Lingxi Cui, Huan Li, Ke Chen, Lidan Shou, and Gang Chen. 2024. Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai. *arXiv preprint arXiv:2407.21523*.

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*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models (llms) on tabular data: Prediction, generation, and understanding–a survey. *arXiv preprint arXiv:2402.17944*.

Mohamed Gueye, Yazid Attabi, and Maxime Dumas. 2023. RC-TGAN: Row conditional-tabular gan for generating synthetic relational databases. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes, Greece.

Manbir Gulati and Paul Roysdon. 2023. Tabmt: Generating tabular data with masked transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 46245–46254, New Orleans, USA.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pages 5549–5581. PMLR.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1439–1449.

Madelon Hulsebos, Çagatay Demiralp, and Paul Groth. 2023. GitTables: A large-scale corpus of relational tables. *Proceedings of the ACM on Management of Data*, 1(1):1–17.

Lasse Skovgaard Jensen and Ali Gürcan Özkil. 2018. Identifying challenges in crowdfunded product development: a review of kickstarter projects. *Design Science*, 4:e18.

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, Hawaii, USA. PMLR.

Chaejeong Lee, Jayoung Kim, and Noseong Park. 2023. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pages 18940–18956, Hawaii, USA. PMLR.

Xiaofeng Lin, Chenheng Xu, Matthew Yang, and Guang Cheng. 2024. Ctsyn: A foundational model for cross tabular data generation. *arXiv preprint arXiv:2406.04619*.

Tongyu Liu, Ju Fan, Nan Tang, Guoliang Li, and Xiaoyong Du. 2024. Controllable tabular data synthesis using diffusion models. *Proceedings of the ACM on Management of Data*, 2(1):1–29.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. 2022. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction (MAKE)*, 4(2):488–501.

Eduardo Reis, Mohamed Abdelaal, and Carsten Binnig. 2024. Generalizable data cleaning of tabular data in latent space. *Proceedings of the VLDB Endowment*, 17(13):4786–4798.

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, Vienna, Austria.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. 2024. Tabdiff: a unified diffusion model for multi-modal tabular data generation. In *NeurIPS 2024 Third Table Representation Learning Workshop*, Vancouver, Canada.

Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alexander J Smola. 2021. Benchmarking multimodal automl for tabular data with text fields. *arXiv preprint arXiv:2111.02705*.

Aivin V Solatorio and Olivier Dupriez. 2023. Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

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.

In *Advances in Neural Information Processing Systems*, volume 37, pages 100196–100212, Vancouver, Canada.

Xiaofeng Wu, Alan Ritter, and Wei Xu. 2025. Tabular data understanding with LLMs: A survey of recent advances and challenges. *arXiv preprint arXiv:2508.00217*.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.

Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*.

June Yong Yang, Geondo Park, Joowon Kim, Hyeongwon Jang, and Eunho Yang. 2024a. Language-interfaced tabular oversampling via progressive imputation and self-authentication. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39.

Shuo Yang, Chenchen Yuan, Yao Rong, Felix Steinbauer, and Gjergji Kasneci. 2024b. P-ta: Using proximal policy optimization to enhance tabular data augmentation via large language models. In *Findings of the Association for Computational Linguistics ACL*, pages 248–264, Bangkok, Thailand.

Haowei Zhang, Shengyun Si, Yilun Zhao, and 1 others. 2024. OpenT2T: An open-source toolkit for table-to-text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–269.

Zilong Zhao, Robert Birke, and Lydia Y. Chen. 2022. Fct-gan: Enhancing table synthesis via fourier transform. *arXiv preprint arXiv:2210.06239*. ArXiv:2210.06239.

Zilong Zhao, Robert Birke, and Lydia Y. Chen. 2025. Tabula: Harnessing language models for tabular data synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 247–259, Sydney, Australia. Springer Nature Singapore.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. 2021. CTAB-GAN: Effective table data synthesizing. In *Proceedings of the 13th Asian Conference on Machine Learning (ACML 2021)*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112, Virtual.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Zynicide. 2017. Wine reviews.

# A Appendix

## A.1 Dataset Details

In this study, "Long-Text" columns primarily refer to "paragraph-level" text fields, which contain significantly more tokens than typical short-text fields. While actual long-form documents like medical reports and legal contracts possess substantial token lengths, they are rarely stored natively within tabular datasets and typically exist as independent files, falling outside the primary scope of this work.

**(1) Kickstarter (Jensen and Özkil, 2018):** A binary classification dataset predicting crowdfunding success. It integrates numerical attributes (e.g., goal, backers) with project descriptions, serving as a standard for evaluating heterogeneous data modeling.

**(2) Cloth (Agarap, 2018):** Consisting of over 23,000 women's clothing reviews, this dataset targets sentiment prediction (regression). It combines unstructured reviews with structured demographics (age) and product categories, providing a realistic scenario for NLP-tabular fusion.

**(3) Wine (Zynicide, 2017):** A large-scale dataset with ∼130k wine reviews connecting unstructured descriptions to structured attributes (price, points, variety). It supports both regression and multi-class classification tasks, testing the model's ability to capture fine-grained product quality signals.

## A.2 Evaluation Metrics

### A.2.1 Wasserstein Distance (WD)

Wasserstein distance (Arjovsky et al., 2017), also known as Earth Mover's Distance (EMD), measures the minimum "cost" required to transform one distribution's probability mass into another, where the cost is defined as the amount of mass moved times the moving distance. In the one-dimensional case, it can be intuitively expressed as the integral of the area between the two cumulative distribution functions (CDFs) $F_{real}(x)$ and $F_{syn}(x)$:

$$W_1(F_{real}, F_{syn}) = \int_{-\infty}^{\infty} |F_{real}(x) - F_{syn}(x)|\, dx \quad (9)$$

A WD of 0 indicates that the two distributions are identical; larger values imply greater distributional discrepancy.

### A.2.2 Jensen–Shannon Divergence (JSD)

Jensen–Shannon Divergence (Menéndez et al., 1997) is a symmetric, smoothed, and bounded variant of KL divergence, which mitigates KL's asymmetry and potential to become infinite. Given two discrete probability distributions $P$ and $Q$, we first define their mixture distribution as $M = \frac{1}{2}(P+Q)$. The KL divergence is:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{10}$$

Then the JSD is defined as the average of the KL divergences from $P$ and $Q$ to $M$:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \tag{11}$$

A JSD of 0 indicates perfect agreement ($P = Q$); larger values correspond to larger divergence between the two distributions.

### A.2.3 Semantic Similarity

To evaluate the overall similarity between text columns in the synthetic data and those in the real data, we measure semantic proximity by computing cosine similarity between their embedding representations. Cosine similarity is defined as

$$\text{CosineSim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \tag{12}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the embeddings of a real-text instance and its synthetic counterpart, respectively. Equivalently, letting $x_i$ and $y_i$ be the $i$-th dimensions of $\mathbf{x}$ and $\mathbf{y}$, we have

$$\text{CosineSim}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{13}$$

Higher cosine similarity indicates stronger semantic consistency between synthetic and real text columns.

## A.3 Implementation Details

Our framework is built upon the LLaMA-Factory (Zheng et al., 2024) library, leveraging its efficient pipelines for instruction tuning and reinforcement learning. To ensure training stability in our distributed environment, we explicitly disable NCCL P2P and IB communication. For optimization, the SFT phase utilizes a learning rate of $2 \times 10^{-5}$, while the PPO phase employs a reduced learning rate of $1 \times 10^{-6}$ with a batch size of 64.

## A.4 Complete TCCF Results

Tables 6–8 report the column-wise TCCF scores on Kickstarter, Wine, and Cloth. The results show that Qwen2.5-7B+AFT-TAB achieves consistently strong cross-modal consistency across heterogeneous attributes, matching the trend of the aggregated metrics in Figure 3. On Kickstarter (Table 6), AFT-TAB maintains high TCCF across both categorical fields (e.g., `final_status`, `currency`) and the numerical attribute `goal`, whereas hybrid baselines such as CTGAN/TabDDPM exhibit larger per-column variance, suggesting weaker global semantic coupling after decoupled generation. On Wine (Table 7), AFT-TAB preserves near-perfect consistency on semantically rich attributes like `variety` while remaining competitive on `points` and `price`; in contrast, the purely generative baseline GReaT suffers a substantial drop on `variety`, which dominates its overall degradation. On Cloth (Table 8), AFT-TAB delivers balanced improvements on hierarchical semantic attributes (e.g., `Class Name`, `Department Name`, `Division Name`) where serialization-based or purely generative methods often degrade, indicating that the RL-driven consistency reward effectively mitigates attribute-text misalignment under long-range dependencies.

## A.5 Computational Complexity

Table 5 presents a comparison of the computational complexity and efficiency of our AFT-Tab framework against various baseline methods on the Wine dataset. The experiments were conducted in a 4-GPU A800 (80G) environment. While memory usage is largely consistent across methods, AFT-Tab requires more training time due to the PPO policy optimization stage. However, this additional time cost is acceptable considering the substantial performance gains in cross-modal consistency (TCCF).

Table 5: Training overhead comparison on the Wine dataset.

| Method | GPUs | Training Time | TCCF |
|---|---|---|---|
| CTGAN+Qwen2.5-7B | 4 (A800 80G) | 57 min | 0.6552 |
| TabDDPM+Qwen2.5-7B | 4 (A800 80G) | 1 hour 15 min | 0.6582 |
| GReaT+Qwen2.5-7B | 4 (A800 80G) | 1 hour 26 min | 0.4440 |
| **AFT-Tab+Qwen2.5-7B** | 4 (A800 80G) | 2 hours 37 min | **0.9841** |

## A.6 Synthetic Data Samples

To provide a granular assessment of the generative fidelity, Table 9 presents a side-by-side comparison

of review samples from the real dataset versus those produced by our generative model. These samples span a diverse range of user attributes (e.g., age), metadata (e.g., ratings, titles), and unstructured text content.

A close inspection reveals that the synthetic reviews exhibit a remarkable degree of semantic and stylistic alignment with the ground truth. The generated texts successfully replicate the natural flow of colloquial language and reasonable subjective judgments characteristic of authentic user feedback. Notably, the model captures fine-grained contextual details such as specific body measurements, fit recommendations, and tactile descriptions of materials while also preserving complex emotional nuances, including ambivalence (e.g., expressing disappointment while retaining the item). This high level of coherence and specificity makes the synthetic samples challenging for humans to distinguish from real reviews, demonstrating the model's capability to effectively model the multidimensional joint distribution of structured attributes and unstructured text.

Table 6: TCCF results on the Kickstarter dataset

| | Method | final_status | country | currency | disable_communication | goal | Average |
|---|---|---|---|---|---|---|---|
| Kickstarter | Qwen2.5-7B+CLLM | 0.8837 | 0.7021 | 0.7828 | 0.6849 | 0.8430 | 0.7793 |
| | Qwen2.5-7B+CTGAN | 0.8157 | 0.9109 | 0.8973 | 0.8179 | 0.9571 | 0.8798 |
| | Qwen2.5-7B+TabDDPM | 0.7914 | 0.9032 | 0.9182 | 0.6279 | 0.9782 | 0.8438 |
| | Qwen2.5-7B+GReaT | 0.8173 | 0.7574 | 0.7726 | 0.7062 | 0.9635 | 0.8034 |
| | Qwen2.5-7B+AFT-TAB | 0.9359 | 0.8853 | 0.8784 | 0.7591 | 0.9592 | **0.8836** |

Table 7: TCCF results on the Wine dataset

| | Method | variety | country | points | price | Average |
|---|---|---|---|---|---|---|
| Wine | Qwen2.5-7B+CLLM | 0.9240 | 0.9253 | 0.9911 | 0.6314 | 0.8680 |
| | Qwen2.5-7B+CTGAN | 0.6552 | 0.9502 | 0.9492 | 0.8119 | 0.8416 |
| | Qwen2.5-7B+TabDDPM | 0.6582 | 0.9211 | 0.9337 | 0.9689 | 0.8705 |
| | Qwen2.5-7B+GReaT | 0.4440 | – | 0.8735 | 0.7263 | 0.6813 |
| | Qwen2.5-7B+AFT-TAB | 0.9841 | 0.7639 | 0.9825 | 0.9894 | **0.9300** |

Table 8: TCCF results on the Cloth dataset

| | Method | Rating | Age | Class Name | Department Name | Division Name | Positive Feedback Count | Average |
|---|---|---|---|---|---|---|---|---|
| Cloth | Qwen2.5-7B+CLLM | 0.7087 | 0.9241 | 0.8026 | 0.9368 | 0.9867 | 0.8875 | 0.8744 |
| | Qwen2.5-7B+CTGAN | 0.6505 | 0.9114 | 0.7463 | 0.6974 | 0.8887 | 0.9104 | 0.8008 |
| | Qwen2.5-7B+TabDDPM | 0.6505 | 0.8432 | 0.6841 | 0.7568 | 0.8242 | 0.8894 | 0.7747 |
| | Qwen2.5-7B+GReaT | 0.4115 | 0.8879 | 0.4445 | 0.3831 | 0.8188 | 0.9469 | 0.6488 |
| | Qwen2.5-7B+AFT-TAB | 0.9673 | 0.8697 | 0.9019 | 0.9121 | 0.9936 | 0.8681 | **0.9188** |

Table 9: Qualitative comparison of real vs. synthetic reviews. The synthetic samples demonstrate high fidelity in linguistic style, emotional coherence, and specific attribute details (e.g., fit, material).

| Age | Rating | Title | Review Text |
|---|---|---|---|
| ***Panel A: Real Data Samples*** | | | |
| 22 | 5 | Cannot get enough of this! | This is beautiful and fits perfect. It's even prettier in person, and extremely comfortable and high quality. I'm 5'3" and it hits right above the knee. The straps are adjustable and fit true to size. I bought an XS and it fits perfect. Must buy for a sweet little night gown! |
| 32 | 1 | Ugly design | So I bought this, even after seeing the other reviews because I love the designer... This fits just as others describe. It was not flattering to me at all, 5"2. Had to go back. |
| 85 | 3 | I'm a lil disappointed | Wish someone warned of the navy one being not as pictured!! I ordered the navy because the crocheted part appeared to be more intricate than the ivory. In reality, the navy knit is just like the ivory... I'm not a fan of tops that require a camisole but this top is a very "retailer" piece so most likely I will keep it. |
| ***Panel B: Synthetic Data Samples (Ours)*** | | | |
| 40 | 5 | Adorable pants | These are so cute! I am 5'1" and got the petite size. I would have needed the regular size in another pant I tried on at store. These fit nicely but not too tight and were not too short. They could go from work to out with friends. You do get a ton of fabric for this price point. |
| 31 | 1 | The fabric is too thin | This piece of clothing was extremely uncomfortable because of the very thin fabric. It would have been better if they had made it with silk or a sheer material but still covered in sequins. |
| 82 | 3 | Not as pictured | This dress is beautiful in the pictures. The color is rich but more of a purple or rose color then the picture shows. The top looks as shown on. The skirt is not at all as it is shown. It isn't flowered, it's just a dark floral print and it doesn't have pockets like shown. |