COLLAFUSE: NAVIGATING LIMITED RESOURCES AND PRIVACY IN COLLABORATIVE GENERATIVE AI

Short Paper

Domenique Zipperling, University of Bayreuth & Fraunhofer FIT

Simeon Allmendinger, University of Bayreuth & Fraunhofer FIT

Lukas Struppek, Technical University of Darmstadt & German Center for Artificial Intelligence

Niklas Kühl, University of Bayreuth & Fraunhofer FIT

Abstract

In the landscape of generative artificial intelligence, diffusion-based models present challenges for socio-technical systems in data requirements and privacy. Traditional approaches like federated learning distribute the learning process but strain individual clients, especially with constrained resources (e.g., edge devices). In response to these challenges, we introduce COLLAFUSE*, a novel framework inspired by split learning. Tailored for efficient and collaborative use of denoising diffusion probabilistic models,* COLLAFUSE *enables shared server training and inference, alleviating client computational burdens. This is achieved by retaining data and computationally inexpensive GPU processes locally at each client while outsourcing the computationally expensive processes to the shared server. Demonstrated in a healthcare context,* COLLAFUSE *enhances privacy by highly reducing the need for sensitive information sharing. These capabilities hold the potential to impact various application areas, such as the design of edge computing solutions, healthcare research, or autonomous driving. In essence, our work advances distributed machine learning, shaping the future of collaborative GenAI networks.*

Keywords: Generative Models, Distributed Learning, Split Learning, Diffusion Model.

1 Introduction

In the realm of Information Systems Research, the emergence of generative artificial intelligence (GenAI) technologies like ChatGPT and DALL-E has marked a significant milestone. These advancements, as detailed by Feuerriegel et al. (2023), have broadened public access to GenAI and catalyzed its integration into diverse sectors. Among GenAI innovations, the denoising diffusion probabilistic model (DDPM) stands out for its ability to generate high-quality images through an advanced denoising process, outperforming earlier methods such as generative adversarial networks (GANs) (Goodfellow et al., 2020) or variational autoencoders (VAEs) (Kingma and Welling, 2014) in terms of diversity and convergence guarantees (Dhariwal and Nichol, 2021; Nichol and Dhariwal, 2021). However, the implementation of DDPMs in business analytics and other fields is not without challenges. These models demand extensive data sets and computational resources (Z. Wang et al., 2023), which are often limited, especially in decentralized systems (Hirt and Kühl, 2018). The healthcare sector exemplifies these constraints, where data scarcity, privacy concerns, and high costs of data collection are present (Veeraragavan and Nygård, 2023; Y. Wang T. G. and Choo, 2023). To address these challenges, researchers have explored various strategies, including patch-wise training (Z. Wang et al., 2023), few-shot learning (Lu et al., 2023; Ruiz et al., 2023; Zhang et al., 2023), and notably, federated learning (FL) (Fan and Liu, 2020; McMahan et al.,

2017). FL enhances data accessibility, yet raises privacy concerns (Shokri et al., 2017; Zhu et al., 2019) and the need for significant local computational capabilities.
In response, we introduce **CollaFuse**, a new collaborative learning and inference framework for DDPMs, inspired by split learning (Gupta and Raskar, 2018). CollaFuse aims to balance the computationally intensive denoising process between local clients and a shared server, with a focus on optimizing the trade-off between performance, privacy, and resource utilization—which are crucial requirements for real-world information systems implementations. This framework transforms the optimization challenge into a multi-criteria problem, addressing the core requirements of such applications in practice. Building upon these criteria, our research investigates the impact of different degrees of collaboration defined by a cut-ratio $c \in [0,1]$, formulating two key hypotheses:

Hypothesis 1. In our framework CollaFuse, collaborative learning of DDPMs positively influences the fidelity of generated images compared to non-collaborative local training ($c = 1$).

Hypothesis 2. Increasing collaborative effort ($c \downarrow$) improves performance (a) and the amount of disclosed information (b) while implicitly reducing the locally consumed GPU energy (c).

Our initial analysis, including a healthcare-focused experiment with MRI brain scans, supports these hypotheses. As a consequence, CollaFuse holds promise for applicants such as small medical institutions or even individual practitioners with edge devices to engage in collaborative model training and inference, e.g., medical training. Beyond healthcare, the framework exhibits potential in domains like autonomous driving, where edge computing resources are constrained, yet computational demands and privacy considerations are high. Moving forward, our research will delve deeper into the CollaFuse framework, analyzing its performance in terms of image fidelity and diversity, assessing potential privacy risks, and exploring resource efficiency in additional scenarios. In pursuit of a more comprehensive understanding and depth our scenarios are characterized by a parameter grid: number of clients, data domain, absolute amount of client data, conditional inference, as well as client-dependent variance scheduler and cut-ratio. This comprehensive investigation aims to advance our understanding of (distributive) GenAI within socio-technical systems as demanded by related literature (Abbasi et al., 2023; Feuerriegel et al., 2023). Consequently, our ongoing research on CollaFuse can offer guidelines on how to apply DDPMs collaboratively and work as a blueprint for future GenAI networks designs in various domains.

2 Background

The foundations of our work are built by the collaborative concepts of distributed learning as well as the architectural principles of diffusion models. *Federated Learning* (FL) was first introduced by McMahan et al. (2017) as a solution to make distributed data accessible for training without storing it centrally. Since then, FL has been on a triumphant march and received a lot of attention (Hard et al., 2019; Karnebogen et al., 2023). At its core, FL iteratively composes locally trained models into a global model requiring clients to share model updates and gradients, thereby increasing the privacy risk of, e.g., membership inference (Shokri et al., 2017) or reconstruction attacks (Zhu et al., 2019). Furthermore, FL comes with high computational requirements at the client-side (Thapa et al., 2022).
Another paradigm for collaborative learning is *Split Learning* (SL), which was introduced by Gupta and Raskar (2018) exploiting the sequencing of operations in neural networks. In general, SL splits a neural network among multiple clients and a shared server. This is especially intriguing as clients can use a server to train most of the model while keeping the data and labels locally.
In 2020, research advanced diffusion models with the introduction of *Denoising Diffusion Probabilistic Models* (DDPM) (Ho et al., 2020), offering an alternative to GANs (Goodfellow et al., 2020) for image generation. DDPMs involve two processes: diffusion and denoising. The diffusion process adds noise from a Gaussian distribution incrementally to an image over T steps. In the denoising process, the model estimates the noise added at each step $t \in [0,T]$. While training the model, weights are updated, calculating

the loss based on the difference between true and estimated noise. Accordingly, the initial training image is not necessary for the denoising process. On this basis, DDPMs are able to generate new images from pure Gaussian noise, which closely resemble the images of the training data set. Fidelity metrics, like the Kernel-Inception distance (KID) (Binkowski et al., 2018), gauge diffusion model performance by quantifying the difference between the distributions of real and generated images.

3 Related Work

With *Generative Adversarial Networks (GANs)* being composed of two components, a generator and discriminator, research demonstrates different integration of the two components within the collaboration process. Hardy et al. (2019), being among the first to use distributed data sets to train GANs, propose MD-GAN. The approach consists of one central generator while clients hold discriminators locally. In contrast, J. Wang et al. (2023) apply FL for GANs to cross-modality brain image synthesis. Using differential privacy gradient descent, only the generators of each client are aggregated. The discriminators remain local. Fan and Liu (2020) offer empirical results for FL-GANs indicating that synchronizing discriminator and generator across clients yields the best results for two data sets. On this basis, W. Li et al. (2022) offer IFL-GAN, an improved version of the FL-GAN, enabling local GANs to hold different weights resulting in faster convergence of the global model. As FL exhibits privacy concerns, there have been efforts to address these issues within collaborative training of GANs. Augenstein et al. (2020) apply an algorithm for differentially private federated GANs to effectively tackle commonly occurring data issues. Veeraragavan and Nygård (2023) address trust-related weaknesses of existing federated GAN solutions by combining three building blocks federated GANs, consortium blockchains, and Shamir's Secret Sharing algorithm enabling generation of synthetic data in decentralized settings. Additionally, Yin et al. (2023) propose a hybrid federated split learning framework for wireless networks and analyze the trade-off between training time and energy consumption, showing efforts to combine FL and SL.

Research on collaborative training methods is still scarce in the domain of *Diffusion Models*. Jothiraj and Mashhadi (2023) introduce the Phoenix technique for training unconditional diffusion models in a horizontal FL setting. The objective is to address mode coverage issues often seen in non-independent and identically distributed datasets. The data-sharing approach achieves a performance boost by sharing only 4-5% of the data among clients, minimizing communication overhead. Personalization and threshold filtering techniques outperform comparison methods in terms of precision and recall but fall short in image quality compared to the proposed technique. The paper suggests further exploration to enhance image quality in future work.

By mainly focusing on FL, current literature on collaborative GenAI, especially for DDPMs, neglects benefits from different collaborative paradigms. Adapting principles from SL can bring typical advantages such as reduced local resources and increased privacy to collaborative GenAI. In proposing COLLAFUSE we want to tap into these advantages and push GenAI to its next evolutionary level.

4 Framework

We propose COLLAFUSE, a framework facilitating collaborative GenAI for image generation with DDPMs across clients. Inspired by SL, the less resource-intensive diffusion process is computed locally by each client, whereas the computationally intensive denoising process is strategically split at step $t_c = (1-c)T$ during both training and inference. This results in a shared model (backbone) hosted on a shared server, coupled with individual local models for each client. The unique design of COLLAFUSE allows clients to retain potentially sensitive data locally while outsourcing the majority of computationally intensive denoising operations to a centralized server. The split is governed by the cut-ratio c, dictating the computational load for each client and the extent of disclosed information. For example, if $T = 50$ and $c = 0.8$, 20% of the denoising process (from $t = 1$ to $t_c = 10$) is trained and stored on the shared server, while the remaining 40 denoising steps (80%) are trained locally, ensuring privacy of these steps.

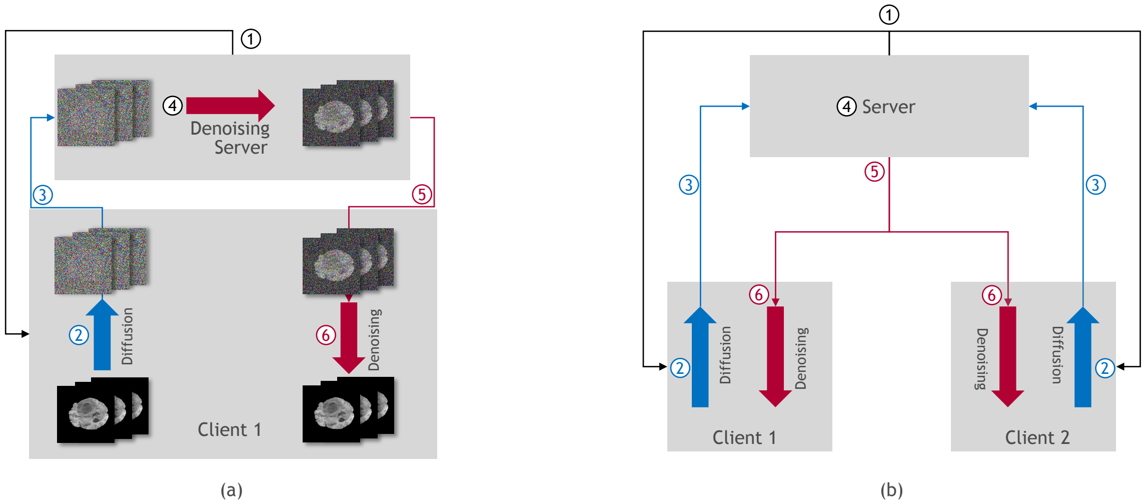

Figure 1. The training procedure of COLLAFUSE *comprises six steps from client perspective (a) and system (b): Server triggers diffusion process of clients (1), clients apply diffusion (2), clients send diffused images and noise to server (3), server denoises the images until t_c (4), server sends the partially denoised images to client (5), and clients locally finish denoising process (6).*

As illustrated in Figure 1a, from the client's perspective, the training sequence orchestrated by COLLAFUSE comprises six key steps. The server initiates the client's diffusion process (1), after which the client computes its image data's diffusion process batch-wise (2) and forwards the resulting noised images, along with the corresponding added random noise $\varepsilon = (\varepsilon_0, ..., \varepsilon_{t_c})$ for each image, back to the server (3). Utilizing the noised images, the server undertakes the initial phase of the denoising process, computing the loss between the estimated and provided random noise (4). The partially denoised images are subsequently transmitted back to the client (5), where the remaining denoising process is executed (6).
The multi-client perspective is depicted in Figure 1b, showcasing two clients providing noised images to the shared server and receiving partially denoised images for further training. Similarly, each client has the capability to generate new images by bypassing the diffusion process of existing images and solely denoising them from pure random noise.

5 Experimental Evaluation

To assess our framework, COLLAFUSE, we simulate a healthcare-related scenario involving three clients and one server. Every client data set is independent comprising 4,920 MRI scans from 123 patients each. The hold-out test data set contains 5,000 images from 125 further patients (Bakas et al., 2017). The applied DDPM employs an identical cosine variance scheduler, $T = 100$ steps, and maintains a fixed image size of (128×128) across clients. The training process spans 300 epochs with a fixed learning rate of 0.001 and a batch size of 150.
Our experimental investigation delves into the impact of the cut-ratio c on the trade-off between performance, disclosed information, and GPU energy consumption considering cut-ratio values $c \in [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$. Performance is assessed using the common fidelity metric KID (Binkowski et al., 2018) on both the client-dependent training and hold-out data sets, employing the feature extractor from the clean-fid library (Parmar et al., 2022). GPU energy consumption is measured using the *codecarbon* Python package [1]. Disclosed information for each client is approximated through the mean squared

[1] https://mlco2.github.io/codecarbon/

error (MSE) for a pixel-by-pixel comparison and KID scores between partially denoised images at the split step t_c and real images of clients. To effectively simulate a distributed healthcare-related scenario, the experiment runs on a cluster with four NVIDIA A100-SXM4-40GB GPUs. Computation tasks are distributed across clients and the server, with each client utilizing a single GPU.

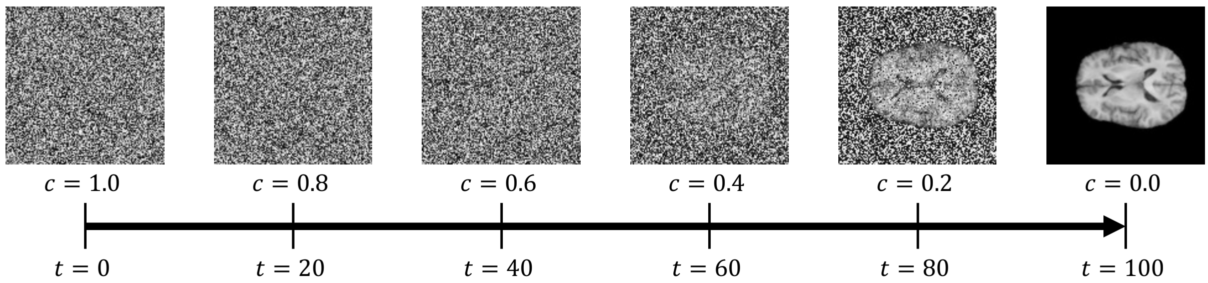

Figure 2. Illustration of the denoising process in DDPMs: Exemplary images generated at denoising step t for various cut-ratios c. The distinguishing features of the generated images remain effectively concealed behind noise during the majority of denoising steps.

6 Results

Our findings delve into the nuanced trade-off among all three dimensions: the performance, gauged by the fidelity of generated images; the extent of disclosed information, verifying whether images at the split point equal original images within respective client data sets; and the GPU power usage across clients and server. Figure 3 illustrates the trade-off between performance and disclosed information. Both are presented featuring stacked KID scores composed of all three clients. Additionally, disclosed information is assessed with the MSE metric. Lower KID scores signify enhanced performance, underscoring fidelity in generated images. Conversely, the investigation of disclosed information operates in the opposite direction. Here, the objective is to maximize KID and MSE scores, ensuring the revelation of as few characteristics as possible on the shared server to uphold stringent data privacy.

The analysis of Figure 3 unveils two significant observations. Firstly, the stacked KID scores of performance exhibit a U-shaped pattern concerning the cut-ratio, particularly evident for client data. Consequently, collaborative efforts in our experiment lead to a reduced aggregated KID score compared to the scenario where clients exclusively train the model locally (100%). Interestingly, the performance is reduced, if a large part of the global denoising process is conducted on the shared server (0% – 40%), thereby contributing to the observed U-shaped trend. With regard to the disclosed information, both pixel-wise comparison and the KID score imply that, despite conducting up to 80% of computationally intensive denoising steps on the server, a substantial portion of information associated with the images remains concealed in comparison to total global denoising (0%). Figure 2 exemplary illustrates the denoising process for different cut-ratios, which increases the comprehension of our results. Moreover, GPU power usage of the diffusion process exhibits limited computational intensity in the experiment, and the relocation of denoising steps to the server correlates positively with reduced local GPU energy demand.

Overall, our experiment supports Hypothesis 1, showing that collaborative learning with COLLAFUSE improves image fidelity compared to non-collaborative local training (c = 1). Hypothesis 2 is evident in disclosed information, and reduced local computational intensity when denoising steps are moved to the

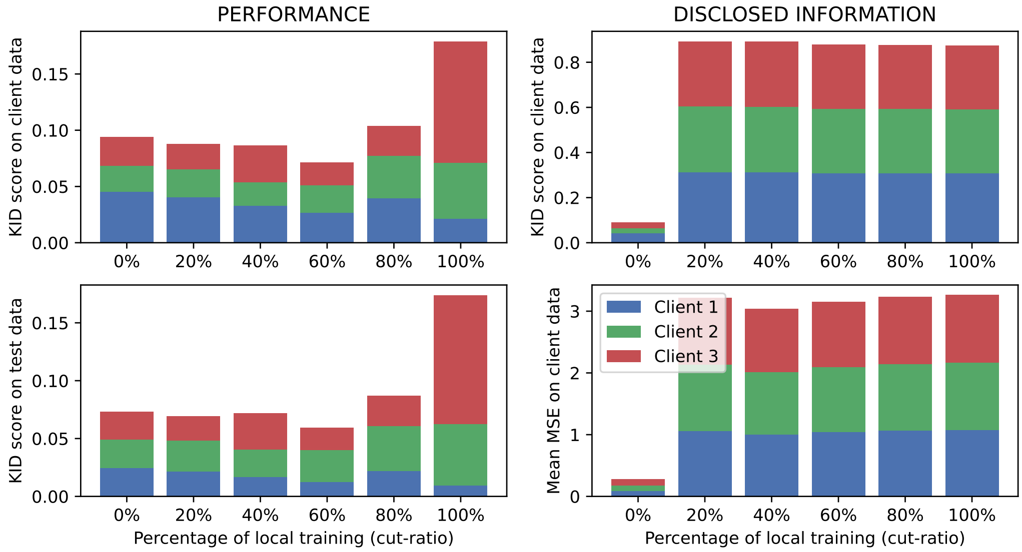

Figure 3. Experimental outcomes for performance and disclosed information in the CollaFuse *framework based on our healthcare-related experiment: Figures include KID scores and MSE from pixel-wise comparisons. Collaborative efforts, illustrated by the U-shaped pattern of stacked KID scores, lead to better performances compared to exclusive local or global model training. Despite conducting up to 80% of the computationally intensive denoising steps on the server, a significant amount of information in the images remains concealed.*

server. Fascinatingly, concerning performance, the hypothesis only holds to some extent as the results indicate a tipping point where further collaboration does not lead to an increase in performance.

7 Conclusion and Outlook

In this paper, we introduce CollaFuse, an innovative collaborative learning and inference framework designed for denoising diffusion probabilistic models. The primary objective is to address the trade-off between performance, privacy, and resource utilization—an imperative aspect for the practical implementation of information systems in real-world scenarios. Drawing inspiration from split learning (Gupta and Raskar, 2018), CollaFuse aims to balance the computationally intensive denoising process across local clients and a shared server. The framework is particularly beneficial in domains where data is scarce, private, and computational resources on local devices are limited at the client level, e.g., edge devices. This especially includes scenarios of industry (S. Li et al., 2022) and healthcare (Sivarajah et al., 2023). CollaFuse innovatively partitions the computationally extensive denoising process into two independently trainable components. The latter remains with the client, ensuring data privacy, while the initial part is shared among clients and collaboratively learned on a server, amplifying the amount of data utilized for training and inference. Our experiment demonstrates that clients can execute numerous denoising steps on the server before client data is disclosed. The findings further indicate that a decreasing cut-ratio c effectively shifts computational effort to a shared server backbone, enhancing performance generalizability. In summary, our experiment provides initial evidence supporting the advantages of collaborative learning within CollaFuse across performance, disclosed information, and local GPU energy. This highlights CollaFuse as a practical solution to the challenges of collaborative training and inference. Looking ahead, our research roadmap involves a more in-depth analysis of CollaFuse. We plan to explore the performance, encompassing both fidelity and diversity in the generated images. Additionally, our analysis includes challenging disclosed information using threat models, incorporating potential attacks like reconstruction inference. Extending our investigation, we will examine resource-related aspects, measuring floating-point operations and training durations. This thorough examination will contribute to a deeper understanding of the transformative potential of CollaFuse influencing applications in industry and healthcare and paving the way for future advancements in collaborative generative AI applications.

References

Abbasi, A., R. H. Chiang, and J. Xu (2023). "Data Science for Social Good." *Journal of the Association for Information Systems* 24 (6), 1439–1458.

Augenstein, S., H. B. McMahan, D. Ramage, S. Ramaswamy, P. Kairouz, M. Chen, R. Mathews, and B. A. y Arcas (2020). "Generative Models for Effective ML on Private, Decentralized Datasets." In: *International Conference on Learning Representations (ICLR)*.

Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos (Sept. 2017). "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features." *Scientific Data* 4 (1).

Binkowski, M., D. J. Sutherland, M. Arbel, and A. Gretton (2018). "Demystifying MMD GANs." In: *International Conference on Learning Representations (ICLR)*.

Dhariwal, P. and A. Q. Nichol (2021). "Diffusion Models Beat GANs on Image Synthesis." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8780–8794.

Fan, C. and P. Liu (2020). "Federated Generative Adversarial Learning." In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 3–15.

Feuerriegel, S., J. Hartmann, C. Janiesch, and P. Zschech (2023). "Generative AI." *Business and Information Systems Engineering*.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio (2020). "Generative adversarial networks." *Communications of the ACM* 63 (11), 139–144.

Gupta, O. and R. Raskar (2018). "Distributed learning of deep neural network over multiple agents." *Journal of Network and Computer Applications* 116, 1–8.

Hard, A., K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage (2019). "Federated Learning for Mobile Keyboard Prediction." *arXiv preprint* arXiv:1811.03604.

Hardy, C., E. Le Merrer, and B. Sericola (2019). "MD-GAN: Multi-Discriminator Generative Adversarial Networks for Distributed Datasets." In: *International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 866–877.

Hirt, R. and N. Kühl (2018). "Cognition in the Era of Smart Service Systems: Inter-organizational Analytics through Meta and Transfer Learning." *International Conference on Information Systems (ICIS)*.

Ho, J., A. Jain, and P. Abbeel (2020). "Denoising Diffusion Probabilistic Models." In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33, pp. 6840–6851.

Jothiraj, F. V. S. and A. Mashhadi (2023). "Phoenix: A Federated Generative Diffusion Model." *arXiv preprint* arXiv:2306.04098.

Karnebogen, P., C. Kaymakci, L. Willburger, B. Häckel, and A. Sauer (2023). "Structuring Federated Learning Applications : A Literature Analysis and Taxonomy." In: *European Conference on Information Systems (ECIS)*.

Kingma, D. P. and M. Welling (2014). "Auto-Encoding Variational Bayes." In: *International Conference on Learning Representations (ICLR)*.

Li, S., M. Iqbal, and N. Saxena (2022). "Future industry internet of things with zero-trust security." *Information Systems Frontiers*, 1–14.

Li, W., J. Chen, Z. Wang, Z. Shen, C. Ma, and X. Cui (2022). "IFL-GAN: Improved Federated Learning Generative Adversarial Network With Maximum Mean Discrepancy Model Aggregation." *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Lu, H., H. Tunanyan, K. Wang, S. Navasardyan, Z. Wang, and H. Shi (2023). "Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Image Diffusion Models to Learn Any Unseen Style." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14267–14276.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas (2017). "Communication-Efficient Learning of Deep Networks from Decentralized Data." In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282.

Nichol, A. Q. and P. Dhariwal (2021). "Improved Denoising Diffusion Probabilistic Models." In: *International Conference on Machine Learning (ICML)*, pp. 8162–8171.

Parmar, G., R. Zhang, and J. Zhu (June 2022). "On Aliased Resizing and Surprising Subtleties in GAN Evaluation." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 11400–11410.

Ruiz, N., Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman (2023). "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov (2017). "Membership Inference Attacks Against Machine Learning Models." In: *Symposium on Security and Privacy (S&P)*, pp. 3–18.

Sivarajah, U., Y. Wang, H. Olya, and S. Mathew (2023). "Responsible Artificial Intelligence (AI) for Digital Health and Medical Analytics." *Information Systems Frontiers*, 1–6.

Thapa, C., P. C. Mahawaga Arachchige, S. Camtepe, and L. Sun (June 2022). "SplitFed: When Federated Learning Meets Split Learning." *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (8), 8485–8493.

Veeraragavan, N. R. and J. F. Nygård (2023). "Securing Federated GANs: Enabling Synthetic Data Generation for Health Registry Consortiums." In: *International Conference on Availability, Reliability and Security (ARES)*, 89:1–89:9.

Wang, J., G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, and Y. Jin (2023). "FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis." *Neurocomputing (Amsterdam)* 546, 126282.

Wang Tongnian; Guo, Y. and K.-K. R. Choo (2023). "Enabling Privacy-Preserving Prediction for Length of Stay in ICU - A Multimodal Federated-Learning-based Approach." In: *European Conference on Information Systems (ECIS)*.

Wang, Z., Y. Jiang, H. Zheng, P. Wang, P. He, Z. Wang, W. Chen, and M. Zhou (2023). "Patch Diffusion: Faster and More Data-Efficient Training of Diffusion Models." In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Yin, B., Z. Chen, and M. Tao (2023). "Predictive GAN-Powered Multi-Objective Optimization for Hybrid Federated Split Learning." *IEEE Transactions on Communications* 71 (8), 4544–4560.

Zhang, Z., L. Han, A. Ghosh, D. N. Metaxas, and J. Ren (2023). "SINE: SINgle Image Editing with Text-to-Image Diffusion Models." In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6027–6037.

Zhu, L., Z. Liu, and S. Han (2019). "Deep Leakage from Gradients." In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14747–14756.