ICML
International Conference
On Machine Learning

Understanding and Mitigating Memorization in Diffusion Models for Tabular Data

Zhengyu Fang*, Zhimeng Jiang*, Huiyuan Chen, Xiao Li, Jing Li

Motivation

- **Key question:** Does memorization occur in **tabular** diffusion models, and if so, how can it be effectively mitigated?

- **Our contribution:**
- 1) Conducting the first comprehensive investigation into memorization behaviors within tabular diffusion models
- 2) Introduce **TabCutMixPlus**, a simple yet effective data augmentation technique to mitigate memorization

Metrics – Memorization Ratio

- **Distance**

$$d(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{M}\left(\text{norm}\left(\sqrt{\sum_{i \in \mathcal{F}_{num}} (\boldsymbol{x}_i - \boldsymbol{x}'_i)^2}\right) + \sum_{j \in \mathcal{F}_{cat}} \mathbf{1}(\boldsymbol{x}_j \neq \boldsymbol{x}'_j)\right).$$

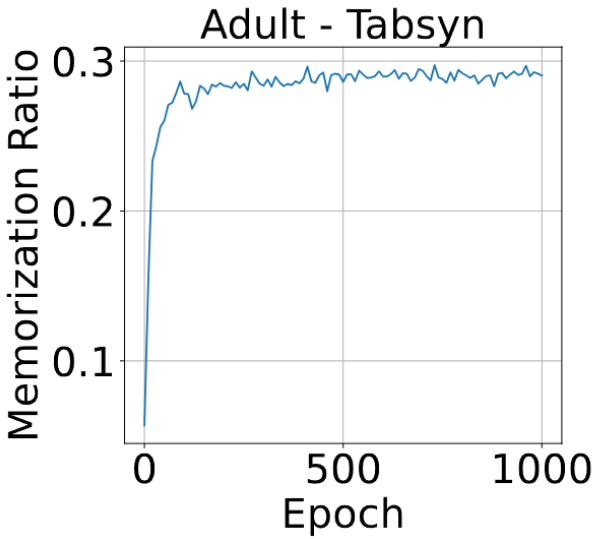
- **Distance Ratio**

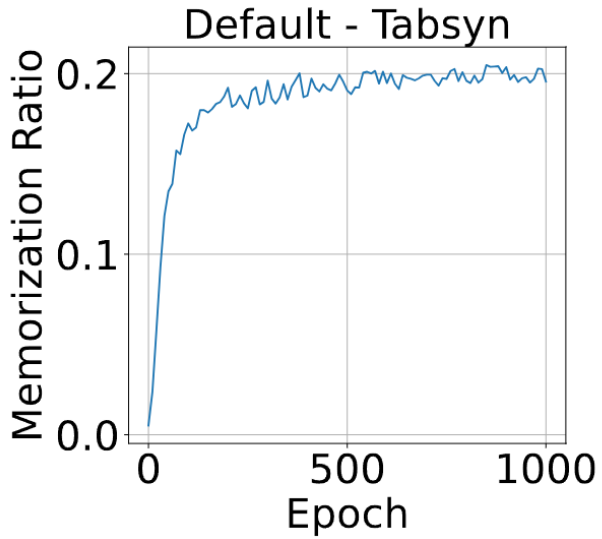
$$r(x) = \frac{d(x, \text{NN}_1(x, \mathcal{D}))}{d(x, \text{NN}_2(x, \mathcal{D}))}$$

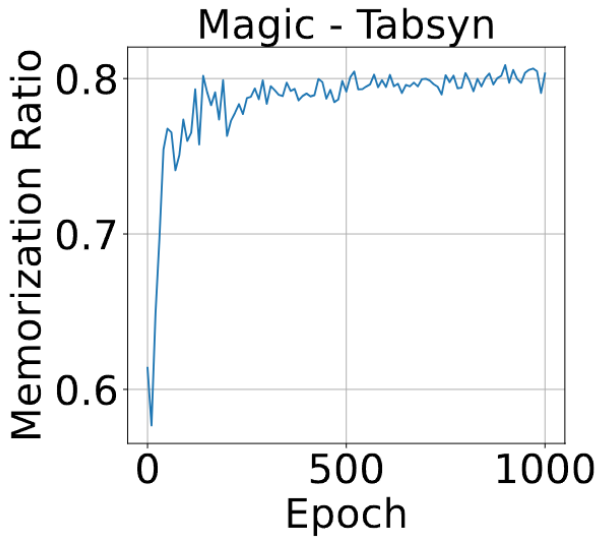
- **Memorization Ratio**

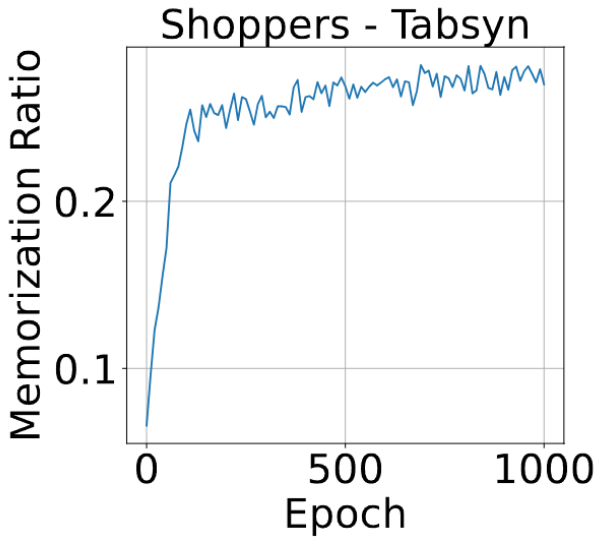
$$\text{Mem. Ratio} = \frac{1}{|\mathcal{G}|} \sum_{x \in \mathcal{G}} \mathbb{I}(r(x) < \frac{1}{3})$$

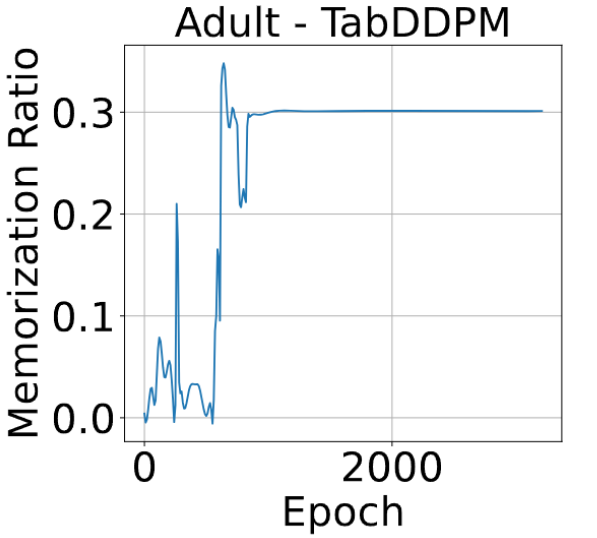
Preliminary

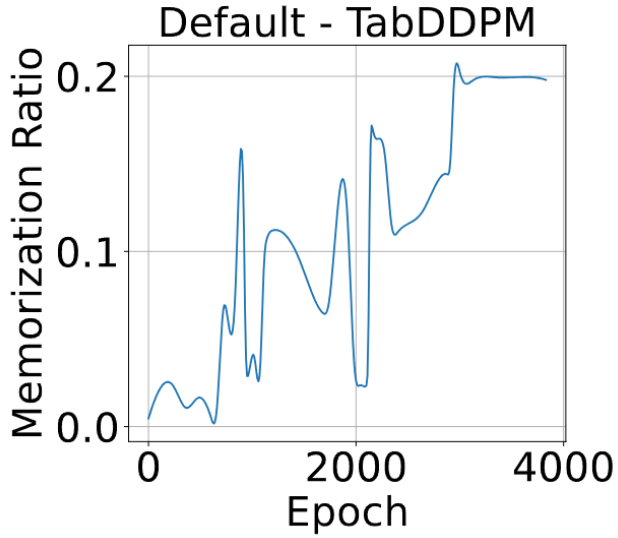


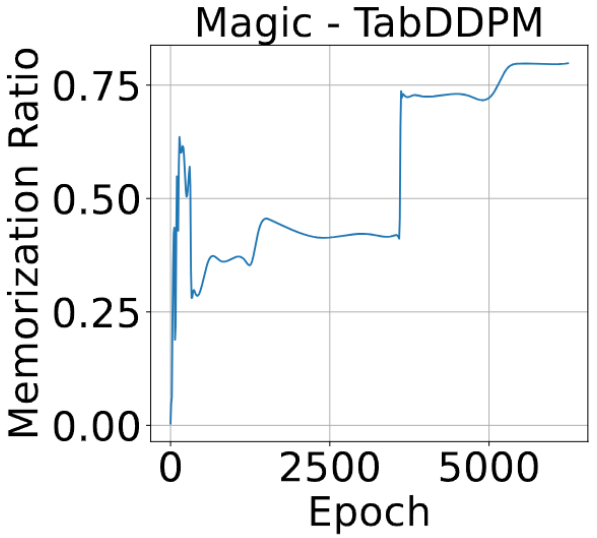


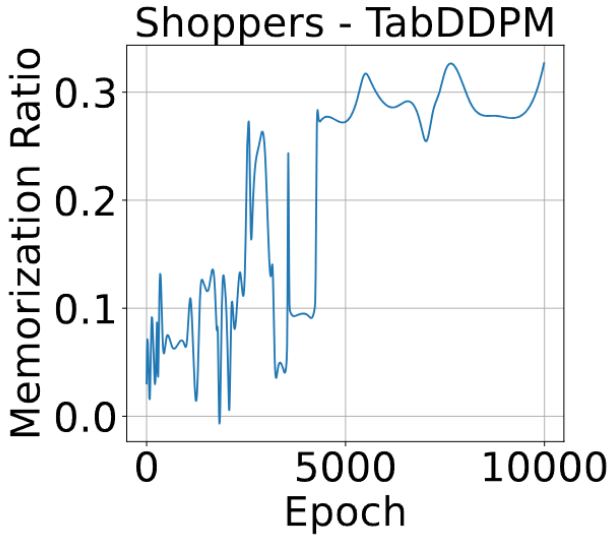




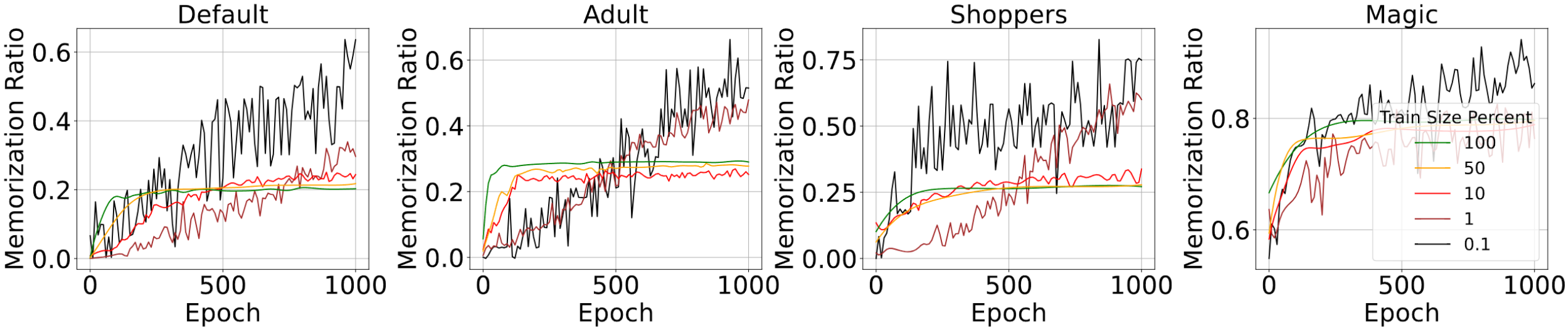


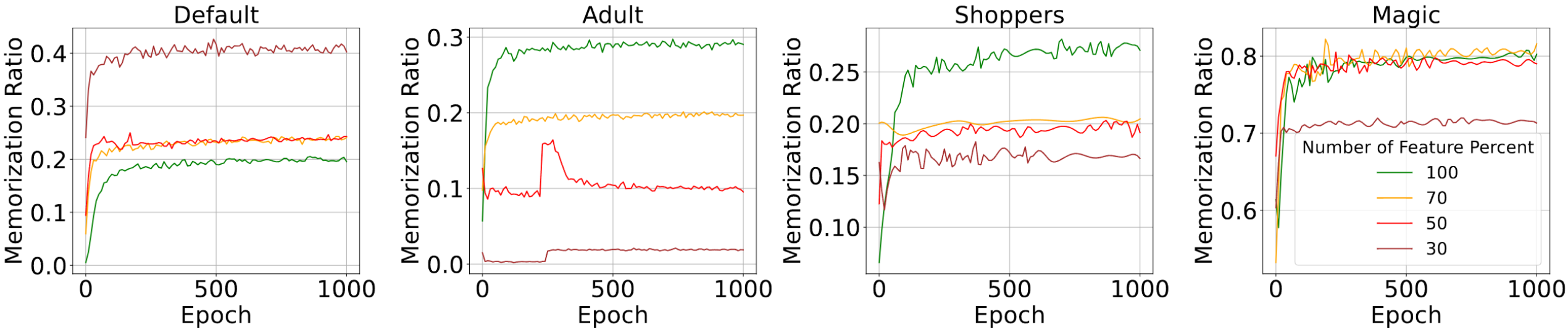






Preliminary





TabCutMixPlus

Algorithm 2 Pseudo-code of TabCutMixPlus

Require: Training set $\mathcal{D}$, Number of samples N

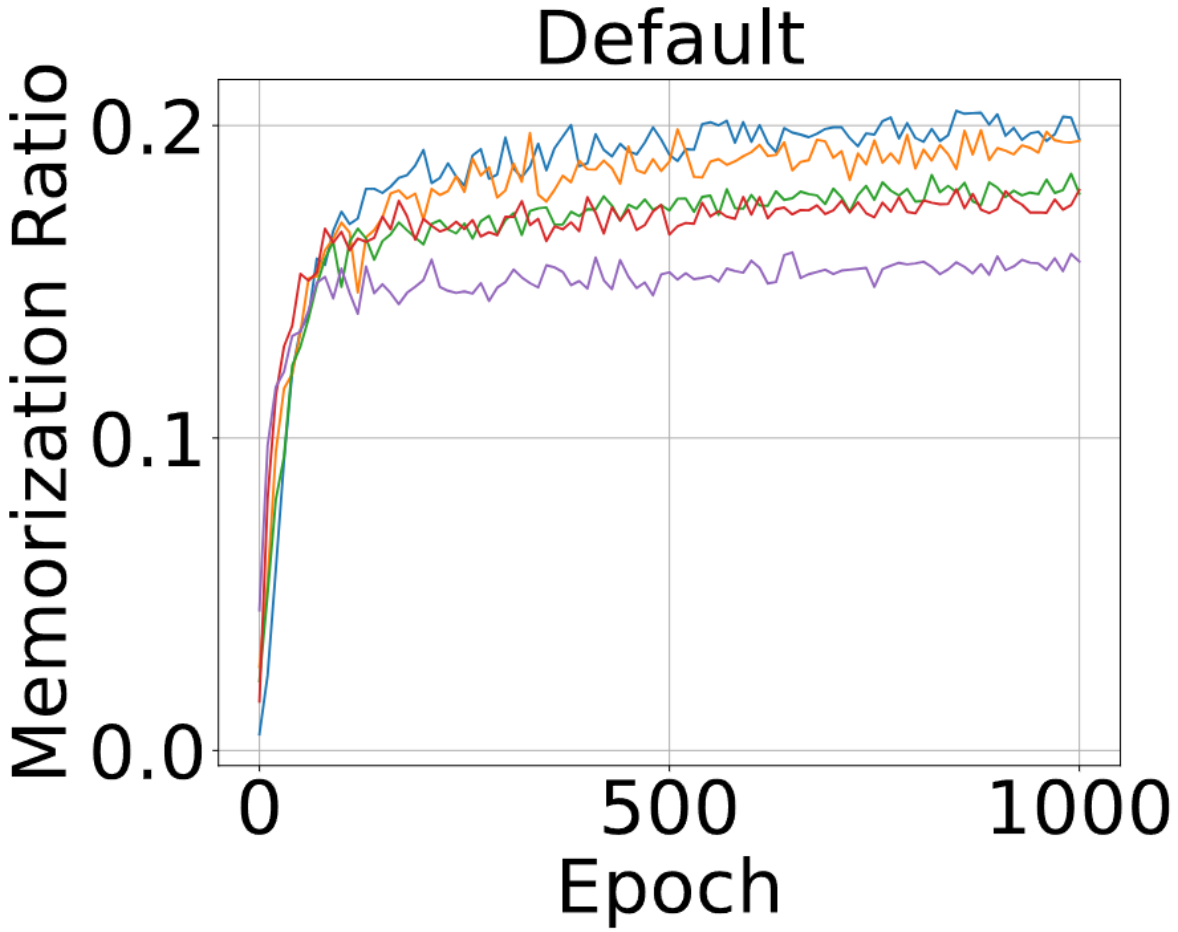
Augmented sample set $\tilde{\mathcal{D}} = \emptyset$
for $i = 1$ to N **do**
  Sample class c from $\{1, \cdots, C\}$ with prior class distribution; ▷ Keep class ratio after augmentation.
  Sample $(\boldsymbol{x}_A, y_A)$ and $(\boldsymbol{x}_B, y_B)$ from class c in $\mathcal{D}$; ▷ Randomly select two training samples from the same class.
  Calculate correlation metrics for features: (a) Pearson correlation coefficient for numerical feature; (b) Cramér's V based on contingency tables for categorical features; (c) ETA coefficient for numerical-categorical pairs.
  Perform hierarchical clustering on features using correlation metrics; ▷ Group features based on similarity.
  for each cluster k **do**
    Sample $\lambda \sim \text{Unif}(0, 1)$ and sampling binary mask M_k with Bernoulli distribution $\text{Bern}(\lambda)$; ▷ Proportion of features to exchange within cluster k.
    $\tilde{\boldsymbol{x}}_k \leftarrow M_k \odot \boldsymbol{x}_{A,k} + (1 - M_k) \odot \boldsymbol{x}_{B,k}$; ▷ Mix features in cluster k based on binary mask M_k.
    Add $\tilde{\boldsymbol{x}}_k$ to $\tilde{\boldsymbol{x}}$;
  end for
  $\tilde{y} \leftarrow c$; ▷ Assign the label of the new sample.
  $\tilde{\mathcal{D}} = \tilde{\mathcal{D}} \cup (\tilde{\boldsymbol{x}}, \tilde{y})$; ▷ Save the augmented sample.
end for
return New Training Set $\mathcal{D} \cup \tilde{\mathcal{D}}$

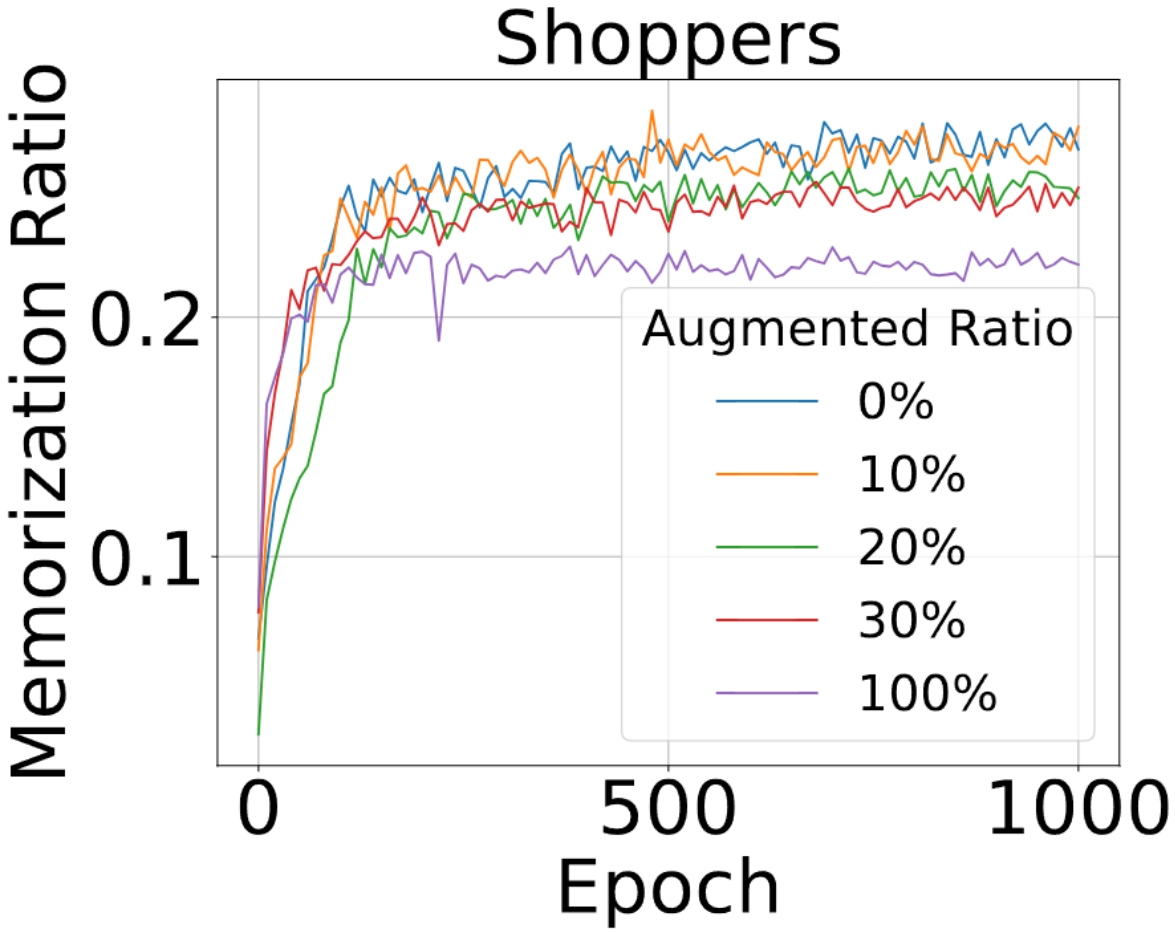
Experiments

Table 1. The overview performance comparison for tabular diffusion models on more datasets. "TCM" represents our proposed **TabCut-Mix** and "TCMP" represents **TabCutMixPlus**. "Mem. Ratio" represents memorization ratio. "Improv" represents the improvement ratio on memorization.

	Methods	Mem. Ratio (%) ↓	Improv.	MLE (%)↑	α-Precision(%)↑	β-Recall(%)↑	Shape Score(%)↑	Trend Score(%)↑	C2ST(%)↑	DCR(%)
Adult	STaSy	26.02 ± 0.89	-	90.54 ± 0.17	85.79 ± 7.85	34.35 ± 2.46	89.14 ± 2.29	86.00 ± 2.97	51.89 ± 14.87	50.46 ± 0.39
	STaSy+**Mixup**	24.89 ± 1.30	**4.37%** ↓	90.74 ± 0.06	90.00 ± 1.91	34.24 ± 2.47	90.28 ± 1.69	87.56 ± 1.06	52.61 ± 6.52	50.08 ± 0.59
	STaSy+**SMOTE**	22.92 ± 3.77	**11.91%** ↓	90.50 ± 0.24	85.81 ± 11.39	32.11 ± 5.13	86.91 ± 0.81	84.36 ± 2.36	45.12 ± 8.82	50.46 ± 0.20
	STaSy+**TCM**	20.89 ± 1.33	**19.71%** ↓	90.45 ± 0.30	85.39 ± 1.61	31.24 ± 0.97	88.33 ± 3.63	85.39 ± 4.03	45.49 ± 4.78	50.92 ± 0.39
	STaSy+**TCMP**	21.45 ± 2.60	**17.59%** ↓	90.72 ± 0.06	86.71 ± 4.12	32.63 ± 1.81	89.62 ± 1.55	86.05 ± 2.44	49.12 ± 9.95	50.75± 0.59
	TabDDPM	31.01 ± 0.18	-	91.09 ± 0.07	93.58 ± 1.99	51.52 ± 2.29	98.84 ± 0.03	97.78 ± 0.07	94.63 ± 1.19	51.56 ± 0.34
	TabDDPM+**Mixup**	30.04 ± 0.41	**3.14%** ↓	90.82 ± 0.12	95.78 ± 0.68	47.65 ± 1.35	98.02 ± 1.08	96.78 ± 1.33	93.65 ± 3.59	50.86 ± 0.86
	TabDDPM+**SMOTE**	28.98 ± 0.78	**6.56%** ↓	90.41 ± 0.36	94.93 ± 1.72	46.10 ± 0.65	93.40 ± 1.12	90.76 ± 1.76	80.75 ± 0.84	51.82 ± 0.56
	TabDDPM+**TCM**	27.55 ± 0.19	**11.16%** ↓	91.15 ± 0.06	94.97 ± 0.06	47.43 ± 1.46	98.65 ± 0.03	97.75 ± 0.07	85.61 ± 16.03	50.99 ± 0.65
	TabDDPM+**TCMP**	26.10 ± 2.11	**15.83%** ↓	90.54 ± 0.17	92.26 ± 6.97	43.49 ± 3.74	95.10 ± 4.27	91.50 ± 6.53	84.76 ± 10.12	50.68 ± 0.89
	TabSyn	29.26 ± 0.23	-	91.13 ± 0.09	99.31 ± 0.39	48.00 ± 0.22	99.33 ± 0.09	98.19 ± 0.50	98.68 ± 0.41	50.42 ± 0.27
	TabSyn+**Mixup**	28.29 ± 0.28	**3.30%** ↓	90.75 ± 0.24	98.63 ± 0.81	45.73 ± 2.67	98.30 ± 0.90	97.91 ± 0.12	98.05 ± 2.22	50.97 ± 1.10
	TabSyn+**SMOTE**	27.10 ± 0.15	**7.36%** ↓	89.97 ± 0.76	98.60 ± 0.50	44.72 ± 0.45	94.47 ± 0.57	91.74 ± 0.42	82.55 ± 0.71	48.42 ± 0.78
	TabSyn+**TCM**	27.03 ± 0.22	**7.60%** ↓	91.09 ± 0.17	99.04 ± 0.42	44.95 ± 0.42	99.40 ± 0.07	98.51 ± 0.08	89.18 ± 1.94	50.67 ± 0.11
	TabSyn+**TCMP**	25.99 ± 0.52	**11.17%** ↓	90.96 ± 0.16	98.43 ± 1.04	43.23 ± 2.96	98.38 ± 0.91	96.53 ± 1.47	93.39 ± 6.01	50.30 ± 0.78

Experiments – Different Augmented Ratio





ICML
International Conference
On Machine Learning

Thank you!

- https://github.com/fangzy96/TabCutMix