MMGIA: Gradient Inversion Attack Against Multimodal Federated Learning via Intermodal Correlation

Lele Zheng[1,2], **Yang Cao**[2], **Leo Yu Zhang**[3], **Wei Wang**[4*], **Yulong Shen**[1] and **Xiaochun Cao**[4]
[1]School of Computer Science and Technology, Xidian University
[2]Department of Computer Science, Institute of Science Tokyo
[3]School of Information and Communication Technology, Griffith University
[4]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
llzhengstu@gmail.com, wangwei29@mail.sysu.edu.cn

Abstract

Multimodal federated learning (MMFL) enables collaborative model training across multiple modalities, such as images and text, without requiring direct data sharing. However, the inherent correlations between modalities introduce new privacy vulnerabilities, making MMFL more susceptible to gradient inversion attacks. In this work, we propose MMGIA, an intermodal correlation-driven gradient inversion attack that systematically exploits multimodal correlation to enhance data reconstruction quality. MMGIA consists of a two-stage optimization framework: the first stage independently reconstructs each modality using traditional gradient inversion techniques, while the second stage refines these reconstructions through pre-trained feature extractors to align modalities in a shared latent space. To further improve reconstruction accuracy, we introduce a quality-weighted fusion strategy, which dynamically integrates multimodal embeddings into a global fused representation that serves as a guiding signal for refining each modality's reconstruction. This ensures that high-quality reconstructions contribute more to the optimization process, preventing degradation in well-reconstructed modalities while enhancing weaker ones. We conduct extensive experiments on multiple multimodal scenarios, demonstrating that MMGIA outperforms both the only existing multimodal attack and state-of-the-art single-modal attacks, revealing the heightened privacy risks in MMFL.

1 Introduction

Multimodal learning integrates data from diverse modalities (e.g., text, images, audio) to develop models capable of understanding and leveraging information from multiple sources [Che *et al.*, 2023]. By combining complementary features across modalities, multimodal learning achieves significant improvements in tasks such as emotion recognition [Feng and Narayanan, 2022; Chen and Zhang, 2022], vision–language interaction [Liu *et al.*, 2020; Guo *et al.*, 2023], and medical diagnosis [Bernecker *et al.*, 2022; Agbley *et al.*, 2021]. These models capture richer contextual information and semantic relationships, enhancing both performance and generalization. However, privacy concerns remain a critical challenge, severely hindering the broader adoption and development of multimodal learning.

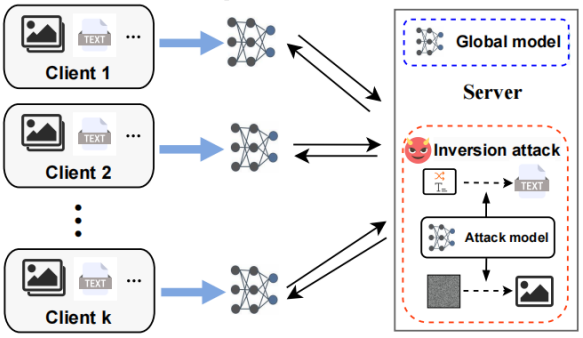


Figure 1: Multimodal federated learning and gradient inversion attack.

Federated Learning (FL) is a distributed machine learning paradigm that restricts model training to local clients while enabling collaborative learning through the exchange of gradients or model updates [Kairouz *et al.*, 2021]. By eliminating the need for centralized data storage and transmission, FL significantly mitigates privacy risks while harnessing the potential of distributed data from diverse sources. Recently, FL has been extended to multimodal scenarios, resulting in the development of multimodal federated learning (MMFL) [Feng *et al.*, 2023; Lin *et al.*, 2023]. MMFL enables clients to train models on local multimodal datasets, such as those combining text and image data, and share modality-specific gradients with a central server. The server aggregates these gradients to update a global model, as illustrated in Fig. 1. By integrating the strengths of federated learning and multimodal processing, MMFL facilitates collaborative training across diverse data types, addressing the growing demand for multimodal solutions in real-world applications, including healthcare, e-commerce, and social media analytics.

Although FL mitigates the risk of direct data leakage by sharing only gradients, studies have demonstrated that gradients can still encode substantial information, potentially re-

*Corresponding author

vealing sensitive details about the original data. Gradient inversion attacks (GIAs), a widely recognized threat, enable attackers to reconstruct training data by minimizing the distance between shared gradients and those generated from candidate inputs [Zhu *et al.*, 2019; Geiping *et al.*, 2020]. These risks are further amplified in multimodal federated learning due to the correlation between modalities. For example, gradients from the text modality may encode semantic clues that facilitate the reconstruction of visual data, while visual gradients can reveal information useful for recovering text. This intermodal leakage significantly increases privacy risks in MMFL, making it more vulnerable to attacks compared to single-modal scenarios. While researchers have begun exploring this issue, the only existing work relies on unrealistic assumptions, rendering it ineffective in practice.

Existing Attack. Research on gradient inversion attacks in MMFL is limited, with MGIS [Liu *et al.*, 2024] being the only existing method. However, its practicality is constrained by unrealistic assumptions and methodological issues. MGIS assumes that the adversary can determine in advance which modality achieves better reconstruction by comparing reconstructed images and text using metrics like PSNR for images and word matching rates for text. Based on this, the adversary uses the label from the better-performing modality to assist reconstruction of the weaker one. However, this assumption may not hold in practice, as adversaries typically lack access to the original data, making such evaluations infeasible. Moreover, these metrics are modality-specific and not directly comparable, complicating the identification of the better-performing modality. In their experiments, MGIS circumvents these limitations by assuming that the text modality produces superior results and using text labels to assist image reconstruction. Thus, effectively leveraging intermodal correlations to enhance gradient inversion attacks remains an open challenge, underscoring the importance of our work.

Our Contributions. In this work, we propose MMGIA, a novel intermodal correlation-driven gradient inversion attack for multimodal federated learning. MMGIA employs a two-stage framework to systematically reconstruct multimodal data. The first stage independently reconstructs each modality using traditional gradient inversion techniques, while the second refines the reconstructions by leveraging pre-trained feature extractors to map the reconstructed data into a shared latent space, explicitly capturing intermodal relationships. A dynamic quality-weighted fusion strategy is applied to integrate multimodal embeddings into a global fused representation, which serves as a guiding signal to enhance the reconstruction of each modality by balancing contributions from stronger and weaker modalities. Additionally, MMGIA incorporates an iterative optimization mechanism, where the outputs of the second stage are reintroduced as inputs to the first, with iterations continuing until the similarity between modality embeddings in the shared latent space exceeds a predefined threshold, ensuring high reconstruction accuracy and semantic alignment. Through extensive experiments on diverse multimodal datasets, we demonstrate that MMGIA significantly outperforms existing multimodal and single-modal attacks, revealing heightened privacy risks in MMFL and underscoring the need for privacy-preserving mechanisms.

2 Related Work

2.1 Multimodal Federated Learning

Multimodal federated learning extends traditional FL to handle multiple modalities (e.g., text, images, audio), leveraging their complementary characteristics to build global models. Early work by [Saeed *et al.*, 2020] introduced scalogram-signal correlation learning, a self-supervised method for robust multimodal representation learning in FL. More recently, [Xiong *et al.*, 2022] proposed MMFed, a multimodal FL framework using cross-attention, and [Chen and Zhang, 2022] introduced FedMSplit to address missing modalities. CreamFL [Yu *et al.*, 2023] developed a contrastive representation-level ensemble to aggregate heterogeneous multimodal clients. [Feng *et al.*, 2023] proposed the first comprehensive benchmark dataset and evaluation framework specifically designed for multimodal federated learning. [Li *et al.*, 2024] proposed MuEP, a multimodal benchmark for embodied planning addressing diversity, modality limitations, and coarse metrics.

2.2 Gradient Inversion Attack

Gradient inversion attacks have emerged as a critical threat to the privacy of federated learning, where adversaries aim to reconstruct client data from shared gradients. Existing studies have primarily focused on gradient inversion in single-modal FL, such as reconstructing images or text.

Image. [Zhu *et al.*, 2019] demonstrated in their pioneering work that sharing gradients can leak data privacy, introducing DLG to reconstruct images by minimizing the Euclidean distance between true and dummy gradients. However, DLG performs best with small images and batch sizes. Building on this, [Zhao *et al.*, 2020] proposed iDLG, which extracts labels from gradients but is limited to the case where the batch size is one. [Geiping *et al.*, 2020] improved image reconstruction by replacing Euclidean with cosine distance and applying total variation for denoising. CPL [Wei *et al.*, 2020] enhanced attacks further with L_2 distance and label-based regularization. Recently, GIFD [Fang *et al.*, 2023] uses the GAN prior via optimizing the feature domain of the generative model to generate stable and high-fidelity inversion results.

Text. Most GIAS focus on image reconstruction, with recent work exploring textual data. DLG [Zhu *et al.*, 2019] pioneered reconstruction from transformer gradients, while TAG [Deng *et al.*, 2021] improved it with an L_1 loss term. LAMP [Balunovic *et al.*, 2022] incorporated language model priors and discrete-continuous optimization to recover text, and FILM [Gupta *et al.*, 2022] extended gradient-based text leakage in LMs to settings with larger batch sizes. Fowl et al. [Fowl *et al.*, 2022] proposed another attack targeting NLP tasks, but their approach relied on a stronger assumption that the server could send malicious updates to the client.

Multimodal data. As stated in Section 1, MGIS [Liu *et al.*, 2024] is the only gradient inversion attack method for MMFL, but its practicality is limited.

3 Threat Model

Threat Model: Multimodal data, due to interdependencies between modalities, reveals additional information and in-

creases the risk of attackers reconstructing sensitive data. We consider a horizontal multimodal federated learning setting with two widely used modalities: image and text, common in applications like medical diagnosis and social recommendation. Each client trains local models on multimodal data and independently shares the gradients of different modalities with the central server [Chen and Li, 2022]. Since different modalities use distinct architectures, fused gradients (e.g., concatenation-based) can still be easily separated by modality. The adversary leverages exchanged gradients to reconstruct private data.

Adversarial Goal: In multimodal gradient inversion attacks, the adversary's goal is to reconstruct the original multimodal input data, including texts and images, from the shared gradients exchanged during training. By leveraging the modality-specific information contained in the gradients and exploiting intermodal correlations, the adversary iteratively refines the reconstruction process. This approach combines the stability of single-modality reconstruction with the complementarity of cross-modal alignment, aiming to accurately recover the original text and image inputs, thereby exposing the vulnerabilities in multimodal federated learning.

Adversary's Knowledge: As shown in Fig. 1, we assume the adversary to be an honest-but-curious server with specific knowledge that facilitates the attack. First, the adversary has access to the shared gradients exchanged between clients and the central server during training. Second, the adversary can utilize pre-trained auxiliary models to extract embeddings from the reconstructed data of each modality and map them into a shared latent space, effectively capturing intermodal correlations. This assumption is plausible, as many publicly available models, such as CLIP, can perform this task. The mapping enables cross-modal alignment and refinement, which are crucial for enhancing reconstruction quality. Finally, the adversary is aware of the global model architecture, including its structure and loss function.

4 Design of MMGIA

Fig. 2 illustrates the framework of our proposed methodology, which consists of two main stages: modality-specific reconstruction and correlation-driven refinement. In the first stage, traditional gradient inversion techniques are applied to reconstruct preliminary results for each modality from the shared gradients. The second stage refines these results by exploiting latent consistency constraints between modalities.

4.1 Modality-Specific Reconstruction

In the first stage, we utilize the shared gradients for each modality and independently apply gradient inversion attacks to the image and text modalities to generate initial reconstruction results. The single-modal gradient inversion attack is formulated as an optimization problem that reconstructs the original input by minimizing the difference between the true gradients and the gradients computed from the reconstructed data. The general objective is:

$$\hat{x}^*, \hat{y}^* = \arg\min_{\hat{x},\hat{y}} \mathcal{L}_{\text{grad}}(\nabla W(x,y), \nabla W(\hat{x},\hat{y})), \quad (1)$$

where $\mathcal{L}_{\text{grad}}$ is a distance function and $(\hat{x},\hat{y})$ denote dummy data optimized using gradient descent to have similar gradients $\nabla W(\hat{x},\hat{y})$ to true data (x,y).

Gradient Inversion for Image Modality: For the image modality, we can define the optimization objective as the L_2 distance between the shared gradients and the generated gradients. To further enhance reconstruction quality, we incorporate additional regularization terms, i.e., total variation (TV) regularization, which promotes smoothness in the reconstructed image. The reconstruction error is:

$$\mathcal{L}^I_{\text{grad}} = \sum_{l=1}^{L} \left\| \nabla_{W_{(l)}} \mathcal{L}(\hat{x}_I, \hat{y}_I) - \nabla W_{(l)} \right\|_2 + \text{TV}(\hat{x}_I), \quad (2)$$

where L denotes the total layers. After multiple iterations, this approach produces a reconstructed image that approximates the original input.

Gradient Inversion for Text Modality: For the text modality, we can use a cosine reconstruction loss to quantify the similarity between the reconstructed and shared gradients, expressed as:

$$\mathcal{L}^T_{\text{grad}} = 1 - \frac{1}{L} \sum_{l=1}^{L} \frac{\nabla_{W_{(l)}} \mathcal{L}(\hat{x}_T, \hat{y}_T) \cdot \nabla W_{(l)}}{\|\nabla_{W_{(l)}} \mathcal{L}(\hat{x}_T, \hat{y}_T)\|_2 \|\nabla W_{(l)}\|_2}, \quad (3)$$

where L denotes the total layers. This choice is informed by the findings of [Balunovic *et al.*, 2022], which highlight that cosine loss provides superior performance in text-based attacks. After obtaining the initial reconstruction results $\hat{x}^*_I$ and $\hat{x}^*_T$ for the image and text modalities, they are passed to the second stage for further optimization. Importantly, the first stage of our framework directly employs existing single-modal gradient inversion attacks in a plug-and-play manner. This modular design allows MMGIA to seamlessly incorporate both current and future single-modal techniques without modifying the overall architecture. Consequently, MMGIA remains broadly compatible with ongoing advancements in gradient inversion research, ensuring long-term extensibility and adaptability.

4.2 Correlation-Driven Refinement

The second stage leverages intermodal correlations to refine and complement the first-stage reconstructions. Due to gradient sparsity, the initial reconstructions often lack detail and semantic completeness. To address this, the second stage refines initial reconstructions by exploiting latent consistency constraints between modalities. For convenience, we denote the first-stage reconstruction result for the image modality, $\hat{x}^*_I$, as x_{img}, and that for the text modality, $\hat{x}^*_T$, as x_{text}.

Extracting Latent Features.

The first step of this stage involves mapping the initial reconstruction results from the first stage into a shared latent feature space to explicitly capture the correlations between modalities. To achieve this, we leverage CLIP's pre-trained visual and text encoders, which are specifically designed to align image and text representations within a common embedding space. The reconstructed image and text data are

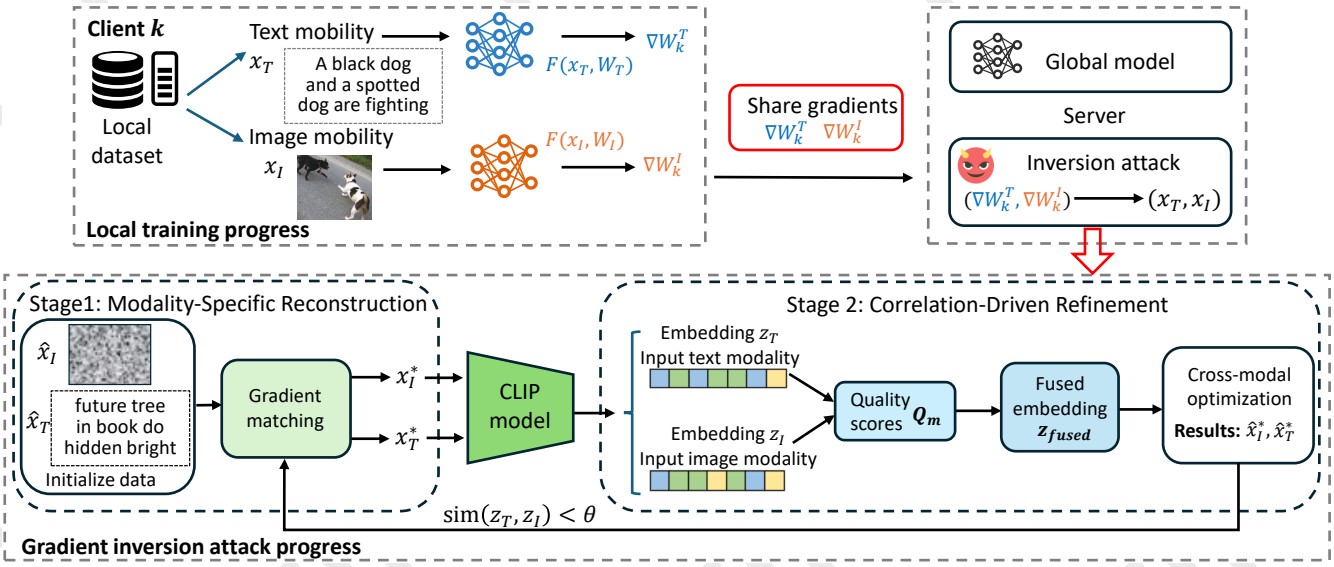


Figure 2: Framework of the proposed MMGIA. MMGIA consists of two stages. In the first stage, single-modality gradient inversion is performed independently for each modality, reconstructing the initial image and text data based on their respective gradients. In the second stage, we utilize CLIP to extract feature embeddings from the initial reconstruction results, compute a fused embedding as a unified target, and align each modality's embedding with the fused embedding to iteratively refine the reconstruction.

processed through CLIP's respective encoders, generating latent feature embeddings that preserve high-level semantic relationships across modalities. Specifically,

$$z_{\text{img}} = \text{CLIP}(x_{\text{img}}), z_{\text{text}} = \text{CLIP}(x_{\text{text}}). \tag{4}$$

Both modalities are mapped into a unified latent feature space, enabling their embeddings to be directly compared and optimized. This shared latent space serves as the basis for subsequent cross-modal optimization and joint refinement.

Cross-Modal Optimization.
We first propose a **direct cross-modal alignment optimization method** that leverages the inherent correlations between modalities to enhance reconstruction quality. This approach is based on the assumption that, in a shared latent space, the latent feature embeddings of different modalities should exhibit consistency. To achieve this, we define the following alignment loss function:

$$\mathcal{L}_{\text{align}} = \|z_{\text{img}} - z_{\text{text}}\|^2. \tag{5}$$

Here, z_{img} and z_{text} represent the latent feature embeddings of the reconstructed results for the image and text modalities, respectively. By minimizing this loss, we explicitly align the feature embeddings of both modalities in the shared space, enhancing reconstruction consistency and improving the reconstructed data quality. We employ gradient descent to iteratively update the reconstructed results x_{img} and x_{text}, ensuring that their feature embeddings in the shared space become progressively closer.

Although this method improves cross-modal alignment and enhances reconstruction quality, we observe that it may negatively impact certain modalities in practice. For instance, the alignment process might degrade the quality of a modality that initially has better reconstruction, as it forces alignment with a weaker modality. To address this issue, we further propose an improved method to dynamically balance the influence of each modality during optimization, ensuring that alignment does not lead to performance degradation while enhancing overall reconstruction quality.

Quality-Weighted Cross-Modal Optimization. To address the challenge of intermodal interference observed in direct cross-modal alignment, we propose a quality-weighted multimodal optimization method. This approach incorporates a global fused embedding as a guiding signal to direct the reconstruction of individual modalities while minimizing negative interference between them. By dynamically accounting for reconstruction quality, the method balances the influence of each modality during optimization, ensuring robust and consistent improvements in reconstruction accuracy.

We begin by evaluating the reconstruction quality of each modality in the first stage using a scoring function $Q(\cdot)$, which quantifies the alignment between the shared gradient and the gradient generated by the reconstructed data. For modality m, the quality score is defined as:

$$Q(x_m) = 1 - \frac{\|\nabla\mathcal{L}_m(x_m) - \nabla\mathcal{L}_{m,\text{true}}\|}{\|\nabla\mathcal{L}_{m,\text{true}}\|}, \tag{6}$$

where $\nabla\mathcal{L}_{m,\text{true}}$ denotes the shared gradient corresponding to modality m, $\nabla\mathcal{L}_m(x_m)$ represents the gradient generated by the current reconstruction x_m. Higher scores indicate a stronger alignment between the reconstructed data and the shared gradients, reflecting superior reconstruction quality.

Based on the quality scores, dynamic weights are assigned to each modality to determine its contribution to the global fused embedding. The dynamic weights are computed as:

$$w_m = \frac{Q(x_m)}{\sum_{m'} Q(x_{m'})}. \tag{7}$$

The normalization ensures that all modalities contribute proportionally to global optimization. Modalities with higher reconstruction quality receive greater weights, amplifying their influence on the fused embedding, while weaker modalities have reduced impact. Using the dynamic weights, the latent feature embeddings of all modalities are combined to generate a global fused embedding:

$$z_{\text{fused}} = \sum_m w_m z_m, \tag{8}$$

where z_m represents the latent feature embedding of different modalities, and z_{fused} is the global fused embedding that integrates multimodal information. This fused embedding balances the strengths of high-quality modalities while reducing the negative impact of weaker modalities, providing a robust global signal for optimization. The global fused embedding z_{fused} serves as a guiding signal for optimizing the reconstruction of individual modalities. For each modality, the optimization objective minimizes the distance between its embedding z_m and the fused embedding z_{fused}:

$$\mathcal{L}_{\text{fused},m} = \|z_m - z_{\text{fused}}\|^2. \tag{9}$$

Reconstruction updates are performed iteratively using gradient descent:

$$x_m^{(t+1)} = x_m^{(t)} - \eta \frac{\partial \mathcal{L}_{\text{fused},m}}{\partial x_m^{(t)}}, \tag{10}$$

where $x_m^{(t)}$ is the reconstruction of modality m at iteration t. η is the learning rate. This process ensures that each modality's reconstruction is guided by the global fused embedding, improving consistency and reducing inter-modal interference.

The quality-weighted cross-modal optimization dynamically adjusts weights based on reconstruction quality, using the global fused embedding as a unified target to enhance accuracy. This method prevents quality degradation of stronger modalities during alignment, ensuring a stable and robust optimization for multimodal gradient inversion attacks.

4.3 Iterative Reconstruction Optimization

After completing the second stage, the reconstruction results are reintroduced into the first stage for further optimization. This iterative process continues until the similarity between modality embeddings in the shared latent space exceeds a predefined threshold θ, indicating convergence. The similarity is calculated as:

$$\text{sim}(z_{\text{img}}, z_{\text{txt}}) = z_{\text{img}} \cdot z_{\text{txt}} / \left(\|z_{\text{img}}\| \|z_{\text{txt}}\|\right). \tag{11}$$

Each iteration refines reconstruction quality by stabilizing single-modality optimization and enhancing cross-modal alignment, leveraging multimodal correlations to progressively improve reconstruction accuracy in multimodal gradient inversion attacks.

5 Experiments

In this section, we evaluate MMGIA by comparing it with state-of-the-art single-modal and multimodal attack methods. We also investigate the impact of different architectures, auxiliary models, convergence thresholds, and defense methods on its performance.

5.1 Experimental Setup

Datasets. We evaluate the proposed MMGIA on three multimodal datasets, i.e., CIFAR100 [Krizhevsky *et al.*, 2009], OpenI [Demner-Fushman *et al.*, 2016], and Flickr30K [Plummer *et al.*, 2015]. CIFAR100 is a widely used dataset for image classification, consisting of 60,000 images from 100 object categories. Since CIFAR100 does not contain text descriptions, we generate text for each image by employing a simple template-based approach: "A [class] in the image.", where [class] is replaced with the corresponding category label of the image. OpenI is a multimodal medical dataset that contains 7,470 chest X-ray images along with corresponding radiology reports. Flickr30K is a multimodal dataset containing 30,000 images, each paired with five human-annotated textual descriptions. These datasets collectively enable a comprehensive evaluation of MMGIA in different multimodal scenarios, including general image-text associations, medical image-text relations, and natural image captioning.

Evaluation Metrics. To evaluate the effectiveness of our method, we use traditional single-modal evaluation metrics. For the image modality, we employ SSIM to assess the quality of reconstructed images by measuring their structural similarity to the original inputs [Wang *et al.*, 2004]. For the text modality, we use ROUGE metrics, reporting aggregated F-scores for ROUGE-1, ROUGE-2, and ROUGE-L, which measure unigrams, bigrams, and longest matching subsequences, respectively [Lin, 2004]. We randomly sample 100 images from each dataset to evaluate attack performance. These metrics provide a standardized evaluation of MMGIA's ability to recover high-quality content and enable rigorous comparisons with existing gradient inversion methods.

Baselines. We compare MMGIA with the only existing multimodal attack, MGIS [Liu *et al.*, 2024], which cannot be directly used due to its unrealistic assumption. Thus, we have to follow its experimental setting, using the text modality to assist in attacking the image modality. Additionally, we benchmark against single-modal attacks, including DLG [Zhu *et al.*, 2019], IG [Geiping *et al.*, 2020], and GIFD [Fang *et al.*, 2023] for image modality, and DLG [Zhu *et al.*, 2019], TAG [Deng *et al.*, 2021], and LAMP [Balunovic *et al.*, 2022] for text modality, to comprehensively evaluate MMGIA's effectiveness. We also compare MMGIA (d), a direct modality alignment method defined by Eq. (5), which aligns image and text modalities in a shared latent space.

Local Models. We conduct experiments using various architectures for both text and image modalities. For text, we compare TinyBERT$_4$, BERT$_{\text{BASE}}$, and BERT$_{\text{LARGE}}$, and for images, we evaluate LeNet-5, ResNet-18, and ResNet-50. These models vary in complexity and feature extraction capabilities. Unless otherwise specified, the local models used in our experiments are ResNet-50 for the image modality and BERT$_{\text{LARGE}}$ for the text modality.

Auxiliary Models. We use pre-trained visual and text encoders from the CLIP model to compute embeddings for images and text, enabling cross-modal alignment in our attack. Specifically, we benchmark our method using two CLIP-based models: OpenAI CLIP [Radford *et al.*, 2021], which is pre-trained on 400 million image-text pairs from the Internet, providing a strong multimodal representation in the general

	Attack	LeNet-5 SSIM	ResNet-18 SSIM	ResNet-50 SSIM
CIFAR100	MGIS	0.748	0.834	0.869
	DLG	0.696	0.822	0.853
	IG	0.625	0.791	0.814
	GIFD	0.742	0.858	0.876
	MMGIA (d)	0.764	0.907	0.912
	MMGIA	**0.771**	**0.915**	**0.927**
OpenI	MGIS	0.209	0.253	0.264
	DLG	0.203	0.245	0.257
	IG	0.237	0.274	0.306
	GIFD	0.261	0.319	0.379
	MMGIA (d)	0.279	0.376	0.435
	MMGIA	**0.295**	**0.422**	**0.479**
Flickr30K	MGIS	0.213	0.271	0.290
	DLG	0.211	0.263	0.278
	IG	0.247	0.317	0.329
	GIFD	0.286	0.385	0.446
	MMGIA (d)	0.305	0.459	0.485
	MMGIA	**0.320**	**0.498**	**0.517**

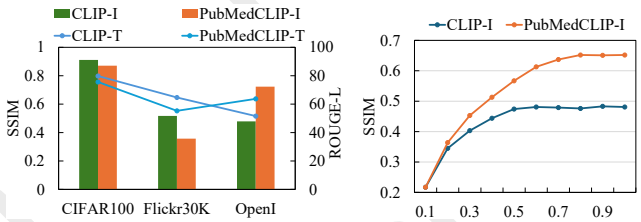
Table 1: Main results of image reconstruction from gradients with different attacks, for various datasets and architectures in the setting with batch size equal to 1.

domain, and PubMedCLIP [Eslami *et al.*, 2021], a domain-specific variant fine-tuned on 80,000 radiology image-text pairs, enhancing its ability in medical applications.

5.2 Attack Performance Comparison

Attack Performance on Image Modality. Table 1 shows the reconstruction results for the image modality, where MMGIA achieves the highest SSIM score, demonstrating its effectiveness in reconstructing high-quality images. Compared to MGIS, MMGIA significantly enhances reconstruction quality, thanks to its two-stage optimization strategy: the first stage ensures independent reconstruction of each modality, while the second stage refines results by leveraging inter-modal correlations. Unlike single-modal attacks, which rely solely on image gradients, MMGIA utilizes complementary semantic cues from the text modality to reduce ambiguity and improve fine-grained details. Furthermore, compared to MMGIA (d), which directly aligns modalities, MMGIA's quality-weighted fusion strategy ensures high-quality reconstructions guide weaker ones without degrading strong modalities, resulting in dynamic and robust optimization.

Attack Performance on Text Modality. Table 2 presents the reconstruction results for the text modality, where MMGIA also achieves the highest ROUGE scores, outperforming all baselines, including MGIS and single-modal attacks (DLG, TAG, LAMP). This improvement is primarily due to our quality-weighted fused embedding optimization, which dynamically balances the influence of the image and text modalities during the reconstruction process. Unlike MGIS, which only focuses on using the text modality to improve image reconstruction, MMGIA explicitly integrates multimodal relationships bidirectionally, enabling both modalities to benefit from cross-modal information. We omit DLG's attack results as its performance in text reconstruction is identical to that of MGIS. This is because MGIS only uses the text modality to assist image recovery, without introducing any additional optimization techniques for text reconstruction itself. Instead, it simply adopts the L_2 loss from DLG for text reconstruction, limiting its potential for improvement.



(a) Impact of auxiliary model. (b) Impact of threshold θ.

Figure 3: Performance of MMGIA with different auxiliary models and thresholds.

Impact of Local Model. We also evaluate the impact of different network architectures on attack performance, as shown in Table 1 and Table 2. For the image modality, LeNet-5 shows the weakest reconstruction, ResNet-18 performs better, and ResNet-50 achieves the highest SSIM scores due to its richer feature information, while LeNet-5's simplicity and limited gradients result in poorer performance. For text modality, TinyBERT$_4$ produces the weakest reconstructions, BERT$_{\text{BASE}}$ performs moderately better, and BERT$_{\text{LARGE}}$ achieves the highest ROUGE scores. The differences stem from model size and representational capacity: TinyBERT$_4$'s limited layers struggle with complex linguistic structures, while BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$, with deeper architectures, provide progressively better contextual representations. These findings demonstrate that deeper architectures with more expressive gradient information are more susceptible to gradient inversion attacks, whereas simpler models naturally provide stronger resistance due to their weaker gradient representations.

Impact of Auxiliary Model. Fig. 3(a) shows the impact of different auxiliary models on attack performance. We compare the reconstruction results using OpenAI CLIP and PubMedCLIP across CIFAR100, Flickr30K, and OpenI datasets. The results show a clear distinction: on CIFAR100 and Flickr30K, directly using OpenAI CLIP leads to better reconstruction quality, while on the OpenI dataset, using PubMedCLIP achieves superior performance. This difference arises because OpenI is a medical dataset. OpenAI CLIP, trained on 400 million general-domain image-text pairs, provides stronger representations for natural image datasets like CIFAR100 and Flickr30K but lacks the specialized knowledge required for medical. In contrast, PubMedCLIP, fine-tuned on 80,000 radiology image-text pairs, is better suited for medical feature extraction, resulting in improved reconstruction on OpenI. These findings highlight the importance of selecting domain-specific auxiliary models for optimizing multimodal gradient inversion attacks, as feature extractor suitability significantly affects reconstruction quality.

Impact of Threshold θ. Fig. 3(b) illustrates the impact of varying embedding similarity thresholds θ, as defined by Eq. (11), on the attack performance for image modality in the OpenI dataset, comparing the effectiveness of two auxiliary models, CLIP and PubMedCLIP. The results show that attack quality improves as θ increases but stabilizes at differ-

	Attack	TinyBERT_4			$\text{BERT}_{\text{BASE}}$			$\text{BERT}_{\text{LARGE}}$		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
CIFAR100	MGIS	63.9	10.2	43.7	71.1	11.4	37.4	74.9	18.1	52.6
	TAG	73.5	13.1	51.3	80.2	18.6	57.4	82.4	23.7	57.1
	LAMP	79.7	46.5	69.7	84.6	51.3	61.2	86.7	55.4	62.4
	MMGIA (d)	86.4	55.7	73.2	88.3	59.7	73.9	89.2	63.1	73.5
	MMGIA	**87.3**	**58.4**	**75.9**	**91.4**	**64.5**	**78.4**	**93.7**	**67.4**	**79.8**
OpenI	MGIS	27.4	1.9	17.6	32.1	0.8	11.6	42.7	2.7	26.4
	TAG	33.8	4.5	21.3	45.9	1.6	22.4	55.9	5.9	35.7
	LAMP	51.7	11.3	33.7	62.7	21.9	43.2	67.1	16.4	44.1
	MMGIA (d)	60.5	13.7	39.4	62.3	15.1	40.7	69.4	15.9	47.8
	MMGIA	**62.6**	**17.4**	**42.1**	**64.3**	**23.8**	**44.5**	**73.9**	**21.7**	**51.6**
Flickr30K	MGIS	32.7	2.4	19.1	43.5	2.7	23.2	54.9	9.8	29.8
	TAG	37.5	3.9	21.4	45.8	3.7	25.7	60.4	17.1	36.3
	LAMP	67.6	44.1	49.7	70.4	47.3	51.2	76.2	42.6	51.9
	MMGIA (d)	64.3	43.9	46.2	72.9	46.4	55.6	79.7	46.8	52.3
	MMGIA	**71.6**	**47.5**	**58.2**	**76.5**	**52.9**	**60.3**	**85.9**	**59.3**	**60.7**

Table 2: Main results of text reconstruction from gradients with different attacks, for various datasets and architectures in the setting with batch size equal to 1. R-1, R-2, and R-L, denote ROUGE-1, ROUGE-2 and ROUGE-L scores respectively. Bold indicates the best result, and underline indicates the suboptimal result.

ent thresholds for the two models. With CLIP, performance saturates at $\theta \approx 0.5$, reflecting its limited semantic alignment capacity for medical data. In contrast, PubMedCLIP achieves optimal performance at $\theta \approx 0.7$, demonstrating its superior ability to capture semantic relationships between medical text and images. This discrepancy arises from the distinct pretraining objectives of the models: CLIP, a general-purpose model, reaches its alignment capacity at lower thresholds, while PubMedCLIP, designed specifically for the medical domain, continues to improve at higher thresholds.

Method	Metric	B=1	B=2	B=4
MGIS	SSIM	0.264	0.148	0.091
	Rouge-L	25.9	21.3	16.5
MMGIA	SSIM	0.652	0.531	0.317
	Rouge-L	51.6	40.7	33.8

Table 3: Impact of different batch sizes (B).

Impact of Batch Size. Table 3 shows the attack performance on the OpenI dataset with batch sizes greater than 1. The auxiliary model is PubMedCLIP, while the local model is $\text{BERT}_{\text{LARGE}}$. Our findings indicate that as the batch size increases, reconstruction quality declines. This decline results from the gradient aggregation effect, reducing the sample-specific information encoded in the shared gradients. Nonetheless, MMGIA effectively leverages intermodal correlations to generate meaningful reconstructions, albeit with reduced detail and accuracy compared to the $B = 1$ setting.

Defense Performance of Differential Privacy. Differential privacy (DP) is a widely used method to protect data privacy. Here, we evaluate the impact of DP on the performance of MMGIA. Specifically, we employ the commonly used DPSGD (Differentially Private Stochastic Gradient Descent) algorithm [Abadi *et al.*, 2016], which ensures (ϵ, δ)-DP by adding Gaussian noise to the clipped gradients during each training iteration. In the experiments, δ is fixed at 10^{-5}, and the noise scale is adjusted to achieve target privacy budgets ϵ ranging from 1.0 to 10.0. The performance of MMGIA under different privacy budgets is shown in Table 4. As the privacy budget decreases (i.e., as ϵ becomes smaller), the attack performance of MMGIA significantly degrades. However, this degradation in attack performance comes at the cost of reduced model utility.

Method	SSIM	ROUGE-L	ACC
$\epsilon = 1.0$	0.214	16.7	0.53
$\epsilon = 5.0$	0.495	48.9	0.74
$\epsilon = 10.0$	0.637	64.5	0.87
no DP	0.652	67.2	0.90

Table 4: Comparison of SSIM (Image) and ROUGE-L (Text) under different ϵ values and without DP, under the OpenI dataset and PubMedCLIP auxiliary model. ACC is the model prediction accuracy.

6 Conclusion

In this paper, we proposed MMGIA, an intermodal correlation-driven gradient inversion attack, to expose the heightened privacy risks inherent in multimodal federated learning. By leveraging a two-stage optimization framework, MMGIA systematically reconstructs multimodal data, utilizing a quality-weighted fusion strategy to align embeddings and enhance reconstruction quality. Our experiments demonstrated that MMGIA significantly outperforms the only existing multimodal attack and state-of-the-art single-modal attacks, showcasing its ability to exploit the inherent correlations in multimodal data for high-fidelity reconstruction. These findings reveal the greater privacy vulnerabilities in multimodal data, where intermodal dependencies exacerbate the risks of gradient leakage. Our work not only exposes critical privacy risks in MMFL but also lays the foundation for further exploration of vulnerabilities and defense mechanisms in multimodal federated learning systems.

Acknowledgments

This research was supported in part by the National Key R & D Program of China (No. 2023YFB3107500), in part by the National Natural Science Foundation of China (No. 62220106004, 62411540034, 62306343), in part by the Technology Innovation Leading Program of Shaanxi (No. 2023KXJ-033), in part by the Innovation Capability Support Program of Shaanxi (No. 2023-CX-TD-02), in part by the China Scholarship Council, in part by the China Postdoctoral Science Foundation (No. 2024M753741), in part by the Fundamental Research Funds for the Central Universities (No. ZDRC2202), and in part by JSPS KAKENHI JP23K24851, JST PRESTO JPMJPR23P5, and JST CREST JPMJCR21M2.

References

[Abadi *et al.*, 2016] 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*, pages 308–318, 2016.

[Agbley *et al.*, 2021] Bless Lord Y Agbley, Jianping Li, Amin Ul Haq, Edem Kwedzo Bankas, Sultan Ahmad, Isaac Osei Agyemang, Delanyo Kulevome, Waldiodio David Ndiaye, Bernard Cobbinah, and Shoistamo Latipova. Multimodal melanoma detection with federated learning. In *2021 18th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, pages 238–244. IEEE, 2021.

[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.

[Bernecker *et al.*, 2022] Tobias Bernecker, Annette Peters, Christopher L Schlett, Fabian Bamberg, Fabian Theis, Daniel Rueckert, Jakob Weiß, and Shadi Albarqouni. Fednorm: Modality-based normalization in federated learning for multi-modal liver segmentation. *arXiv preprint arXiv:2205.11096*, 2022.

[Che *et al.*, 2023] Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23(15):6986, 2023.

[Chen and Li, 2022] Sijia Chen and Baochun Li. Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 1469–1478. IEEE, 2022.

[Chen and Zhang, 2022] Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 87–96, 2022.

[Demner-Fushman *et al.*, 2016] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[Deng *et al.*, 2021] Jieren Deng, Yijue Wang, Ji Li, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. Tag: Gradient attack on transformer-based language models. *arXiv preprint arXiv:2103.06819*, 2021.

[Eslami *et al.*, 2021] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

[Fang *et al.*, 2023] Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. Gifd: A generative gradient inversion method with feature domain optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4976, 2023.

[Feng and Narayanan, 2022] Tiantian Feng and Shrikanth Narayanan. Semi-fedser: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling. *arXiv preprint arXiv:2203.08810*, 2022.

[Feng *et al.*, 2023] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4035–4045, 2023.

[Fowl *et al.*, 2022] Liam Fowl, Jonas Geiping, Steven Reich, Yuxin Wen, Wojtek Czaja, Micah Goldblum, and Tom Goldstein. Decepticons: Corrupted transformers breach privacy in federated learning for language models. *arXiv preprint arXiv:2201.12675*, 2022.

[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.

[Guo *et al.*, 2023] Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, 2023.

[Gupta *et al.*, 2022] Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. Recovering private text in federated learning of language models. *Advances in neural information processing systems*, 35:8130–8143, 2022.

[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.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2024] Kanxue Li, Baosheng Yu, Qi Zheng, Yibing Zhan, Yuhui Zhang, Tianle Zhang, Yijun Yang, Yue Chen, Lei Sun, Qiong Cao, et al. Muep: A multimodal benchmark for embodied planning with foundation models [c]. In *Intemational Joint Conferences on Artificial Intelligence. IJCAI*, pages 129–138, 2024.

[Lin *et al.*, 2023] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4):539–553, 2023.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Liu *et al.*, 2020] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Federated learning for vision-and-language grounding problems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11572–11579, 2020.

[Liu *et al.*, 2024] Xuan Liu, Siqi Cai, Renjie He, and Jingling Yuan. Mutual gradient inversion: Unveiling privacy risks of federated learning on multi-modal signals. *IEEE Signal Processing Letters*, 2024.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Saeed *et al.*, 2020] Aaqib Saeed, Flora D Salim, Tanir Ozcelebi, and Johan Lukkien. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal*, 8(2):1030–1040, 2020.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[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 gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*, 2020.

[Xiong *et al.*, 2022] Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multimodal federated learning. *Neurocomputing*, 480:110–118, 2022.

[Yu *et al.*, 2023] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023.

[Zhao *et al.*, 2020] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

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