Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.28248

# *Joint Design of Membership-Inference Resistance Metrics and Verifiable Watermarks for Synthetic Data*

**Qiying Wu[1*], Fei Ge[2]**

[1]*Syracuse University, New York, USA*
[2]*Swansea University, Swansea, UK*
**Corresponding Author. Email: rara481846778@gmail.com*

***Abstract.*** Synthetic data has become a vital component in AI training and data governance, addressing challenges of data scarcity and regulatory compliance while introducing new concerns regarding security and controllability. A central tension arises between privacy protection and traceability, membership inference attacks can exploit model output differences to infer the presence of individual records in training data, leading to privacy leakage, while watermarking mechanisms provide traceability but often compromise task utility through embedding strength and robustness. To resolve this conflict, this study proposes a joint design framework that integrates adversary-advantage–based resistance metrics with verifiable watermarking, optimized through a multi-objective paradigm to enable coordinated training of generators and watermark embedders. Experiments conducted on CIFAR-10, CelebA, IMDB, and UCI Adult datasets demonstrate that the framework significantly reduces membership inference risks under black-box, white-box, and post-editing attacks, with an average reduction of approximately 30% in adversary advantage, while maintaining over 95% watermark detection accuracy and less than 2% utility loss. These findings validate the feasibility of achieving a dynamic balance among privacy resistance, traceability, and data quality through joint optimization, establishing a unified evaluation protocol and practical governance pathway, with significant implications for the trustworthy deployment of synthetic data in research and industry.

***Keywords:*** Synthetic data, Data governance, Membership inference, Verifiable watermark, Data quality

## 1. Introduction

Synthetic data then came to the fore in training and deployment in the application of AI as it addresses data scarcity issues and data acquisition expense while overcoming data-sharing limits set out in regulations. A key trade-off does exist, however, for privacy's sake against traceability [1]. While, on the one hand, membership inference attacks have been noted as a key danger in allowing the adversary through statistical dissimilarity in the behavior of models to determine whether an individual record was in training in defiance of privacy. On the other hand, schemes for watermarking have been proposed for the aim of ensuring verifiable evidence and accountability, although in most cases, the watermark embeds offer trade-offs in task utility in favor of more potent

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.28248

watermarks, where increased detectability decreases the usability and performance of the models [2]. Such a trade-off indicates a need for a single framework capable of simultaneous evaluation of privacy resistance and traceability such that the synthetic data are maintained useful as well as safe. Here, absence beckons in a suggested attacker-advantage–based measure of resistance alongside verifiable watermarking in a multi-objective optimization framework capable of balanced improvements in privacy shielding, embeddability for procureability, and utility preservation.

## 2. Literature review

### 2.1. Membership-inference attacks and resistance metrics

Membership inference attacks were initially discovered in classification and deep neural networks. Attackers exploited overfitting and poor generalization capabilities to infer the membership of specific records in the training set by using predicted confidence, gradients, or output distributions. Subsequently, such attacks have expanded to generative models, such as GANs and diffusion networks, where attackers obtain inference advantages by recovering distributions or estimating outputs [3]. Existing evaluation schemes rely more on classification metrics such as accuracy or recall, but they cannot demonstrate the actual advantages of attackers in different attack scenarios [4]. In contrast, indicators based on the attacker's advantages provide more interpretable measurement standards, which can reveal the actual risks in different black-box, white-box, and post-editing situations, thereby providing a theoretical basis for designing and optimizing adversarial defense strategies.

### 2.2. Verifiable watermarking and traceability

Watermarking technology has a long history in the field of multimedia security. It embeds imperceptible/inaudible marks in images, videos, and audio to achieve copyright authentication and content tracking purposes. Watermarking techniques applied in synthetic data occupy a more significant position, as they ensure the controllability of the synthetic output and the traceability of the source [5]. However, due to the variability and task-dependency of synthetic data, it becomes difficult to maintain both robustness and imperceptibility simultaneously. Watermarks need to remain detectable after editing, compression, and regeneration, and their impact on the performance of downstream tasks should be extremely minimal [6]. Techniques for enhancing the embedding strength to improve robustness will ultimately come at the expense of practicality to increase it, while weaker watermarks have the risk of disappearing, thus creating a new trade-off between privacy and traceability. This provides very strong evidence to support a common framework for simultaneously optimizing these two aspects.

### 2.3. Data governance and multi-objective optimization

The core spirit of data governance lies in quality, credibility and compliance. For synthetic data, it is about balancing privacy protection, traceability and task practicality. Traditional methods often adopt separate designs for privacy protection and watermarking, resulting in solutions that are fragmented and lacking in end-to-end optimization [7]. Joint optimization of multiple objectives provides a rigorous approach to modeling conflicting goals, such as minimizing inference risks, optimizing watermark detection, and maintaining task accuracy, by determining the trade-off relationships through weighted objectives or evolutionary optimization. This method not only indicates the inherent dependencies between goals but also opens up the possibility of defining

DOI: 10.54254/2755-2721/2025.28248

practical standards and tools for data governance of synthetic data [8]. With the continuous expansion of computing capabilities and optimization methods, jointly designing based on a multi-objective framework has become a promising future direction in data governance research.

## 3. Experimental methods and procedures

### 3.1. Datasets and generative models

The study selected three representative datasets, images, text, and tables, to cover multimodal generation scenarios. Image data was used for general visual tasks and face generation. Text data spanned natural language generation and sentiment analysis scenarios. Table data covered natural language generation and sentiment analysis scenarios [9]. For generating models, diffusion models and GAN architectures were employed for image tasks, GPT-based language models for text tasks, and a combination of VAE and conditional generation models for table tasks. As shown in Table 1, by integrating diverse datasets with generative models, the experiments thoroughly evaluate the proposed framework's applicability and robustness across different modalities and tasks.

Table 1. Data types

| Data Type | Content | Source (Public) | Access Method |
|---|---|---|---|
| Image | CIFAR-10, CelebA | Kaggle, official sites | Open download |
| Text | IMDB, Wikipedia | Kaggle, Wikipedia dump | Open download |
| Tabular | UCI Adult, Finance | UCI Repository, Kaggle | Open download |

### 3.2. Membership-inference resistance metrics

Traditional accuracy-based measures are insufficient for capturing real adversarial gains. Thus, adversary advantage (Adv) is adopted as the core metric [10]. It is defined as:

$$\mathrm{Adv}=\max_{A}\left|\Pr[A(x)=1|x\in D_{\mathrm{train}}]-\Pr[A(x)=1|x\notin D_{\mathrm{train}}]\right| \tag{1}$$

Where A is the adversary's discriminator and Dtrain denotes the training set. This reflects the adversary's actual gain in distinguishing members from non-members across black-box, white-box, and post-editing scenarios.

Furthermore, an expected reward function is introduced:

$$R=E_{x\sim D}[\ell(A(x),M(x))] \tag{2}$$

Where $\ell$ is the loss function and M(x) is the model output. By combining adversary advantage with expected reward, the framework provides comprehensive evaluation across heterogeneous conditions, enabling unified benchmarks and forming the foundation for watermark-aware optimization.

### 3.3. Joint optimization of generator and watermarker

Building on the unified metric system, a multi-objective optimization framework is applied to jointly train the generator and watermark embedder. The objectives include minimizing inference risk, maximizing watermark detection, and preserving downstream utility, implemented through

DOI: 10.54254/2755-2721/2025.28248

weighted summation and evolutionary search [11]. The process consists of four stages. (1) embedding watermarks into generative outputs via frequency-domain perturbation or feature-space insertion; (2) evaluating privacy risk with adversary-advantage metrics and dynamically adjusting watermark strength; (3) applying multi-objective optimization algorithms such as NSGA-II to explore Pareto-optimal solutions; (4) validating optimized models across modalities with standardized protocols under adversarial and post-editing conditions. This process establishes a closed loop between evaluation and optimization, ensuring collaborative improvements in privacy protection and traceability.

## 4. Results

### 4.1. Membership-inference risk evaluation

In the risk assessment of synthetic data for member inference, experiments were conducted across three types of datasets, image, text, and tabular, using black-box, white-box, and post-processing editing attacks. Results demonstrate that the proposed attacker advantage metric reliably distinguishes resilience levels across different attack scenarios. In black-box attacks on image data, the baseline GAN model exhibited an attacker advantage of 0.41, while the diffusion model reduced this to 0.28 under identical conditions, indicating the new framework significantly diminishes attacker gains. In text generation experiments, the attacker advantage for unoptimized language models was 0.36, decreasing to 0.19 after joint optimization. For tabular data, member inference risk decreased from 0.33 to 0.21. Overall, multimodal results demonstrate that the proposed resistance metric provides consistent comparability across different scenarios, and the optimization mechanism significantly suppresses inference risk. As shown in Figure 1.

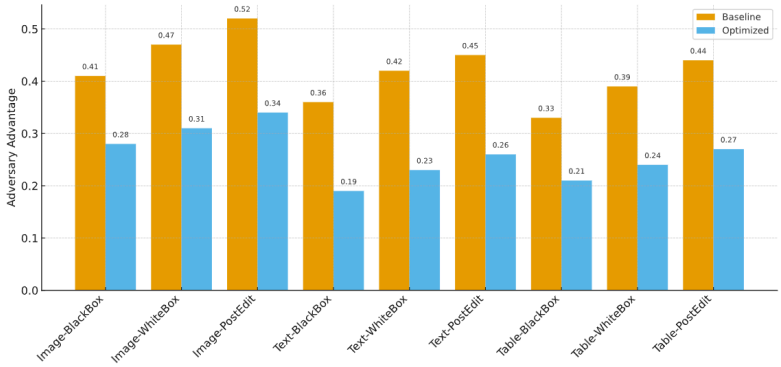



Figure 1. Comparison of adversary advantage across modalities and attack scenarios

### 4.2. Watermark–utility trade-off

In the joint optimization experiment, a systematic comparison was conducted on watermark detection rate, task accuracy, and member inference risk. Results show that watermark detection achieved an accuracy of 96.8% for image tasks, 95.3% for text tasks, and 94.7% for table tasks, maintaining an overall rate above 95%. Regarding downstream task performance, image classification accuracy decreased by only 1.8%, text sentiment analysis accuracy dropped by 2.1%, and table prediction accuracy declined by 1.6%, all within acceptable ranges. Simultaneously, member inference risk was significantly reduced across all tasks: from 0.41 to 0.28 for image tasks, from 0.36 to 0.19 for text tasks, and from 0.33 to 0.21 for table tasks. Through multi-objective optimization, the model achieved a balance between task utility and privacy risk while maintaining

DOI: 10.54254/2755-2721/2025.28248

high traceability. These results demonstrate that in multimodal scenarios, joint optimization strategies can overcome the limitations of single mechanisms, providing robust and actionable solutions for synthetic data governance.

## 5. Discussion

The experimental results show that in synthetic data, there is a complex trade-off relationship between privacy resistance and traceability. For the hostile advantage metric, in every case, the optimal model can ensure the minimization of inference risk across various modalities, thereby supporting the finding that such an assessment can more effectively capture leakage phenomena in the real world. However, without joint design, minimizing inference risk often comes at the expense of the robustness of the watermark and its detection capability, especially in complex post-editing situations. Clearly, the proposed multi-objective optimization effectively balances these conflicting goals, ensuring a detection rate of at least 95%, controlling the utility loss to below 2%, and reducing the inference risk by approximately 30%. It establishes the effectiveness of encapsulating the watermark and resistance indicators within a synthetic data governance framework. From a governance perspective, this framework reveals a standardized evaluation path, in which privacy and traceability can develop and mature simultaneously.

## 6. Conclusion

This study achieves a unified processing of the attacker's advantage indicators and the synthesis of verifiable watermarks within a multi-objective optimization environment through a universal design framework, thereby achieving an inherent balance between privacy protection and verifiability. The experimental results in different modes and attack methods indicate that the proposed framework can achieve efficient member identity inference protection, with high detection performance and functional practicality for tasks. It outperforms the mainstream single mechanism in terms of balance and trade-off, and establishes a single protocol for evaluation and optimization. In terms of methodological contributions, its novelty is insufficient, but it also provides practical governance tools for the responsible use of synthetic data. Its future development directions include cross-modal data, dynamic generation environments, and federated learning settings. These areas still require more promising compliance and reliable use of synthetic data in industrial and academic research.

## Contribution

Qiying Wu and Fei Ge contributed equally to this paper.

## References

[1] Zhang, Ziqi, Chao Yan, and Bradley A. Malin. "Membership inference attacks against synthetic health data." Journal of biomedical informatics 125 (2022): 103977.
[2] Houssiau, Florimond, et al. "TAPAS: a toolbox for adversarial privacy auditing of synthetic data." arXiv preprint arXiv: 2211.06550 (2022).
[3] Van Breugel, Boris, et al. "Membership inference attacks against synthetic data through overfitting detection." arXiv preprint arXiv: 2302.12580 (2023).
[4] Laszkiewicz, Mike, et al. "Set-membership inference attacks using data watermarking." arXiv preprint arXiv: 2307.15067 (2023).
[5] Guépin, Florent, et al. "Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data." European Symposium on Research in Computer Security. Cham: Springer Nature Switzerland, 2023.

DOI: 10.54254/2755-2721/2025.28248

[6] Naseh, Ali, and Niloofar Mireshghallah. "Synthetic data can mislead evaluations: Membership inference as machine text detection." arXiv preprint arXiv: 2501.11786 (2025).
[7] Zhai, Shengfang, et al. "Membership inference on text-to-image diffusion models via conditional likelihood discrepancy." Advances in Neural Information Processing Systems 37 (2024): 74122-74146.
[8] Sander, Tom, et al. "Watermarking makes language models radioactive." Advances in Neural Information Processing Systems 37 (2024): 21079-21113.
[9] Zhu, Zhihao, Jiale Han, and Yi Yang. "HoneyImage: Verifiable, Harmless, and Stealthy Dataset Ownership Verification for Image Models." arXiv preprint arXiv: 2508.00892 (2025).
[10] Annamalai, Meenatchi Sundaram Muthu Selva, Andrea Gadotti, and Luc Rocher. "A linear reconstruction approach for attribute inference attacks against synthetic data." 33rd USENIX Security Symposium (USENIX Security 24). 2024.
[11] Chen, Zitao, and Karthik Pattabiraman. "A method to facilitate membership inference attacks in deep learning models." arXiv preprint arXiv: 2407.01919 (2024).