REPFAIR-QGAN: ALLEVIATING REPRESENTATION BIAS IN QUANTUM GENERATIVE ADVERSARIAL NETWORKS USING GRADIENT CLIPPING

Kamil Sabbagh, Hadi Salloum, & Yaroslav Kholodov
Research Center for Artificial Intelligence
Innopolis University
Innopolis, 420500, Russia
{k.sabbagh, h.salloum, ya.kholodov}@innopolis.ru

ABSTRACT

This study introduces a novel application of Quantum Generative Adversarial Networks (QGANs) by incorporating a new fairness principle, *representational fairness*, which improves equitable representation of various demographic groups in quantum-generated data. We propose a *group-wise* gradient norm clipping technique that constrains the magnitude of discriminator updates for each demographic group, thereby promoting fair data generation. Furthermore, our approach mitigates the issue of mode collapse, which is inherent in both QGANs and classical GANs. Empirical evaluations confirm that this method enhances *representational fairness* while maintaining high-quality sample generation.

1 INTRODUCTION

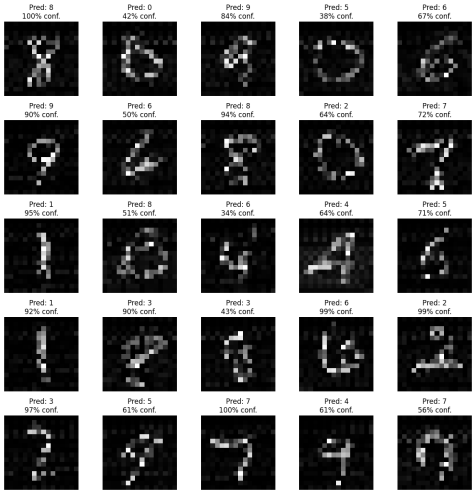
Ensuring fairness in machine learning (ML) has become paramount as ML models increasingly influence decision-making in diverse domains Pessach & Shmueli (2022); Caton & Haas (2024); Mehrabi et al. (2021). Although Generative Adversarial Networks (GANs) produce high-fidelity synthetic data, they often exhibit *representation bias*, misrepresenting or underrepresenting certain demographic groups.

Quantum computing holds the potential to revolutionize deep learning Salloum et al. (2025) and generative AI by leveraging quantum parallelism and entanglement to enhance model expressiveness and training efficiency. However, as quantum-based generative models evolve, it is imperative to ensure that they do not inherit or exacerbate fairness issues that have historically plagued classical AI models. Addressing these biases is crucial to realizing the full potential of quantum-enhanced generative AI in practical applications Perrier (2021).

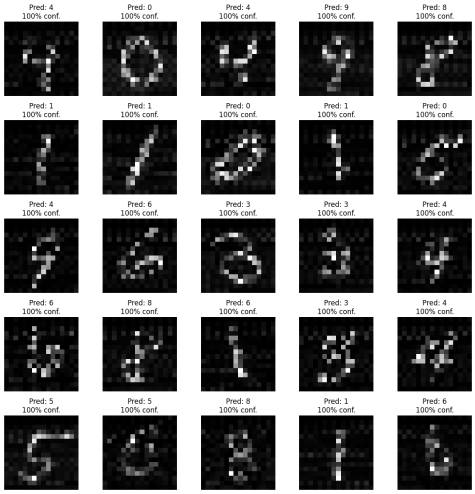
We address this issue by building on *RepFair-GAN* Sabbagh et al., which introduces *group-wise gradient norm clipping* to control the magnitude of each group's gradient during discriminator training. This method improves *representational fairness* without compromising sample quality. Our work extends RepFair-GAN to Quantum Generative Adversarial Networks (QGANs), leading to *RepFair-QGAN*, which utilzies quantum computational advantages to further enhance fairness in synthetic data generation. Figure 1a and Figure 1b show the generated samples from QGAN and RepFair- QGAN, respectively. The qGAN samples exhibit higher variability in classification confidence, while RepFair:qGAN produces more consistently classified digits, supporting its improved fairness in digit generation.

Key Novelty

- We introduce *group-wise gradient norm clipping* for mitigating representation bias, ensuring each demographic group's gradient updates are uniformly controlled.
- Our method inherently tackles *mode collapse*, a well-known challenge in both classical and quantum GANs.



(a) Generated digit samples using QGAN. Predicted digits and classifier confidence scores are displayed. Variability in confidence indicates potential biases in generation.



(b) Generated digit samples using RepFair-QGAN. Predictions show higher confidence and consistency, demonstrating improved fairness in generative modeling.

Figure 1: Comparison of digit samples generated using qGAN and RepFair-QGAN. The confidence and consistency in predictions highlight the fairness improvements achieved by RepFair-QGAN.

- We extend our approach to QGANs, leveraging quantum computational advantages to enhance representational fairness while retaining high-quality sample generation.

The remainder of this paper details the theoretical foundations of our gradient-clipping strategy, provides empirical evidence of its effectiveness in promoting uniform data generation, and discusses its broader implications for fairness in generative modeling.

2 METHODOLOGY

In this section, we outline our methodological framework for designing and training generative models. We first discuss the fundamentals of QGANs, underscoring their potential for learning complex distributions. We then examine *representational fairness*, a separate but important concern that focuses on ensuring sensitive attributes are uniformly represented in generated data.

2.1 QGANS

QGANs extend classical GAN architectures by embedding quantum resources into the data generation process. Formally, let $U(\theta)$ be a parameterized quantum circuit acting on an n-qubit initial state $|0\rangle^{\otimes n}$. The quantum generator then produces the state

$$|\psi_\theta\rangle \;=\; U(\theta)\,|0\rangle^{\otimes n}, \tag{1}$$

whose measurement outcomes approximate samples drawn from the target distribution P_{data}.

A classical or quantum discriminator D_ϕ evaluates whether a given sample is real or generated. Denoting the generator's output distribution by $G_\theta(z)$, where z is drawn from some latent distribution P_z, one can write a typical adversarial loss function as

$$\mathcal{L}(D_\phi) \;=\; \sum_x P_{\text{data}}(x)\,\log\big(D_\phi(x)\big) \;+\; \sum_z P_z(z)\,\log\big(1 - D_\phi(G_\theta(z))\big), \tag{2}$$

which the discriminator maximizes with respect to ϕ. The generator's parameters θ are trained to minimize the same objective, resulting in the minimax optimization:

$$\min_\theta \max_\phi \; \mathcal{L}(\theta, \phi). \tag{3}$$

By leveraging quantum parallelism, QGANs can encode high-dimensional distributions with fewer parameters than purely classical networks. However, current quantum hardware limitations, such as qubit noise, restricted circuit depth, and limited qubit counts, can hamper practical performance. While the broader subject of fairness is important in machine learning, we do not incorporate fairness constraints into QGAN training in this work.

2.2 REPRESENTATIONAL FAIRNESS

Previous works Tan et al. (2020); Choi et al. (2020) have examined biases in classical GANs that produce non-uniform distributions over sensitive attributes. Achieving balanced group representation alone does not necessarily prevent such biases Kenfack et al. (2021). Here, we consider a generic methodology for promoting equitable outcomes in generative models by examining how data from sensitive groups are generated.

Consider a dataset $D = \{X, S\}$, where $\mathcal{X} = \{x_i\}_{i=1}^N$ is drawn from $P_{\text{data}}(X)$ and $\mathcal{S} = \{s_i\}_{i=1}^N$ is a binary sensitive attribute. The generator

$$g_\theta : \mathbb{R}^d \to \mathbb{R}^{n \times n} \tag{4}$$

transforms random noise $Z \sim P_z$ into synthetic data

$$D_\theta = g_\theta(Z). \tag{5}$$

By applying a function $h : X \to S$, one obtains a distribution

$$P\big(h(g_\theta(Z))\big), \tag{6}$$

which captures how frequently each sensitive attribute appears in the generated samples.

Definition (Representational Fairness). A generator g_θ is said to be ε-representationally fair if it produces a distribution of sensitive attributes that is, up to ε, indistinguishable from the uniform distribution $U(S)$. Formally,

$$\operatorname{dist}\Big(P\big(h(g_\theta(Z))\big), U(S)\Big) \;\leq\; \varepsilon. \tag{7}$$

Minimizing the Kullback–Leibler divergence between these two distributions is one approach to achieving

$$\operatorname{KL}\Big(P\big(h(g_\theta(Z))\big), U(S)\Big) \;=\; \varepsilon. \tag{8}$$

When $\varepsilon = 0$, the generator perfectly matches the uniform distribution of the sensitive attribute.

In principle, one could explore how representational fairness constraints might interact with a quantum generator. However, in this work, we focus on classical fairness considerations and do not explicitly integrate fairness objectives into our QGAN setup.

3 EXPERIMENTS AND RESULTS

We conducted our experiments using the MNIST dataset LeCun et al. (1998), divided into black and white background subsets. A Conditional Generative Adversarial Network (CGAN) Mirza & Osindero (2014) was trained to generate digits while considering both variations. A pre-trained classifier with 100% accuracy identified the digit and background color of each generated sample, enabling analysis of representation distribution and generative biases. Experiments were run on the Pennylane QPU simulator for efficient evaluation.

Figure 2 shows the distribution of generated digits for RepFair-QGAN and qGAN. The mean count across classes is around 200, but qGAN exhibits imbalances, overrepresenting some digits (e.g., digit 3 with 383 samples) and underrepresenting others (e.g., digit 9 with 71 samples). RRepFair-QGAN achieves a more uniform distribution, enhancing fairness.

To quantify fairness, Figure 3 presents variance and standard deviation measurements. QGAN's variance is 9496.89, compared to 3658.89 for RepFair-QGAN, while its standard deviation is 97.45, reduced to 60.49 for RepFair:qGAN. Lower values indicate better balance in generated samples.

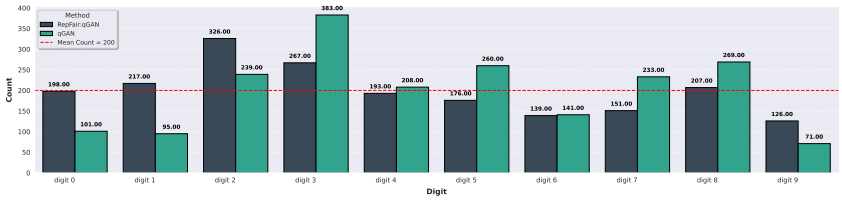


Figure 2: Distribution of generated digits across methods. The x-axis represents digit classes (0-9), and the y-axis shows the sample count. RepFair-QGAN achieves a more balanced representation than QGAN.

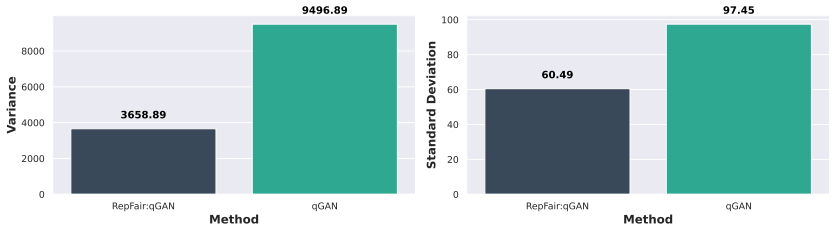


Figure 3: Fairness evaluation: Variance and standard deviation of digit distributions for RepFair-QGAN and QGAN. Lower values indicate improved balance.

Our results demonstrate that RepFair-QGAN enhances fairness by reducing representation imbalances, as indicated by lower variance and standard deviation. Future work may extend this approach to more complex datasets and further investigate the impact of fairness constraints on generative modeling.

4 Limitations

Due to the limited number of qubits available, this study focuses exclusively on the MNIST dataset. MNIST offers a multi-class setting necessary to examine fairness issues in representations; however, more complex datasets cannot be effectively learned with the current qubit constraints, and simpler datasets would not provide enough class diversity. Extending these techniques to more complex datasets will require additional qubits or more efficient quantum architectures, which remains a key direction for future work.

5 Conclusion

Our findings demonstrate that group-wise gradient clipping successfully extends to QANs, enabling the exploration of fairness in emerging quantum gate-based machine learning. We observed that biases commonly inherited from classical deep learning indeed transfer to the quantum domain, as shown by skewed image generation on MNIST. By incorporating group-wise gradient clipping, these biases were mitigated, yielding more balanced representations across classes. While MNIST was chosen due to qubit limitations and its multi-class nature, future research could expand these techniques to more complex datasets as quantum hardware evolves. This work highlights the importance of fairness considerations in quantum models and lays the foundation for further studies on mitigating representation bias in quantum machine learning.

Acknowledgments

This work was supported by the Research Center for Artificial Intelligence at Innopolis University.

REFERENCES

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Patrik Joslin Kenfack, Daniil Dmitrievich Arapov, Rasheed Hussain, SM Ahsan Kazmi, and Adil Khan. On the fairness of generative adversarial networks (gans). In *2021 International Conference" Nonlinearity, Information and Robotics"(NIR)*, pp. 1–7. IEEE, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Elija Perrier. Quantum fair machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 843–853, 2021.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Kamil Sabbagh, Patrik Joslin Kenfack, Adín Ramírez Rivera, and Adil Khan. Repfair-gan: Mitigating representation bias in gans using gradient clipping.

Hadi Salloum, Kamil Sabbagh, Vladislav Savchuk, Ruslan Lukin, Osama Orabi, Marat Isangulov, and Manuel Mazzara. Performance of quantum annealing machine learning classification models on admet datasets. *IEEE Access*, 2025.

Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.