Quantum-Aware Image Encoding and Adversarial Perturbation

by

Naziba Tasnim
21201320
SK Saif Ibna Ezhar Arko
21241010
MD Adnan Hossain Sadi
22299314
Lail Fadlin
21201204
Omar Nasif Chowdhury
21301035

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Naziba Tasnim

21201320

SK Saif Ibna Ezhar Arko

21241010

MD Adnan Hossain Sadi

22299314

Lail Fadlin

21201204

Omar Nasif Chowdhury

21301035

Approval

The thesis titled "Quantum-Aware Image Encoding and Adversarial Perturbation" submitted by

1. Naziba Tasnim (21201320)
2. SK Saif Ibna Ezhar Arko (21241010)
3. MD Adnan Hossain Sadi (22299314)
4. Lail Fadlin (21201204)
5. Omar Nasif Chowdhury (21301035)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 31st January, 2026.

Examining Committee:

Supervisor:
(Member)

Ipshita Bonhi Upoma

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

All images involved in this study are sourced from publicly available datasets in Kaggle with permissions for research included.

Abstract

The emergence of generative AI poses problems for people in several industries, including artists and other content creators whose work can be trained on without their consent, along with public figures who can be impersonated. As generative AI is now being researched at the quantum level, adversarial AI in quantum remains a largely unexplored topic. With present NISQ architecture and current hardware limitations for quantum image processing in mind, we present a hybrid quantum adversarial AI system using the QPIXL++ library for investigating image perturbations with quantum encoding using Flexible Representation of Quantum Images (FRQI) and Sparse Fast Walsh Hadamard Transform (SFWHT) compression. Topics for investigation in this area include pixel-domain vs frequency-domain perturbations under quantum compression, optimal perturbation strength, quantum superposition, working around quantum hardware limitations and applying optimization techniques with models as DINOv2. This project aims to establish a foundation and launch new research directions in hybrid quantum machine learning and adversarial AI. After completion of the study, it was found that the mean preserved perturbation value for the images was 99.99%, meaning that most of the classically introduced perturbations survived quantum encoding. Additionally, the mean SSIM vaule was close to 1 and the mean PSNR value was 32.30 dB which is above the accepted threshold for indistinguishable images to human eyes.

Keywords: Quantum Image Processing, Image Perturbation, Quantum Encoding, Quantum Machine Learning, Adversarial AI

Acknowledgement

This journey would not have been taught to occur without the blessings of the Almighty and the encouragement of our well-wishers. We owe a great gratitude towards supervisor Ipshita Bonhi Upoma ma'am because of her tutoring and strict guidance.
We would also like to express our sincere gratitude to the Department of Computer Science and Engineering at Brac University for providing the academic environment and resources necessary to complete this research. We also appreciate the assistance of the library staff for their support in accessing the required literature and digital repositories throughout the course of this work.
Furthermore, we are grateful to the open-source quantum computing community, whose tools and documentation were instrumental in the development and simulation of the quantum-aware algorithms presented in this thesis. In particular, we would like to acknowledge the use of QPiXL, which played a vital role in the image encoding and processing frameworks explored in this study. Finally, we owe our families and friends that have always come to our assistance. Thank you to the people who constantly supported us in some way or another, to being a part of this excellent journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

First of all, with the introduction of large scale Latent Diffusion Models and transformer based vision models, the media sphere and digital art have been fundamentally redefined. Hence, they are some revolutionary but even more so relying on the consumption of large datasets that may have been gathered without the glaring approval and compensation of the original author, this has steered to a flourishing area of adversarial defense against digital intellectual property. In addition to that, it is characterized by the development of models like Glaze and Nightshade which apply unnoticeable pixel-level perturbations to cloak the stylistic aspects of an image from AI encoders. Conversely, QML and it can be translated to mean Quantum Machine Learning has emerged as a paradigm shift and it is providing the possibility of a quantum advantage in solving a problem of data. Various techniques in the area of Quantum Image Processing such as FRQI and QPIXL techniques represent the encoding of classical images into quantum superpositions. At the same time, the existing studies provide that the essential probabilistic character of the quantum states combined with certain compression algorithms such as SFWHT implying that Sparse Fast Walsh Hadamard Transform is applicable to introduce the unique and non-linear adversarial noise that is essentially different than the pixel domain perturbations.

1.2 Rational of the Study or Motivation

The key driving force of this study is once again due to the vituperative arms race between content generators and generative AI developers. Furthermore, classical adversarial perturbations are gradually vulnerable to being skipped by adaptive denoising filters and solid feature extractors like DINOv2, the computational overhead of generating high-fidelity classical masks can be restrictive for individual artists. Initially, quantum-aware encoding advocates a unique opportunity to indicate these gaps by manipulating images at the quantum level specifically by minimizing rotation gates in the QPIXL method. As for example, one can achieve almost undetectable but mathematically significant alterations in pixel values. Moreover, this research is notable because it explores the largely unexplored territory of Quantum Adversarial AI by investigating how pixel-domain and frequency-domain perturbations works under quantum compression. Therefore, finding an optimal balance

between protection strength and visual fidelity within the constraints of NISQ which means Noisy Intermediate Scale Quantum and this architecture is important for creating the next generation of digital sovereignty tools.
Quantum machine learning and image processing is an emerging field, with significant investment from established technology corporations. As current classical adversarial attack models are not safe against possible countermeasures, and with the possibility of quantum generative AI models arising once fault tolerant quantum computers arise, this pipeline aims to prepare for such a future as the adversarial ML research we see today was developed mainly as a response years after the emergence of generative AI models, with no opportunity for preventative measures.

1.3 Problem Statement

The rapid transition of generative AI has inflamed an arms race between artists and AI developers where orthodox classical digital cloaking tools like Glaze and Nightshade serve as the front line of shield. Besides, a massive research gap prevails in the resilience of these State-of-the-Art methods. However, pre-existing classical models are fundamentally deterministic and risingly vulnerable to adaptive de-noising filters and advanced feature extractors such as DINOv2. In addition to that, state of the art apparatus very often request substantial computational power like increasing the capabilities of standard consumer GPUs. Furthermore, exhibit performance deterioration against non-VAE pixel-space methods like Deep Floyd. In the quantum field, while famous encoding tools like Flexible Representation of Quantum Images propose high compression but they remain unfeasible for Noisy Intermediate Scale Quantum devices for to their intense sensitivity to noise and high circuit complexity. In light of this, the pivot problem highlighted by this research is the absence of a hardware efficient, hybrid quantum and classical merged framework which is capable of generating non-linear and quantum aware adversarial perturbations. Ultimately, intellectual property remains crucially vulnerable to unauthorized feature extraction by the next generation of quantum enhanced AI systems without such a unique and hybrid pipeline.

1.4 Objective

The fundamental objective of this research is to analyze and design a Hybrid Quantum Adversarial system that applies quantum-aware encoding to protect digital imagery from unauthorized duplication. The precise goals are:

- to design a hardware-efficient image encoding pipeline by using the QPIXL++ library and incorporating FRQI and SFWHT compression to operate within NISQ hardware restrictions.
- to establish and compare pixel-domain and frequency-domain adversarial perturbations to examine which modality offers exceptional protection against feature extraction while maintaining high structural similarity (SSIM).
- to demonstrate the practicality and scalability of the model with the evidence of experimenting the model with the images that have been secured against the state of the art vision model including DINOv2 and latent diffusion models.

- to observably examine the scalability and resource efficiency of the proposed quantum aware models measured against leading classical adversarial defense systems.

1.5 Methodology in Brief

The methodology starts within the QPIXL framework using a dataset of approximately 15,000 64x64 RGB cat images sourced from Kaggle, which were then converted to PGM format for quantum encoding in 8x8 patches with OpenQASM 2.0 circuits (Hadamard, controlled-RY, CNOT gates) for more accurate statevector simulations matching current quantum hardware capabilities. DINOv2 distance is the primary metric for measuring perturbation, and we will also compare pixel and SFWHT perturbation domains across compression percentages to further measure how well classical perturbations survive quantum encoding.

1.6 Scopes and Challenges

This study dealt with one of the very niche fields of Quantum Machine Learning, Quantum Adversarial Attack, and because of the hardware limitations, many theoretical ideas could not be verified through actual experiments on real quantum hardware. These restrictions of the Noisy Intermediate-Sate Quantum (NISQ) devices limited the study in many ways including not being able to work with higher resolution images and colored images. Additionally, unavailability of open source Quantum Generative Adversarial Network models resulted in this study to not being able to test the performance of the adversarial attack proposed on an actual quantum generative algorithm. In addition, the nature of current classical models require relatively large computational power not accessible to most of the general public.

1.7 Thesis Organizations

The objective of this study was to explore the field of quantum adversarial attack and come up with a pipeline to protect original artworks from future quantum generative models. The construction of this thesis report is coordinated into six chapters to furnish a detailed exploration of quantum aware adversarial AI.

The report starts with chapter 1 which establishes the context of generative AI hazards, the incentive for quantum level protection and the research gap, research objectives and a detailed summary of the methodology. In the second place, chapter 2 vastly explains literature review which provides a critical analysis of state of the art adversarial attacks with quantum image encoding techniques like FRQI and QPIXL with detailed formula explanation and quantum GANs. Lastly, the constrains of current modern classical defense pipelines. After that, chapter 3 which is requirements, impacts and constraints showcased the functional and non-functional standards of the offered system alongside its societal, environmental and ethical aspects. Following that, proposed Methodology is in chapter 4, illustrate the architecture of the hybrid quantum-classical pipeline which include data collection from

Kaggle, the PGM conversion and the execution of QPIXL++ with SFWHT compression techniques. Moreover, chapter 5 explains result analysis and it presents the performance evaluation of the system using few metrics such as Structured Similarity Index Measure (SSIM), Peak Signal to Noise Ratio (PSNR) and DINOv2 distance followed by a mathematical, statistical analysis and comparative discussion with detailed explanation. Finally, chapter 6 covers conclusion which recaps the key findings, indicates novel contributions to the territory of Quantum Machine Learning. Lastly, it discusses the proposed recommendations for future research in post-NISQ systems.

Chapter 2

Literature Review

2.1 Review of Existing Research

Adversarial perturbation

This study [3], [4] highlights some critical vulnerabilities regarding deep neural networks in "The Limitations of Deep Learning in Adversarial Settings". A newly developed method of generating adversarial samples uncovers that targeted misclassification can be induced by modifying only 4.02% of the input features. The authors develop an efficient attack optimization process using forward derivatives and adversarial saliency maps. This approach outperforms traditional gradient based attacks since it allows for fine grained control over the misclassification outcome with minimal perturbation magnitude. The author formalizes white box attack models with full access to the model design and black box attack to the oracle only. Computes adversarial saliency maps by F(X), where F is the derivative of the forward pass of a neural network, providing regular input feature sensitivities. Moreover, the empirical validation of MNIST data set has an attack success rate of 97 per cent at 4.02 per cent on average distortion of all inputs. Besides, the researcher illustrate the ineffectiveness of the current defenses by gradually tightening adversarial examples to evade defenses. Still, it is more systematic in their approach to the adversarial samples generation when calculating forward derivatives and saliency mapping, both using the L-BFGS algorithm and the FGSM schemes [4]

2.1.1 Quantum Image Encoding

Within Quantum Image Processing territory, several encoding techniques have showcased specific effectiveness, most prominently the Flexible Representation of Quantum Images (FRQI) and Quantum Pixel Representation (QPIXL) techniques. Moreover, these models transformed classical image data into quantum states, enabling compact storage, data compression and highly parallelized computation.

FRQI: The FRQI method specifically encodes both pixel color and position concurrently within quantum superposition, utilizing amplitude and phase encoding. For an image of size $2^n \times 2^n$, the representation is expressed as

$$|I\rangle = \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} (\cos\theta_i|0\rangle + \sin\theta_i|1\rangle)|i\rangle, \tag{2.1}$$

where θ_i symbolizes the intensity of i-th pixel. Secondly, this particular scheme obliges $2n+1$ qubits. As for example, encoding a 32×32 image with $n = 5$ requires eleven qubits. In addition to this, FRQI offers data compression as well as parallel processing services. Nonetheless, it is highly susceptible to noise as well as quantum state preparation remains complex [12], [13].

QPIXL: By taking inspiration from quantum image representation models like FRQI and NEQR, the QPIXL framework proposed a uniform image representation which was pixel oriented along with reducing the gate complexity of previous representaion techniques. It is implemented using R_y and CNOT gates, achieving linear gate scaling while eliminating the need for ancillary qubits. Each pixel is represented as

$$|I\rangle = \sum_i \alpha_i |c_i\rangle |p_i\rangle, \tag{2.2}$$

where $|c_i\rangle$ encodes the pixel color and $|p_i\rangle$ specifies its position. Approximately ten to twelve qubits suffice to encode a 32×32 image. By balancing computational efficiency with resilience to noise, QPIXL is particularly well suited for NISQ era quantum devices [14].

2.1.2 Parameterized Quantum Noise Model

Fast Gradient Estimation

With the Bayesian Linear Gradient Estimator for VQAs, this paper [17] illustrates the huge potential of measurement reductions in NISQ devices. The BLGE leverages circuit structure priors in order to balance the bias-variance trade off arising between systematic errors, due to finite sampling, and statistical errors stemming from quantum measurements, within a Bayesian framework.

Key Technical Takeaways:

- Outperforms PSR and finite difference methods on small/moderate measurement budgets
- Realizes $10\times$ measurement reductions in QAOA setups
- Empirical shallow and spectral deep circuit depth adaptation
- Uses parameterized quantum gate frequency components for estimation.

Currently, it is restricted to periodic unitaries. This method performs better in resource constrained settings compared to PSR. This method works really well for practical VQA optimization under constrained measurement budgets. Non-periodic unitaries, barren plateau interactions, adaptive priors, and Hessian based optimization are some other directions that are left for future work.
Comparative measurements confirm that BLGE retains accuracy advantages for different measurement budgets while reducing quantum circuit evaluations and, therefore, is especially useful in NISQ era implementations.

Fast Gradient Sign Method (FGSM): FGSM is a single-step, white-box adversarial attack that produces small but adversarially effective input perturbations by leveraging the gradient of the loss function with respect to the input. For a model M_θ, an input x with true label y, and a loss function $\mathcal{L}$, FGSM generates an adversarial example as

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}\big(\nabla_x \mathcal{L}\big(M_\theta(x), y\big)\big)\,, \tag{2.3}$$

where ϵ constrains the perturbation magnitude under the ℓ_∞ norm. The sign operation adjusts each input dimension in the direction that most rapidly increases the loss, yielding a worst case perturbation under a first order approximation. Owing to its low computational cost and empirical transferability across models, FGSM is commonly used in adversarial training and robustness evaluation, and is often treated as a baseline approximation to more powerful iterative attacks, including **Parameterized Quantum Adversarial Noise Channel (PQANC)** [5].

Parameterized Quantum Adversarial Noise Channel (PQANC): PQANC unifies quantum classical noise mechanisms by tuning Kraus operators and weight parameters to get desired result.

$$\mathcal{E}_{PQANC}(\rho) = \sum_i \lambda_i(\boldsymbol{\theta}) K_i(\boldsymbol{\theta}) \rho K_i^\dagger(\boldsymbol{\theta}), \tag{2.4}$$

In the above equation $\boldsymbol{\theta}$ denotes learnable parameters optimized adversarially to maximize model error under the following cost function

$$C_{adv} = -\|\mathcal{M}(\mathcal{E}_{PQANC}(\rho)) - y_{true}\|, \tag{2.5}$$

This channel effectively emulates complex decoherence processes while enabling inverse-channel noise mitigation, PQANC requires high circuit depth [27].

2.1.3 Quantum GAN

The generation of features from **quantum encoded noisy images**, including those encoded using FRQI, NEQR, and QPIXL methods, has led to the development of complex quantum and hybrid GAN models that can efficiently simulate amplitude, phase, and entanglement noise. Three of them are **HQCGAN**, **QGAN HE**, and **QWAN** which are distinct means of describing the processes of quantum noise and the distributions of features that occur.
The **Hybrid Quantum Classical GAN (HQCGAN)**, developed by Zoufal, Lucchi, and Woerner [10], combines a quantum generator and a classical discriminator. The parameterized unitary circuit $U_G(\boldsymbol{\theta})$ simulates the probability distributions of noisy quantum states, generating samples according to $|\psi_G\rangle = U_G(\boldsymbol{\theta})|0\rangle$, with the loss function calculated based on feedback from the classical discriminator:

$$L_G = -E_{z\sim P_z}[D(G(z))], \tag{2.6}$$

By exploiting quantum effects such as superposition and decoherence, HQCGAN successfully simulates complex amplitude and phase noise patterns that are infeasible for classical GANs. Experiments on 4 to 8 qubit IBM Q systems showed high fidelity reconstruction of Gaussian-distributed FRQI and NEQR encodings, with fidelity scores above $F > 0.92$ under moderate decoherence. Although HQCGAN is highly effective in noise adaptive feature distribution learning with low hardware complexity, its classical discriminator limits its ability to detect entanglement featured noise [10].

The **Quantum Hybrid Entanglement GAN (QGAN-HE)**, proposed by Zhao et al. [26], further enhances noise robust quantum image generation by incorporating entanglement awareness into both the generator and discriminator. The generator generates entangled image states $|\psi_{gen}\rangle$, while the discriminator evaluates entanglement fidelity using a quantum kernel:

$$F_Q = |\langle\psi_{gen}|\psi_{true}\rangle|^2, \tag{2.7}$$

By optimizing directly for quantum fidelity instead of classical loss measures, QGAN-HE retains spatial and phase correlations in noisy FRQI and QPIXL encodings. For 8×8 images under depolarizing and dephasing channels, the model obtained $F_Q \approx 0.95$ with 12 to 16 qubits on IBM Q hardware. Its key strength is in maintaining coherence and accurately modeling entangled noise patterns, although this is achieved at the cost of more complex circuit depth and higher computational complexity [26].

The **Quantum Wasserstein Adversarial Network (QWAN)**, proposed by Ramakrishnan and Banchi [8], uses a geometric loss function with the quantum Wasserstein distance to guarantee perceptually valid noise reconstruction. The generator minimizes

$$\min_{\theta} W_2(\rho, \mathcal{E}_{noise}(\rho_\theta)), \tag{2.8}$$

where the target noisy quantum image is represented by ρ and $\mathcal{E}_{noise}$ is a quantum noise channel such as amplitude damping or dephasing. This approach enables the modeling of continuous noise manifolds with perceptual consistency, surpassing FID-like metrics in quantum image similarity assessment. Simulations on 10 qubit systems showed small geometric distances and high structural similarity ($\text{SSIM}_Q > 0.93$), although the computational cost of Wasserstein gradient estimation limits scalability [8].

2.1.4 Classical pipeline

Glaze

This project [21] is a subject of distinct interest in this work. Glaze is a software developed by a team from University of Chicago which works to prevent such style mimicry. The working principle of glazed is based on digital perturbations. Firstly, the software gathers and isolates specific features of the target image using a feature extractor function. This solution seeks to identify the style specific features of art pieces leveraging style transfer by transferring the original artworks to a different artistic style and then comparing the feature space of the original and transferred image. Then it introduces cloak perturbations to the identified style specific features within the perturbation budget so that the AI agent trying to mimic the art style fails without disrupting the quality of the original artwork noticeably. Glaze claims that about 92.1% times artists were convinced that Glaze was successful in protecting their art styles against real world style mimicry services like scenario.gg. However, this approach faces some challenges. It requires significant compute power which is beyond the scope of lower end GPUs that artists may work on (glazefaq), and there exists a ratio between the amount of perturbation to image quality. It is also weaker against img2img tools and some non-VAE models working in Pixel space such as Deep Floyd.

The newest version of Glaze acknowledges the limitations of using CLIP distance as a perturbation quality measurement metric [22], so for the purposes of this study, we chose to utilize DINOv2 distance in our experiments.

Nightshade

Related to the Glaze project, Nightshade [25] is more of a targeted data poisoning attack that prevents accurate prompt generation, effectively confusing the model about the image itself. It selects random poison text prompts, creates anchor images based on different concepts, adds perturbations to the target image, and creates a poison image using the LPIPS algorithm. In this approach, changes are more visible on artwork with flat colors, and there exists no mimicry protection - it exists as more of a supplement to the previously mentioned software above.

Mist

Mist is an open-source software [19] designed to protect against img2img attacks. The approach this software takes is similar to Glaze, however, the creators of Glaze themselves theorize that for defense against img2img attacks, the settings for the perturbations this software applies is of a very high setting, about 5 times higher than the maximum setting applied through Glaze, which may not be usable for most artists.

Anti Dreambooth

This paper [23] is written with the online GenAI web application DreamBooth in mind, the main incentive being the creation of a defense tool against the potential misuse of images of real people uploaded on the DreamBooth server by malicious actors.
Here, adversarial perturbations are applied to the photos before they are uploaded to the server. Two fundamental techniques are explored here: Fully-trained Surrogate Model Guidance (FSMG) and Alternating Surrogate and Perturbation Learning (ASPL). The former uses a fixed surrogate model with the training dataset consisting of clean images, while the latter takes on a more iterative reverse-engineering approach by updating a surrogate DreamBooth model, an approach that can mirror that of the attackers.
This study uses the VGGFace2 and CelebA HQ datasets. Anti Dreambooth has been showed to degrade the quality of the generated images even with variations in prompts and model versions.
However, this approach may be vulnerable to attackers having access to the unmodified target images beforehand, and due to the targeted nature of this approach, specific counterattacks may be taken by the models present on the server.

2.1.5 Stable Diffusion

Stable Diffusion is a *latent diffusion model* (LDM) developed for text to image generation. Rather than performing diffusion directly in pixel space, it operates within a *compressed latent space*, substantially reducing computational requirements while maintaining high image fidelity [28]. The approach integrates *score based diffusion* with *denoising autoencoders*, enabling the model to progressively transform Gaussian noise into semantically coherent images conditioned on textual prompts.

Architecture: Stable Diffusion uses a *Variational Autoencoder (VAE)* to encode input images x into latent representations $z \sim q(z|x)$ and to decode these representations back into pixel space. A *U Net denoiser* predicts the injected noise $\epsilon_\theta(z_t, t, y)$ at each diffusion timestep t, conditioned on text embeddings y (e.g., derived from CLIP). Cross attention mechanisms integrate textual information into intermediate feature maps. During the forward diffusion process, Gaussian noise is added according to

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2.9}$$

while the reverse process iteratively removes noise to recover an estimate of the original latent representation z_0.

Training: The model is trained by minimizing a noise-prediction objective defined in latent space:

$$\mathcal{L} = E_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|^2\right], \tag{2.10}$$

Conditioning on paired image text data aligns visual outputs with textual prompts, and large-scale pretraining supports generalization across diverse styles and concepts [18].

Features and Applications: Notable features include computational efficiency due to latent space diffusion, flexible conditioning on multiple modalities (e.g., text, reference images, depth maps), and strong image quality. As a result, Stable Diffusion is applied to tasks such as text-to-image synthesis, inpainting, super resolution, and parameter-efficient fine-tuning using methods such as LoRA or DreamBooth [20], [24].

2.1.6 Simulation platform: QClab++

QClab++ is a C++ framework for quantum computing designed to support simulation, algorithm prototyping, and hybrid quantum classical development. Building on the foundational principles of QCL, or Quantum Computation Language, it emphasizes computational efficiency, contemporary programming abstractions, and flexible cross platform architectures. The framework is particularly well suited for research in quantum machine learning, quantum generative modeling, and adversarial quantum experiments [11], [15].

It offers the basic abstractions of quantum registers, parameterized quantum circuits and user defined unitary gate operations. Both state density matrix and state vector. there are simulations, and thus the amplitude, phase and en-in can be modelled. tanglement dynamics. This framework also has noise modeling that includes noise modeling. standard procedures, such as amplitude damping and depolarization and adversarial perturbation algorithms such as FGSM or PQANC applied on FRQI, NEQR, or. QPIXL encoded image states. The structure has classical optimisation algorithms, including gradient descent and Adam, which are practiced to prepare variational quantum circuits. These HQCGAN algorithms are used as feedback loops of hybrid quantum classical architectures. Besides, QClab++ provides measure. routines and measures of fidelity permitting the quantitative measurements of quantum states. and the productions of the circuits. QClab++ capabilities of combining variational circuits with parameters are the main procedures. explicit noise modeling, hybrid classical quantum workflows and extensibility to. custom QGAN architectures. The capabilities make the framework all these. applicable to quantum generative modelling and adversarial. quantum machine learning to teaching and development variational algorithms. prototyping.

2.1.7 Evaluation Metrics

The **Structural Similarity Index Measure (SSIM)** [2], assesses perceptual similarity by simultaneously considering luminance, contrast, and structural information in two images:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2.11}$$

where μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance between x and y. SSIM values range from -1 to 1, with 1 indicating identical structural content. Its sensitivity to perceptual differences has made it widely used in image restoration, compression, and super resolution research.
The **Learned Perceptual Image Patch Similarity (LPIPS)** [7], evaluates perceptual differences using deep feature representations rather than raw pixel comparisons:

$$LPIPS(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{x}^l_{hw})\|_2^2, \tag{2.12}$$

where l indexes the layers of a pretrained network and w_l denotes learned layer-specific weights. Lower LPIPS values indicate stronger perceptual similarity, whereas higher scores reflect more noticeable differences. LPIPS is widely adopted for benchmarking generative models, especially GANs and other perceptually driven synthesis tasks.
Quantum Fidelity [1], quantifies the overlap between two quantum states or density matrices:

$$F(\rho, \sigma) = \left(\mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]\right)^2, \tag{2.13}$$

This metric ranges from 0, representing orthogonal states, to 1, indicating identical states. Quantum fidelity is a fundamental measure of similarity in quantum communication, machine learning, and generative quantum modeling.
The **Fréchet Inception Distance (FID)** [6], compares the statistical distributions of real and generated images:

$$FID = ||\mu_r - \mu_g||_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \tag{2.14}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the means and covariances of real and generated feature distributions. Lower FID scores indicate that generated images are more realistic and diverse, making it a standard metric for evaluating GANs and diffusion based generative models.
Finally, the **Peak Signal to Noise Ratio (PSNR)** is a logarithmic measure of image fidelity that quantifies distortion between a reference image I and a reconstructed or perturbed image $\hat{I}$ based on the mean squared error (MSE):

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \tag{2.15}$$

where MAX is the maximum possible pixel intensity. Expressed in decibels, higher PSNR values correspond to lower distortion and greater similarity between images. Although PSNR is straightforward to compute and analytically convenient, it does not always align with human visual perception [2].

2.2 Summary of Key Findings

We have several key areas of research for the development of the hybrid quantum adversarial pipeline. The second edition of the Glaze project utilizes the DINOv2 metric for better image protection as opposed to CLIP distance, therefore becoming integral for our proposed system. The one pixel attack model, while demonstrating potential, is compute intensive in combination with DINOv2 and it requires re adjusting the current pipeline. QPIXL/FRQI for encoding Fast gradient sign method/PQANC for adversarial noise promises a solid quantum adversarial pipeline.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The functional requirements for this project include support for multiple image formats, successful conversion of those images to PGM (portable grey map) format, generation of FRQI circuits from images, application of adversarial perturbation, a classical simulation of quantum circuits, and batch processing of multiple images through simulation. Non-functional requirements include processing speed, memory usage, scalability and reproducibility.

3.2 Societal Impact

This study aims to contribute to visual creator empowerment and to the growing scope of the real-world impact of hybrid-quantum systems.

3.3 Environmental Impact

There are no immediate health or safety risks or any biohazards associated with this study.

3.4 Ethical Issues

No AI-generated images were utilized in this project, and more research is required when it comes to the detection of AI generated images in software.

3.5 Project Management Plan

The project scope includes measurement of quality metrics through transformer models and automated batch processing with plans for testing on IBM Quantum. At this stage, the project is not designed with GUI application and mobile platform support in mind.

3.6 Risk Management

The success of this project is inherently tied to the development and scalability of quantum processing power. As a significant amount of funding is being allocated towards the development of quantum computers in mainstream applications, there has been a growing interest in the development of quantum machine learning. This project aims to provide some groundwork into ethical adversarial AI practices for media and identity protection in a post-NISQ world.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

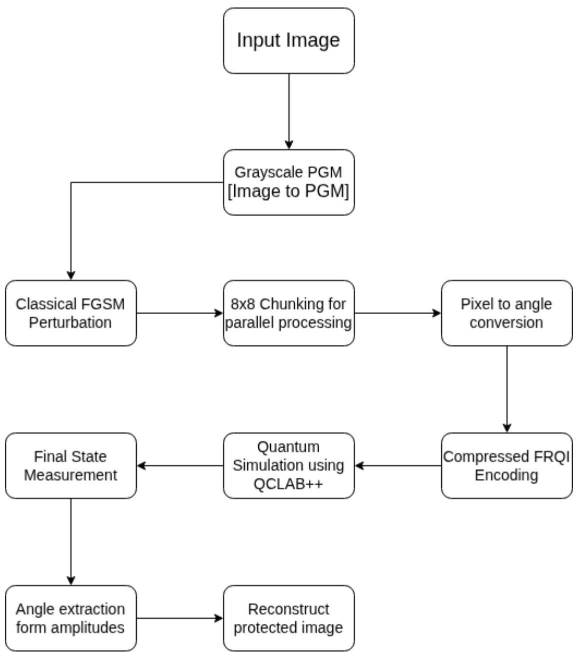


Figure 4.1: Methodology overview

The methodology in this study follows a structure which starts with the collection of data. The study expects 64-pixel images as input which are collected from publicly available datasets. After that, the images go through the pre-processing step where the input images are converted to PGM format. Following that, those images are converted to a quantum image representation format through the Flexible Representation of Quantum Images (FRQI) using the QPIXL method. During this step, this study takes advantage of the compression mechanism of the QPIXL method which originally was implemented to reduce the number of gates required to convert an image data to a quantum state by the authors of Quantum pixel representations and compression for N-dimensional images [16]. This compression technique alters the image in such a way that the change in pixel information cannot be detected by human eyes however the underlying color values of the pixels are altered. This is achieved by reducing the rotation gates on qubits where the rotation values will

be negligible thus slightly changing the color values of the pixels. This alteration works as a perturbation on the original image and can be used to corrupt training data in generative AI moddels. As this method generates quantum level perturbed images, theoretically this adversarial attack should be effective on Quantum Generative Adversarial Networks (QGANs) as well. However, due to the unavailability of any open source QGAN models, this claim is yet to be verified.
Additionally, this study implements classical approaches for perturbations using the Fast Gradient Sign Method (FGSM) framework and used the DINOv2 Validator class to validate perturbation performance. Key parameters include a perturbation strength (epsilon) range of 4-20, a Peak Signal-to-Noise (PSNR) ratio of over 30 dB (theorized to be imperceptible to the human eye), and a Structural Similarity Index Measure (SSIM) of over 0.95 to ensure that the perturbed image looks similar to the original.

4.2 Preliminary Design or Design (Model) Specification

In the earlier design, a few potential solutions have been taken into account to solve the issue of quantum-awareness of adversarial perturbation and be computationally feasible in NISQ hardware constraints. The encoding schemes, perturbation domains and optimization strategies formed the major design aspects. Three quantum image encoding schemes were initially taken into account: Flexible Representation of Quantum Images (FRQI), Novel Enhanced Quantum Representation (NEQR), and Quantum Pixel Representation (QPIXL). FRQI was selected as the main encoding technique due to its effective encoding of amplitude of pixel values according to rotation angle that is an intrinsic compression compress based perturbation. Even though NEQR is more faithful to the grayscale representation, its required alacrity is significantly higher than that required by modern NISQ instruments. QPIXL was not used as the primary encoding system, but it was applied on the basis of its compression mechanism despite having a linear gate scaling. Two perturbation domains were compared in this study: pixel-domain and frequency-domain (SFWHT). The pixels values in the spatial domain are manipulated by pixel-domain perturbations and can be interpreted as results and easily implemented directly. Perturbations of the frequency domain, attained by the application of Sparse Fast Walsh-Hadamard Transform, utilize the transform coefficients to incorporate perturbations that may be more robust to certain types of image processing. The two techniques were held in a comparative study to understand which region holds the power of adversaries in a better condition with the help of quantum encoding and decoding process.

Optimization Strategy: When dealing with the classical perturbation stage there were three optimization strategies considered, i.e. single step Fast Gradient Sign Method (FGSM), iterative Fast Gradient Sign Method (I-FGSM) and Momentum Iterative Fast Gradient Sign Method (MI-FGSM). Similar to the case of MI-FGSM, the algorithm has been selected as the primary optimization method since it possesses superior convergence properties and momentum preparation can be used to exit local minima. The method has the best quality of adversarial perturbation over single-step FGSM and is computationally cheap relative to more complex optimisation schedules such as the Carlini-Wagner attacks.

Quantum Simulation Framework: Due to the lack of real quantum hardware, the framework is created with a hybrid formula of a simulation utilizing quantum circuit simulation with QCLAB++. The model involves state vector simulation as an accuracy model and patch based processing to manage complexity of computations. The simulated pixel sizes of individual image tiles are 4×4 or 8×8, a trade-off between quantum circuit depth and the potential depth of classical hardware simulation. This tiling scheme is also indicative of realistic restrictions one would encounter when applied to real quantum processors of small scale, in terms of qubits.

Verification and Validation: Measured validation in terms of a number of metrics is present in the design specification. DINOv2 distance This is the primary measure of the strength of perturbation, and it measures the distance in the feature space of original and perturbed images:

$$d(x, x') = 1 - \frac{f(x) \cdot f(x')}{\|f(x)\| \|f(x')\|}, \tag{4.1}$$

f(x) is the DINOv2 embedding. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are used to assess the quality of preservation of images to ensure that the perturbations are invisible. Pre-quantum and post-quantum distance measurements enable assessment of perturbation survival through the quantum encoding-decoding pipeline. Target thresholds were established as: PSNR $>$ 30 dB for imperceptibility, SSIM $>$ 0.95 for structural preservation, and perturbation preservation $>$ 90% for adversarial effectiveness.

Design Trade-offs: The design proposed gives more priority to practical feasibility over theoretical quantum advantage taking into account the limitations of current NISQ era. The quantum compression parameter gives the control over the trade off between image quality and quantum perturbation performance. Lower compression rates preserve more visual fidelity but provide weaker quantum-level perturbations, while higher compression introduces stronger quantum noise at the cost of potential visual artifacts. The epsilon parameter (4-20) controls classical perturbation strength, with epsilon=8 identified as the optimal balance point during preliminary experiments. This preliminary design was validated through small-scale simulations on subsets of the dataset before full-scale implementation, confirming the feasibility of the proposed pipeline and informing the final parameter selections detailed in subsequent sections.

4.3 Data Collection

We used an open source image dataset from kaggle containing 64x64 sized images of cats named Cat Faces 64x64 [9].

4.3.1 Image to PGM conversion

For quantum processing, it was necessary to convert the images to grayscale PGM because of hardware limitations. For this study, image data was converted to PGM files using the python library named Pillow.

4.3.2 Summary of Preprocessed Data

The final dataset had 15747 images of cats. The size of the images were 64x64 pixels and all the images were in PGM format.

4.4 Implementation of Selected Design

The implementation of the proposed hybrid quantum-classical adversarial system was developed using a modular pipeline architecture that integrates classical perturbation techniques with quantum image encoding. The system was implemented primarily in Python 3.8+ using the QPIXL++ library for quantum image representation and QCLAB++ for quantum circuit simulation. All calculations were done on an NVIDIA RTX 3080 card to utilize hardware assistance in the calculation of state vectors. The implementation pipeline has four major steps:

1. Preprocessing and format conversion,
2. Classical adversarial perturbation,
3. Quantum encoding with FRQI compression,
4. Quantum simulation and decoding.

The stages were formulated in a more modular way to help in determining parameters and comparing the results of the stages with the various parameters. To work efficiently with batch processing images, the system was adopted with the aid of the Python multiprocessing library and patch-based decomposition allows quantum circuit simulation under the current hardware limit. The 64x64 grayscale picture is broken into non overlapping patches (4x4 or 8x8 pixels) that can be built up of quantum circuits and resized. Not only does this solve the problem of the small number of qubits in NISQ devices but it also allows more affordable simulation, as patches are simulated concurrently. Classical perturbation module is trained on Momentum Iterative Fast Gradient Sign Method (MI-FGSM) and DINOv2 is the feature extractor model. The model was DINOv2 (ViT-S/14 version), and it was applied without any fine-tuning because it already has ready-made weights that can be utilized to represent features in a robust manner. Perturbation gradient are calculated on the cosine distance of the DINOv2 feature space and momentum accumulation (decay=0.9) is used to stabilize the optimization process over the iterations. It was managed by a standard information format stream that combined classical and quantum elements. The output after running the modified FGSM algorithm was classically perturbed images of PGM format. The perturbed PGM images were then transformed into normalized pixel intensity arrays, and they were encoded into quantum rotation angles using the formula

$$\theta = \pi \times \frac{\text{intensity}}{255} \tag{4.2}$$

for FRQI representation. Rotation angles bellow a certain threshold were ignored to introduce quantum level perturbations to the image states. The optimization involved the quality assurance measures which were applied across the pipeline such as PSNR and SSIM threshold automated check, circuit depth monitor and gradient norm monitor. All the intermediate results are documented and hence the survival rates of perturbation of quantum encoding-decoding cycle can be statistically examined in detail. Finally, the quantum states were measured to reproduce the classical image data and then this image was compared to the original input image for perturbation performance using DINOv2 feature distance and image quality using PSNR and SSIM. It has around 3500 lines of written Python code in reproducible and extensible modular units to execute the entire process, namely, all the preprocessing programs, quantum circuit generators, and evaluation programs.

4.4.1 Adversarial Perturbation of PGM

The preprocessed PGM images undergo the white-box MI-FGSM perturbation component. The optimization algorithm used in this stage tries to maximize the feature distance of the DINOv2 validator class to ensure maximum perturbation strength. The iterations parameter determined the number of times the momentum iterative optimization algorithm was allowed to run. The maximum allowed alteration on pixel value was bounded by a L-infinity constraint (epsilon) in range between 4 to 20 units, defaulting to 8. The perturbation was achieved using a random gradient with spatial smoothing ranging from 0.5 to 0.99. The output of this state are the classically perturbed PGM images which then go through an evaluation step to ensure acceptable image quality and perturbation performance. Finally images are separated into non overlapping 8x8 blocks for quantum processing because of hardware limitations.

4.4.2 Quantum Encoding

The image blocks from the previous section undergo FRQI (Flexible Representation of Quantum Images) encoding, which encodes pixel intensity as rotation angles in a quantum state. The intensity to rotation conversion is done using the equation 4.2. The quantum state representing the images requires $\log_2(N)$ qubits for denoting all pixels of an nxm image where N $= n \times m$ and one qubit for storing the pixel intensity. Meaning for 8x8 blocks 6 qubits are required to represent all possible pixels and 1 for pixel intensity, so 7 qubits in total. The rotations are achieved by using R_y gates.

$$|I\rangle = \frac{1}{\sqrt{N}} \sum_i [\cos(\theta_i)|0\rangle + \sin(\theta_i)|1\rangle] \otimes |i\rangle, \tag{4.3}$$

The quantum circuit for encoding a 2x2 pixel image data into a quantum state using the QPIXL method is shown in Figure 4.2. Quantum level perturbation is achieved while encoding the image into quantum states taking advantage of the compression mechanism of the QPIXL method, as it slightly alters the color values of the pixels of the original image which may affect the performance of the generative models while being indistinguishable to the human eye.
In addition, the images are encoded under Sparse Fast Walsh-Hadamard Transform (SFWHT) in order to compare the perturbation performance in frequency domain with pixel domain.

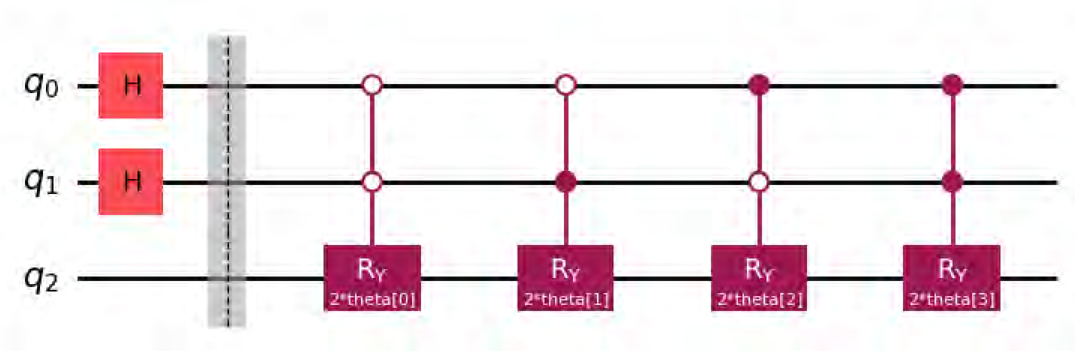


Figure 4.2: Quantum circuit for encoding

4.4.3 Quantum Decoding

Finally, the quantum state's amplitudes are observed from the state vector representation and continuous measurements were not necessary as we simulated the circuits in a classical computer. Using those amplitudes, the PGM image is reconstructed and evaluation metrics like DINOv2 distance, PSNR and SSIM is run on the reconstructed image. If the computation was done on a real quantum computer, we would have to perform quantum tomography to get back the image from the quantum state.

Chapter 5

Result Analysis

5.1 Performance Evaluation

For this study, several evaluation metrics were used to verify the performance of the proposed pipeline for both strength of perturbation and visual similarity between the perturbed output and the original input image. Those metrics were used to evaluate the performance of the system both before the quantum simulation and after. The metrics are as follows

5.1.1 Pre-Quantum Distance

This metric measured the cosine distance of DINOv2 features between the original input image and the perturbed output image before simulating the quantum encoding. This can be used to measure the level of protection in the image, where a higher value of pre-quantum distance means better protection because feature distance of DINOv2 is calculated in such a way where higher distance means the DINOv2 generator model finds it more difficult to reproduce the image.

5.1.2 Post-Quantum Distance

This metric measures the exact same thing as the pre-quantum distance after completing qunatum encoding, quantum perturbation and decoding stages.

5.1.3 Preserved Perturbations

This mectric indicates the amount of protection that remains after quantum processing. The formula used to calculate the metric is

$$\text{perserved\_pct} = \left(\frac{\text{post\_quantum\_dist}}{\text{pre\_quantum\_dist}}\right) \times 100\%, \tag{5.1}$$

5.1.4 PSNR - Peak Signal to Noise Ratio

This metric was used to determine the quality of perturbed images. The unit of PSNR is decibels and the higher the value, the better the image quality. PSNR was calculated using the equation 2.15.

5.1.5 SSIM - Structured Similarity Index Measure

This is also an image quality measure like PSNR but it is often considered to be more correlated with human perception compared to PSNR. This metric considers luminance, contrast and structure of an image as evaluation criterion. It could have been calculated using the equation 2.11 however, for this study we used the skimage.metrics.structural_similarity from the scikit-image python library.

5.1.6 Mean Perturbation

This metric measures the average absolute difference between the original image's pixel intensity and the pixel intensity of the final output image after all the processing. Low value of mean perturbation means less visible change.

5.1.7 Max Perturbation

This measures the maximum absolute change in the intensity of any pixel of an image. The value cannot go over the limit set by the epsilon value provided while running the system.

5.2 Analysis of Design Solutions

This study expects that introduced perturbations will remain intact after quantum processing so that quantum generative adversarial networks cannot efficiently process the images generated through the proposed system. For this reason, the desired value of Preserved Perturbations is around 100 %. An image is considered well protected if the cosine distance of DINOv2 features is higher than 0.5. This study labels image quality of images with PSNR value more than 35 dB to be of excellent quality. dB 30 to 35 is considered good as PSNR 28 to 30 dB means virtually indistinguishable images for human eyes. dB 25 to 30 is considered Acceptable and bellow 25 dB marked as poor quality image. The values of SSIM ranges from -1 to 1 and 1 means identical images. So, the preferable value of SSIM is close to 1. Taking all these evaluation metrics into account, this study experimented with several configuration of the system and attempted to find the best possible performance and image quality.

5.3 Configuration Parameters

This study predominantly used four parameters to affect the performance of the model

5.3.1 Epsilon

This parameter limited the maximum allowed change in any pixel value of an image. This ensures that the perturbed images do not completely change the original image. Higher value of epsilon will result in stronger protection, however will reduce the image quality compared to the original image.

5.3.2 Iterations

This parameter controls the number of steps the optimization algorithm takes to find the best result. This study uses momentum based iterative method as its optimization algorithm instead of using a neural network solution.

5.3.3 Patch Size

This determines the block size of the clusters used during quantum processing. As it is not possible to process the whole image in one iteration because of the lower number of qubits in NISQ era quantum hardware. This clustering also helps during quantum simulation as simulation a lager number of quantum bits in a classical computer is not feasible.

5.3.4 Quantum Compression Percentage

This parameter determines the compression rate used during quantum encoding using QPIXL to achieve adversarial property by controlling the ignorable rotation threshold on the quantum state.

5.4 Statistical Analysis

Because of computational limitations, we initially worked with a small sample size to optimize the parameters before running the model on the whole dataset. During this phase we managed to narrow down the variables for our parameter and concluded that epsilon value of 8 provided the best balance for both perturbation performance and image quality. We also found that patch size of 8 is feasible to process and compression of 5% provided acceptable image quality. Finally, we processed all images within the dataset using QCLAB++ on Nvidia RTX 3080 with these optimized values for different number of optimization iterations. Following figures show the difference between the results when epsilon is set to 8 vs 15 while other parameters were constant.

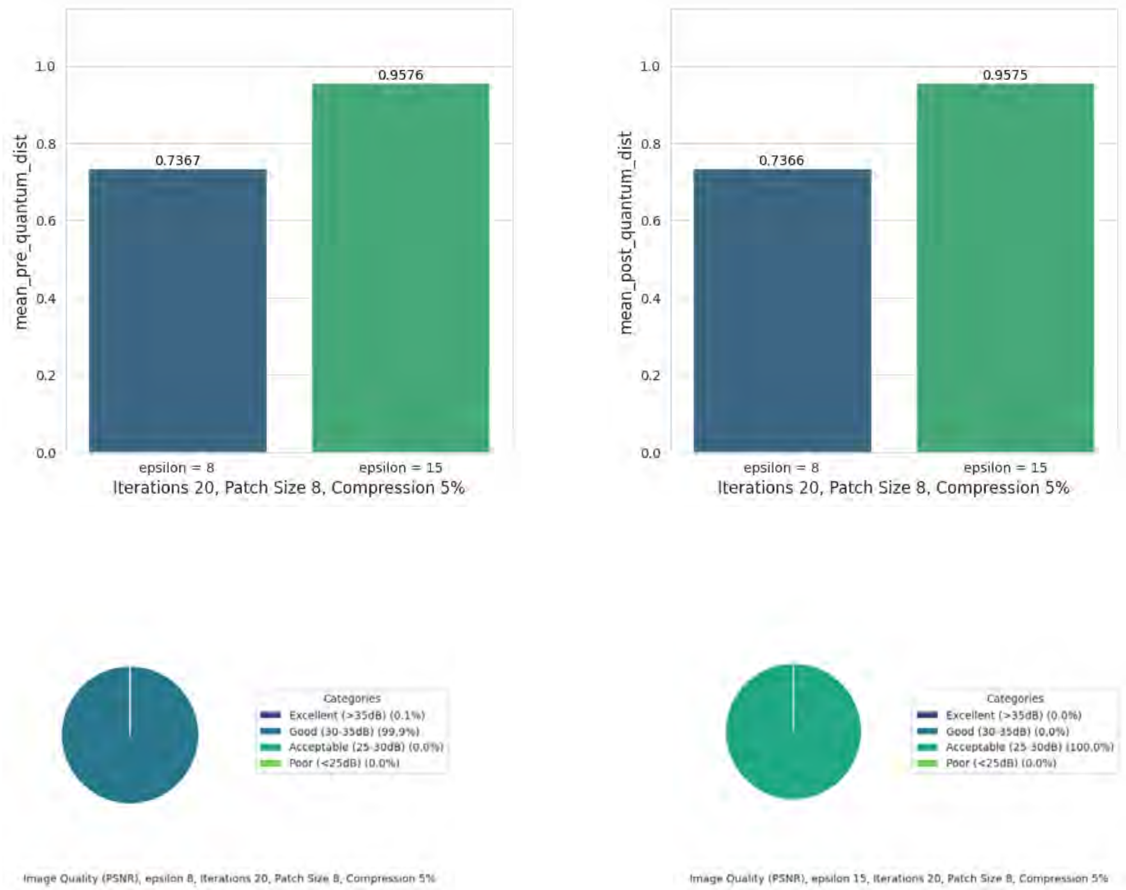


Figure 5.1: Comparison of Epsilon 15 vs 8

From Figure 5.1 it can be concluded that although the performance of the perturbations is better when epsilon value is 15, which is to be expected, the image quality reduces significantly. For this reason, we chose the epsilon value to be 8 to balance both perturbation performance and image quality.

5.5 Comparisons and Relationships

In the final stage of the study, we compared results between interations value 20 and 50 keeping other parameters constant and expectedly found that both perturbation performance and image quality to be sightly better when the optimization function is allowed to be run for more iterations. Figure 5.2 & 5.3 show the results for different iteration values.

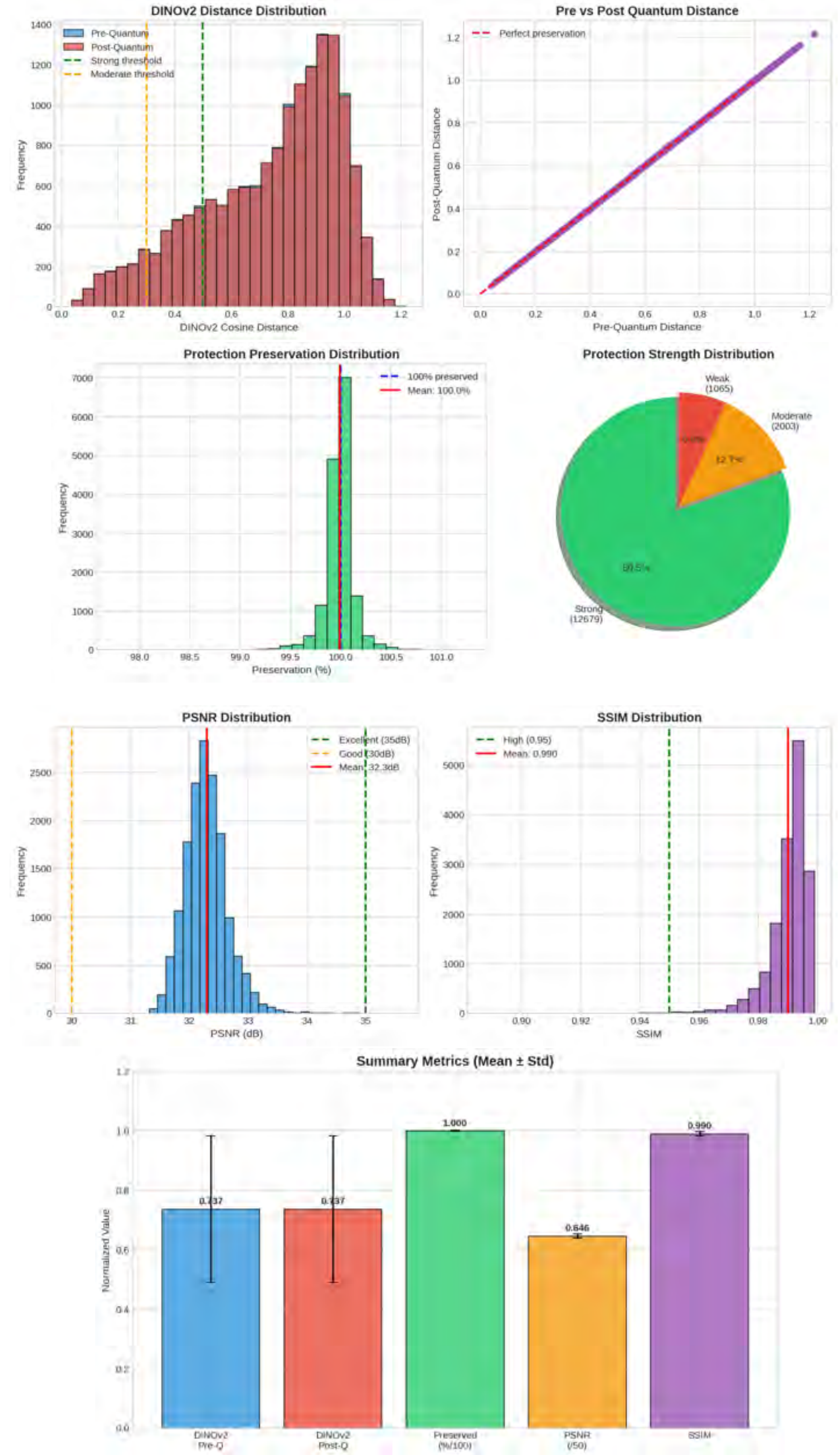


Figure 5.2: Iterations 20, Epsilon 8, Patch Size 8, Compression 5%

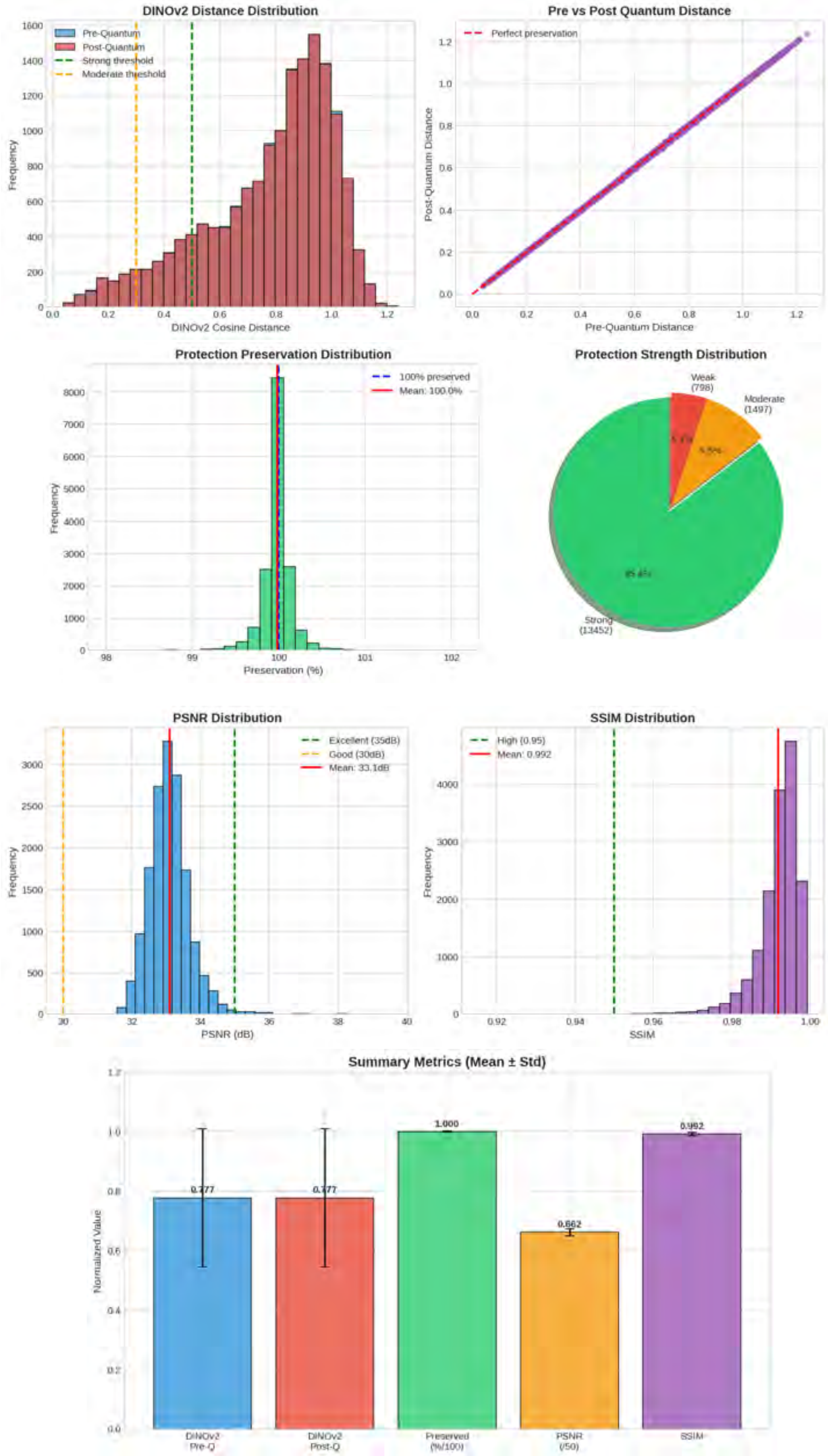


Figure 5.3: Iterations 50, Epsilon 8, Patch Size 8, Compression 5%

From Figure 5.2 & 5.3 we can see that majority of the DINOv2 Distance distribution fall above the strong threshold of 0.5 and preserved perturbation to be near 100%. Additionally PSNR value was greater than 30 dB for majority of the images and SSIM was close to 1. So, it can be concluded that our introduced classical perturbation servived the quantum processing and the introduced perturbation not only performed well but also did not degrade the images by a noticeable margin.

We also compared the system proposed in this study with an alternative classical adversarial algorithm and found that the model performs similar to the alternative if not better. Figure 5.4 show the comparison between the alternative and method proposed in this study.

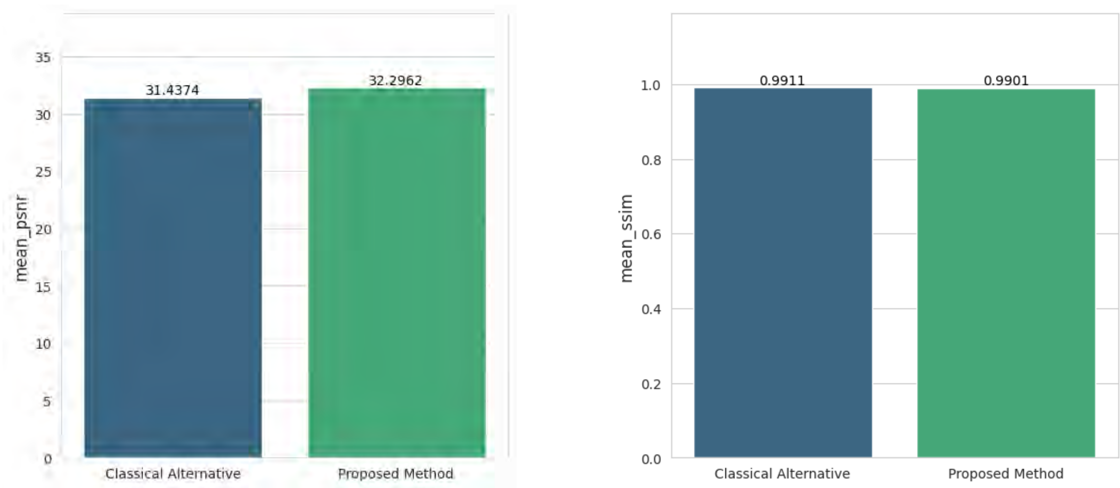


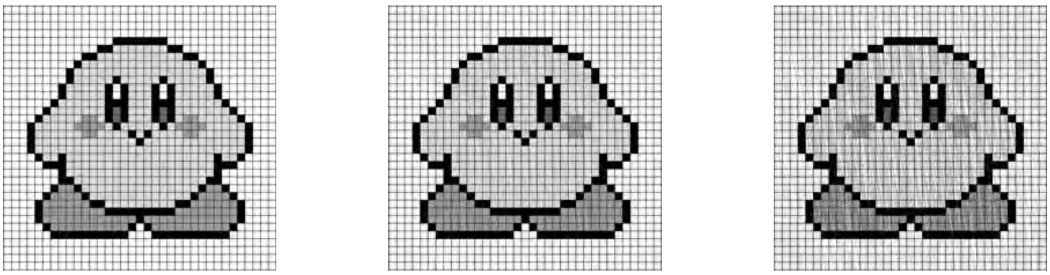
Figure 5.4: Classical alternative vs Proposed method (Epsilon = 8)

Compression %	Pixel-Domain Survival	SFWHT-Domain Survival
0%	100%	100%
50%	96.6%	73.6%
90%	66.4%	36.4%

Table 5.1: Domain comparison across compression percentages

In addition, we tested pixel-domain vs frequency-domain (SFWHT) domain in the process of compression. FRQI utilizes SFWHT for internal compression, however, we found that pixel domain performed consistently better. In both cases, we have found that classical perturbations can survive quantum encoding. Table 5.1 shows the results for different compression rates on Pixel and Frequency domains.

Finally, Figure 5.5 shows the visual difference between Input and Output images for 0% and 20% compression rate. It can be seen that there is virtually no difference between the input image and the classically perturbed image without quantum compression. Yet, this provided DINOv2 feature distance of 0.53 and PSNR value of 36.9 dB. This indicates that our introduced perturbation was successful to provide strong performance without sacrificing noticeable visual quality. The image with classical and quantum adversarial properties is noticeably different from the original image but it provided even more protection with feature distance of 0.57. The perturbation performance is not drastically better after quantum compression because the method used in this study for introducing quantum adversarial property adds model independent random perturbations while the classical adversarial attack specifically targets the generator model and adds perturbation to maximize the confusion for the generator model.



Input Image | Output without Compression | Output with 20% Compression

Figure 5.5: Comparison of Image quality

5.6 Discussions

Although we had a general satisfactory result, our quantum perturbation did not make much difference. This is likely due to simulation experiments with additional encoding overhead. As the classically introduced perturbations were present in the quantum encoded state validated by the preserved perturbations metric being close to 100%, this attack should also be effective against any quantum generative adversarial network in the future. However, it would have been preferable if we could come up with quantum adversarial algorithms with better perturbation performance, like the ones used classically. Although there have been studies on adversarial attacks for quantum classifiers, no previous study attempted to work with quantum adversarial properties for generative models. This study also explores the survival rates of classical adversarial qualities after quantum processing on images, which has never been done before. Finally, the compression mechanism introduced by QPIXL was never tested as an adversarial property prior to this study.

Chapter 6

Conclusion

6.1 Summary of Findings

The results demonstrate that there is sufficient potential with the exploration of hybrid quantum image perturbations. With the given image sizes and constraints, the results have consistently demonstrated high PSNR values with strong perturbation strengths against a feature extraction model. The results infer that classical perturbations can survive quantum encoding.

6.2 Contributions to the Field

This study is not aimed at demonstrating immediate quantum advantage, which is not feasible for current QML pipelines given current hardware limits. It is also not designed to replace classical methods completely, rather, it is aiming to set initial infrastructure for further research into the intersection of Quantum Imaging and Adversarial AI. Current adversarial AI systems have been released much later than generative AI research, and it may take at least a few years for those systems to be usable for the general public without a relatively powerful GPU or cloud access, both of which may be expensive to maintain. Therefore, this research aims to contribute to preventative measures through the given pipeline in the hopes that it may inspire more works in both domains in the future.

6.3 Recommendations for Future Work

As described in previous sections, this study was limited by the Noisy Intermediate-State Quantum (NISQ) hardware but the recent advancement of the hardware technology from Google and IBM along with other frontliners in quantum hardware research promises capable quantum hardware within the next few years. Access to those quantum computers will make it possible to work with higher resolution images and colored images. Additionally, developing several other quantum adversarial algorithms to compare with the one proposed in this study could be another path to advance the field of study and prepare for a future when efficient quantum hardware will be available.

Bibliography

[1] R. Jozsa, "Fidelity for mixed quantum states," *Journal of Modern Optics*, vol. 41, no. 12, pp. 2315–2323, 1994. DOI: 10.1080/09500349414552171.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861.

[3] C. Szegedy et al., "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018. arXiv: 1706.08500 `[cs.LG]`. [Online]. Available: https://arxiv.org/abs/1706.08500.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. arXiv: 1801.03924 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1801.03924.

[8] S. Chakrabarti, Y. Huang, T. Li, S. Feizi, and X. Wu, "Quantum wasserstein generative adversarial networks," 2019. arXiv: 1911.00111 `[quant-ph]`. [Online]. Available: https://arxiv.org/abs/1911.00111.

[9] S. Ghosh. "Cats faces 64x64 (for generative models)," Kaggle, Accessed: Jan. 26, 2026. [Online]. Available: https://www.kaggle.com/datasets/spandan2/cats-faces-64x64-for-generative-models.

[10] C. Zoufal, A. Lucchi, and S. Woerner, "Quantum generative adversarial networks for learning and loading random distributions," *npj Quantum Information*, vol. 5, no. 103, 2019. DOI: 10.1038/s41534-019-0227-2.

[11] C. Zoufal, A. Lucchi, and S. Woerner, "Quantum generative adversarial networks for learning and loading random distributions," *npj Quantum Information*, vol. 5, no. 1, p. 103, 2019.

[12] N. Iyengar et al., "Frqi encoding for quantum image processing," *Quantum Information Processing*, vol. 19, p. 120, 2020. DOI: 10.1007/s11128-020-02708-0.

[13] M. Mastriani, "Quantum image processing: A review," *Information*, vol. 11, no. 12, p. 610, 2020. DOI: 10.3390/info11120610.

[14] K. Amankwah et al., "Qpixl: Efficient quantum image encoding for nisq devices," *Quantum Information Processing*, vol. 20, no. 1, p. 45, 2021. DOI: 10.1007/s11128-020-02864-w.

[15] M. Benedetti, M. Fiorentini, and M. Lubasch, "Quantum generative models for learning and loading probability distributions," *Quantum Science and Technology*, vol. 6, no. 4, p. 045 024, 2021.

[16] M. G. Amankwah, D. Camps, E. W. Bethel, R. Van Beeumen, and T. Perciano, "Quantum pixel representations and compression for n-dimensional images," *Scientific Reports*, vol. 12, 2022. DOI: 10.1038/s41598-022-11024-y.

[17] L. Bittel and M. Kliesch, "Fast gradient estimation for variational quantum algorithms," *Quantum*, vol. 6, p. 745, 2022. DOI: 10.22331/q-2022-07-14-745.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR Workshops*, 2022.

[19] Anonymous, *Mist: Open-source protection against img2img ai attacks*, Available online, 2023.

[20] E. Hertz et al., "Lora: Low-rank adaptation of large language models," in *NeurIPS 2023*, 2023.

[21] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," 2023. arXiv: 2302.04222 `[cs.CR]`. [Online]. Available: https://arxiv.org/abs/2302.04222.

[22] S. Shan, S. Wu, H. Zheng, and B. Y. Zhao, "A response to glaze purification via impress," 2023. arXiv: 2312.07731 `[cs.CR]`. [Online]. Available: https://arxiv.org/abs/2312.07731.

[23] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. Tran, and A. Tran, "Anti-dreambooth: Protecting users from personalized text-to-image synthesis," 2023. arXiv: 2303.15433 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/2303.15433.

[24] J. Zhu et al., *Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation*, arXiv:2208.12242, 2023.

[25] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," 2024. arXiv: 2310.13828 `[cs.CR]`. [Online]. Available: https://arxiv.org/abs/2310.13828.

[26] T. Zhao, L. Banchi, and F. Caruso, "Quantum hybrid entanglement gans for noise-resilient image generation," *arXiv preprint arXiv:2406.08915*, 2024.

[27] L. Smith, "Noise-resilient hybrid quantum–classical algorithms for near-term quantum devices (nisq era)," 2025.

[28] psyker-team, *psyker-team/mist: Watermark tool to protect images from unauthorized diffusion mimicry*, GitHub repository, GPL-3.0 licensed software, 2026. [Online]. Available: https://github.com/psyker-team/mist.