SECRET-PROTECTED EVOLUTION FOR DIFFERENTIALLY PRIVATE SYNTHETIC TEXT GENERATION

Tianze Wang [1,2] Zhaoyu Chen [1] Jian Du [1] Yingtai Xiao [1] Linjun Zhang [2] Qiang Yan [1]

[1]TikTok [2]Department of Statistics, Rutgers University

Introduction

Text data has become extremely valuable for large language models (LLMs). However, a lot of high-quality text in the real world is private and cannot be freely used due to privacy concerns.

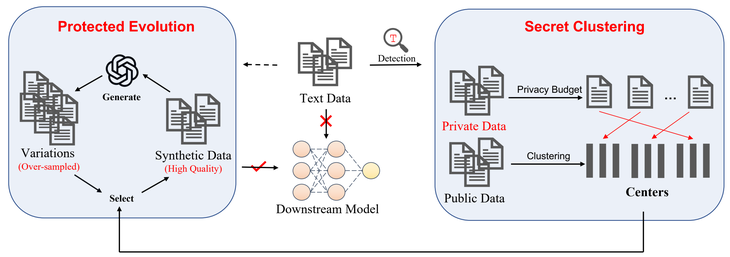
To address this, differentially private (DP) synthetic text generation has been proposed. Yet, existing DP generation imposes uniform guarantees that often overprotect non-sensitive content, resulting in substantial utility loss and computational overhead. Furthermore, existing Private Evolution (PE) [3] pipelines are highly inefficient due to redundant pairwise similarity computations and iterative data processing.

Secret-Protected Evolution

We adopt the secret protection concept in [2], which provides privacy guarantees tailored to specific sensitive information rather than uniform dataset membership. With predefined secrets, non-sensitive public data is leveraged freely without protection. This significantly enhances efficiency, reserving the privacy budget exclusively for secret-related adjustments rather than applying it uniformly across the dataset.

Protection is formulated by calibrating against the adversary's prior, directly bounding the reconstruction success probability. This conceptually relaxes Gaussian Differential Privacy (GDP) [1] by enforcing protection only at a specific point on the trade-off curve. The relaxation weakens the privacy constraint but yields higher utility and practical reconstruction protection.

Building on these principles, we introduce the **Secret-Protected Evolution (SecPE)** framework. As illustrated in Figure 1, SecPE consists of two key components: (1) *Secret Clustering*, which detects sensitive attributes and forms representative centers by updating public clusters with noisy private data; and (2) *Protected Evolution*, which iteratively samples variations from high-quality synthetic data, evaluates them against the noisy representatives, and selects the best candidates. This architecture preserves the practicality of Private Evolution while effectively shifting the privacy focus toward secrets.



Key Contributions

- we propose a private synthetic data generation framework that emphasizes *secret protection* rather than canonical DP, thereby improving utility by reducing the noise typically required under DP;
- we develop a secret-protected clustering method that substantially reduces runtime complexity compared to the PE approach, enabling scalability to larger datasets while maintaining competitive performance;
- we empirically demonstrate that SecPE achieves higher efficiency, lower Fréchet Inception Distance (FID), and better downstream accuracy than μ-GDP–based PE baselines under the same reconstruction guarantees.

Theoretical Foundation: Secret Protection

Definition (Secret Protection): Let D be a training dataset where each sample may contain secrets from $S = \{s_1, \ldots, s_m\}$. For a secret $s_j \in S$, let π_j denote a prior distribution such that $Pr(D_j^k) \leq p_j$. A randomized mechanism $\mathcal{A}$ satisfies $(\mathbf{p}, \mathbf{r})$-secret protection if, for any reconstruction attack B:

$$Pr_{D_j \sim \pi_j, \mathcal{A}}[B(\mathcal{A}(D_j)) = s_j] \leq r_j \quad \forall j$$

Here, $\mathbf{p}$ and $\mathbf{r}$ are vectors, allowing us to set tailored prior and posterior bounds for each distinct secret.

Relaxation of GDP: We align the neighboring relation so that $D \simeq D_j$ differ only by a single element: specifically, $s_j \in x \in D$ and $s_j \notin x' \in D_j$. Consequently, any μ-GDP mechanism provides $(\mathbf{p}, \mathbf{r})$-secret protection, where:

$$r_j = 1 - \Phi\left(\Phi^{-1}(1 - p_j) - \mu\right)$$

The SecPE Pipeline

SecPE modifies the traditional PE pipeline in two fundamental ways to achieve efficiency and secret-level guarantees:

1. **Noise Calibration:** Instead of canonical DP sensitivity noise, we compute a tailored noise parameter σ via a linear program that bounds the blow-up function for the target $(\mathbf{p}, \mathbf{r})$.
2. **Representative Voting:** We replace the direct pairwise similarity voting with clustering, evaluating candidates only against a small set of noisy representative centers.

Algorithm 1 Secret-Protected Evolution (SecPE)

Input: Dataset $D_{pri} \cup D_{pub}$, secrets S_{sec}, embedding model Ψ, N_{syn}, L.
Initialize: $S_0 \leftarrow \text{RANDOM}(N_{syn} \times L)$.
for $t = 0, 1, \ldots, T-1$ **do**
1. **Secret Clustering:** Obtain noisy cluster centers and sizes $\{(\tilde{e}_k, \tilde{n}_k)\}_{k=1}^{K}$ using D_{pub} and D_{pri}.
2. **Similarity Voting:** Embed candidates $S_t \to E_t$. Assign each $e_i \in E_t$ to the nearest noisy center $\tilde{e}_k$ and update vote histograms.
3. **Selection:** $S_{syn}^{t+1} \leftarrow$ Top N_{syn} samples according to the histogram.
4. **Variation:** $S_{t+1} \leftarrow [\text{VARIATION}(S_{syn}^{t+1}, L), S_{syn}^{t+1}]$.

end for
Output: Synthetic text dataset S_{syn}^{T}.

Computational Efficiency

SecPE leverages Secret Clustering with K anchors to reduce traditional PE complexity from $O(|D_{pri}|N_{syn})$ to $O(KN_{syn})$, where $K \ll |D_{pri}|$. This yields massive runtime savings, as illustrated in the following table:

Table 1. LLM generation time and histogram computation time (seconds) for one epoch.

Time (sec)	OpenReview		PubMed		Yelp	
	LLM	Histogram	LLM	Histogram	LLM	Histogram
Aug-PE	1698.7	126.9	828.5	32.2	347.1	30126.4
SecPE	1693.1	**1.5**	830.8	**0.5**	347.6	**2.3**

Downstream Task Performance

We evaluate synthetic text utility by fine-tuning a classifier to predict OpenReview research areas and ratings. As Table 2 shows, SecPE consistently outperforms the Aug-PE baseline in accuracy across privacy budgets $(\mathbf{r}/\mathbf{p})$, demonstrating superior utility under equal reconstruction guarantees.

Table 2. Performance comparison of downstream tasks (classification accuracy) on OpenReview.

LLM	Method	$r/p = 2$		$r/p = 10$		$r/p = 50$		$r/p = \infty$	
		Area	Rating	Area	Rating	Area	Rating	Area	Rating
GPT2	Aug-PE	29.06	25.70	27.94	27.12	32.48	27.88	**41.06**	28.70
	SecPE_{15}	**30.77**	30.26	31.88	**30.70**	30.34	28.27	39.02	28.38
	SecPE_{20}	28.98	**31.38**	32.67	28.23	30.30	29.56	38.74	**30.49**
	SecPE_{25}	30.34	29.24	**34.81**	30.66	32.48	**30.31**	38.60	**30.49**
Qwen-2.5-1.5B	Aug-PE	32.70	25.55	32.23	25.80	36.49	28.52	40.20	28.09
	SecPE_{15}	38.34	**27.73**	**38.67**	26.02	36.09	**30.95**	36.03	**32.03**
	SecPE_{20}	37.17	26.94	36.95	26.44	35.85	29.52	39.63	28.30
	SecPE_{25}	**38.92**	27.66	37.03	**27.82**	**40.24**	28.81	**40.24**	28.86

Real-Synthetic Similarity

SecPE requires less noise to attain the same level of reconstruction protection. Consequently, it achieves a lower FID than GDP-based baselines under strict privacy settings, generating higher-fidelity synthetic text.

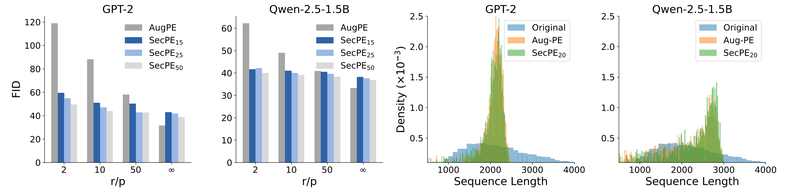


Figure 1. Left: FID relative to original data under varying $\mathbf{r}/\mathbf{p}$. Right: Synthetic sequence-length distributions for SecPE compared with original data.

Conclusion

- **Tighter Trade-offs:** SecPE calibrates protection at the secret level, relaxing Gaussian DP and improving text fidelity.
- **Highly Scalable:** Secret clustering dramatically accelerates the PE pipeline.
- **Practical Privacy:** Unlocks more effective privacy-preserving generation when sensitive content is sparse but highly consequential.

References

[1] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy, 2019.

[2] Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions, 2022.

[3] Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model apis 1: Images, 2025.

jian.du@tiktok.com