Multi-Split Quantum–Classical Normalizing Flows via Analytically Invertible Quantum Transforms

Achal Jain, Abhijith Raju Nair, Dinesh Singh
Vision Intelligence and Machine Learning (VIML) Group
School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, India
{b22080, b22079}@students.iitmandi.ac.in, dineshsingh@iitmandi.ac.in

***Abstract*—Normalizing flows are a powerful class of generative models that offer exact likelihood estimation and tractable sampling via invertible transformations. Simultaneously, quantum machine learning (QML) promises to enhance expressivity through high-dimensional Hilbert space mappings. However, integrating quantum neural networks (QNNs) into normalizing flows presents a fundamental theoretical conflict, the *Measurement Paradox*. Standard quantum measurement is a non-invertible operation, breaking the strict diffeomorphic constraints required by flows. Existing literature relies on approximate reconstruction (e.g., quantum variational autoencoders), which sacrifices the exact log-likelihood guarantee that defines normalizing flows. In this work, we propose the multi-split quantum-classical normalizing flow (MSQCNF), a hybrid architecture designed to mitigate the effects of this paradox. We introduce two novel contributions: 1) The *Cosine Chain*, a mathematical formulation that directly addresses the measurement paradox by constructing quantum ansatzes whose post-measurement mappings remain analytically invertible with a tractable Jacobian, thereby recovering analytic invertibility under ideal expectation-value assumptions. 2) A *multi-split* coupling architecture that introduces auxiliary assistant pathways to stabilize gradient flow and mitigate the barren plateaus common in hybrid training. Theoretical analysis reveals that deep quantum circuits in this context converge to truncated fourier series, offering a pathway for quantum-inspired classical transforms. Experimental validation on tabular benchmarks and image generation tasks demonstrates that optimized multi-split configurations achieve performance competitive with classical baselines on image density modeling, while remaining less effective on tabular benchmarks—highlighting both the promise and current limitations of the proposed transform.**

***Index Terms*—Quantum Neural Networks, Quantum Machine Learning, Generative AI Models, Normalizing Flows, Hybrid Quantum–Classical Models, Variational Quantum Circuits**

I. Introduction

Generative modeling remains a cornerstone of unsupervised learning, with normalizing flows [1] distinguishing themselves through their ability to perform exact log-likelihood estimation and efficient sampling. Unlike generative adversarial networks (GANs) [2], which require adversarial training and suffer from mode collapse [3], or diffusion models [4], which need hundreds of denoising steps, flows rely on a strict mathematical requirement: the mapping from data space to latent space must be a *diffeomorphism*—a mapping that is both invertible and differentiable [5]. This enables single-pass generation and exact density evaluation, making flows particularly suited for applications requiring likelihood-based model comparison or outlier detection [1]. Zhai *et al.* [6] demonstrated that when combined with modern architectures like transformers, normalizing flows can achieve generation quality comparable to diffusion models while retaining their computational advantages.

Simultaneously, quantum machine learning (QML) is a promising paradigm for enhancing classical models [7]. Variational quantum circuits can represent functions in exponentially large Hilbert spaces, theoretically offering expressivity advantages over classical networks of comparable parameter count [8]. The unitary nature of quantum evolution implies inherent reversibility, suggesting quantum circuits as natural candidates for flow-based generative modeling.

However, a fundamental incompatibility arises at the quantum–classical interface. While unitary evolution is reversible ($U^{-1} = U^{\dagger}$), quantum measurement is a projection that collapses the wavefunction. The mapping from a quantum state $|\psi\rangle$ to a classical expectation value $\langle M\rangle = \langle\psi|M|\psi\rangle$ is inherently many-to-one—infinitely many states can yield identical measurements. This loss of information breaks the bijectivity required by diffeomorphic flows, rendering exact likelihood computation impossible. We refer to this incompatibility as the *Measurement Paradox*.

We hypothesize that it is possible to construct a hybrid quantum-classical architecture that strictly preserves analytic invertibility without relying on approximation bounds. Operating under the standard expectation-value abstraction (assuming infinite measurement shots and noiseless execution), this is achieved by analyzing a common variational ansatz and showing that the resulting measurement-induced mapping admits a closed-form inverse and Jacobian. Separately, to address optimization challenges in hybrid training, we incorporate auxiliary classical pathways that help maintain stable gradient flow. In this work, our contributions include

1) **The Cosine Chain Protocol:** A derived measurement formulation and ansatz structure designed to mitigate the measurement paradox by enabling exact analytical reconstruction of input states post-measurement. By mapping quantum evolution to invertible cosine-based transformations, we recover the diffeomorphic structure required for normalizing flows.
2) **Multi-Split Architecture:** A coupling and autoregressive topology designed to mitigate the optimization difficulties inherent to QNNs. By introducing assistant splits, we create a stable classical gradient highway

that bypasses potential barren plateaus. This allows the model to converge reliably even when the quantum component struggles with optimization landscapes, effectively balancing quantum expressivity with classical trainability.

We emphasize that our goal is not to claim near-term hardware advantage, but rather to identify ansatz structures whose measurement maps are analytically tractable. Such constructions expand the design space of quantum and quantum-inspired invertible models and may become increasingly relevant as fault-tolerant quantum devices mature.

II. Related Work

The intersection of quantum computing and generative modeling has produced several distinct approaches, primarily categorized into quantum-native models and hybrid architectures. We contextualize our work against these existing paradigms to highlight the distinctiveness of the MSQCNF framework.

A. Quantum Generative Models

Early quantum generative efforts focused on quantum circuit born machines (QCBMs) [7] and quantum GANs (QGANs) [9]. While QCBMs exploit the probabilistic nature of quantum measurement directly, they lack tractable likelihood evaluation. QGANs, while expressive, inherit the training instabilities of classical GANs, such as mode collapse and sensitive hyperparameter tuning. These limitations motivate the exploration of Quantum Normalizing Flows, aiming for the stability and exact likelihood estimation characteristic of flow-based models.

B. The Challenge of Hybrid Flows

Constructing a normalizing flow on quantum hardware requires a unitary operator that maps a simple base distribution to the target data distribution.

Continuous Variable (CV) Approaches: Some works utilize continuous-variable quantum computing (photonic systems) to construct flows [10]. While theoretically elegant, these require specialized photonic hardware and do not translate directly to the qubit-based (gate-model) processors that currently dominate the NISQ landscape.

Qubit-Based Hybrid Flows: On gate-based devices, the primary challenge is the non-invertibility of measurement. Existing solutions have largely relied on approximation:

- *Approximate Reconstruction:* Rousselot and Spannowsky [11] propose a model that learns an Inverse State Preparation (ISP) routine. This functions similarly to a Variational Autoencoder (VAE), where the inverse is learned via a reconstruction loss. Unlike our approach, this forfeits the strict bijectivity guarantee of normalizing flows; the validity of the likelihood estimation depends entirely on how well the ISP network converges.
- *Implicit Quantum Invertibility Assumptions:* Zhang and Cui [12] integrate parameterized quantum circuits into normalizing flow layers and exploit their unitary nature to argue invertibility at the level of the quantum transformation. However, the overall flow necessarily includes classical-to-quantum encoding and quantum-to-classical decoding steps involving measurement or expectation-value readout, which are inherently many-to-one and thus non-invertible. These interfaces are not formally incorporated into the flow's bijection or Jacobian, leaving the invertibility of the full classical-to-classical mapping only implicitly assumed. As a result, the likelihood computation relies on an idealized treatment of the quantum block as a perfectly invertible operator, rather than a physically realizable quantum channel.

In contrast, our *Cosine Chain Protocol* provides a mathematically rigorous, exact analytical inverse for the quantum measurement process on gate-based devices. This exactness guarantees the strict diffeomorphism required for exact likelihood computation, resolving the paradox more fundamentally than relying on reconstruction error bounds.

C. Gradient Stabilization in Quantum Architectures

The training of parameterized quantum circuits is plagued by Barren Plateaus [13], where gradients vanish exponentially with qubit count. Standard mitigation strategies involve shallow circuits or specific initialization strategies [14].

Our work introduces a structural mitigation via a multi-split architecture. Unlike standard coupling layers (e.g., RealNVP [15]) which typically split features 50/50, or the 3-split hybrid attempt by Zhang and Cui [12], we systematically analyze classical pathways. These pathways ensure that even if the quantum gradient contribution vanishes, the classical gradient signal propagates, preventing the training stagnation observed in pure quantum flows.

III. Background

Notation: Scalars are denoted by non-bold symbols (e.g., x, w, α), vectors by bold lowercase letters (e.g., $\mathbf{x}, \mathbf{w}$), and matrices/operators by bold uppercase letters (e.g., $\mathbf{J}, \mathbf{U}$). Random variables are written in uppercase (e.g., $\mathbf{X}, \mathbf{Z}$) and their realizations in lowercase (e.g., $\mathbf{x}, \mathbf{z}$). The i-th component of a vector $\mathbf{x}$ is denoted by x_i.

For a scalar a, $|a|$ denotes absolute value. For a square matrix $\mathbf{A}$, $|\det(\mathbf{A})|$ denotes the magnitude of its determinant.

A. Normalizing Flows

Normalizing Flows are a class of generative models that learn complex probability distributions through a sequence of invertible transformations [1]. Given a simple base distribution $p_{\mathbf{Z}}(\mathbf{z})$ (typically $\mathcal{N}(0, I)$), a flow constructs a bijective mapping $f = f_K \circ \cdots \circ f_1$ to model $p_x(x)$. By the change of variables formula:

$$\log p_{\mathbf{X}}(\mathbf{x}) = \log p_{\mathbf{Z}}(\mathbf{z}) + \sum_{i=1}^{K} \log |\det(\mathbf{J}_i)| \tag{1}$$

where $\mathbf{z} = f^{-1}(\mathbf{x})$ and $\mathbf{J}_i = \frac{\partial f_i}{\partial \mathbf{z}_{i-1}}$ is the Jacobian of the i-th layer.

Coupling Layers: To ensure tractable Jacobian computation, architectures like RealNVP [15] employ coupling layers that partition input $\mathbf{x}$ into two halves $(\mathbf{x}_{1:d}, \mathbf{x}_{d+1:D})$:

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d} \tag{2}$$

$$\mathbf{y}_{d+1:D} = \mathbf{x}_{d+1:D} \odot \exp\big(s(\mathbf{x}_{1:d})\big) + t(\mathbf{x}_{1:d}) \tag{3}$$

where $s(\cdot)$ and $t(\cdot)$ are arbitrary neural networks. This yields a triangular Jacobian with $\mathcal{O}(D)$ determinant computation. Neural Spline Flows [16] extend this with monotonic rational-quadratic splines, providing greater expressivity per parameter.

B. Quantum Computing Preliminaries

A qubit exists in a superposition $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Computation proceeds via unitary operators U ($U^\dagger U = I$). We utilize standard parameterized rotation gates $R_P(\theta) = e^{-i\frac{\theta}{2}P}$ for $P \in \{X, Y, Z\}$, and entanglement via CNOT gates. Classical data $\mathbf{x}$ is embedded via *Angle encoding*, where x_i parameterizes rotation gates, or *Amplitude encoding* into state amplitudes.

C. Variational Quantum Circuits

Variational quantum circuits (VQCs), also called parameterized quantum circuits (PQCs), form the backbone of near-term quantum machine learning [7]. A VQC consists of:

1) **Encoding layer:** Maps a classical input vector $\mathbf{x}$ to a quantum state $|\psi_{\mathbf{x}}\rangle$ in Hilbert space via a feature embedding circuit
2) **Variational ansatz:** Applies a sequence of parameterized unitary operations $U(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the set of trainable parameters
3) **Measurement:** Extracts classical information by evaluating the expectation value $\langle M\rangle = \langle\psi_{\mathbf{x}}|U^\dagger(\boldsymbol{\theta})MU(\boldsymbol{\theta})|\psi_{\mathbf{x}}\rangle$ of an observable M

The ansatz structure—comprising rotation gates and entangling operations—determines circuit expressibility. Training proceeds via classical optimization of $\boldsymbol{\theta}$ using gradient-based or gradient-free methods. When gradients are employed, they are often computed using techniques such as the parameter-shift rule or adjoint differentiation.

IV. Methodology

A. Analytic Invertibility: The Cosine Chain

To resolve the measurement paradox, we derive a specific class of quantum ansatzes that are analytically invertible post measurement. We analyze a serial chain of n qubits, as illustrated in Fig. 1.

Circuit Structure: We adopt a hardware-efficient ansatz [17], a widely-used VQC structure suitable for near-term quantum devices. Each qubit i is initialized to $|0\rangle$. Classical input x_i is encoded via angle encoding using $R_X(x_i)$ rotations. The variational ansatz applies trainable rotations $R_Y(w_i)$ parameterized by weights $\mathbf{w}$, with CNOT gates providing entanglement between adjacent qubits. The observable for each qubit is the Pauli-Z operator, yielding expectation values $y_i = \langle Z_i\rangle \in [-1, 1]$.

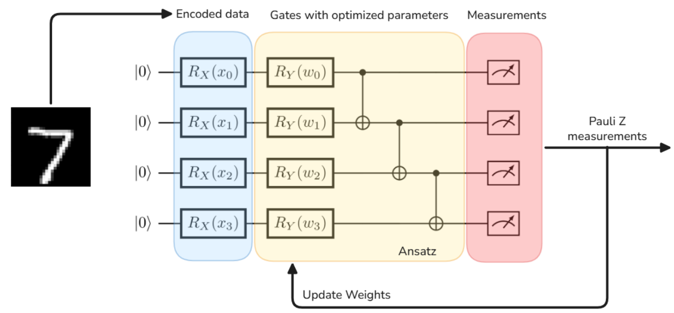

Fig. 1. Parameterized quantum circuit structure for the Cosine Chain. Data is encoded via $R_X(x_i)$ rotations (left), followed by variational layers with trainable $R_Y(w_i)$ gates and CNOT entanglers (center). Pauli Z measurements (right) yield expectation values which are used to update the weights for the $R_Y(w_i)$ gates.

Derivation: We derive the closed-form input–output relation using the Heisenberg picture. The full circuit unitary is written as a composition of an encoding layer, a variational rotation layer, and an entangling layer:

$$\mathbf{U} = \mathbf{U}_{\mathrm{CNOT}}\,\mathbf{U}_Y(\mathbf{w})\,\mathbf{U}_X(\mathbf{x}). \tag{4}$$

The data-encoding layer applies independent rotations on each qubit:

$$\mathbf{U}_X(\mathbf{x}) = \bigotimes_{i=0}^{n-1} R_X^{(i)}(x_i), \tag{5}$$

and the variational layer is

$$\mathbf{U}_Y(\mathbf{w}) = \bigotimes_{i=0}^{n-1} R_Y^{(i)}(w_i). \tag{6}$$

The entangling layer consists of a nearest-neighbor CNOT chain:

$$\mathbf{U}_{\mathrm{CNOT}} = \prod_{i=0}^{n-2} \mathrm{CNOT}_{i,i+1}, \tag{7}$$

where $\mathrm{CNOT}_{i,i+1}$ denotes a CNOT gate with control qubit i and target qubit $i+1$, applied sequentially from $i = 0$ to $n-2$.

We define the measurement outcome on qubit i as

$$y_i = \langle 0^{\otimes n}|\mathbf{U}^\dagger Z_i \mathbf{U}|0^{\otimes n}\rangle. \tag{8}$$

a) Single-qubit rotations: Using the Pauli conjugation identities,

$$R_X^\dagger(x)\,Z\,R_X(x) = Z\cos x + Y\sin x, \tag{9}$$

$$R_Y^\dagger(w)\,Z\,R_Y(w) = Z\cos w + X\sin w, \tag{10}$$

$$R_Y^\dagger(w)\,Y\,R_Y(w) = Y, \tag{11}$$

we apply the rotations sequentially.

After conjugation by $R_X(x_i)$ using (9), the observable becomes

$$Z\cos x_i + Y\sin x_i. \tag{12}$$

Upon subsequent conjugation with $R_Y(w_i)$ using (10) and (11), the observable becomes

$$(Z\cos w_i + X\sin w_i)\cos x_i + Y\sin x_i. \tag{13}$$

Taking the expectation on $|0\rangle$ in (13), and using $\langle 0|X|0\rangle = \langle 0|Y|0\rangle = 0$, only the Z component contributes, yielding

$$\langle Z_i\rangle = \cos(w_i)\cos(x_i). \tag{14}$$

b) Effect of the CNOT chain: For a CNOT gate with control c and target t,

$$\mathrm{CNOT}^\dagger(Z_t)\mathrm{CNOT} = Z_c Z_t, \qquad \mathrm{CNOT}^\dagger(Z_c)\mathrm{CNOT} = Z_c. \tag{15}$$

Propagating Z_i backwards through the serial CNOT chain using (15) yields

$$Z_i \longrightarrow \prod_{k=0}^{i} Z_k. \tag{16}$$

Substituting (16) into the measurement definition (8) gives

$$y_i = \Big\langle \prod_{k=0}^{i} Z_k \Big\rangle = \prod_{k=0}^{i} \langle Z_k \rangle. \tag{17}$$

c) Final expression: Substituting the single-qubit expectation (14) into (17) gives

$$y_i = \prod_{k=0}^{i} \cos(w_k)\cos(x_k), \tag{18}$$

or equivalently,

$$y_0 = \cos(w_0)\cos(x_0), \tag{19}$$

$$y_i = \cos(w_i)\cos(x_i)\, y_{i-1}, \qquad i > 0. \tag{20}$$

d) Analytic Invertibility: The serial transformation is analytically invertible. From (20), we can recover $\mathbf{x}$ from $\mathbf{y}$ recursively:

$$x_0 = \arccos\left(\frac{y_0}{\cos(w_0)}\right), \tag{21}$$

$$x_i = \arccos\left(\frac{y_i}{\cos(w_i)\, y_{i-1}}\right), \quad i > 0. \tag{22}$$

This formulation allows us to compute the exact Jacobian determinant required for the flow loss function. The determinant is tractable due to the recursive structure, which induces a lower-triangular Jacobian:

$$|\det(\mathbf{J})| = \left(\prod_{i=0}^{n-1} \sin(x_i)\right) \prod_{i=0}^{n-1} (\cos w_i)^{n-i} (\cos x_i)^{n-1-i}. \tag{23}$$

The analytic invertibility derived above is established under the standard expectation-value abstraction commonly used in variational quantum algorithm analysis, corresponding to the infinite-shot limit. In practical hardware settings, finite-shot sampling and device noise introduce stochasticity that may perturb the inverse mapping and log-determinant computation. Our objective in this work is algorithmic: to demonstrate that measurement-induced non-invertibility is not fundamentally unavoidable, but instead depends on ansatz design. Evaluating robustness under realistic noise models is an important direction for future work.

Generalization to Other Entanglement Topologies: The serial CNOT chain determines the multiplicative coupling structure in the above equations. Alternative entanglement patterns (e.g., circular topologies) modify which qubit outputs are coupled, but the resulting expressions remain products of trigonometric terms derived from Pauli conjugation. For example, a circular topology where qubit n connects back to qubit 0 introduces additional multiplicative dependencies between endpoints.

The analytic invertibility derived above is preserved as long as the effective dependency graph remains acyclic during the forward pass, allowing the ansatz to be adapted to hardware connectivity constraints.

A cyclic chain can still be invertible when used within a coupling-layer architecture. However, such cyclic entanglement is incompatible with an autoregressive formulation, since sequential inversion would require parameters associated with yet-unreconstructed dimensions. This violates the autoregressive inversion order and prevents exact reconstruction.

Practical Considerations: The cosine chain derivation reveals several important implications:

- *Vanishing signals*: Since $|\cos(\cdot)| \leq 1$, the forward transform produces values that diminish as the length of the cosine chain increases, causing vanishing gradients. We mitigate this by limiting the chain depth through blockwise resets every k qubits.
- *Volume contraction of the transformation*: From the closed-form Jacobian, it follows that $|\det(\mathbf{J})| < 1$, since $|\sin(\cdot)| \leq 1$ and $|\cos(\cdot)| \leq 1$. Thus, the cosine chain implements a strictly volume-contracting mapping in input space. In the context of normalizing flows, consistently contractive mappings bias the model toward density concentration and limit its ability to represent distributions that require local volume expansion [1], [15]. Although this contraction can be advantageous for modeling sharply peaked distributions, it restricts expressivity and may require stacking with complementary expansive transformations to achieve flexible density modeling.
- *Sequential dependency*: The recursive structure $y_i = f(y_{i-1})$ creates inherent sequential dependencies, limiting GPU parallelization. While classical simulation suffers from this bottleneck, execution on actual quantum hardware would compute all qubit measurements simultaneously.
- *Numerical conditioning*: Ill-conditioning may arise when $\cos(w_i)$, $\cos(x_i)$ or $\sin(x_i)$ approaches zero. We mitigate this by constraining rotation parameters away from degenerate regions via tanh reparameterization, ensuring stable Jacobian values throughout training.

B. Multi-Split Coupling Architecture

Motivation: While the cosine chain provides analytical invertibility, preliminary experiments showed that purely quantum flows struggle to model complex distributions effectively. As shown in Table I, a flow using only quantum transformations achieves significantly worse likelihood than classical baselines. We attribute this gap to two practical factors: (1) vanishing signals that reduce effective representational capacity, and (2) slower convergence during training. These observations are consistent with prior work [12], which proposed augmenting quantum transforms with classical assistant pathways.

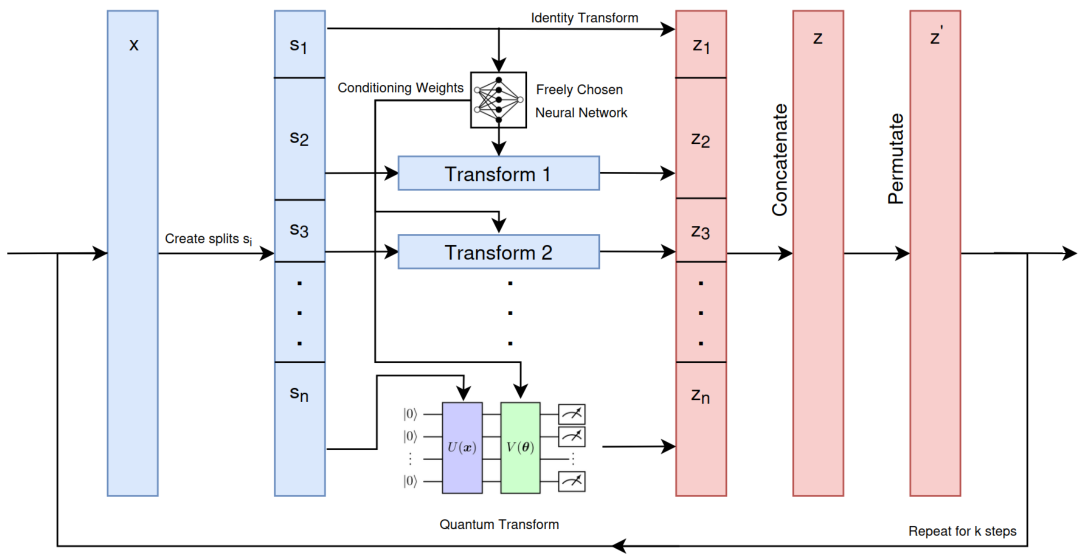

Fig. 2. Transformation layer of the multi-split quantum–classical normalizing flow (coupling). The input features are first partitioned into multiple splits (left). One split follows an identity path and serves as a shared conditioning context for a classical neural network, which generates parameters for the remaining transformation modules. These modules may be classical or quantum; the quantum block implements the cosine-chain or Fourier-inspired transform. The transformed outputs are concatenated, permuted, and the procedure is repeated for k layers (right)

Beyond Three Splits: Inspired by the 3-split architecture of [12], we investigate how the number of splits affects model capacity. Rather than adopting the 3-split design directly, we systematically evaluate multi-split topologies with $k \in \{3, 4, 5, 6, 7\}$ splits, varying both the number and types of transforms applied to each partition.

Coupling Architecture: We generalize the standard coupling layer by partitioning the input vector $\mathbf{x} \in \mathbb{R}^d$ into $k+1$ disjoint feature groups (see Fig. 2):

$$\mathbf{x} = (\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}).$$

The first group $\mathbf{x}^{(0)}$ is left unchanged and acts as a shared conditioning context, while each of the remaining groups is transformed in parallel, each conditioned on $\mathbf{x}^{(0)}$:

$$\mathbf{y}^{(0)} = \mathbf{x}^{(0)}, \tag{24}$$

$$\mathbf{y}^{(j)} = T_j\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(0)}\right), \qquad j = 1, \dots, k. \tag{25}$$

Here, each T_j denotes an invertible transformation module, which may be either classical (e.g., rational–quadratic spline) or quantum (e.g., cosine chain). The assignment of features to groups and the choice of transformation type are fixed per layer and define the multi-split topology.

Since each transformed group depends only on its own inputs and the shared conditioning group $\mathbf{x}^{(0)}$, the transformations can be evaluated in parallel. Treating the conditioning inputs as fixed, the Jacobian is block-triangular, with determinant

$$\det(\mathbf{J}) = \prod_{j=1}^{k} \det\left(\frac{\partial \mathbf{y}^{(j)}}{\partial \mathbf{x}^{(j)}}\right).$$

Autoregressive Architecture: We also extend masked autoregressive flows (MAF) [18] to support heterogeneous classical–quantum transformations.

Autoregressive flows impose an ordering over input dimensions and transform each feature sequentially conditioned on all preceding ones. For an ordering π,

$$y_{\pi(i)} = T_{\pi(i)}\left(x_{\pi(i)} \mid x_{\pi(1)}, \dots, x_{\pi(i-1)}\right), \qquad i = 1, \dots, d, \tag{26}$$

where each $T_{\pi(i)}$ is invertible.

Each $T_{\pi(i)}$ may be classical (e.g., RQ-spline) or quantum. The assignment defines the *architecture pattern*. We consider four patterns:

- *Cyclic:* classical and quantum transforms alternate along the ordering,
- *Random:* the number of classical and quantum transforms is fixed, but their positions in the ordering are assigned randomly,
- *Quantum-first:* early dimensions use quantum transforms,
- *Quantum-second:* later dimensions use quantum transforms.

Later dimensions receive richer conditioning context and therefore allow more context-dependent transformations, making module placement an important design choice.

C. Theoretical Connection: Trigonometric Expansion and Fourier Structure

The analytic invertibility derived in Section IV-A holds strictly for a *single* ansatz layer. When stacking multiple layers—each consisting of R_Y rotations and CNOT chains—the resulting mapping becomes a composition of trigonometric functions and products. Unlike the single-layer case, these compositions are generally oscillatory and non-monotone. Since trigonometric functions are periodic and non-injective, multiple distinct inputs can produce identical expectation values. Consequently, global invertibility of the post-measurement mapping $X \mapsto Y$ cannot be guaranteed for deep stacks.

However, it is crucial to clarify that in our multi-split architecture, the Cosine Chain is utilized as a single transformation block within a coupling layer. Because invertibility is analytically guaranteed within this isolated block, the sequence of multiple coupling layers remains globally invertible by construction, successfully circumventing the non-injectivity of isolated deep quantum stacks.

Despite the isolated depth limitation, the functional form of deeper compositions exhibits useful structure. Conjugation by rotation gates produces linear combinations of Pauli operators whose coefficients are products of $\sin(\cdot)$ and $\cos(\cdot)$ terms. Repeated application therefore yields expectation values that are sums of products of trigonometric functions of the inputs. Using standard product-to-sum identities, such expressions can be rewritten as trigonometric polynomials.

For a single input dimension x, the output admits a Fourier-series representation of the form

$$f(x) = a_0 + \sum_{m=1}^{M} \left(a_m \cos(mx) + b_m \sin(mx)\right), \tag{27}$$

where the highest frequency M increases with circuit depth and the number of data re-uploadings. This follows from the Fourier characterization of parameterized quantum circuits, where repeated trigonometric interactions generate higher-frequency components through product-to-sum identities [19], [20].

More generally, for multi-dimensional inputs $X \in \mathbb{R}^d$, each measured expectation value can be expressed as

$$y_i(X) = \sum_{m \in \mathcal{F}_i} \left(c_m \cos(m^\top X) + d_m \sin(m^\top X)\right), \tag{28}$$

where $\mathcal{F}_i \subset \mathbb{Z}^d$ is a finite set of frequency vectors determined by the circuit structure and parameters.

Thus, while deep cosine-chain circuits are not bijective in general, they exhibit a clear *spectral inductive bias*: increasing depth enlarges the set of accessible frequency components. In effect, deeper circuits behave as structured trigonometric feature generators rather than invertible maps.

This observation motivates separating two roles of the quantum ansatz:

- **Single-layer:** an analytically invertible building block suitable for flows,
- **Multi-layer:** a nonlinear Fourier feature generator that improves expressivity at the cost of invertibility.

This distinction motivates the design of an alternative transformation that preserves invertibility while explicitly capturing the additive spectral structure implied by the quantum circuit.

D. Fourier-Inspired Quantum Transform

Motivated by the Fourier convergence observation, we designed a simplified "Fourier" transform that directly implements an additive cosine structure:

$$Y[0] = W_0 \cdot \cos(X_0) \tag{29}$$

$$Y[i] = W_i \cdot \cos(X_i) + Y[i-1] \quad \text{for } i > 0 \tag{30}$$

where W_i are learnable weights. This additive (rather than multiplicative) structure avoids the vanishing signal problem of the cosine chain. To prevent unbounded growth, we reset the cumulative sum every k features (block size), yielding a blocked Fourier series. The inverse is:

$$X_0 = \arccos(Y_0/W_0) \tag{31}$$

$$X_i = \arccos((Y_i - Y_{i-1})/W_i) \quad \text{for } i > 0 \tag{32}$$

This quantum-inspired transform is completely classical and hardware-independent, yet retains a similar spectral inductive bias to that observed in actual quantum circuits. This bridges the gap between quantum theory and classical modeling, justifying the value of translating quantum structures into robust classical forms.

V. Experimental Results

We base our implementation on the reference implementation of neural spline flows [16] and retain their default configuration where not specified below. The following paragraphs describe our architectural choices, quantum transformation design, and practical safeguards introduced to preserve analytic invertibility in the hybrid setting. Unless otherwise stated, baseline results are taken directly from the literature and are reported for comparison under identical evaluation protocols.

Flow architecture: All models use 10 flow layers. Each layer is followed by an invertible linear transform parameterized via LU decomposition. For rational–quadratic spline transforms (as in Neural Spline Flows), we use 4 bins and a tail bound of $[-3, 3]$ with linear tails outside this range. The conditioner networks are 2-block ResNets with 64 hidden units per block, ReLU activations and dropout probability 0.2. Masks in coupling layers are cyclically rotated across layers to ensure that all features are transformed over the stack. For autoregressive transforms, we evaluate different masking patterns (Cyclic, Random, Quantum-first, and Quantum-second).

Invertibility Safeguards: The analytic Jacobian in (23) requires non-zero trigonometric factors. To prevent singularities, we strictly restrict data angles x_i and weights w_i to the interval $(\alpha, \pi/2 - \alpha)$ via a $\tanh$ reparameterization:

$$\mathbf{v} = \left(\tfrac{\pi}{4} - \alpha\right)\tanh(\tilde{\mathbf{v}}) + \tfrac{\pi}{4}, \quad \text{for } \mathbf{v} \in \{\mathbf{x}, \mathbf{w}\}, \tag{33}$$

where $\alpha = 10^{-3}$. This guarantees non-degenerate Jacobians. Furthermore, we implement two stability measures: 1) **Depth Blocking:** To prevent exponential signal attenuation, we reset the product accumulator every $k = 4$ qubits; and 2) **Range Adaptation:** We map quantum outputs $\mathbf{y} \in [0, 1]$ to $[-3, 3]$ using a learned affine transform followed by a scaled $\tanh$, ensuring compatibility with subsequent classical flow layers.

A. Datasets and Evaluation Protocol

Tabular datasets: We evaluate density estimation performance on five standard benchmarks: *POWER*, *GAS*, *MINIBOONE*, *HEPMASS*, and *BSDS300*. These datasets and preprocessing protocols follow the setup introduced in Masked Autoregressive Flow (MAF) [18] and subsequently adopted in modern flow-based density estimation works [1], [16].

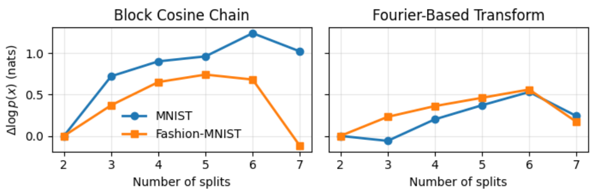

Fig. 3. Effect of the number of splits on density estimation performance. Results are shown as improvements in $\log p(x)$ relative to the 2-split baseline for MNIST and Fashion-MNIST. Increasing the number of splits consistently improves performance up to a moderate depth, after which gains degrade.

Performance is reported in terms of test log-likelihood in Table I.

Image datasets: We further evaluate on MNIST, Fashion-MNIST. We adopt the same VAE+flow evaluation protocol used in Neural Spline Flows [16], enabling direct comparison. We report the evidence lower bound (ELBO) and importance-weighted estimates of $\log p(x)$, with error bars corresponding to two standard deviations.

All results are averaged over multiple runs. Unless otherwise stated, baseline results are taken directly from the literature—primarily Neural Spline Flows [16] and Masked Autoregressive Flow [18]—and are reported for comparison under identical evaluation protocols.

B. Pure Quantum Flow Performance

We first evaluate flows constructed purely from cosine-chain quantum transforms (both coupling and autoregressive). The goal of this experiment is to isolate the representational capacity of the quantum transform alone. Results across tabular and image datasets consistently show that pure cosine-chain flows underperform classical baselines (Table I). While analytically invertible, their multiplicative structure leads to signal attenuation and limited effective capacity. This confirms that analytic invertibility alone is insufficient for competitive density modeling.

C. Effect of Multi-Split Architectures

We next study the effect of increasing the number of splits in the multi-split architecture. We evaluate $k \in \{2, 3, 4, 5, 6, 7\}$ splits using cosine-chain transforms.

Across MNIST and Fashion-MNIST, increasing the number of splits consistently improves ELBO and log-likelihood up to a moderate number of splits (Figure 3). Beyond this point, performance degrades, likely due to reduced conditioning context per split. These results support the hypothesis that assistant classical pathways stabilize training and compensate for quantum gradient degradation.

D. Effect of Transform Depth

We further analyze the effect of increasing the depth of the cosine-chain and Fourier transforms within each hybrid layer. For both transforms, we observe a consistent trend: performance improves as the chain depth increases, reaches an optimal regime, and then degrades beyond that point. Empirically, we identify a depth of 4 as the optimal configuration.

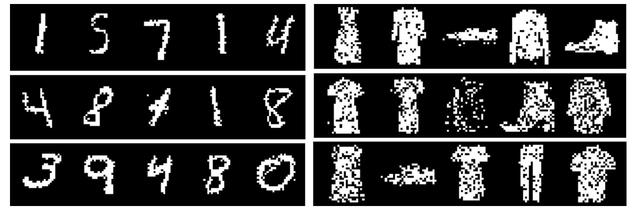

Fig. 4. Qualitative samples from MNIST and Fashion-MNIST. Each block shows real data (top), samples from the classical RQ-NSF baseline (middle), and samples from our Fourier multi-split model (bottom).

This behavior can be attributed to competing factors. Greater depth increases representational capacity by enabling richer trigonometric interactions, but simultaneously amplifies signal attenuation (in cosine chains) or cumulative bias (in additive Fourier blocks), ultimately harming optimization.

Throughout these experiments, the split sizes assigned to each transformation were kept constant to isolate the effect of depth alone. These findings suggest that moderate transform depth provides the best balance between expressivity and numerical stability.

E. Fourier-Based Quantum Transform

Replacing the multiplicative cosine chain with the additive Fourier-inspired transform yields consistent improvements across datasets (Table I and II). The additive structure alleviates vanishing signals while preserving the spectral inductive bias suggested by the theoretical analysis. Notably, Fourier-based hybrids close much of the gap to classical RQ-NSF baselines while maintaining analytic invertibility.

F. Qualitative Results

Figure 4 shows samples generated from MNIST and Fashion-MNIST. Hybrid models produce visually sharp and diverse samples comparable to classical spline flows.

VI. Conclusion and Future Work

We introduced the multi-split quantum-classical normalizing flow (MSQCNF), resolving the measurement paradox via structured, analytically invertible transforms such as the Cosine Chain and the Fourier-inspired additive module. Crucially, operating under the expectation-value abstraction via classical simulation enables the realization of ideal, noiseless quantum transformations on classical hardware, allowing stable and tractable likelihood-based training. In this regime, our hybrid multi-split configurations stabilize optimization and achieve density modeling performance competitive with strong classical baselines on image tasks.

While these results demonstrate that quantum-inspired transforms can serve as powerful and analytically tractable components within modern flow architectures, deployment on physical Quantum Processing Units (QPUs) will introduce finite sampling (shot noise) and device-level noise. Understanding how these stochastic perturbations affect the stability of the inverse mapping and log-determinant computation remains an important open problem. Future work can therefore focus on evaluating robustness under realistic

TABLE I

TEST LOG-LIKELIHOOD (IN NATS) ON UCI DATASETS AND BSDS300. ERROR BARS CORRESPOND TO TWO STANDARD DEVIATIONS. MODELS MARKED WITH † REPORT RESULTS FROM PRIOR WORK, WHILE UNMARKED ROWS CORRESPOND TO OUR IMPLEMENTATIONS. HIGHER VALUES INDICATE BETTER DENSITY ESTIMATES.

Model	POWER ↑	GAS ↑	HEPMASS ↑	MINIBOONE ↑	BSDS300 ↑
FFJORD†	0.46 ± 0.01	8.59 ± 0.12	-14.92 ± 0.08	-10.43 ± 0.04	157.40 ± 0.19
NAF†	0.62 ± 0.01	11.96 ± 0.33	-15.09 ± 0.40	$-\mathbf{8.86} \pm 0.15$	$\mathbf{157.73} \pm 0.04$
Block-NAF†	0.61 ± 0.01	12.06 ± 0.09	-14.71 ± 0.38	-8.95 ± 0.07	157.36 ± 0.03
SOS†	0.60 ± 0.01	11.99 ± 0.41	-15.15 ± 0.10	-8.90 ± 0.11	157.48 ± 0.41
Glow†	0.42 ± 0.01	12.24 ± 0.03	-16.99 ± 0.02	-10.55 ± 0.45	156.95 ± 0.28
RQ-NSF (C)†	0.64 ± 0.01	$\mathbf{13.09} \pm 0.02$	-14.75 ± 0.03	-9.67 ± 0.47	157.54 ± 0.28
RQ-NSF (AR)†	$\mathbf{0.66} \pm 0.01$	$\mathbf{13.09} \pm 0.02$	$-\mathbf{14.01} \pm 0.03$	-9.22 ± 0.48	157.31 ± 0.28
Quantum-Based Models (Ours)					
Blocked Cosine Chain (C)	0.05 ± 0.01	$\mathbf{8.16} \pm 0.02$	-19.61 ± 0.02	-10.92 ± 0.46	152.08 ± 0.28
Blocked Cosine Chain (AR)	-0.21 ± 0.01	6.57 ± 0.02	-22.44 ± 0.02	-12.71 ± 0.46	149.58 ± 0.29
Cosine Multi-Split (C)	0.06 ± 0.01	7.34 ± 0.02	-19.94 ± 0.02	-11.36 ± 0.45	153.14 ± 0.28
Cosine Multi-Split (AR)	0.32 ± 0.01	7.96 ± 0.02	-19.07 ± 0.02	-10.23 ± 0.45	152.86 ± 0.28
Fourier Multi-Split (C)	0.24 ± 0.01	7.68 ± 0.02	-19.34 ± 0.02	-9.60 ± 0.44	$\mathbf{154.01} \pm 0.28$
Fourier Multi-Split (AR)	$\mathbf{0.43} \pm 0.01$	8.10 ± 0.02	$-\mathbf{18.99} \pm 0.02$	$-\mathbf{9.52} \pm 0.43$	153.48 ± 0.28

TABLE II

VARIATIONAL AUTOENCODER TEST-SET RESULTS (IN NATS) REPORTING THE EVIDENCE LOWER BOUND (ELBO) AND IMPORTANCE-WEIGHTED ESTIMATES OF $\log p(x)$. ERROR BARS CORRESPOND TO TWO STANDARD DEVIATIONS. MODELS MARKED WITH † REPORT RESULTS FROM PRIOR WORK. LARGER VALUES INDICATE BETTER PERFORMANCE.

	MNIST		Fashion-MNIST	
Model	ELBO ↑	$\log p(x)$ ↑	ELBO ↑	$\log p(x)$ ↑
RQ-NSF (C)†	$-\mathbf{82.08} \pm 0.46$	$-\mathbf{79.63} \pm 0.42$	-226.59 ± 1.52	-224.54 ± 1.51
RQ-NSF (AR)†	-82.14 ± 0.47	-79.71 ± 0.43	$-\mathbf{226.34} \pm 1.52$	$-\mathbf{224.47} \pm 1.51$
Quantum-Based Models (Ours)				
Cosine Chain (C)	-86.15 ± 0.47	-83.35 ± 0.44	-231.40 ± 1.52	-229.19 ± 1.52
Cosine Chain (AR)	-89.55 ± 0.47	-85.73 ± 0.44	-234.56 ± 1.52	-231.47 ± 1.52
Cosine Multi-Split (AR)	-82.12 ± 0.45	-79.94 ± 0.42	-227.76 ± 1.52	-225.93 ± 1.52
Fourier Multi-Split (AR)	$-\mathbf{82.07} \pm 0.45$	$-\mathbf{79.81} \pm 0.42$	$-\mathbf{227.60} \pm 1.52$	$-\mathbf{225.66} \pm 1.52$

hardware constraints, incorporating generative quality metrics such as Fréchet Inception Distance, exploring other ansatz architectures, and conducting systematic comparisons against purely classical trigonometric parameterizations and alternative generative paradigms such as diffusion models.

REFERENCES

[1] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *JMLR*, vol. 22, 2021.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[3] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[5] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[6] S. Zhai, R. Zhang, P. Nakkiran, D. Berthelot, J. Gu, H. Zheng, T. Chen, M. A. Bautista, N. Jaitly, and J. Susskind, "Normalizing flows are capable generative models," *arXiv preprint arXiv:2412.06329*, 2024.

[7] M. Benedetti, E. Bermejo, A. Acín, J. I. Latorre, and A. Perdomo-Ortiz, "Parameterized quantum circuits as machine learning models," *Quantum Sci. Technol.*, vol. 4, 2019.

[8] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning," *Nat. Commun.*, vol. 12, 2021.

[9] S. Lloyd and C. Weedbrook, "Quantum generative adversarial learning," *Phys. Rev. Lett.*, vol. 121, 2018.

[10] N. Killoran, T. R. Bromley, J. M. Arrazola, M. Schuld, N. Quesada, and S. Lloyd, "Continuous-variable quantum neural networks," *Phys. Rev. Research*, vol. 1, 2019.

[11] A. Rousselot and M. Spannowsky, "Generative invertible quantum neural networks," *SciPost Phys.*, vol. 16, 2024.

[12] A. Zhang and W. Cui, "Hybrid quantum-classical normalizing flow," *arXiv preprint arXiv:2405.13808*, 2024.

[13] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nat. Commun.*, vol. 9, 2018.

[14] E. Grant, L. Wossnig, M. Ostaszewski, and M. Benedetti, "An initialization strategy for addressing barren plateaus," *Quantum*, vol. 3, 2019.

[15] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," in *ICLR*, 2017.

[16] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "Neural spline flows," in *NeurIPS*, 2019.

[17] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, "Hardware-efficient variational quantum eigensolver," *Nature*, vol. 549, 2017.

[18] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," in *NeurIPS*, 2017.

[19] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum models," *Phys. Rev. A*, vol. 103, 2021.

[20] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, "Data re-uploading for a universal quantum classifier," *Quantum*, vol. 4, 2020.