GUIDED DIFFUSION FOR TRUSTWORTHY SYNTHETIC TEXT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid evolution of Large Language Models is fueled by massive datasets, yet this reliance poses severe privacy risks, as models can inadvertently memorize and even directly output sensitive training data. Synthetic text generation offers a promising solution to this challenge; however, existing methods struggle to reconcile rigorous privacy guarantees with high data utility. In this paper, we propose a new paradigm for private synthetic text generation: Guided Diffusion. Leveraging the unique steerability of Diffusion Language Models, we introduce novel sampling and re-masking procedures that actively guide the iterative denoising steps. These mechanisms steer the generative process to reconstruct fluent text that adheres to a target semantic distribution. We demonstrate that our framework is naturally compatible with Differential Privacy (DP): once the semantic embeddings of the private corpus are protected, the DP guarantee carries over to the generated text by post-processing. Empirically, we show that our approach generates high-quality synthetic text with precise semantic alignment. Furthermore, we validate its effectiveness in DP synthetic text generation: across two datasets and two diffusion backbones, it outperforms Private Evolution and DP-SGD-based fine-tuning in distributional fidelity and downstream utility, while generating significantly faster than Private Evolution.

1 INTRODUCTION

The rapid evolution of Large Language Models (LLMs) has fundamentally reshaped Natural Language Processing, a paradigm shift driven primarily by the massive scale of training data. However, this reliance on web-scale data accumulation comes with a steep risk of exposing sensitive information. Real-world datasets, ranging from private medical records and financial reports to personal communication logs, are invaluable for training robust models, but they often contain sensitive personal information that must not be exposed. Recent studies have demonstrated that generative models are prone to memorization (Carlini et al., 2021), where they can inadvertently regenerate training samples verbatim. This vulnerability poses severe ethical concerns and violates strict privacy regulations such as GDPR (Voigt & Von dem Bussche, 2017) and CCPA (California State Legislature, 2018).

To address this, synthetic text generation has emerged as a promising solution. Ideally, a synthetic dataset mirrors the statistical properties and semantic richness of the real distribution while ensuring that no individual data point is leaked. However, existing approaches typically struggle to balance these objectives, falling into two broad categories. The first relies on *heuristic sanitization*, such as entity masking or scrubbing (Lison et al., 2021). While these methods preserve linguistic fluency, they lack formal guarantees and have been repeatedly shown to fail against sophisticated reconstruction attacks (Carlini et al., 2021; Balle et al., 2022; Zhou et al., 2024). The second category employs *rigorous formalisms* like Differential Privacy (DP), often applied via DP-SGD (Abadi et al., 2016). While provably privacy-preserving, these methods frequently result in a catastrophic loss of utility when applied to text, yielding outputs that are repetitive or incoherent.

One fundamental limitation behind such heavy utility compromise is that provable privacy guarantees such as DP rely on noise to ensure *statistical closeness*. For example, DP requires statistical indistinguishability between the leakage produced by any two adjacent datasets for individual privacy. To support this, existing noise mechanisms require calibration to the absolute divergence or distance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, text representations in vocabulary space have extremely high variance. For instance, the sentences "The movie is excellent" and "This is a superb film" are semantically identical yet possess disjoint vocabulary representations (orthogonal one-hot vectors). Standard distance metrics in this space fail to capture semantic proximity, and the required noise scale to mask these differences is often prohibitively large, destroying the linguistic structure of the output.

In this paper, we propose a novel framework for synthetic text generation that seamlessly integrates with off-the-shelf DP clustering algorithms to provide rigorous privacy guarantees. Our core insight is that, while raw text resides in a high-dimensional, discrete space, its semantic embedding can be captured on a low-dimensional, continuous manifold. We observe that DP mechanisms can achieve significantly higher utility if applied within this compact embedding space instead of the sparse vocabulary space. The challenge then shifts to a generation problem:

How to convert privatized, low-dimensional semantic embeddings into coherent, discrete text?

While latent-conditioned generation is well-established in computer vision (e.g., image editing (Rombach et al., 2022; Ye et al., 2023)), transferring this success to text is challenging due to the discrete nature of language. Auto-regressive (AR) LLMs, which generate tokens sequentially, are notoriously difficult to steer globally using a fixed embedding vector. To bridge this gap, we leverage the unique properties of Diffusion Language Models (DLMs; Li et al., 2022a; Sahoo et al., 2024; Nie et al., 2025; You et al., 2025; Song et al., 2025; Yang et al., 2025b). Unlike AR models, DLMs predict blocks of tokens (sentences) through an iterative denoising process. This paradigm allows us to effectively steer the generation by minimizing the distance between the generated sequence and the target semantic embedding at each diffusion step.

Specifically, we introduce a Guided Diffusion framework that generates coherent text conditioned on semantic embeddings. To enable precise control, we propose two novel components for masked DLMs: *semantic-aware sampling* and *semantic-aware re-masking*. They allow the DLM to align its discrete outputs with the continuous embedding guidance signal. By integrating DP mechanisms directly into the embedding stage, we present the first framework capable of utilizing DLMs for synthetic data generation with rigorous privacy protection. In summary, we make the following contributions:

- **Guided Diffusion Framework**: We propose a novel framework that, for the first time, exploits the steering capabilities of DLMs to generate coherent text conditioned on semantic embeddings.
- **DP Integration**: Our Guided Diffusion framework naturally accommodates off-the-shelf DP clustering mechanisms in the embedding space, and the DP guarantee carries over to the synthetic text by post-processing.
- **Empirical Effectiveness**: We validate our framework through extensive experiments, demonstrating superior synthetic data quality and strong downstream utility in DP synthetic text generation: across two datasets and two diffusion backbones, it outperforms Private Evolution and DP-SGD-based baselines in distributional fidelity and downstream utility, while generating substantially faster than Private Evolution. Code for reproducing our results is provided in the supplementary material.

2 PRELIMINARIES

Diffusion Language Models. Diffusion Language Models (DLMs) are non-autoregressive generative models that synthesize text through an iterative denoising process. Unlike AR LLMs that generate tokens sequentially from left to right, DLMs generate a sequence of L tokens in parallel by progressively refining a noisy input state.

We focus on the denoising process, illustrated in Figure 1. Let $\boldsymbol{x}_t$ denote the token sequence at step t, where t goes from T down to 0. The process begins with $\boldsymbol{x}_T$, a fully masked sequence consisting entirely of the special token `[MASK]`. The core component is a mask predictor model $\mathcal{M}_\theta$, which takes a partially masked sequence $\boldsymbol{x}_t$ as input and outputs a probability matrix $\mathbf{P}_t \in \mathbb{R}^{L\times|\mathcal{V}|}$ for L positions over the vocabulary $\mathcal{V}$. The transition from step t to $t-1$ involves two critical operations:

Sampling. Given the current state $\boldsymbol{x}_t$ (e.g., "This paper `[MASK]` a `[MASK]` approach..."), the model $\mathcal{M}_\theta$ predicts the probability distribution for every token position. From this probability matrix $\mathbf{P}_t$, we

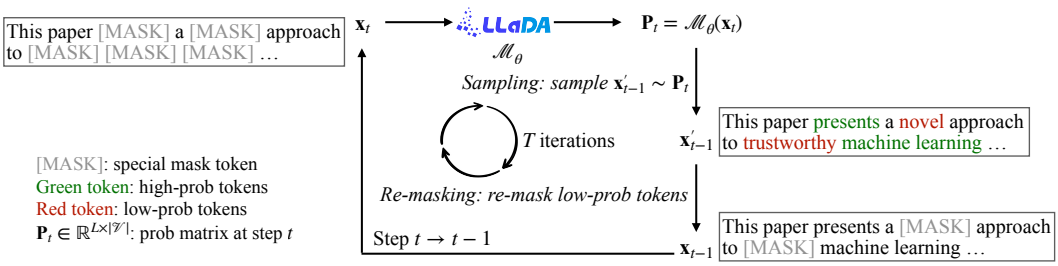

Figure 1: High-level denoising pipeline for DLMs. At each step t, the model predicts token probabilities for the masked input $\boldsymbol{x}_t$. A candidate sequence $\boldsymbol{x}'_{t-1}$ is sampled, and low-confidence tokens are re-masked to produce the refined state $\boldsymbol{x}_{t-1}$.

independently sample a candidate token for each masked position to form a fully populated candidate sequence $\boldsymbol{x}'_{t-1}$, as an intermediate state (see Figure 1).

Re-Masking. Since early predictions may be inaccurate, we do not accept all sampled tokens immediately. We obtain a confidence score (probability) for each token in $\boldsymbol{x}'_{t-1}$. As shown in Figure 1, tokens with high confidence (depicted in green) are retained, while tokens with low confidence (depicted in red) are reverted to the `[MASK]` token. This produces the input state $\boldsymbol{x}_{t-1}$ for the next iteration. This cycle repeats until $t = 0$ and all masks are removed, resulting in the final coherent sentence.

Differential Privacy. Differential Privacy (DP; Dwork et al., 2006) bounds the influence of any single record on a randomized mechanism's output. We present the formal definition for (ε, δ)-DP.

Definition 2.1 ((ε, δ)-DP)**.** A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if for all neighboring datasets $\mathcal{X}, \mathcal{X}'$ differing in a single record and all measurable sets S,

$$\Pr[\mathcal{M}(\mathcal{X}) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{X}') \in S] + \delta.$$

Smaller ε and δ correspond to stronger privacy guarantees: an adversary distinguishing two neighboring datasets must accept a higher Type-I (or Type-II) error rate at any fixed Type-II (or Type-I) budget. Our framework relies on two standard properties.

(i) The *Gaussian mechanism* (Balle & Wang, 2018) $\mathcal{M}(\mathcal{X}) = f(\mathcal{X}) + \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$ on a vector-valued query $f : \mathcal{X}^n \to \mathbb{R}^d$ with L_2 sensitivity Δ achieves (ε, δ)-DP with

$$\delta = \Phi\left(\tfrac{\Delta}{2\sigma} - \tfrac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\,\Phi\left(-\tfrac{\Delta}{2\sigma} - \tfrac{\varepsilon\sigma}{\Delta}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

(ii) *Post-processing invariance*: any function applied to the output of an (ε, δ)-DP mechanism that does not access the private input incurs no additional privacy loss. In other words, once the noisy embeddings are released under DP, any text generated from them alone, without touching the private data again, inherits the same DP guarantee.

3 Guided Diffusion Language Models

In this section, we present our framework of Guided Diffusion. The primary objective is to steer the generative process such that the resulting text $\boldsymbol{x}$ achieves high similarity to a target semantic embedding E_{tgt}, while preserving linguistic fluency. To achieve this, we introduce two novel mechanisms – *Semantic-Aware Sampling* and *Semantic-Aware Re-Masking* – and integrate them directly into the iterative denoising pipeline of standard DLMs.

Let $\mathcal{M}_\theta$ denote the diffusion mask predictor. At each diffusion step t, the model receives a partially masked sequence $\boldsymbol{x}_t$ and outputs a probability matrix $\mathbf{P}_t \in \mathbb{R}^{L\times|\mathcal{V}|}$, where each row represents the probability distribution over the vocabulary $\mathcal{V}$ for the corresponding token position. To output a coherent token block of length L, DLMs typically employ sampling and re-masking steps and iteratively remove the noise, which is the `[MASK]` token, in the input. Now we introduce our Semantic-Aware Sampling and Semantic-Aware Re-Masking designs.

3.1 SEMANTIC-AWARE SAMPLING

Standard DLMs employ independent token sampling, where a candidate sequence is formed by sampling from the probability distribution at each token position independently. While efficient, this approach is agnostic to the global semantic trajectory required for our specific guidance task.

To enforce semantic alignment, we propose a two-stage rejection-sampling approach that actively selects candidates balancing exploration, fluency, and semantic distance. We present the algorithm in Algorithm 1. Specifically, at diffusion step t, given the probability matrix $\mathbf{P}_t$, we first generate a pool of αN candidate sequences $\{\hat{\boldsymbol{x}}'^{(k)}\}_{k=1}^{\alpha N}$ via independent sampling: $\hat{\boldsymbol{x}}'^{(k)} \sim \mathbf{P}_t$.

We enforce that all αN candidate sequences differ by at least one token. We then evaluate these candidates using a hybrid scoring function to filter for quality and diversity:

$$S_k = a \cdot \mathcal{L}(\hat{\boldsymbol{x}}'^{(k)}) + b \cdot \mathcal{D}(\hat{\boldsymbol{x}}'^{(k)}), \quad k \in [\alpha N],$$

where $\mathcal{L}(\cdot)$ denotes the log-probability of the sequence under the model $\mathcal{M}_\theta$, and $\mathcal{D}(\cdot)$ measures the Hamming distance between $\hat{\boldsymbol{x}}'^{(k)}$ and the highest likelihood sequence from $\mathbf{P}_t$. a and b are scalar weights for the two scoring terms; the values used throughout our experiments are reported in Appendix C, and Appendix D shows that generation quality is insensitive to the ratio b/a over the tested range. We sort the candidates by S_k and retain the top N sequences, denoted as $\{\hat{\boldsymbol{x}}^{(k)}\}_{k=1}^{N}$.

Algorithm 1 Semantic-Aware Sampling

Require: Prob. matrix $\mathbf{P}_t$, target embedding E_{tgt}, constants α, a, b, N, embedding model $\mathcal{E}$
1: Generate pool $\mathcal{C} = \{\hat{\boldsymbol{x}}'^{(k)}\}_{k=1}^{\alpha N}$ via $\hat{\boldsymbol{x}}'^{(k)} \sim \mathbf{P}_t$
2: Compute $S_k = a \cdot \mathcal{L}(\hat{\boldsymbol{x}}'^{(k)}) + b \cdot \mathcal{D}(\hat{\boldsymbol{x}}'^{(k)}), k \in [\alpha N]$
3: Filter $\mathcal{C}$ to top N candidates $\{\hat{\boldsymbol{x}}^{(k)}\}$ based on S_k
4: Compute $E^{(k)} = \mathcal{E}(\hat{\boldsymbol{x}}^{(k)})$ for filtered candidates
5: Select $\hat{\boldsymbol{x}}^* = \hat{\boldsymbol{x}}^{(k^*)}$, $k^* = \arg\min_k ||E^{(k)} - E_{\text{tgt}}||_2$
6: **Return** $\hat{\boldsymbol{x}}^*$

Next, we select the candidate that best aligns with the target semantics. For each retained candidate $\hat{\boldsymbol{x}}^{(k)}$, we compute its semantic embedding $E^{(k)} = \mathcal{E}(\hat{\boldsymbol{x}}^{(k)})$, where $\mathcal{E}(\cdot)$ is the pre-trained semantic encoder. We evaluate the alignment with the target embedding E_{tgt} using the Euclidean distance $d_k = ||E^{(k)} - E_{\text{tgt}}||_2$. The optimal candidate $\hat{\boldsymbol{x}}^*$ is selected as the minimizer of this distance:

$$\hat{\boldsymbol{x}}^* = \hat{\boldsymbol{x}}^{(k^*)} \quad \text{where} \quad k^* = \arg\min_{k \in [N]} d_k$$

This selected sequence $\hat{\boldsymbol{x}}^*$ serves as the base estimation for the current step, effectively steering the generation manifold toward the target semantics before the re-masking phase.

3.2 SEMANTIC-AWARE RE-MASKING

After obtaining the candidate sequence $\hat{\boldsymbol{x}}^*$, the denoising process requires a re-masking step to transition to the next noisy state $\boldsymbol{x}_{t-1}$. Standard strategies typically mask tokens based on low model confidence (low prediction probability) or simple random selection. However, these heuristics may inadvertently discard tokens that are critical for semantic alignment or retain tokens that steer the text away from the target semantic embedding E_{tgt}.

To address this issue, we propose a contribution-based re-masking algorithm. Our insight is to strictly retain tokens that maximally contribute to minimizing the distance to the target semantic embedding, while masking those that are less relevant.

We present the algorithm in Algorithm 2. Let $w_1, w_2, \cdots, w_L$ be the L tokens in $\hat{\boldsymbol{x}}^*$: $\hat{\boldsymbol{x}}^* = [w_1, w_2, \ldots, w_L]$ and let $E_{\text{base}} = \mathcal{E}(\hat{\boldsymbol{x}}^*)$ denote its embedding. To quantify the contribution of the i-th token w_i, we perform a granular ablation: we replace w_i with the special `[MASK]` token to create a perturbed sequence $\hat{\boldsymbol{x}}^*_{\setminus i}$ and compute its embedding $E_{\setminus i} = \mathcal{E}(\hat{\boldsymbol{x}}^*_{\setminus i})$. We define the importance score Δ_i as the *change* in distance to the target induced by ablating w_i:

$$\Delta_i = \left\|E_{\setminus i} - E_{\text{tgt}}\right\|_2 - \|E_{\text{base}} - E_{\text{tgt}}\|_2, \quad i \in [L]$$

Algorithm 2 Semantic-Aware Re-Masking

Require: Sequence $\hat{\boldsymbol{x}}^*$, target embedding E_{tgt}, embedding model $\mathcal{E}$, denoise schedule ρ_t
1: Compute base embedding $E_{\text{base}} = \mathcal{E}(\hat{\boldsymbol{x}}^*)$
2: **for** position $i = 1$ **to** L **do**
3: Ablate token $w_i \rightarrow$ `[MASK]` to get $\hat{\boldsymbol{x}}^*_{\setminus i}$
4: Compute importance $\Delta_i = ||\mathcal{E}(\hat{\boldsymbol{x}}^*_{\setminus i}) - E_{\text{tgt}}||_2 - ||E_{\text{base}} - E_{\text{tgt}}||_2$
5: Identify indices $\mathcal{I}_{\text{mask}}$ corresponding to the $\rho_t/2$ smallest scores Δ_i, where $|\mathcal{I}_{\text{mask}}| = \rho_t/2$
6: Update $\boldsymbol{x}_{t-1}$ by setting $w_i \leftarrow$ `[MASK]` for all $i \in \mathcal{I}_{\text{mask}}$
7: Identify indices $\mathcal{I}_{\text{low-prob}}$ corresponding to the $\rho_t/2$ lowest prediction probabilities, where $|\mathcal{I}_{\text{low-prob}}| = \rho_t/2$ and $\mathcal{I}_{\text{mask}} \cap \mathcal{I}_{\text{low-prob}} = \emptyset$
8: Update $\boldsymbol{x}_{t-1}$ by setting $w_i \leftarrow$ `[MASK]` for all $i \in \mathcal{I}_{\text{low-prob}}$
9: **Return** $\boldsymbol{x}_{t-1}$

A large positive Δ_i means masking w_i moves the sequence away from the target, indicating that w_i is critical to alignment. A near-zero Δ_i means w_i is irrelevant to alignment, and a negative Δ_i means masking w_i actually *improves* alignment, so the token is beneficial to mask. In accordance with the diffusion noise schedule, we retain the tokens with the largest Δ_i and re-mask the remaining tokens. Specifically, at step t, when the diffusion noise schedule calls for masking ρ_t tokens, we first re-mask the $\rho_t/2$ tokens with the smallest Δ_i and then re-mask an additional $\rho_t/2$ low-probability tokens. This simultaneously ensures semantic alignment and effective denoising.

By iterating these semantic-aware sampling and re-masking steps, we ensure that the generative process converges to a coherent sentence where the corresponding semantic embedding is close to the target reference embedding vector E_{tgt}.

Necessity of Both Components. Semantic-aware sampling and re-masking play complementary roles: sampling steers candidate selection toward the target manifold, while re-masking preserves semantically critical tokens. We verify that disabling either component substantially degrades both semantic alignment (§ 3.3) and downstream DP utility (§ 4.2), confirming both stages are essential.

3.3 Semantic Alignment and Quality Preservation of Guided Diffusion

To demonstrate that Guided Diffusion can effectively generate high-quality sentences while closely aligning with a target semantic embedding, we conduct a comprehensive evaluation on standard benchmarks.

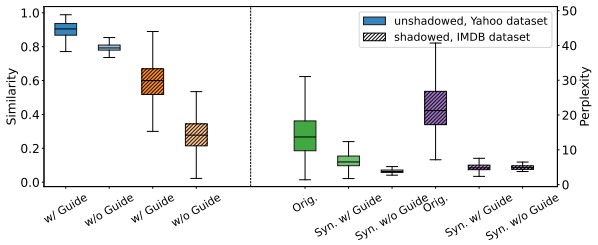

Figure 2: Evaluation of semantic alignment and generation quality. **Left:** Cosine similarity between generated and target embeddings. Guided Diffusion significantly improves alignment compared to unguided generation. **Right:** Perplexity (PPL) distribution. Our method maintains high fluency (low PPL), comparable to unguided generation and superior to the high-variance original human text.

Experimental Setup. We use two datasets covering topical and sentiment-based steering: `Yahoo Answers Topics` (Zhang et al., 2015) (10 topic categories) and `IMDB` (Maas et al., 2011) (rich-sentiment movie reviews); from each we randomly draw 1,000 reference sentences exceeding 256 tokens as semantic targets. We employ `LLaDA-8B-Instruct` (Nie et al., 2025) as the backbone DLM, with target embeddings produced by `sentence-t5-base` (Ni et al., 2022) for `Yahoo` and `stsb-roberta-base-v2` (Reimers & Gurevych, 2019) for `IMDB`. We report (i) *semantic alignment* as cosine similarity between generated and target embeddings, and (ii) *generation quality* as perplexity (PPL) computed with Qwen3-8B (Yang et al., 2025a). We further include an ablation isolating the contribution of semantic-aware sampling and re-masking.

Evaluation Results. The results are presented in Figure 2 and Table 1. Across both datasets, Guided Diffusion achieves strong semantic alignment with the target embeddings without sacrificing much on fluency, and both semantic-aware sampling and re-masking are necessary to reach this quality.

Semantic Alignment. Figure 2 (Left) shows that our Guided Diffusion achieves high fidelity to the target embeddings. The median cosine similarity between synthetic and original embeddings is 0.91 for `Yahoo` and 0.60 for `IMDB`. In contrast, unguided generation results in significantly lower semantic alignment (e.g., a median of 0.28 for `IMDB`), confirming that our Guided Diffusion can effectively preserve the high fidelity to the target embedding.

Generation Quality. In Figure 2 (Right), we demonstrate that Guided Diffusion preserves the linguistic fluency. The synthetic samples exhibit excellent quality, with median PPL scores of 6.51 (`Yahoo`) and 4.83 (`IMDB`). Notably, these scores are comparable to unguided generation (e.g., 4.89 for `IMDB`), indicating that our sampling and re-masking interventions do not degrade linguistic coherence.

Ablation Study. To isolate the contribution of each guidance mechanism, we ablate semantic-aware sampling and semantic-aware re-masking on `Yahoo` and `IMDB`. When disabling one component, we replace it with its standard unguided counterpart (independent token sampling or low-confidence re-masking) while leaving the other component active, and report the mean cosine similarity over $1{,}000$ samples in Table 1.

Table 1: Ablation of semantic-aware sampling and re-masking. We report the cosine similarity between generated and target embeddings on `Yahoo` and `IMDB` as mean$_{\pm\text{std}}$; bold marks the best result. Disabling either substantially degrades alignment, confirming both stages are essential.

Method	`Yahoo`	`IMDB`
Guided Diffusion (Full)	$\mathbf{0.90}_{\pm 0.04}$	$\mathbf{0.59}_{\pm 0.11}$
w/o semantic-aware re-masking	$0.79_{\pm 0.02}$	$0.35_{\pm 0.10}$
w/o semantic-aware sampling	$0.80_{\pm 0.02}$	$0.34_{\pm 0.10}$

The full framework achieves the best alignment on both datasets, and disabling either causes a sharp drop. The effect is most pronounced on `IMDB`, where the partial variants (0.35 and 0.34) collapse toward the fully unguided baseline (see Figure 2). The two components yield comparable scores in isolation (e.g., 0.79 vs. 0.80 on `Yahoo`), indicating that they capture complementary signals: semantic-aware sampling steers the candidate selection toward the target manifold, while semantic-aware re-masking preserves semantically critical tokens across diffusion steps. Combining both restores the full alignment quality, confirming that the two stages are essential and reinforcing.

4 Differentially Private Synthetic Data

In this section, we present our framework for generating differentially private (DP) synthetic text. The DP synthetic data generation pipeline is illustrated in Figure 3.

The fundamental challenge in applying DP to natural language is that text is discrete and symbolic; unlike numerical data, one cannot simply compute the "average" of multiple sentences and add DP noise to obfuscate individual contributions. Consequently, prior works have been forced to rely on complex and unstable methods like DP-SGD (Yu et al., 2022) or use an LLM to synthesize a huge corpus of synthetic data and select according to private samples iteratively (Xie et al., 2024).

The utility of the latter approach, Private Evolution (PE; Xie et al., 2024), is limited mainly by its generation pipeline rather than by the DP mechanism. If DP noise were the main limitation, removing it would raise utility substantially. Instead, on `PubMed`, PE improves only from 27.90% at $\varepsilon = 1$ to 28.60% without DP noise ($\varepsilon = \infty$; Table 4). PE's own results show the same pattern: moving from $\varepsilon = 1$ to $\varepsilon = \infty$ changes most metrics only modestly, leaves some unchanged or lower, and does not close the gap to training on the original data (Table 2 in Xie et al. (2024)). A weaker privacy guarantee therefore does not remove PE's proposal-and-selection bottleneck. This pipeline is also costly: PE generates a large pool of candidate texts in every iteration and then selects among them, so with the same backbone a *single PE iteration* already takes $1.53\times$ as long as Guided Diffusion (§ 4.2). These limitations motivate a generation paradigm that is both more efficient and achieves higher utility for DP synthetic text.

Our framework circumvents these limitations by shifting the privacy mechanism to the semantic embedding space. While embeddings allow for straightforward DP aggregation (e.g., computing noisy centroids), converting these abstract vectors back into coherent text has historically been an ill-posed inverse problem. Our key insight is that our Guided Diffusion algorithm (Algorithms 1 and 2) effectively solves this "inversion gap," acting as a robust decoder that reconstructs high-fidelity

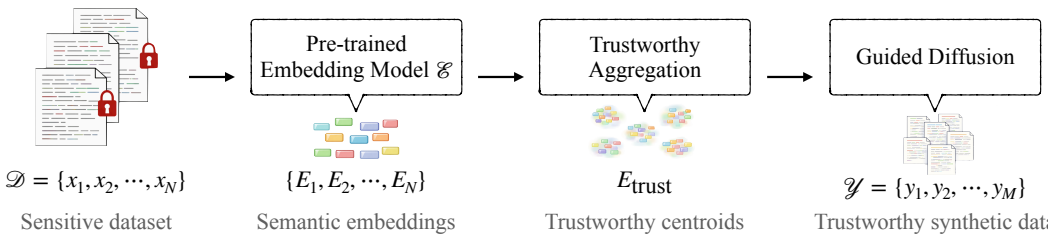

Figure 3: DP synthetic data generation with Guided Diffusion. The input is a private dataset, and the trustworthy aggregation is (ε, δ)-DP aggregation.

text conditioned on the DP embeddings. This enables a streamlined "embed-privatize-generate" workflow that was previously infeasible.

4.1 Framework for DP Synthetic Text Generation with Guided Diffusion

As depicted in Figure 3, our framework contains three distinct stages. We assume that the base DLM and the text embedding model contain no information about the private samples, i.e., they are not trained on the private samples. To ensure that this assumption holds in our evaluation, we construct `ICLR26` from reviews published after these models were released (§ 4.2). We now introduce the three stages of DP synthetic text generation with Guided Diffusion:

Semantic Embedding: We project the sensitive private dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ into a semantic space using a pre-trained text embedding model $\mathcal{E}$. This yields a set of continuous embedding vectors $\{E_1, \ldots, E_N\}$. These vectors are normalized with L_2 norm equal to 1, simplifying the sensitivity bound of the aggregated embedding vectors.

DP Aggregation: We apply a DP aggregation mechanism to these vectors. To avoid collapsing the diverse semantic information of the dataset into a single mean, we employ the practical DP clustering algorithm (Chang & Kamath, 2021). This algorithm partitions the high-dimensional embeddings into semantically coherent groups and computes a set of representative noisy centroids E_{dp} satisfying (ε, δ)-DP. For *labeled* corpora (e.g., `ICLR26` in § 4.2, where each review carries a score), DP clustering is run *class-conditionally*: separately within each label class on its embedding subset, yielding per-class noisy centroids $\{E_{dp,c}\}$. Each synthetic sample inherits the class label of its source per-class centroid, and the per-class sample budget is fixed in advance as a public hyperparameter. Since the per-class subsets are disjoint, the joint release of all centroids is end-to-end (ε, δ)-DP. These centroids effectively capture the distribution of the private text while ensuring privacy protection for individual samples.

Guided Diffusion Generation: Finally, we utilize the DP embedding E_{dp} as the guidance signal (E_{tgt}) for our Guided Diffusion algorithm with DLM $\mathcal{M}_\theta$. The model iteratively denoises a random sequence, steered by our semantic-aware sampling and re-masking strategies (§ 3), to generate an arbitrarily large set of synthetic texts $\mathcal{Y}$ that are semantically aligned with E_{dp}. The original private samples are rigorously protected with (ε, δ)-DP guarantees.

Privacy Guarantee. Under standard conditions of public pretrained $\mathcal{E}$ and $\mathcal{M}_\theta$, class-conditional DP aggregation via the DP clustering algorithm (Chang & Kamath, 2021), and a fixed generation procedure that does not access $\mathcal{D}$, the released synthetic dataset $\mathcal{Y}$ is (ε, δ)-DP with respect to $\mathcal{D}$ by post-processing invariance. The formal statement (Proposition E.1) and proof are in Appendix E.

4.2 Evaluation: DP-Compliant Synthetic Text Generation

To demonstrate the efficacy of our DP synthetic text paradigm with Guided Diffusion, we generate synthetic text under (ε, δ)-DP guarantees and evaluate its downstream utility and its distributional fidelity to the private data.

Experimental Setup. To isolate the generation algorithm from the capability of the underlying generator, we run every method on the same DLM backbone, either of two widely used DLMs, `LLaDA-8B-Instruct` (Nie et al., 2025) or `Dream-7B` (Ye et al., 2025), and compare against PE (Xie et al., 2024) and `DP-SGD+LoRA` (Li et al., 2024). That is, PE and `DP-SGD+LoRA` also use

Table 2: DP synthetic text generation at $(\varepsilon, \delta) = (1, 10^{-6})$ with 2,000 synthetic samples, reported as mean$_{\pm \text{std}}$ over four seeds. Utility is macro-F1 on `ICLR26` and causal next-token accuracy on `PubMed`. Bold marks the best result per dataset and backbone.

			Distributional fidelity			
Backbone	**Method**	**Utility** (%) ↑	FD ↓	JSD ↓	Dist-2 ↑	MAUVE ↑
`ICLR26` (utility: macro-F1)						
LLaDA-8B-Instruct	PE	$13.34_{\pm 1.38}$	$0.7753_{\pm 0.0032}$	$0.3401_{\pm 0.0006}$	0.0730	$0.6366_{\pm 0.0054}$
	DP-SGD+LoRA	$20.12_{\pm 4.16}$	$0.6026_{\pm 0.0015}$	$0.3523_{\pm 0.0007}$	0.0431	$0.6475_{\pm 0.0031}$
	Guided Diffusion	$\mathbf{24.85}_{\pm 2.80}$	$\mathbf{0.3995}_{\pm 0.0020}$	$\mathbf{0.2787}_{\pm 0.0008}$	**0.0891**	$\mathbf{0.6923}_{\pm 0.0042}$
Dream-7B	PE	$12.49_{\pm 0.57}$	$0.7531_{\pm 0.0041}$	$0.3060_{\pm 0.0005}$	0.0785	$0.6409_{\pm 0.0079}$
	DP-SGD+LoRA	$\mathbf{29.23}_{\pm 3.37}$	$0.6848_{\pm 0.0020}$	$0.3278_{\pm 0.0006}$	0.0398	$0.6281_{\pm 0.0012}$
	Guided Diffusion	$28.68_{\pm 2.40}$	$\mathbf{0.4333}_{\pm 0.0013}$	$\mathbf{0.2620}_{\pm 0.0006}$	**0.0984**	$\mathbf{0.6848}_{\pm 0.0045}$
`PubMed` (utility: accuracy)						
LLaDA-8B-Instruct	PE	$22.15_{\pm 0.08}$	$0.5464_{\pm 0.0039}$	$0.3388_{\pm 0.0005}$	0.1221	$0.7318_{\pm 0.0040}$
	DP-SGD+LoRA	$25.45_{\pm 0.11}$	$0.9021_{\pm 0.0070}$	$0.3432_{\pm 0.0011}$	0.0786	$0.6779_{\pm 0.0021}$
	Guided Diffusion	$\mathbf{32.68}_{\pm 0.08}$	$\mathbf{0.3169}_{\pm 0.0057}$	$\mathbf{0.2061}_{\pm 0.0007}$	**0.2737**	$\mathbf{0.8637}_{\pm 0.0017}$
Dream-7B	PE	$20.46_{\pm 0.18}$	$0.5801_{\pm 0.0037}$	$0.3144_{\pm 0.0003}$	0.1282	$0.7319_{\pm 0.0071}$
	DP-SGD+LoRA	$25.16_{\pm 0.14}$	$0.9445_{\pm 0.0052}$	$0.3287_{\pm 0.0008}$	0.0723	$0.6804_{\pm 0.0017}$
	Guided Diffusion	$\mathbf{32.89}_{\pm 0.08}$	$\mathbf{0.2387}_{\pm 0.0039}$	$\mathbf{0.1569}_{\pm 0.0004}$	**0.3014**	$\mathbf{0.8764}_{\pm 0.0060}$

the same DLM to generate synthetic text. For each backbone, all of the methods use the same private data, privacy parameters, synthetic-data size, prompts, and downstream evaluation.

We evaluate on two private datasets. `PubMed` (Yu et al., 2023) contains 75,316 medical abstracts; following the official PE protocol, we fine-tune `BERT-small` (Turc et al., 2019) on the synthetic DP samples and report causal next-token prediction accuracy on real held-out abstracts in `PubMed`'s test split. To rule out the possibility that the training data of the evaluated models, including the DLMs and the embedding models, contains the private samples, we construct a new dataset, `ICLR26`, of paper reviews from ICLR 2026, which were released after all evaluated models. We follow the setup and train/test split of the ICLR 2023 review dataset used by PE (Xie et al., 2024). Specifically, we train a `roberta-base` (Liu et al., 2019) classifier on the synthetic reviews to predict the five-class review score and report macro-F1 on real held-out reviews in the test split. Consistent with PE, we use the embedding model `sentence-t5-base` for `PubMed` and `stsb-roberta-base-v2` for `ICLR26`. We run PE for four proposal-and-selection iterations, which already takes more than $4\times$ computational cost compared with Guided Diffusion.

Every method generates 2,000 synthetic samples at $(\varepsilon, \delta) = (1, 10^{-6})$, and results are averaged over four seeds. Beyond downstream utility, we also measure distributional fidelity to the real data with four metrics: the Fréchet distance (FD) between MPNet (Song et al., 2020) embeddings of synthetic and real texts, the Jensen–Shannon divergence (JSD) between their unigram distributions, Distinct-2 (Li et al., 2016) for lexical diversity, and MAUVE (Pillutla et al., 2021), which compares the two distributions in a language-model embedding space and penalizes both synthetic texts that are unlike real ones and real texts that the synthetic set fails to cover.

Evaluation Results. In Table 2, Guided Diffusion achieves the best score on every fidelity metric in all four dataset–backbone settings and the best utility in three of them; on `ICLR26` with `Dream-7B`, its macro-F1 is comparable with `DP-SGD+LoRA` (within standard deviation).

Downstream Utility. On `ICLR26` with `LLaDA-8B-Instruct`, Guided Diffusion reaches 24.85% macro-F1, compared with 13.34% for PE and 20.12% for `DP-SGD+LoRA`. With `Dream-7B`, its macro-F1 (28.68%) is 16.19 points above PE and 0.55 points below `DP-SGD+LoRA`. PE's macro-F1 stays near 13% with both backbones. On `PubMed`, Guided Diffusion improves causal next-token accuracy over PE by 10.53 points with `LLaDA-8B-Instruct` (32.68% vs. 22.15%) and by 12.43 points with `Dream-7B` (32.89% vs. 20.46%), and over `DP-SGD+LoRA` by more than 7 points with both backbones. The performance of Guided Diffusion also exceeds the results of PE's official implementation, where PE and `DP-SGD-FT` (Yue et al., 2023) on `GPT-2 Large` use their default generators with many more iterations (27.90% and 31.10%; Table 4). Our result approaches the non-private upper bound obtained by fine-tuning on the real private corpus (34.60%).

Distributional Fidelity. Guided Diffusion produces synthetic text that is closer to the private data. It attains the lowest FD and unigram JSD and the highest Distinct-2 and MAUVE in every setting. The largest gap is in FD: with `LLaDA-8B-Instruct`, Guided Diffusion reduces FD from 0.78 (PE) and 0.60 (`DP-SGD+LoRA`) to 0.40 on `ICLR26`, and from 0.55 and 0.90 to 0.32 on `PubMed`. The higher Distinct-2 shows that this fidelity does not come from collapsing to a few prototypical texts.

Computational Efficiency. On eight A100 GPUs, Guided Diffusion generates the $2{,}000$ `ICLR26` samples in 2.74 hours. End-to-end PE with the same backbone takes 4.19 hours for *a single* private-selection iteration and 12.20 hours for the four iterations used in Table 2, so Guided Diffusion is $4.45\times$ faster. The difference is structural: PE repeatedly queries the generator for candidate variants and selects among them with a DP histogram, whereas our framework aligns each sample within a single denoising run. In our implementation, DLM forward passes account for 47.6% of the runtime, encoder scoring for 29.0%, and candidate construction for 17.0%. We do not include DP-SGD-based methods here, since their cost is dominated by private fine-tuning with per-sample gradient clipping and PE reports DP fine-tuning to be more than $10\times$ slower than PE (Xie et al., 2024).

Hyperparameter Sensitivity. Guided Diffusion is insensitive to the privacy budget: on `PubMed` with `LLaDA-8B-Instruct`, its accuracy is 32.66% / 32.65% / 32.48% at $\varepsilon = 1$ / 2 / 4, so the strict budget $\varepsilon = 1$ costs little utility, and it stays above every DP baseline at each budget (Table 4). Generation quality is insensitive to the candidate pool size α and the weight ratio b/a in Algorithm 1. See Appendix D for more results.

Ablation Study. To verify that the gains observed in § 3 carry over to the downstream settings, we ablate each guidance mechanism on `PubMed` with `LLaDA-8B-Instruct` while keeping the same DP embedding pipeline. As shown in Table 3, removing either component substantially degrades downstream accuracy: without semantic-aware sampling, accuracy collapses from 32.66% to 18.01%, and without semantic-aware re-masking it drops to 26.73%. This confirms that both components remain essential in DP synthetic text generation.

Table 3: Ablation of semantic-aware sampling and re-masking on `PubMed` with `LLaDA-8B-Instruct` under $\varepsilon = 1$. Both components remain essential when generation is conditioned on DP embeddings.

Method	Accuracy (%)
Guided Diffusion (Full)	**32.66**
w/o semantic-aware sampling	18.01
w/o semantic-aware re-masking	26.73

5 Related Work

Prior DP synthetic text generation falls into two streams: *DP-SGD-based fine-tuning* (Yoon et al., 2019; Yu et al., 2021; Yue et al., 2023) and *Private Evolution (PE) variants* (Lin et al., 2024; Xie et al., 2024) that filter LLM candidates via DP histograms; both inherit structural limits from per-sample clipping or rejection sampling. The closest concurrent work, Zhou et al. (2025), reconstructs DP-protected *images* from privatized embeddings via diffusion-based image editing (Rombach et al., 2022); translating this "embedding-to-data" approach to text is non-trivial because text is discrete and no off-the-shelf guided text-diffusion pipeline exists. Our framework bridges this gap via DLMs. A detailed elaboration is given in Appendix B.

6 Conclusion & Limitations

We propose a novel framework that converts semantic embeddings into high-fidelity text via Guided Diffusion, establishing a new paradigm for private synthetic text generation. Our approach naturally integrates Differential Privacy, ensuring precise semantic control while providing rigorous privacy guarantees with negligible utility cost. Our empirical results demonstrate the superiority of Guided Diffusion over prior DP synthetic text generation baselines across datasets and diffusion backbones.

Limitations and Future Work. Several extensions remain open. Semantic-aware re-masking evaluates one ablated embedding per token; sparse or batched ablations could amortize this cost for very long sequences. Our generation pipeline operates on fixed-length token blocks; extending to longer documents fits naturally with hierarchical or sliding-window schemes that we leave for future work. Empirically, we focus on English short-form generation; transferring to multilingual corpora, code, and structured outputs is a promising direction.

AI USE STATEMENT

We used generative AI tools to assist with the writing of this paper, mainly to check grammar and to polish the wording and readability of author-written text. The tools were also used to format LaTeX tables and figures. All research ideas, the Guided Diffusion method and its technical components, the privacy analysis, the experimental design and implementation, and the interpretation of results were developed by the authors. We did not use generative AI to generate datasets, develop the method or conceptual framework, formulate or prove mathematical claims, design experiments, or interpret results. The authors reviewed and revised all AI-polished text. The authors take responsibility for all the final content of this work.

ETHICS STATEMENT

The authors strictly adhere to the ICLR Code of Ethics. This work aims to reduce the privacy risks in synthetic text generation. The privacy guarantee holds under the assumptions stated in Proposition E.1; practitioners should verify these assumptions before deployment. DP bounds the influence of each individual record but does not prevent inferences about population-level properties of the private corpus. Our experiments use publicly available data: `PubMed` abstracts, `Yahoo Answers Topics`, `IMDB` reviews, and ICLR 2026 paper reviews publicly posted on OpenReview. We use the reviews only to evaluate synthetic data generation and make no attempt to identify reviewers. Our study involves no human-subject experiments. We discuss broader impacts in Appendix A.

REPRODUCIBILITY STATEMENT

We describe Guided Diffusion in § 3, with pseudocode in Algorithms 1 and 2, and the DP synthetic text framework in § 4. Appendix C lists all hyperparameters, the backbone and embedding models, and the DP clustering configuration, and Appendix D reports the sensitivity to the privacy budget, the number of synthetic samples, and the generation hyperparameters. All backbone DLMs, embedding models, and baseline implementations are publicly available, and § 4.2 describes the evaluation datasets and protocols; `ICLR26` follows the setup and train/test split of the ICLR 2023 review dataset used by PE (Xie et al., 2024). Our code is provided in the supplementary material.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156. IEEE, 2022.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=pzpWBbnwiJ`.

California State Legislature. California Consumer Privacy Act of 2018. `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375`, 2018.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Alisa Chang and Pritish Kamath. Differentially private clustering in google's differential privacy library. https://github.com/google/differential-privacy/tree/main/learning/clustering, 2021. Accompanying article available at https://research.google/blog/practical-differentially-private-clustering/.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPpQk7FJXF.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PUIqjT4rzq7.

Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

Frederik Harder, Milad Jalali, Danica J. Sutherland, and Mijung Park. Pre-trained perceptual features improve differentially private image generation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=R6W7zkMz0P.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. PrE-text: Training language models on private federated data in the age of LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19043–19061. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hou24c.html.

Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. Sok: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4696–4713. IEEE, 2024.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. Llm-pbe: Assessing data privacy in large language models. *Proceedings of the VLDB Endowment*, 17(11):3201–3214, 2024.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35: 4328–4343, 2022a.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=bVuP3ltATMz.

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YEhQs8POIo.

Zinan Lin, Tadas Baltrusaitis, and Sergey Yekhanin. Differentially private synthetic data via APIs 3: Using simulators instead of foundation model. In *Will Synthetic Data Finally Solve the Data Access Problem?*, 2025. URL https://openreview.net/forum?id=7AvUO96Pmk.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.323. URL https://aclanthology.org/2021.acl-long.323/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the association for computational linguistics: ACL 2022*, pp. 1864–1874, 2022.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

Bowen Tan, Zheng Xu, Eric P. Xing, Zhiting Hu, and Shanshan Wu. Synthesizing privacy-preserving text data via finetuning *without* finetuning billion-scale LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=FCm4laCLiH.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model apis 2: text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Da Yu, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Janardhan Kulkarni, Zinan Lin, Chulin Xie, Huishuai Zhang, and Wanrong Zhang. Training private and efficient language models with synthetic data from llms. In *Socially Responsible Language Modelling Research*, 2023.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023.

Jianqing Zhang, Yang Liu, JIE FU, Yang Hua, Tianyuan Zou, Jian Cao, and Qiang Yang. PCE-volve: Private contrastive evolution for synthetic dataset generation via few-shot private data and generative APIs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IKCfxWtTsu.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Felix Zhou, Samson Zhou, Vahab Mirrokni, Alessandro Epasto, and Vincent Cohen-Addad. Private training & data generation by clustering embeddings. *arXiv preprint arXiv:2506.16661*, 2025.

Zhanke Zhou, Jianing Zhu, Fengfei Yu, Xuan Li, Xiong Peng, Tongliang Liu, and Bo Han. Model inversion attacks: A survey of approaches and countermeasures. *arXiv preprint arXiv:2411.10023*, 2024.

A Broader Impact

This paper presents work whose goal is to advance the field of privacy-preserving machine learning. Our framework for private synthetic text generation has significant potential to benefit society by enabling the safe utilization of sensitive datasets, such as medical records, financial reports, and private communications, that are currently inaccessible due to privacy regulations (e.g., GDPR). By reconciling high-fidelity generation with rigorous DP guarantees, our approach mitigates critical risks associated with Large Language Models, including data memorization and leakage from training corpora.

While advancements in generative AI carry inherent risks of misuse, such as the generation of misinformation or impersonating content, our research focuses specifically on defensive mechanisms that certify safety. The DP guarantee bounds the influence of any single record on the released synthetic data, providing a quantifiable foundation for responsible release of synthetic corpora derived from sensitive sources. We hope this work contributes to the responsible and transparent deployment of AI systems in sensitive domains.

B Detailed Related Work

Guided Diffusion Models. Image diffusion models have achieved remarkable success in controllable synthesis, using embedding-based guidance to steer generation (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021; Rombach et al., 2022). While extensive research optimizes this guidance for images (Ramesh et al., 2022; Ye et al., 2023; Zhang et al., 2023; Feng et al., 2023; Bansal et al., 2024), its application to DLMs remains underexplored. Unlike the continuous image domain, text is discrete, and the dominance of prompt-based AR LLMs has overshadowed the development of embedding-based control. To the best of our knowledge, we are the first to effectively guide DLMs in converting semantic embeddings into coherent, high-fidelity text.

Differentially Private Synthetic Data Generation. Prior research on generating synthetic data with DP guarantees spans both image and text domains (Hu et al., 2024), generally bifurcating into two primary streams: 1) *DP-SGD-based methods* (Yoon et al., 2019; Yu et al., 2021; Li et al., 2022b; Harder et al., 2023; Yue et al., 2023; Ghalebikesabi et al., 2023; Dockhorn et al., 2023; Tan et al., 2025) and 2) *Private Evolution (PE) variants* (Lin et al., 2024; Xie et al., 2024; Hou et al., 2024; Zhang et al., 2025; Lin et al., 2025).

DP-SGD approaches involve training/fine-tuning a generative model. While foundational, these methods suffer from significant computational overhead due to per-sample gradient clipping and often yield suboptimal utility when scaling to LLMs. In contrast, PE variants circumvent direct training by iteratively querying public LLMs to generate a massive pool of synthetic candidates. A DP selection mechanism, which adds Gaussian noise to distance histograms, is then employed to filter for samples that align with the private data distribution. While PE offers improved efficiency and utility-privacy tradeoffs compared to DP-SGD, it remains heavily dependent on the quality of manual sampling prompts and implicitly constrains the diversity and volume of synthetic samples due to the rejection-sampling bottleneck.

Recently, Zhou et al. (2025) introduced a novel paradigm for *image* synthesis, demonstrating that high-fidelity images can be reconstructed from DP embeddings using diffusion-based image editing (Rombach et al., 2022). However, translating this "embedding-to-data" approach to the text domain is non-trivial due to the discrete nature of language and the absence of established Guided Diffusion pipelines for text. Our work bridges this gap: we propose an alternative approach that, for the first time, exploits the steerable properties of DLMs to efficiently convert semantic embeddings into high-utility synthetic text. Crucially, this framework naturally accommodates DP mechanisms, providing rigorous privacy guarantees alongside superior generation quality.

C Implementation Details

We list the concrete hyperparameter values used throughout our experiments. All settings are held fixed unless stated otherwise. All experiments are run on A100 GPUs.

C.1 Diffusion Backbone

We use `LLaDA-8B-Instruct` (Nie et al., 2025) and `Dream-7B` (Ye et al., 2025) as the backbone masked diffusion language models. For these models, every generation uses the settings below.

- Sequence length: $L = 448$ tokens. Note that LLaDA takes a semi-autoregressive approach, with default block size 32.
- We use the default number of diffusion (denoising) steps: $T = L = 448$.
- Denoise schedule ρ_t: We take the default linear schedule used by `LLaDA-8B-Instruct` and `Dream-7B`.

C.2 Semantic-Aware Sampling (Algorithm 1)

- Candidate pool factor: $\alpha = 20$.
- Top-N filter size: $N = 32$.
- Log-probability weight: $a = 1$.
- Hamming-distance weight: $b = 0.1$.

We did not tune a and b extensively; Appendix D shows that varying $b/a \in \{0.1, 0.2, 0.5\}$ and $\alpha \in \{5, 10, 20\}$ has little effect on generation quality.

C.3 Semantic-Aware Re-Masking (Algorithm 2)

- Denoise schedule ρ_t: We take the same default linear schedule used by the backbone `LLaDA-8B-Instruct` and `Dream-7B`.

C.4 DP Aggregation

- DP clustering: Chang & Kamath (2021)'s implementation with default hyperparameters; the privacy budget (ε, δ) is specified per experiment. The number of clusters is set to 100. For each class, the number of synthetic samples generated from each released centroid is fixed uniformly in advance. We do not use private cluster sizes.

C.5 Embedding Models

We deliberately use two different semantic embedding models chosen to match the dominant semantic axis of each dataset. `sentence-t5-base` (Ni et al., 2022), trained on a broad mixture of sentence-level tasks, produces embeddings that emphasize topical and factual content; we therefore use it for `PubMed` (medical abstracts grouped by topic area) and `Yahoo` (10 explicit topic categories). `stsb-roberta-base-v2` (Reimers & Gurevych, 2019), fine-tuned on the STS-Benchmark sentence-similarity task that emphasizes evaluative and stylistic similarity, captures sentiment- and style-driven semantics more reliably, so we use it for the review-heavy datasets `ICLR26` and `IMDB`, where rating and stylistic register dominate the relevant signal.

- `ICLR26`, `IMDB`: `stsb-roberta-base-v2` (Reimers & Gurevych, 2019).
- `PubMed`, `Yahoo`: `sentence-t5-base` (Ni et al., 2022).

D Sensitivity Analyses

Privacy Budget. Table 4 varies $\varepsilon \in \{1, 2, 4\}$ with $\delta = 10^{-6}$ using `LLaDA-8B-Instruct` on `PubMed`. The accuracy of Guided Diffusion changes by less than 0.2 points, so the strict budget $\varepsilon = 1$ used in our main results costs little utility relative to looser budgets. At $\varepsilon = 1$, Guided Diffusion is within 2 points of the non-private upper bound of fine-tuning on the real private corpus (34.60%). For comparison, the table also lists the baselines across budgets: PE and `DP-SGD-FT` (Yue et al., 2023) on `GPT-2 Large` from PE's official implementation (Xie et al., 2024), which use their default generators, and `DP-SGD+LoRA` on `Llama-3-8B` (Li et al., 2024). Guided Diffusion exceeds every DP baseline at each tested budget, and even PE without DP noise ($\varepsilon = \infty$, 28.60%). Scaling DP-SGD

Table 4: Effect of the privacy budget on `PubMed` accuracy, with $\delta = 10^{-6}$. Guided Diffusion uses `LLaDA-8B-Instruct`. †Results from PE's official implementation, where each baseline uses its default generator.

Method	ε	Accuracy (%)
Without fine-tuning		0.07
Fine-tuning on real data (non-DP)		34.60
PE†	1	27.90
	∞	28.60
DP-SGD-FT (GPT-2 Large)†	1	31.10
	∞	33.10
DP-SGD+LoRA (Llama-3-8B)	1	24.88
	2	24.70
	4	24.37
Guided Diffusion	1	**32.66**
	2	**32.65**
	4	**32.48**

Table 5: Effect of the number of synthetic samples on `ICLR26` (`LLaDA-8B-Instruct`, $\varepsilon = 1$), reported as mean$_{\pm\text{std}}$.

Synthetic samples	Macro-F1 (%)
100	$13.13_{\pm 5.59}$
500	$20.46_{\pm 6.30}$
1,000	$20.83_{\pm 5.71}$
2,000	$\mathbf{24.85}_{\pm 2.80}$

Table 6: Effect of generation hyperparameters on `ICLR26`. Each row changes one hyperparameter from the default.

Setting	Guide sim. ↑	MPNet sim. ↑	PPL ↓
Default ($T = 448$, $\alpha = 20$, $b/a = 0.1$)	0.6920	0.5396	4.218
$T = 224$	0.6756	0.5335	7.948
$\alpha = 5$	0.6929	0.5345	4.482
$\alpha = 10$	0.6930	0.5405	4.352
$b/a = 0.2$	0.6919	0.5385	4.247
$b/a = 0.5$	0.6910	0.5415	4.203

to `Llama-3-8B` with LoRA does not help: its accuracy stays within 24.4%–24.9% across budgets, below `DP-SGD-FT` on `GPT-2 Large` at $\varepsilon = 1$.

Number of Synthetic Samples. Since the synthetic texts are a post-processing of the released DP centroids, generating more of them incurs no additional privacy cost. Table 5 shows that utility improves as the synthetic training set grows from 100 to 2,000 samples: macro-F1 rises from 13.13% to 24.85% on `ICLR26` with `LLaDA-8B-Instruct`.

Generation Hyperparameters. Table 6 varies the candidate pool factor α, the weight ratio b/a of the sampling score, and the number of denoising steps T. We report the cosine similarity between each generated text and its target embedding under the guidance encoder (Guide sim.) and under an independent MPNet encoder (MPNet sim.), together with perplexity (PPL). Pool size and weighting have little effect: across $\alpha \in \{5, 10, 20\}$ and $b/a \in \{0.1, 0.2, 0.5\}$, guide similarity stays within 0.691–0.693 and perplexity within 4.20–4.48. Halving the denoising steps to $T = 224$ lowers guide similarity from 0.692 to 0.676 and nearly doubles perplexity, from 4.22 to 7.95. This degradation with fewer denoising steps is expected: it is a property of the DLM rather than of our guidance design. We therefore recommend using the default number of denoising steps of each DLM.

E Formal Privacy Guarantee and Proof Sketch

Proposition E.1 (DP Guarantee). *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the private corpus, where y_i is an optional class label, and let $\mathcal{Y}$ denote the labeled synthetic dataset produced by the DP Guided Diffusion framework. Suppose the following conditions hold:*

1. ***Public pretrained models.*** *The embedding model $\mathcal{E}$ and the diffusion language model $\mathcal{M}_\theta$ are pretrained on a corpus disjoint from $\mathcal{D}$ and are treated as public.*
2. ***Class-conditional DP aggregation.*** *The DP clustering algorithm (Chang & Kamath, 2021) is run separately within each label c, releasing noisy centroids $E_{dp,c}$ that are (ε, δ)-DP with respect to record-level adjacency on the per-class subset $\mathcal{D}_c = \{x_i : y_i = c\}$. (For an unlabeled corpus, this reduces to a single class.)*
3. ***Fixed generation procedure.*** *Guided Diffusion (Algorithms 1 and 2) is a randomized function of $\{E_{dp,c}\}$ and the public models alone, and does not access $\mathcal{D}$. Each synthetic sample inherits the class label c of its source per-class centroid.*

Then the released synthetic dataset $\mathcal{Y} = \{(\tilde{x}_j, c_j)\}$ is (ε, δ)-DP with respect to $\mathcal{D}$.

Proof sketch of Proposition E.1 (DP Guarantee). By condition (2), each per-class centroid release $E_{dp,c}$ is (ε, δ)-DP w.r.t. $\mathcal{D}_c$. Since the per-class subsets $\{\mathcal{D}_c\}$ form a disjoint partition of $\mathcal{D}$, parallel composition yields (ε, δ)-DP for the joint release $\{E_{dp,c}\}$ w.r.t. $\mathcal{D}$. Conditions (1) and (3) ensure that $\mathcal{Y}$ is a randomized function of $\{E_{dp,c}\}$ and public model components only, with no further dependence on $\mathcal{D}$. Post-processing invariance (§ 2) preserves the (ε, δ)-DP guarantee on the final synthetic data release. □