CDO-GIA: A Robust Textual Gradient Inversion Attack Against Federated Language Models via Continuous-Discrete Optimization

Jiajie Wang[1,2†], **Weibo Xu**[1†], **Ruichen Xia**[3,4*] and **Jiahao Nie**[1*]
[1]School of Information Technology and Artificial Intelligence,
Zhejiang University of Finance and Economics, Hangzhou 310018, China
[2]Raina Technology, Shaoxing 312000, China
[3]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[4]Institute of Computing Innovation, Zhejiang University, Hangzhou 310058, China
wangjiajie@raina.tech, xuweibo5610@zufe.edu.cn, rcxia@zju.edu.cn, jhnie@hdu.edu.cn

Abstract

Gradient inversion attacks (GIAs) have shown that shared gradients in federated learning leak private training data. However, current textual GIAs fail in large batch size, as simply increasing batch size can serve as a stable defense against such attacks. In this paper, we propose CDO-GIA, a novel textual gradient inversion attack via continuous-discrete optimization that extends the maximum effective batch size from 4 to 8 for reconstructing text from gradient. In continuous optimization phase, CDO-GIA incorporates a dynamic weight decay mechanism to alleviate embedding homogenization caused by embedding regularization at large batch sizes, thereby enhancing unigrams reconstruction accuracy. Moreover, our method introduces a tabu beam search mechanism guided by the language model prior during discrete optimization. This design facilitates fine-grained exploration of high-dimensional token spaces via precise token order adjustment, thus reconstructing semantically coherent sequences. By alternating optimization between continuous and discrete phases, CDO-GIA effectively doubling practical attack limit of prior textual GIAs. Extensive experiments are conducted on three binary text classification datasets. The experimental results indicate that CDO-GIA surpasses all baseline methods. With a batch size of 8, it achieves performance gains of 25.01%, 155.64%, and 22.72% over the baselines in terms of Rouge-1, Rouge-2 and Rouge-L scores.

1 Introduction

Federated learning (FL) [McMahan *et al.*, 2017] has emerged as a privacy-preserving paradigm for distributed machine learning, where clients collaboratively train models by sharing only gradient updates rather than raw training data [Kairouz *et al.*, 2021]. This architecture promises enhanced data privacy by keeping sensitive information localized on client devices, making it particularly valuable for applications involving personal text data, medical records, and proprietary information. However, this privacy guarantee is fundamentally undermined by gradient inversion attacks (GIAs), which enable the reconstruction of original training samples from shared gradient updates. As illustrated in Fig. 1, although clients compute gradients locally on their private data and transmit only model updates to the central server, a honest-but-curious server that can exploit these gradients to recover the raw training inputs, thereby directly exposing users' sensitive personal information. The capability to reconstruct raw training inputs from shared gradients poses severe privacy risks, particularly for language models that process highly sensitive textual data, such as personal communications, medical records, or financial documents.

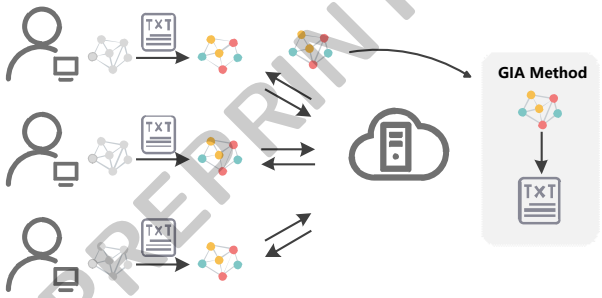


Figure 1: Threat of GIAs in FL. (a) Standard FL workflow: server distributes global model to clients; clients compute local gradients from private data and upload gradients to server. (b) Gradient inversion attack: GIA method uses these gradients to reconstruct raw text inputs from aggregated updates.

The foundational work of [Zhu *et al.*, 2019] first formalized GIAs as an optimization problem: given observed gradients, one seeks dummy inputs whose gradients closely match the observed ones under a chosen distance metric. This approach has achieved remarkable success in the computer vision domain, [Ye *et al.*, 2024] demonstrated high-fidelity reconstruction of ImageNet-scale images (224×224 pixels) [Deng *et al.*, 2009] even with batch sizes as large as 64. Subsequent advances have further enhanced the scalability and robustness of GIAs through architectural insights and refined optimization strategies, establishing GIAs as a critical threat for image-based FL systems.

*Corresponding author.

In stark contrast, GIAs for textual data remains underdeveloped. While high-fidelity reconstruction of image inputs from gradients is now effectively achievable, recovering text from gradients presents substantially greater challenges, with current methods exhibiting severe performance degradation as batch size increases. This limitation stems from two aspects: 1) Text semantic representation is highly context-dependent, requiring exact recovery of unigrams for meaningful privacy leakage, unlike images that approximate reconstructions often suffice for privacy compromise. 2) Text data possesses non-Euclidean attributes, leading its vector representations typically to being high-dimensional and sparse. This results in a solution space that is significantly larger than that of other data types, implying that reconstructing training data involves solving a higher-dimensional inverse problem. Consequently, existing textual GIA methods fail beyond batch size of 4, as simply increasing batch size can serve as a defense against textual GIAs. This limitation not only hinders privacy risk assessment but also impedes the development of effective defenses for text-based federated systems.

To tackle these challenges, we propose CDO-GIA, a novel gradient inversion attack via continuous-discrete optimization that extends the maximum effective batch size from 4 to 8 for text reconstruction from gradient. In the continuous optimization phase, we observe that conventional embedding regularization mechanisms [Balunovic *et al.*, 2022] under large-batch setting tend to enforce embeddings of all reconstructed texts to converge toward a uniform norm. This homogenization effect compromises the diversity of reconstructed texts and degrades the reconstruction accuracy of unigrams. To resolve this, we propose a dynamic weight decay mechanism that dynamically adjusts the weight coefficient of the embedding regularization term in the optimization process. Specifically, the higher weights are first applied to ensure optimization aligns with linguistic distributions, then later stages reduce weights to emphasize gradient matching precision. As a result, our mechanism retains the natural variation of the embedding length distribution and improves reconstruction performance of the unigrams. After mapping continuous embeddings to discrete token sequences via an Embed2Seq mapper, CDO-GIA introduces a language model prior-guided Tabu Beam Search (TBS) mechanism to achieve precise token order adjustment in the discrete optimization phase. Specifically, the mechanism generates candidate sequences by iteratively applying transformations to token sequences during the search. To evaluate these candidates, it employs a joint heuristic combining gradient reconstruction loss and language model perplexity, alongside a tabu list designed to preclude the revisiting of explored sequence. This design avoids local optima and enables fine-grained exploration of the high-dimensional sparse token space, thereby reconstructing more semantically coherent sequences.

Our main contribution are summarized as follows:

- We propose CDO-GIA, a novel textual gradient inversion attack via continuous-discrete optimization that extends the maximum effective batch size from 4 to 8 for text reconstruction from gradient.
- We propose a dynamic weight decay mechanism in continuous optimization to counteract embedding homogenization induced by embedding regularization and enhance unigrams reconstruction accuracy.
- We propose a TBS mechanism guided by language model priors during discrete optimization, which facilitates fine-grained exploration of high-dimensional token spaces via precise token order adjustment to reconstruct semantically coherent sequences.
- Extensive experiments validate the superiority of CDO-GIA and indicate that simply increasing the batch size is insufficient as a stable defense strategy. The code and models are available at https://github.com/weber-xuu/CDO-GIA.

2 Related Work

2.1 Privacy Attack

As FL gains widespread attention for its privacy-preserving advantages, an increasing number of studies focus on its security, particularly its vulnerabilities to attacks and potential threats, to ensure the system remains reliable while safeguarding privacy [Mothukuri *et al.*, 2021]. Early study on privacy attacks have focused on some inference attacks, such as membership inference attacks [Truex *et al.*, 2019; Wang *et al.*, 2024] and feature inference attacks [Luo *et al.*, 2021]. Membership inference attacks refer to the disclosure of data privacy information by speculating whether or not a particular training sample is included in the training set. [Aono *et al.*, 2017] theoretically demonstrates that even a portion of the gradient can reveal crucial information about the training data. [Nasr *et al.*, 2019] proposes a comprehensive framework for privacy analysis in FL, introducing passive and active attack methods for membership inference. [Melis *et al.*, 2019] demonstrates that the adversary can infer certain attributes of the data set by training a property classifier. [Luo *et al.*, 2021] conducts a feature inference attack during the model prediction phase of vertical FL. They developed a general attack method that leverages multiple prediction outputs for both neural network and random forest models. However, the above inference attacks are primarily applicable to shallow networks and can only perform partial inferences on the dataset, without being able to precisely reconstruct the training data.

2.2 Gradient Inversion Attacks

Recently, GIAs have become a significant threat to FL systems due to its ability to accurately reconstruct training data for deep neural networks using gradient updates. [Phong *et al.*, 2017] provides theoretical insights for GIAs by demonstrating the feasibility of data reconstruction for a single neuron or a single-layer network. [Zhu *et al.*, 2019] first formulated GIAs as an optimization problem by minimizing the distance between gradients from dummy inputs and true gradients, enabling data reconstruction through iterative approximation. [Geiping *et al.*, 2020] demonstrated that inputs to any fully connected layer can be analytically recovered regardless of the rest of the network. They introduced cosine similarity

as the gradient distance metric and total variation as a regularizer, enabling high-resolution image reconstruction. [Geiping *et al.*, 2020] introduced cosine similarity as the gradient distance metric and total variation as a regularizer, enabling high-resolution image reconstruction. [Wei *et al.*, 2020] proposes a framework to evaluate and compare different GIAs, and analyzed how different configurations in training can affect reconstruction performance. [Yin *et al.*, 2021] introduces a group consistency regularization term that further improves attack efficiency and reconstruction performance. [Li *et al.*, 2023] proposes a two-stage inversion method to reconstruct speech data from gradients. Their method first reconstructs speech features from the gradient and then reconstructs the speech waveform based on the recovered features. More recently, [Wu *et al.*, 2025] leverages prior knowledge of diffusion models to improve reconstruction performance on high-resolution datasets and larger batch sizes.

To evaluate the threat of GIAs on language models. [Deng *et al.*, 2021] explores the application of GIAs to transformer-based language models and analyzed the impact of different model architectures and weight distributions on reconstruction performance. [Balunovic *et al.*, 2022] proposes a continuous-discrete optimization method tailored for text data, and they introduced an embedding regularization term in continuous optimization. It achieves text reconstruction with a batch size of 4. [Gao *et al.*, 2025] proposes an attack method that is entirely based on discrete optimization, which is directly reconstructs the original data from the discrete token space, achieving the high-precision text reconstruction with a batch size of 1. However, these works still fail to efficiently search the high-dimensional token space, and the reconstruction performance is limited by the batch size, with text reconstruction achievable only when the batch size does not exceed 4.

3 Preliminary

3.1 Gradient Inversion Attacks

GIAs refer to a privacy breach in FL where an adversary intercepts the ground-truth gradients $\nabla_\theta \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)$ shared by participants to reconstruct the private training data $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, with $\boldsymbol{y}^*$ denoting the ground-truth labels. Following the formulation in [Balunovic *et al.*, 2022] and assuming $\boldsymbol{y}^*$ is known, GIAs can be modeled as an optimization problem:

$$\arg\min_{\boldsymbol{x}} \mathcal{D}\left(\nabla_\theta \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*), \nabla_\theta \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)\right) \tag{1}$$

where the dummy input $\boldsymbol{x}$ is iteratively optimized to minimize the divergence $\mathcal{D}(\cdot)$. Common choices for $\mathcal{D}(\cdot)$ include the ℓ_1-norm, the ℓ_2-norm, or cosine distance to quantify the mismatch between the dummy and ground-truth gradients.

3.2 Embedding Regularization

A critical challenge in optimizing token embeddings $\boldsymbol{x}$ via gradient descent is the phenomenon of norm divergence. In the absence of constraints, the ℓ_2-norms of the optimized embeddings tend to grow uncontrollably to minimize the gradient matching loss. This causes the embeddings to drift away from the valid latent space of the language model's vocabulary, resulting in the reconstruction of semantically meaningless or illogical text [Balunovic *et al.*, 2022].

To mitigate this, Embedding Regularization $\mathcal{R}_{\text{reg}}(\cdot)$ is employed as a prior to anchor the scale of the optimized embeddings. It enforces the average norm of the candidate embeddings to align with the statistical properties of the pre-trained embedding matrix, as defined in Equation (2):

$$\mathcal{R}_{\text{reg}}(\boldsymbol{x}) = \left(\frac{1}{n} \sum_{i=1}^{n} \|x_i\|_2 - \frac{1}{V} \sum_{j=1}^{V} \|e_j\|_2 \right)^2 \tag{2}$$

where n denotes the sequence length of the dummy embeddings $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, and V represents the vocabulary size of the embedding matrix $W_{\text{embed}} = \{e_1, \ldots, e_V\}$. This regularization serves as a foundational component in the continuous optimization phase of our proposed CDO-GIA. However, simply applying Equation (2) with a fixed weight leads to embedding homogenization in large-batch settings. Therefore, CDO-GIA adapts this term via a dynamic weight decay mechanism to balance structural constraints with reconstruction diversity.

3.3 Language Model Prior

Minimizing the gradient distance solely is often insufficient for textual reconstruction, as the optimization landscape is highly non-convex and under-determined. Ideally, the reconstructed text should not only match the gradients but also adhere to the syntactic and semantic norms of natural language. To enforce this, we incorporate a Language Model Prior as a critical regularization term.

In our framework, the Language Model Prior serves to evaluate the fluency of candidate sequences, guiding the discrete optimization process away from semantically incoherent local optima. Specifically, we utilize a pre-trained model to compute the perplexity $\mathcal{L}_{\text{LLM}}(\cdot)$ of a token sequence. A lower perplexity indicates a higher probability that the sequence aligns with natural linguistic distributions. For a token sequence $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$, the perplexity is defined as:

$$\mathcal{L}_{\text{LLM}}(\boldsymbol{t}) = -\frac{1}{n} \sum_{i=0}^{n-1} \log P(t_{i+1} \mid t_1, t_2, \ldots, t_i) \tag{3}$$

where $P(t_{i+1} \mid \ldots)$ denotes the conditional probability of the next token predicted by the pre-trained model. This prior provides the necessary semantic guidance for our Tabu Beam Search mechanism to effectively explore the discrete token space.

4 Methodology

4.1 Overall Structure

Fig. 2 illustrates the comprehensive framework of CDO-GIA within a standard Federated Learning paradigm. In this setting, clients collaboratively train a model by uploading local gradient updates to a central server. We assume an honest-but-curious server (the adversary) who intercepts these shared gradients and model parameters to reconstruct the private

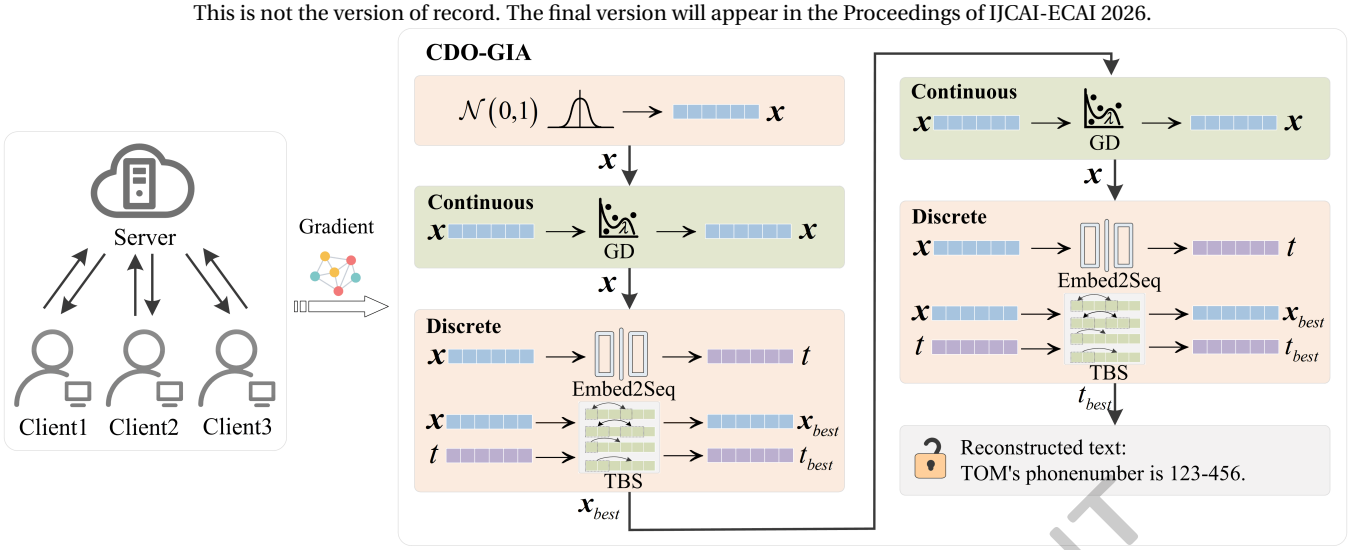


Figure 2: The framework of the proposed CDO-GIA. **Left:** In the FL setting, an honest-but-curious server intercepts gradients uploaded by clients to reconstruct private text. **Right:** The attack process begins with initializing dummy embeddings from a normal distribution $\mathcal{N}(0,1)$. It then executes an alternating optimization loop: the **Continuous Phase** updates embeddings via Gradient Descent (GD) with dynamic weight decay, followed by the **Discrete Phase** which maps embeddings to token sequence (Embed2Seq) and refines the sequence using Tabu Beam Search (TBS).

training text. The attack process is initialized by sampling dummy token embeddings $\boldsymbol{x}$ from a standard normal distribution $\mathcal{N}(0,1)$. Subsequently, CDO-GIA executes an alternating continuous-discrete optimization strategy to progressively refine $\boldsymbol{x}$:

Continuous Optimization Phase. The initialized or updated embedding $\boldsymbol{x}$ is first refined via Gradient Descent (GD) to minimize the reconstruction loss. As depicted in the green block of Fig. 2, this phase incorporates a dynamic weight decay mechanism to mitigate embedding homogenization while matching the ground-truth gradients.

Discrete Optimization Phase. The continuous embedding is then projected onto the discrete token space via an Embed2Seq mapper to generate a token sequence $\boldsymbol{t}$. To bridge the modality gap, the Tabu Beam Search (TBS) mechanism is applied simultaneously to $\boldsymbol{x}$ and $\boldsymbol{t}$. This module performs discrete transformations to identify the optimal pair $(\boldsymbol{x}_{\text{best}}, \boldsymbol{t}_{\text{best}})$ that minimizes a joint heuristic of gradient loss and text perplexity.

4.2 Continuous Optimization

The primary objective of the continuous optimization phase is to recover the token embeddings $\boldsymbol{x}$ by minimizing the discrepancy between the dummy gradients and the ground-truth gradients. We employ the cosine distance as the gradient reconstruction loss $\mathcal{L}_{\text{grad}}(\cdot)$, defined as:

$$\mathcal{L}_{\text{grad}}(\boldsymbol{x}) = 1 - \frac{\langle \nabla_\theta \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*), \nabla_\theta \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*) \rangle}{\|\nabla_\theta \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*)\|_2 \, \|\nabla_\theta \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)\|_2} \tag{4}$$

While embedding regularization $\mathcal{R}_{\text{reg}}(\cdot)$ is essential for preventing norm divergence (as discussed in Section 3.2), we identify a critical limitation when applying it to large-batch settings. In previous work [Balunovic *et al.*, 2022], a fixed regularization weight forces the embeddings of all samples in a batch to converge toward the same average norm. We term this phenomenon embedding homogenization. In practice, however, natural text exhibits significant variance in sentence length and complexity, implying that the optimal embedding norms should vary across samples. Consequently, a rigid regularization constraint suppresses this natural variance, leading to performance degradation as the batch size increases.

To resolve this conflict, we propose a Dynamic Weight Decay mechanism. The core idea is to apply strong regularization in the early stages to constrain the search space and prevent divergence, while progressively relaxing this constraint to allow for fine-grained feature recovery and length diversity. The total reconstruction loss $\mathcal{L}_{\text{rec}}(\cdot)$ is formulated as:

$$\mathcal{L}_{\text{rec}}(\boldsymbol{x}) = (1 - \alpha_{\text{reg}})\mathcal{L}_{\text{grad}}(\boldsymbol{x}) + \alpha_{\text{reg}}\mathcal{R}_{\text{reg}}(\boldsymbol{x}) \tag{5}$$

where $\alpha_{\text{reg}} \in [0,1]$ is the dynamic decay factor. It is initialized at 1 and linearly decays to 0 throughout the optimization iterations. Finally, the token embeddings are updated via gradient descent with a learning rate λ:

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \lambda \nabla_{\boldsymbol{x}} \mathcal{L}_{\text{rec}}(\boldsymbol{x}) \tag{6}$$

4.3 Discrete Optimization

Since the continuous embeddings $\boldsymbol{x}$ optimized in the previous phase do not strictly correspond to discrete tokens, we require a discretization process to map them onto the language model's vocabulary. This is followed by a discrete search to refine the token order and recover local dependencies.

Embed2Seq Mapping. Following [Balunovic *et al.*, 2022], we employ an Embed2Seq mapper to project the continuous embedding $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ onto a discrete token sequence $\boldsymbol{t} = \{t_1, \ldots, t_n\}$. For each embedding vector x_i, the mapper identifies the nearest token t_i in the vocabulary V by maximizing the cosine similarity:

$$t_i = \arg \max_{1 \le j \le |V|} \frac{\langle x_i, e_j \rangle}{\|x_i\|_2 \|e_j\|_2} \tag{7}$$

where e_j denotes the j-th vector in the pre-trained embedding matrix W_{embed}.

Tabu Beam Search (TBS). The discrete sequence obtained from Embed2Seq often contains position errors or local inconsistencies. To address this, we introduce the TBS mechanism.

The process begins by initializing the best solution $(\boldsymbol{x}_{\text{best}}, \boldsymbol{t}_{\text{best}})$ with the output of the Embed2Seq mapper. A priority queue Q is established to maintain the beam of the top-B candidate sequences, ordered by their heuristic values $h(\boldsymbol{x}, \boldsymbol{t})$. To prevent cyclic searching and revisiting suboptimal states, a Tabu list $\mathcal{T}$ is maintained with a maximum capacity L.

The search proceeds iteratively for K rounds. In each iteration, the algorithm retrieves all current candidates from Q. For every candidate, four distinct discrete transformation operators (T_1 to T_4) are applied independently to generate new potential sequences. Each newly generated candidate is then checked against the Tabu list. If it has not been visited, we evaluate it using the heuristic function defined in Equation (8). If a new candidate surpasses the current global best solution, $\boldsymbol{x}_{\text{best}}$ and $\boldsymbol{t}_{\text{best}}$ are updated immediately. Valid candidates are then added to the Tabu list and pushed into the priority queue. Finally, at the end of each iteration, the beam size constraint is enforced by retaining only the top-B candidates in Q.

Heuristic Function. The quality of each candidate is evaluated by balancing gradient consistency with linguistic plausibility:

$$h(\boldsymbol{x}, \boldsymbol{t}) = \mathcal{L}_{\text{grad}}(\boldsymbol{x}) + \alpha_{\text{LLM}} \mathcal{L}_{\text{LLM}}(\boldsymbol{t}) \tag{8}$$

where $\mathcal{L}_{\text{LLM}}$ is the perplexity calculated by the language model prior, and α_{LLM} is a weighting factor.

Discrete Transformations. To efficiently explore the solution space and correct both lexical and phrasal ordering errors, we define four transformation operators. In each search step, these operators are applied to a candidate sequence to generate variations:

- T_1 **(Unigram Swap):** Randomly selects two distinct indices in the sequence and swaps the single tokens at these positions. This operation targets isolated word-level ordering errors.
- T_2 **(Bigram Swap):** Randomly selects two non-overlapping bigrams (pairs of consecutive tokens) and swaps their positions. This operation aims to adjust the order of local phrases or collocations.
- T_3 **(Unigram Shift):** Randomly selects a single token from index p_1 and moves it to the position immediately following index p_2. This facilitates the correction of long-range dependency errors for individual words.
- T_4 **(Bigram Shift):** Randomly selects a bigram (starting at index p_1) and moves this pair of tokens to the position immediately following index p_2. This enables the relocation of entire phrases within the sentence structure.

Algorithm 1 Gradient Inversion Attack via Continuous-Discrete Optimization

Input:
  n_i: Number of iterations
  n_c: Number of continuous optimizations per iteration
  n_0: Number of initializations
  λ: Learning rate
Output:
  $\boldsymbol{t}_{\text{best}}$: Best token sequence
Initialization (Warm-Start):
  Sample set $\mathcal{X} = \{\boldsymbol{x}^{(k)} \sim \mathcal{N}(0, I)\}_{k=1}^{n_0}$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^{(k^*)}$ where $k^* = \arg\min_k \mathcal{L}_{\text{rec}}(\boldsymbol{x}^{(k)})$
for $i = 1$ **to** n_i **do**
  Phase 1: Continuous Optimization
  for $c = 1$ **to** n_c **do**
    Compute gradient $\nabla_{\boldsymbol{x}} \mathcal{L}_{\text{rec}}(\boldsymbol{x})$ {Eq. 5}
    $\boldsymbol{x} \leftarrow \boldsymbol{x} - \lambda \nabla_{\boldsymbol{x}} \mathcal{L}_{\text{rec}}(\boldsymbol{x})$
  end for
  Phase 2: Discrete Optimization
    $\boldsymbol{t} \leftarrow$ Embed2Seq$(\boldsymbol{x})$ {Eq. 7}
    $(\boldsymbol{x}_{\text{best}}, \boldsymbol{t}_{\text{best}}) \leftarrow$ TBS$(\boldsymbol{x}, \boldsymbol{t})$ {Tabu Beam Search}
  Feedback Update
    $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{best}}$
end for
return $\boldsymbol{t}_{\text{best}}$

4.4 Gradient Inversion Attack

The complete workflow of CDO-GIA is formalized in Algorithm 1. To mitigate the sensitivity to initialization inherent in non-convex optimization, we employ a warm-start strategy. Specifically, we sample n_0 candidate embeddings from a standard normal distribution $\mathcal{N}(0, I)$ and select the candidate with the minimal initial reconstruction loss as the starting point $\boldsymbol{x}$.

The core of CDO-GIA lies in the alternating continuous-discrete optimization strategy that cycles through two phases for n_i iterations. In the continuous phase, the embedding $\boldsymbol{x}$ is updated via gradient descent for n_c steps. This process minimizes the reconstruction loss $\mathcal{L}_{\text{rec}}$, leveraging our dynamic weight decay mechanism (Section 4.2) to balance gradient matching with embedding diversity. Subsequently, in the discrete phase, the continuous embedding is projected onto the discrete token space via the Embed2Seq mapper. The Tabu Beam Search mechanism is then invoked to rigorously explore the local token space, identifying the optimal embedding-sequence pair $(\boldsymbol{x}_{\text{best}}, \boldsymbol{t}_{\text{best}})$.

More importantly, the embedding $\boldsymbol{x}_{\text{best}}$ obtained from the discrete search is fed back to initialize the next round of continuous optimization. This feedback loop ensures that the gradient-based refinement proceeds from a semantically valid anchor, effectively bridging the gap between the continuous and discrete manifolds.

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate on three binary text classification benchmarks representing varying sequence lengths:

Method	CoLA			SST-2			RottenTomatoes		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
DLG [Zhu *et al.*, 2019]	40.14	3.06	36.67	43.46	6.10	38.96	17.98	0.12	16.47
TAG [Deng *et al.*, 2021]	60.51	7.52	47.29	63.46	10.48	48.58	27.91	0.65	21.49
LAMP [Balunovic *et al.*, 2022]	59.88	16.65	51.61	58.25	18.42	50.94	24.61	1.55	21.10
FET [Gao *et al.*, 2025]	51.71	19.86	40.03	50.68	17.69	41.64	20.15	0.32	19.82
CDO-GIA (Ours)	**68.04**	**28.69**	**60.14**	**66.21**	**27.13**	**56.89**	**32.58**	**3.12**	**29.91**

Table 1: Experimental results (ROUGE scores) on three datasets (Batch Size = 4). **Bold** denotes the best results.

CoLA [Warstadt *et al.*, 2019] (short, 5–9 words), SST-2 [Socher *et al.*, 2013] (moderate, 3–13 words), and Rotten Tomatoes [Pang and Lee, 2005] (long, 14–27 words).

Baselines. We compare CDO-GIA against four state-of-the-art gradient inversion attacks, covering three distinct optimization paradigms. For continuous optimization, we select DLG [Zhu *et al.*, 2019] and TAG [Deng *et al.*, 2021]. For discrete optimization, we include FET [Gao *et al.*, 2025]. Additionally, we compare against LAMP [Balunovic *et al.*, 2022], a continuous-discrete method similar to our approach.

Metrics. To quantify reconstruction fidelity, we employ the standard ROUGE metrics [Lin, 2004]: ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L), measuring unigram, bigram, and longest common subsequence overlap, respectively.

Model Configurations. Following the protocol in [Balunovic *et al.*, 2022], we employ the pre-trained `BERT-base-uncased` [Devlin *et al.*, 2019] as the federated learning target, and utilize `GPT-2-small` [Radford *et al.*, 2019] as the language model prior to compute perplexity in CDO-GIA.

5.2 Performance Comparison

This section evaluates the performance of the proposed CDO-GIA method against state-of-the-art baselines across multiple datasets and batch sizes. We first compare all methods on three text classification datasets (CoLA, SST-2 and RottenTomatoes) with the batch size of language model set to 4. The quantitative results are summarized in Table 1.

Quantitative Observations. As evidenced by Table 1, CDO-GIA consistently achieves the highest scores across all metrics and datasets, establishing a new state-of-the-art performance for gradient inversion attacks. Specifically, LAMP and CDO-GIA (continuous-discrete method) significantly outperform DLG and TAG (purely continuous method). This substantial margin confirms that discrete optimization is essential for bridging the gap between raw gradients and coherent text. Notably, CDO-GIA still exhibits superior lexical and structural fidelity compared to LAMP. The dominance in ROUGE-1 demonstrates that our dynamic weight decay effectively prioritizes the recovery of individual tokens. Moreover, the substantial gains in ROUGE-2 and ROUGE-L suggest that the proposed TBS mechanism successfully captures local bigrams and long-range dependencies, yielding reconstructed sequences with significantly higher structural integrity.

Scalability to Large Batch Sizes. To verify the robustness of CDO-GIA, we evaluate its performance under varying batch sizes on the CoLA dataset (Table 2). All methods experience performance degradation as the batch size increases—due to the increased complexity of the gradient aggregation, while CDO-GIA exhibits remarkable resilience. Notably, the performance margin between CDO-GIA and the strongest baseline (LAMP) widens significantly in larger batch settings. At a batch size of 8, CDO-GIA outperforms LAMP by 25.01% in Rouge-1, 155.64% in Rouge-2 and 22.72% in Rouge-L. This demonstrates that CDO-GIA is uniquely capable of mitigating the "gradient dilution" effect inherent in large-batch training. While FET [Gao *et al.*, 2025] achieves near-perfect reconstruction at batch size of 1, its performance collapses at batch size exceed 1, suggesting that purely discrete search methods struggle with the exponentially expanded search space of multi-sample gradients. In contrast, CDO-GIA's continuous-discrete optimization strategy ensures both high fidelity and optimization stability.

Reconstruction Stability across Samples. Fig. 3 presents the cumulative reconstruction performance across 100 samples with a batch size of 4, where each point reflects the running average up to that sample. Notably, CDO-GIA consistently achieves the highest scores across all three ROUGE metrics throughout the evaluation, significantly outperforming all baselines by substantial margins. While all methods show initial fluctuations, CDO-GIA demonstrates progressively more stable performance in later samples, particularly for ROUGE-2 and ROUGE-L metrics. This sustained superiority and stability confirm CDO-GIA's robustness in reconstructing diverse text samples.

5.3 Ablation Study

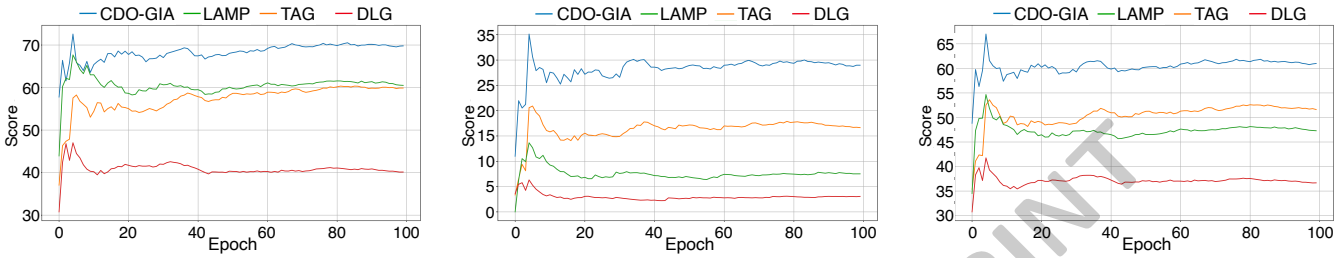
To validate the contribution of each individual component in CDO-GIA, we conduct an ablation study on the CoLA dataset (Batch Size = 4). The results are summarized in Table 3.

Impact of Dynamic Weight Decay. We first evaluate the efficacy of our proposed dynamic mechanism by replacing it with a static embedding regularization (denoted as *w/o Dynamic*). As shown in Table 3, this leads to a sharp performance decline, particularly in ROUGE-1 (dropping from 68.04 to 57.86). This degradation confirms that static regularization induces embedding homogenization in large-batch settings, severely hampering unigram recovery, whereas our dynamic decay effectively preserves lexical diversity.

Impact of Tabu Beam Search. To assess the necessity of discrete optimization, we remove the TBS phase entirely (*w/o*

Method	batch size=1			batch size=2			batch size=4			batch size=8		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
DLG [Zhu *et al.*, 2019]	63.55	7.41	48.20	51.92	6.37	43.31	40.14	3.06	36.67	28.06	0.11	27.81
TAG [Deng *et al.*, 2021]	78.22	9.89	54.85	71.00	9.19	51.99	60.51	7.52	47.29	46.01	2.95	39.38
LAMP [Balunovic *et al.*, 2022]	90.41	53.92	78.58	74.09	29.52	62.34	59.88	16.65	51.61	43.50	6.38	40.09
FET [Gao *et al.*, 2025]	**97.53**	**74.22**	**86.94**	67.60	**52.82**	65.46	51.71	19.86	40.03	36.71	8.89	30.46
CDO-GIA (Ours)	94.37	65.57	84.36	**83.66**	45.01	**73.30**	**69.79**	**28.98**	**61.00**	**54.38**	**16.31**	**49.20**

Table 2: Experimental results (ROUGE scores) on the CoLA dataset under various batch sizes. **Bold** denotes the best results.



(a) ROUGE-1 score (b) ROUGE-2 score (c) ROUGE-L score

Figure 3: Reconstruction performance across 100 test samples (Batch Size = 4).

Method	R-1	R-2	R-L
w/o Dynamic	57.86	16.38	51.10
w/o TBS	67.88	23.01	57.49
w/o LM Prior	67.68	20.74	57.68
CDO-GIA (Full)	**68.04**	**28.69**	**60.14**

Table 3: Ablation study of different components on the CoLA dataset (Batch Size = 4).

Method	R-1	R-2	R-L
LAMP [Balunovic *et al.*, 2022]	59.88	16.65	51.61
LAMP + Dynamic	**67.46**	**25.35**	**57.80**
LAMP + TBS	61.24	20.47	54.06

Table 4: Ablation study of integrating proposed modules into the LAMP baseline on CoLA dataset (Batch Size = 4).

TBS), generating text directly via Embed2Seq mapping. The significant gap in ROUGE-2 and ROUGE-L compared to the full method highlights that while continuous optimization can recover individual tokens, it struggles with ordering. TBS is thus indispensable for refining token order and ensuring structural coherence.

Impact of Language Model Prior. Finally, removing the language model guidance (setting $\alpha_{\text{LLM}} = 0$, denoted as *w/o LM Prior*) results in a noticeable drop in ROUGE-L. This indicates that without perplexity guidance, the reconstruction may satisfy gradient constraints but fail to adhere to natural language grammar. The LM prior serves as a crucial regularizer to enforce linguistic plausibility.

Component Generalizability. To further substantiate the versatility of our proposed modules, we integrate the dynamic weight decay and TBS mechanisms individually into LAMP. As detailed in Table 4, equipping LAMP with Dynamic Weight Decay yields a remarkable performance boost, achieving relative improvements of 12.66% in ROUGE-1 and a striking 52.25% in ROUGE-2. This substantial gain validates that mitigating embedding homogenization is a universal prerequisite for high-fidelity reconstruction, irrespective of the underlying framework. Similarly, the integration of TBS leads to consistent structural refinements, resulting in a 22.94% relative increase in ROUGE-2. These results demonstrate that our components are not merely tailored for CDO-GIA but serve as effective, plug-and-play modules that can significantly enhance existing GIAs.

6 Conclusion

In this paper, we propose CDO-GIA, a continuous-discrete GIAs method that extends the effective attack batch size from 4 to 8. Specifically, we introduce a dynamic weight decay mechanism to mitigate embedding homogenization and a prior-guided Tabu Beam Search to enable precise discrete exploration. Extensive experiments demonstrate that CDO-GIA significantly outperforms state-of-the-art baselines in reconstructing semantically coherent text. Crucially, our findings refute the common assumption that increasing batch size is a sufficient defense, underscoring the urgent need for more rigorous privacy-preserving strategies in federated learning.

Limitation Discussion. While CDO-GIA demonstrates robust reconstruction capabilities, handling extremely long documents remains a common challenge for optimization-based attacks. This is primarily attributed to the exponential expansion of the permutation search space in discrete optimization. Future work could address this by leveraging stronger generative priors to guide the reconstruction of long-range dependencies.

Acknowledgments

We thank all the anonymous reviewers for their constructive feedback. This work was supported by the Zhejiang Province "Lingyan" Key R&D Project of China (No. 2026C02A2002), and Zhejiang Provincial Natural Science Foundation of China (No. LQN26F030003).

Contribution Statement

Jiajie Wang and Weibo Xu contributed equally to this work.

References

[Aono *et al.*, 2017] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security*, 13(5):1333–1345, 2017.

[Balunovic *et al.*, 2022] Mislav Balunovic, Dimitar Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Deng *et al.*, 2021] Jieren Deng, Yijue Wan, Ji Li, Chenghong Wang, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. Tag: Gradient attack on transformer-based language models. In *2021 Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pages 3600–3610. Association for Computational Linguistics (ACL), 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Gao *et al.*, 2025] Ying Gao, Yuxin Xie, Huanghao Deng, and Zukun Zhu. Gradient inversion attack in federated learning: Exposing text data through discrete optimization. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2582–2591, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[Geiping *et al.*, 2020] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

[Kairouz *et al.*, 2021] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[Li *et al.*, 2023] Zhuohang Li, Jiaxin Zhang, and Jian Liu. Speech privacy leakage from shared gradients in distributed learning. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Luo *et al.*, 2021] Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Feature inference attack on model predictions in vertical federated learning. In *2021 IEEE 37th international conference on data engineering (ICDE)*, pages 181–192. IEEE, 2021.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Melis *et al.*, 2019] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[Mothukuri *et al.*, 2021] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[Nasr *et al.*, 2019] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[Pang and Lee, 2005] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, 2005.

[Phong *et al.*, 2017] Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. Privacy-preserving deep learning: Revisited and enhanced. In *Applications and Techniques in Information Security: 8th International Conference, ATIS 2017, Auckland, New Zealand, July 6–7, 2017, Proceedings*, pages 100–110. Springer, 2017.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al.

Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[Truex *et al.*, 2019] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, 14(6):2073–2089, 2019.

[Wang *et al.*, 2024] Xiaodong Wang, Longfei Wu, and Zhitao Guan. Graddiff: Gradient-based membership inference attacks against federated distillation with differential comparison. *Information Sciences*, 658:120068, 2024.

[Warstadt *et al.*, 2019] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[Wei *et al.*, 2020] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating client privacy leakages in federated learning. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*, pages 545–566. Springer, 2020.

[Wu *et al.*, 2025] Liwen Wu, Zhizhi Liu, Bin Pu, Kang Wei, Hangcheng Cao, and Shaowen Yao. Dggi: Deep generative gradient inversion with diffusion model. *Information Fusion*, 113:102620, 2025.

[Ye *et al.*, 2024] Zipeng Ye, Wenjian Luo, Qi Zhou, and Yubo Tang. High-fidelity gradient inversion in distributed learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19983–19991, 2024.

[Yin *et al.*, 2021] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16337–16346, 2021.

[Zhu *et al.*, 2019] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.