Lane-Frame Quantum Multimodal Driving Forecasts for the Trajectory of Autonomous Vehicles

Navneet Singh *Student Member, IEEE* and Shiva Raj Pokhrel *Senior Member, IEEE*

***Abstract*—Trajectory forecasting for autonomous driving must deliver accurate, calibrated multi-modal futures under tight compute and latency constraints. We propose a compact hybrid quantum architecture that aligns quantum inductive bias with road-scene structure by operating in an ego-centric, lane-aligned frame and predicting residual corrections to a kinematic baseline instead of absolute poses. The model combines a transformer-inspired quantum attention encoder (9 qubits), a parameter-lean quantum feedforward stack (64 layers, $\sim$1200 trainable angles), and a Fourier-based decoder that uses shallow entanglement and phase superposition to generate 16 trajectory hypotheses in a single pass, with mode confidences derived from the latent spectrum. All circuit parameters are trained with Simultaneous Perturbation Stochastic Approximation (SPSA), avoiding back-propagation through non-analytic components. In the Waymo Open Motion Dataset, the model achieves minADE (minimum Average Displacement Error) of $1.94\,\mathrm{m}$ and minFDE (minimum Final Displacement Error) of $3.56\,\mathrm{m}$ in the 16 models predicted over the horizon of $2.0\,\mathrm{s}$, consistently outperforming a kinematic baseline with reduced miss rates and strong recall. Ablations confirm that residual learning in the lane frame, truncated Fourier decoding, shallow entanglement, and spectrum-based ranking focus capacity where it matters, yielding stable optimization and reliable multi-modal forecasts from small, shallow quantum circuits on a modern autonomous-driving benchmark.**

***Index Terms*—Autonomous driving, trajectory prediction, quantum machine learning, multi-modal forecasting, residual learning, Waymo Open Motion Dataset**

I. Introduction

ACCURATE short-horizon trajectory forecasting under uncertainty is a central requirement for autonomous driving. An effective forecaster must reason about multiple plausible futures (e.g., straight vs. turn), remain well-calibrated, and operate under strict latency and compute budgets. Classical deep models achieve high accuracy [1], [2] but often at the cost of large parameter counts, multi-pass decoding, and heavy training pipelines [3], [4]. This paper explores a different point in the design space: *can a compact, shallow hybrid quantum model, carefully aligned with the structure of near-term driving, produce useful multi-modal forecasts under tight resource constraints?*

Rather than aiming for state-of-the-art performance or formal quantum advantage, we present an *early, small-scale feasibility study* that examines the applicability of quantum machine learning to vehicle trajectory prediction. Our goal is to understand whether a modest, task-structured quantum circuit, with only a handful of qubits and shallow depth, can participate meaningfully in a modern forecasting pipeline when equipped with strong classical inductive biases.

We introduce a small, task-aligned hybrid quantum architecture that deliberately reduces what the model needs to represent as shown in Fig 1. Each scenario is first expressed in an ego-centric, lane-aligned coordinate frame and paired with a simple kinematic/map-following baseline [5]. The model then predicts only *residual* motion relative to this baseline, shifting learning from full trajectories to smooth, meter-scale corrections that match the statistics of short-horizon driving [6]. This choice contracts the output dynamic range, stabilizes optimization, and focuses capacity where it matters [7].

The model comprises three lightweight quantum components. A *quantum attention encoder* (on 9 qubits) maps query-key-value summaries of the recent motion into an entangled state, providing transformer-inspired mixing of current pose, historical trend, and short-term change using linear/ladder entanglement [8]. A *deep yet parameter-lean quantum feedforward stack* (64 layers, $\sim$1200 trainable angles overall) alternates data re-injection, local rotations, and ring entanglement, with measurements after each layer to form a stable hybrid loop [9]. Finally, a *Fourier-based quantum decoder* uses phase superposition to emit $M{=}16$ coherent trajectory hypotheses in a *single* forward pass; a fixed confidence mapping derived from the latent spectrum ranks these hypotheses without an additional learned head.

Design choices embed strong inductive biases. Residual learning in the lane frame confines the circuit to small, structured deviations from a map/CTRV prior. The truncated Fourier readout imposes a low-frequency smoothness bias appropriate for 2.0 s horizons, while global phase shifts generate diverse, geometrically consistent futures at constant inference cost in K. Shallow entanglement keeps the parameter space small and the Simultaneous Perturbation Stochastic Approximation (SPSA) landscape well-conditioned. Together, these constraints allow a 9-qubit circuit to behave like a larger model *in-distribution*, trading universal expressivity for targeted efficacy.

We train the system end-to-end with SPSA, which is well-suited to parameterized quantum circuits and to our non-differentiable 'min over modes' objective. On the Waymo Open Motion Dataset (WOMD) [3], using standard evaluation on a held-out subset, the model attains meter-scale displacement errors, broad hypothesis coverage, and stable calibration, despite the small qubit count, shallow depth, and fixed-basis decoder. Miss rates and recall/precision-recall trends improve consistently over training, and the optimization trajectory

Authors are with the IoT & Software Engineering Research Lab, School of Information Technology, Deakin University, Geelong, VIC 3125, Australia (e-mail: n.navneetsingh@deakin.edu.au; shiva.pokhrel@deakin.edu.au).

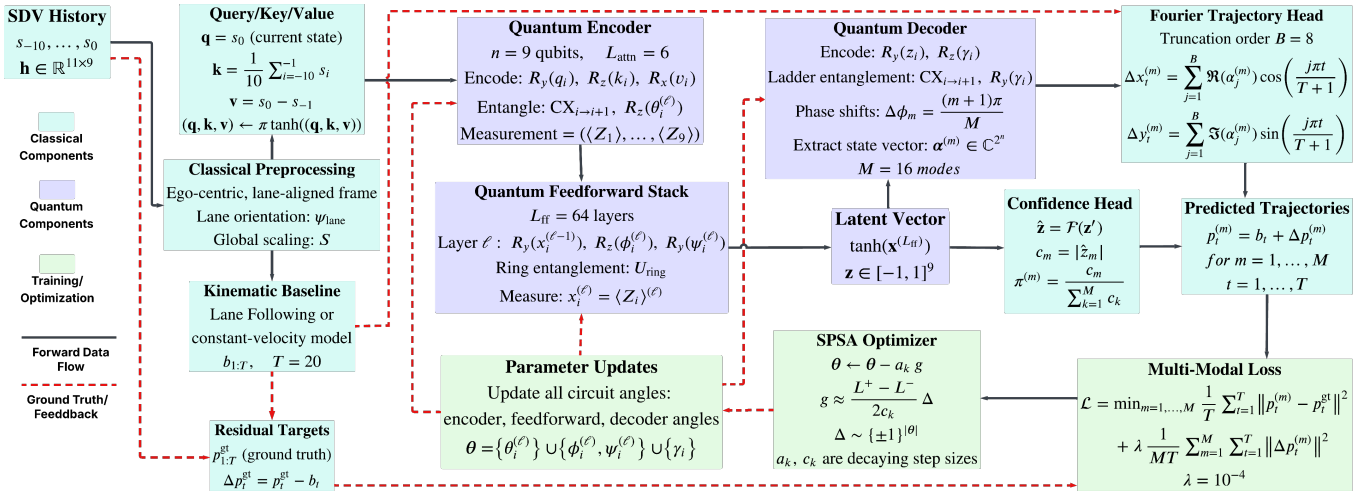


Fig. 1: Overview of the proposed quantum multi-modal trajectory prediction pipeline. SDV history is preprocessed into an ego-centric lane frame and split into a kinematic baseline and query-key-value features, which are encoded by a quantum attention encoder and a deep quantum feedforward stack to produce a latent vector. A quantum decoder, Fourier trajectory head, and confidence head generate multi-modal residual trajectories and mode probabilities that are added to the baseline to obtain final predictions. Solid arrows denote forward data flow, while dashed arrows show construction of residual targets from ground truth and the SPSA-based parameter updates used for training.

exhibits smooth loss reduction, shrinking variance, and a staircase of best-so-far improvements, indicating predictable convergence under SPSA. Compared to a strong kinematic baseline, the model substantially reduces displacement errors while maintaining single-pass multi-modality.

This work sits adjacent to, but distinct from, classical Transformers. Our encoder is *transformer-inspired* (attention-like mixing realized by quantum gates), yet the overall system is not a general-purpose *quantum Transformer*. Instead, it is a narrowly scoped, task-structured hybrid architecture that exploits superposition and interference to generate and rank multiple futures efficiently.

a) Key Contributions:

- *A compact, task-aligned hybrid quantum forecaster.* We combine a quantum attention encoder, a parameter-lean feedforward stack, and a Fourier decoder that leverages phase superposition to produce $M{=}16$ modes in one pass.
- *Residual learning in a lane-aligned frame.* Predicting corrections atop a map/CTRV baseline reduces dynamic range and centers learning on lane-consistent adjustments, improving stability and data efficiency.
- *Single-pass multi-modality with physics-shaped ranking.* Phase-offset decoding yields diverse, smooth hypotheses at constant cost, while a latent-spectrum confidence assigns meaningful ranks without an extra learned head.
- *Reliable gradient-free training for multi-modal objectives.* We demonstrate smooth convergence and shrinking variance when optimizing shallow circuits with SPSA under a non-smooth 'min-over-modes' loss.
- *An early feasibility study of quantum trajectory forecasting on WOMD.* Using only 9 qubits and shallow entanglement, the model achieves meter-scale accuracy on a representative WOMD subset, outperforming a kinematic baseline and illustrating that even very small quantum circuits can contribute to multi-modal vehicle trajectory prediction.

A. Literature

The literature on trajectory prediction is extensive, so we focus on work most closely aligned with our setting and design choices.

Early deep-learning approaches treat motion prediction as sequence modeling over agent states, using LSTMs for pedestrians and vehicles. Social-LSTM [10] explicitly encodes nearby agents via social pooling to forecast human trajectories in crowds, while Lin et al. use LSTMs with attention for human-driven vehicle longitudinal motion in connected environments [11]. Surveys on human and vehicle trajectory forecasting emphasize that realistic predictions must jointly account for agent interactions, map structure and uncertainty, especially for downstream motion planning [12], [13], [14]. Graph-based models such as Trajectron [15] and Trajectron++ [16] represent multi-agent scenes as dynamic spatiotemporal graphs and output probabilistic, multi-modal trajectories that respect vehicle dynamics and can incorporate heterogeneous inputs. Anchor- and goal-based approaches such as TNT [17] and MultiPath++ [18] further improve diversity and accuracy by predicting a compact set of endpoints or latent anchors together with associated trajectories on large-scale driving datasets, including Argoverse and WOMD. More recently, transformer architectures Scene Transformer [19], HiVT [20], Wayformer [21] and MotionLM [22] use global attention over agents, time and HD-map polylines to achieve state-of-the-art performance on WOMD and related benchmarks, but they rely on very deep networks with millions of

parameters and substantial inference latency, which complicates real-time deployment and large-scale calibration.

In parallel, hybrid data–model driven frameworks encode vehicle physics and driver heterogeneity within deep architectures. Bai et al. propose KO-TAFTN, which couples Kepler-based physical modeling of driver-specific preferences with temporal attention over agents and planned CAV trajectories to capture anticipatory interactions in mixed highway traffic [23]. Using variable selection and interpretable attention, KO-TAFTN produces diverse maneuver-level predictions and outperforms purely data-driven baselines, underscoring the value of structured priors in high-capacity sequence models.

Quantum neural networks (QNNs) [24], [25] based on parameterized quantum circuits (PQCs) have emerged as an alternative paradigm for sequence modeling, aiming to obtain high expressivity with relatively few trainable parameters. Chen et al. introduce Quantum Long Short-Term Memory (QLSTM), a hybrid model that embeds variational quantum circuits inside LSTM cells and shows competitive or faster convergence than classical LSTMs on temporal benchmarks with shallow circuits suitable for NISQ devices [26]. Takaki et al. propose a variational quantum recurrent neural network (VQRNN) that learns temporal dependencies via PQCs arranged in a recurrent structure and accurately models classical and quantum dynamical time series with a compact number of qubits and parameters [27].

Existing motion-forecasting systems for autonomous driving therefore fall broadly into: (i) large classical architectures (GNNs, transformers, and hybrid physics–data models such as KO-TAFTN) that achieve high accuracy on WOMD and highway benchmarks but incur significant computational and memory cost, and (ii) emerging quantum and hybrid quantum–classical models that show promise on generic or small-scale time-series tasks yet have not been evaluated on realistic, map-conditioned SDV trajectory prediction benchmarks. To our knowledge, there remains a gap in systematically assessing shallow, few-qubit PQC-based models on short-horizon SDV trajectory forecasting in WOMD, with an emphasis on calibrated multi-hypothesis outputs under tight compute budgets. Our work addresses this gap by formulating lane-aligned residual prediction as a compact quantum sequence modeling problem, designing a low-depth hybrid QNN over 9-dimensional SDV state histories, and empirically comparing its accuracy and calibration against classical baselines in a deployment-oriented setting.

II. Methodology

A. Data Preprocessing

We use the publicly available WOMD, which provides $\sim$100k driving scenarios (20 s at 10 Hz) with tracked trajectories and high-definition maps. Each scenario contains a designated self-driving vehicle (SDV, the ego) and other traffic agents. We focus on predicting the SDV's future trajectory from its past motion and map context. The complete preprocessing pipeline, including lane-aligned coordinates, the kinematic baseline $b_{1:T}$ and residual targets $\Delta p^{\mathrm{gt}}_{1:T}$, is summarized in Algorithm 1.

Algorithm 1: Construction of training examples and residual targets

Input: Waymo scenarios; map; horizon T=20; scale S
Output: triplets $(\mathbf{h}, b_{1:T}, \Delta p^{\mathrm{gt}}_{1:T})$
examples $\leftarrow$ empty list
foreach *scenario and valid SDV index* t_0 **do**
  $(s_{-10}, \ldots, s_0, p^{\mathrm{gt}}_{1:T}) \leftarrow$ ExtractStates(scenario, t_0)
  $\psi_{\text{lane}} \leftarrow$ LaneOrientation(s_0, map)
  $(s_{-10}, \ldots, s_0, p^{\mathrm{gt}}_{1:T}) \leftarrow$ ToLaneFrame$(s_{-10}, \ldots, s_0, p^{\mathrm{gt}}_{1:T}, \psi_{\text{lane}}, S)$
  $v \leftarrow$ Speed(s_0)
  $\omega \leftarrow$ YawRate(s_0)
  if $\|v\| \geq 0.05$ *and* LaneAvailable(*map*) **then**
    $b_{1:T} \leftarrow$ LaneFollowingBaseline(s_0, map, v, T)
  else
    $b_{1:T} \leftarrow$ KinematicBaseline(s_0, v, ω, T)
  for $t \leftarrow 1$ **to** T **do**
    $\Delta p^{\mathrm{gt}}_t \leftarrow p^{\mathrm{gt}}_t - b_t$
  $\mathbf{h} \leftarrow$ StackStates$(s_{-10}, \ldots, s_0)$
  append $(\mathbf{h}, b_{1:T}, \Delta p^{\mathrm{gt}}_{1:T})$ to examples
return *all stored* $(\mathbf{h}, b_{1:T}, \Delta p^{\mathrm{gt}}_{1:T})$

The SDV state is represented by 11 time steps (10 past at 0.1 s plus the current step at time t_0). Each time step has a 9-dimensional feature vector: position (x, y, z), ground-plane velocity (v_x, v_y), yaw rate ω, heading angle, and vehicle dimensions (length and width). We denote the history as $\{s_{-10}, \ldots, s_{-1}, s_0\}$, where s_0 is the current state. The prediction target is the SDV trajectory for the next $T = 20$ steps.

We first transform each scenario into an ego-centric, lane-aligned frame. The SDV's current position becomes the origin, and a reference yaw ψ_{lane} is computed from the map by averaging local road direction vectors within a radius around the SDV. Positions are rotated so that the x-axis aligns with this lane direction and translated so that $(0, 0)$ corresponds to the SDV at t_0. Velocities are transformed similarly.

Next, we construct a simple kinematic baseline trajectory for the SDV. If the SDV speed is non-negligible (≥ 0.05 m/s), we simulate motion for T steps along the map's local road direction at the current speed. If the vehicle is turning (non-zero yaw rate), the baseline follows the road curvature; if reliable map vectors are unavailable, we instead use the current heading. For very low speed or ambiguous lane direction, we propagate the state forward by integrating the last observed speed v and yaw rate ω for T steps (with ω held constant; if $\omega \approx 0$, this reduces to straight-line motion). This yields a baseline trajectory $\{b_1, \ldots, b_T\}$ in the lane-aligned frame. We also apply a global spatial scaling factor S so that positions and residuals lie in a numerically stable range; at evaluation time, metrics are reported in meters by multiplying distances by S.

The model predicts a *residual trajectory* relative to this

Algorithm 2: Quantum encoder (attention mechanism)

Input: $\mathbf{h} \in \mathbb{R}^{11\times 9}$; L_{attn}; $\{\theta_i^{(\ell)}\}$
Output: $\mathbf{x} \in [-1,1]^9$
$(s_{-10},\ldots,s_0) \leftarrow \textsc{ReshapeHistory}(\mathbf{h})$
$\mathbf{q} \leftarrow s_0$
$\mathbf{k} \leftarrow \frac{1}{10}\sum_{i=-10}^{-1} s_i$
$\mathbf{v} \leftarrow s_0 - s_{-1}$
$(\mathbf{q},\mathbf{k},\mathbf{v}) \leftarrow \pi\tanh((\mathbf{q},\mathbf{k},\mathbf{v}))$
$n \leftarrow 9$
prepare $|0\rangle^{\otimes n}$
for $i \leftarrow 1$ **to** n **do**
  apply $R_y(q_i)$ on qubit i
  apply $R_z(k_i)$ on qubit i
  apply $R_x(v_i)$ on qubit i
for $\ell \leftarrow 1$ **to** L_{attn} **do**
  for $i \leftarrow 1$ **to** $n-1$ **do**
    apply $\mathrm{CX}(i \to i{+}1)$
    apply $R_z(\theta_i^{(\ell)})$ on qubit $i{+}1$
    apply $\mathrm{CX}(i \to i{+}1)$
measure Z_i on all qubits and set $\mathbf{x} \leftarrow (x_1,\ldots,x_n)$
return $\mathbf{x}$

Algorithm 3: Quantum feedforward network

Input: $\mathbf{x}^{(0)} \in [-1,1]^9$; L_{ff}; $\{\phi_i^{(\ell)},\psi_i^{(\ell)}\}$
Output: $\mathbf{z} \in [-1,1]^9$
$n \leftarrow 9$
for $\ell \leftarrow 1$ **to** L_{ff} **do**
  prepare $|0\rangle^{\otimes n}$
  for $i \leftarrow 1$ **to** n **do**
    apply $R_y(x_i^{(\ell-1)})$ on qubit i
    apply $R_z(\phi_i^{(\ell)})$ on qubit i
    apply $R_y(\psi_i^{(\ell)})$ on qubit i
  for $i \leftarrow 1$ **to** $n-1$ **do**
    apply $\mathrm{CX}(i \to i{+}1)$
  apply $\mathrm{CX}(n \to 1)$
  measure Z_i on all qubits and set $\mathbf{x}^{(\ell)} \leftarrow (x_1^{(\ell)},\ldots,x_9^{(\ell)})$
$\mathbf{z} \leftarrow \tanh\big(\mathbf{x}^{(L_{\text{ff}})}\big)$
return $\mathbf{z}$

baseline. For each future time step t, we define the ground-truth residual as $\Delta p_t^{\text{gt}} = p_t^{\text{gt}} - b_t$, where p_t^{gt} is the true future position. The model therefore learns to output corrections to a nominal motion (e.g., steering adjustments or acceleration/braking), rather than absolute positions. This focuses learning on deviations from lane-following behavior and reduces dynamic range. Finally, we stack the 11 time-step vectors into an 11×9 array $\mathbf{h} \in \mathbb{R}^{11\times 9}$ (rows $s_{-10},\ldots,s_0$). Each training sample thus consists of $\mathbf{h}$ and a target residual trajectory of length T.

B. Quantum Encoder (Attention Mechanism)

The first stage of our model is a quantum encoder that compresses the history into a latent representation of the SDV's recent motion context. From $\mathbf{h}$ we derive three 9D vectors for an attention-like mechanism: the *query* $\mathbf{q} \in \mathbb{R}^9$ is the current state s_0; the *key* $\mathbf{k} \in \mathbb{R}^9$ is the mean of the 10 past states, summarizing overall history; and the *value* $\mathbf{v} \in \mathbb{R}^9$ is the recent change $s_0 - s_{-1}$. These capture the current configuration, the long-term trend, and the short-term variation. The resulting quantum attention encoder that maps the history $\mathbf{h}$ to the context vector $\mathbf{x}$ is summarized in Algorithm 2.

We encode $(\mathbf{q},\mathbf{k},\mathbf{v})$ into a quantum state using $n = 9$ qubits (one per feature). Starting from $|0\rangle^{\otimes 9}$, we apply a sequence of single-qubit rotations to inject the classical data. For qubit i, we apply $R_y(q_i)$, $R_z(k_i)$, and $R_x(v_i)$ in sequence, where $R_\alpha(\theta)$ is a rotation about axis $\alpha \in \{x,y,z\}$ by angle θ. Before encoding, we apply a $\tanh$ normalization so that $q_i, k_i, v_i \in [-\pi,\pi]$, ensuring angles remain bounded. Denoting the encoding unitary by U_{enc}, the encoded state is

$$|\Psi_{\text{enc}}\rangle = U_{\text{enc}}(\mathbf{q},\mathbf{k},\mathbf{v})\,|0\rangle^{\otimes 9}, \tag{1}$$

$$U_{\text{enc}}(\mathbf{q},\mathbf{k},\mathbf{v}) = \prod_{i=1}^{9}\Big(R_x^{(i)}(v_i)\,R_z^{(i)}(k_i)\,R_y^{(i)}(q_i)\Big). \tag{2}$$

We then introduce entanglement to couple feature dimensions. For each neighboring pair $(i,i+1)$, $i = 1,\ldots,8$, we apply a controlled-NOT gate $\mathrm{CX}(i \to i{+}1)$, a trainable $R_z(\theta_i)$ on qubit $(i+1)$, and another $\mathrm{CX}(i \to i+1)$:

$$|\Psi\rangle \;\mapsto\; \mathrm{CX}_{i\to i+1}\,(I \otimes R_z(\theta_i))_{i+1}\,\mathrm{CX}_{i\to i+1}\,|\Psi\rangle\,. \tag{3}$$

This implements a learned controlled phase rotation that entangles qubits i and $i+1$, allowing adjacent features to influence each other. We refer to data encoding plus these entangling gates as one *quantum attention layer*. We stack $L_{\text{attn}} = 6$ such layers, each with its own parameters $\{\theta_1^{(\ell)},\ldots,\theta_8^{(\ell)}\}$, enabling more complex interactions analogous to multi-layer or multi-head attention. The output of the final attention layer is an entangled multi-qubit state encoding the fused information of $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$. To obtain a classical vector, we measure the expectation value of the Pauli-Z operator on each qubit, yielding $\mathbf{x} \in \mathbb{R}^9$ with

$$x_i = \langle Z_i\rangle = \langle\Psi_{\text{out}}|\,Z_i\,|\Psi_{\text{out}}\rangle\,, \quad x_i \in [-1,1]. \tag{4}$$

This length-9 vector summarizes the motion context after attention-style mixing and serves as input to the quantum feedforward network.

C. Quantum Feedforward Network

The encoder output $\mathbf{x}$ is passed to a deep quantum feedforward network that further refines the latent state. This network has $L_{\text{ff}} = 64$ layers, each a parameterized circuit on the same $n = 9$ qubits. The full encode–rotate–entangle–measure loop that refines $\mathbf{x}$ into the latent vector $\mathbf{z}$ is summarized in Algorithm 3.

In each feedforward layer, we reinitialize the qubits to $|0\rangle^{\otimes 9}$ and encode the current vector $\mathbf{x}$ via rotations: for each qubit i, we apply $R_y(x_i)$ (with x_i scaled into an appropriate range if needed). We then apply two learned single-qubit rotations per qubit, $R_z(\phi_i)$ followed by $R_y(\psi_i)$. The angles $\{\phi_i, \psi_i\}_{i=1}^{9}$ are trainable parameters, playing a role analogous to weights in a fully-connected layer. Each feedforward layer thus has $2n$ parameters (18 when $n = 9$), for about 1152 parameters across all 64 layers. To mix information across qubits, we introduce a ring entanglement pattern:

$$U_{\text{ring}} = \mathrm{CX}_{1\to 2}\,\mathrm{CX}_{2\to 3}\,\cdots\,\mathrm{CX}_{8\to 9}\,\mathrm{CX}_{9\to 1}. \tag{5}$$

This cycle of CNOTs couples all qubits, so that each qubit's state depends on the full vector $\mathbf{x}$. The entangling gates in the feedforward layers are fixed (non-trainable) and serve only to mix data. At the end of each layer ℓ, we measure Pauli-Z expectations to obtain a new 9D vector

$$\mathbf{x}^{(\ell)} = \big(\langle Z_1\rangle^{(\ell)}, \ldots, \langle Z_9\rangle^{(\ell)}\big), \tag{6}$$

which forms the input to the next layer. Repeating this encode-rotate-entangle-measure loop for L_{ff}=64 layers yields a final latent vector $\mathbf{z} \in \mathbb{R}^9$. We apply an elementwise $\tanh$ to $\mathbf{z}$ to keep components in $[-1, 1]$. The vector $\mathbf{z}$ is a compact representation of the scenario, distilled from the SDV's past trajectory and context, and is used by the decoder to generate multi-modal futures.

D. Quantum Decoder for Multi-Modal Prediction

The latent vector $\mathbf{z}$ is fed into a quantum decoder that outputs M future trajectory hypotheses and a confidence for each. We use $M = 16$ modes. The decoder is designed so that a single quantum circuit, together with mode-specific phase shifts, produces all M modes via quantum superposition and interference. Algorithm 4 details the quantum decoder that maps the latent $\mathbf{z}$ to the set of trajectories $\{p_{1:T}^{(m)}\}_{m=1}^{M}$ and their confidences $\{\pi^{(m)}\}_{m=1}^{M}$.

The decoder uses another parameterized n-qubit circuit. We encode $\mathbf{z}$ by applying $R_y(z_i)$ to qubit i. We then apply a trainable $R_z(\gamma_i)$ on each qubit, followed by a sequential "ladder" entanglement pattern: for each $i = 1, \ldots, 8$, we apply $\mathrm{CX}(i \to i+1)$, then $R_y(\gamma_i)$ on qubit $(i+1)$, then another $\mathrm{CX}(i \to i+1)$. Using the same parameter γ_i for both R_z and the ladder coupling allows the circuit to distribute latent information across the amplitudes of different computational basis states. The resulting state $|\Psi(\mathbf{z})\rangle$ is a superposition over all 2^n basis states.

To generate mode-specific trajectories, we exploit global phase shifts. For each mode $m \in \{1, \ldots, M\}$, we start from $|\Psi(\mathbf{z})\rangle$ and apply $R_z(\Delta\phi_m)$ to every qubit, where $\Delta\phi_m = \frac{(m+1)\pi}{M}$. This changes the relative phases but not the basis probabilities, producing a new state $|\Psi^{(m)}\rangle$. We then extract the complex amplitude vector $\alpha^{(m)} \in \mathbb{C}^{2^n}$, where $\alpha_j^{(m)} = \langle j|\Psi^{(m)}\rangle$ for $j = 0, \ldots, 2^n - 1$.

We map $\alpha^{(m)}$ deterministically to a 2D trajectory using a truncated Fourier series. For horizon $T = 20$, we choose truncation order $B = 8$. Ignoring the all-zero basis state ($j = 0$), we use the next B amplitudes as Fourier coefficients. The real parts of these B amplitudes weight cosine basis functions, and the imaginary parts weight sine basis functions. For mode m, the residual trajectory $\Delta p_t^{(m)} = (\Delta x_t^{(m)}, \Delta y_t^{(m)})$ at time t is

$$\Delta x_t^{(m)} = \sum_{j=1}^{B} \mathrm{Re}\big(\alpha_j^{(m)}\big)\,\cos\Big(\frac{j\pi t}{T+1}\Big), \tag{7}$$

$$\Delta y_t^{(m)} = \sum_{j=1}^{B} \mathrm{Im}\big(\alpha_j^{(m)}\big)\,\sin\Big(\frac{j\pi t}{T+1}\Big), \tag{8}$$

for $t = 1, \ldots, T$. We then apply a scale factor $S_r = 1.5$ to match the magnitude of residuals to the normalized data,

$$\Delta p_t^{(m)} = S_r \begin{pmatrix} \Delta x_t^{(m)} \\ \Delta y_t^{(m)} \end{pmatrix}, \tag{9}$$

and add the baseline trajectory to obtain absolute predictions:

$$p_t^{(m)} = b_t + \Delta p_t^{(m)}. \tag{10}$$

Algorithm 4: Quantum decoder for multi-modal prediction

Input: $\mathbf{z} \in [-1, 1]^9$; $b_{1:T}$; T=20; M; B; S_r; $\{\gamma_i\}$
Output: $\{p_{1:T}^{(m)}\}_{m=1}^{M}$; $\{\pi^{(m)}\}_{m=1}^{M}$

$n \leftarrow 9$
prepare $|0\rangle^{\otimes n}$
for $i \leftarrow 1$ **to** n **do**
  apply $R_y(z_i)$ on qubit i
  apply $R_z(\gamma_i)$ on qubit i
for $i \leftarrow 1$ **to** $n-1$ **do**
  apply $\mathrm{CX}(i \to i{+}1)$
  apply $R_y(\gamma_i)$ on qubit $i{+}1$
  apply $\mathrm{CX}(i \to i{+}1)$
$\alpha \leftarrow \textsc{StateVector}() \in \mathbb{C}^{2^n}$
for $m \leftarrow 1$ **to** M **do**
  $\Delta\phi_m \leftarrow \dfrac{(m+1)\pi}{M}$
  $\alpha^{(m)} \leftarrow \textsc{ApplyPhase}(\alpha, \Delta\phi_m)$
  for $t \leftarrow 1$ **to** T **do**
    $\Delta x_t^{(m)} \leftarrow \sum_{j=1}^{B} \mathrm{Re}(\alpha_j^{(m)}) \cos\big(\frac{j\pi t}{T+1}\big)$
    $\Delta y_t^{(m)} \leftarrow \sum_{j=1}^{B} \mathrm{Im}(\alpha_j^{(m)}) \sin\big(\frac{j\pi t}{T+1}\big)$
    $\Delta p_t^{(m)} \leftarrow S_r(\Delta x_t^{(m)}, \Delta y_t^{(m)})$
    $p_t^{(m)} \leftarrow b_t + \Delta p_t^{(m)}$
$\mathbf{z}' \leftarrow \textsc{PadToLength}(\mathbf{z}, M)$
$\hat{\mathbf{z}} \leftarrow \mathcal{F}(\mathbf{z}')$
for $m \leftarrow 1$ **to** M **do**
  $c_m \leftarrow |\hat{z}_m|$
for $m \leftarrow 1$ **to** M **do**
  $\pi^{(m)} \leftarrow c_m / \sum_{k=1}^{M} c_k$
return $\{p_{1:T}^{(m)}\}_{m=1}^{M}$, $\{\pi^{(m)}\}_{m=1}^{M}$

If desired, positions can be mapped back to world coordinates by inverting the lane-aligned transform and scaling.

This decoder yields M distinct trajectories $\{p_{1:T}^{(m)}\}_{m=1}^{M}$ from a single latent $\mathbf{z}$ and a shared set of parameters. Different phase offsets $\Delta\phi_m$ produce diverse interference patterns in the Fourier coefficients and thus diverse, yet smooth, trajectories (e.g., straight vs. turn). Crucially, multi-modality arises from phase variation rather than M separate networks, so all modes are generated in one forward pass.

The decoder also outputs a confidence for each mode. We compute these by analyzing the frequency content of $\mathbf{z}$. Specifically, we take the discrete Fourier transform (DFT) of a zero-padded latent vector $\mathbf{z}' \in \mathbb{R}^{16}$ and use the magnitudes of the M Fourier coefficients as raw scores:

$$c_m = \left|(\mathcal{F}(\mathbf{z}'))_m\right|, \quad m = 1, \ldots, M. \tag{11}$$

Normalizing these yields a confidence distribution $(\pi^{(1)}, \ldots, \pi^{(M)})$:

$$\pi^{(m)} = \frac{c_m}{\sum_{k=1}^{M} c_k}. \tag{12}$$

Although this confidence model is not learned, it provides a geometry-aware ranking: modes that align more strongly with dominant latent frequencies receive higher scores. We use these confidences for ranking-based metrics such as recall and precision.

E. Loss Function and Training Strategy

We train the model by comparing the predicted multi-modal trajectories to the ground truth and minimizing a regression loss that accounts for multiple modes. Given the M predicted trajectories $\{p_{1:T}^{(m)}\}_{m=1}^{M}$ for a sample and the ground-truth trajectory $\{p_t^{\mathrm{gt}}\}_{t=1}^{T}$, we define

$$\begin{aligned}\mathcal{L} := \min_{m=1,\ldots,M} \frac{1}{T}\sum_{t=1}^{T} \left\|p_t^{(m)} - p_t^{\mathrm{gt}}\right\|^2 \\ +\lambda \frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T} \left\|\Delta p_t^{(m)}\right\|^2\end{aligned} \tag{13}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^2$ and λ is a small weighting coefficient. The first term is the mean squared error (MSE) of the best-predicting mode, encouraging at least one trajectory to match the true outcome. The second term penalizes large residuals and discourages overly oscillatory predictions; we use $\lambda = 10^{-4}$ so that this acts as a mild smoothness regularizer. The loss is computed in normalized lane-frame coordinates; for reporting in meters, we rescale distances by S at evaluation time. The SPSA-based training loop used to minimize this loss and update all circuit parameters $\boldsymbol{\theta}$ is summarized in Algorithm 5.

Because the model is a hybrid quantum–classical system with non-analytic components (e.g., the discrete Fourier mapping and $\min$ operator), standard backpropagation through all layers is nontrivial. We therefore use a gradient-free optimizer, SPSA to train all circuit parameters. SPSA is well-suited to high-dimensional optimization with noisy function evaluations. At each iteration, it perturbs the full parameter vector $\boldsymbol{\theta}$ (all trainable angles in the encoder, feedforward stack, and decoder, ~ 1200 parameters) in two opposite random directions and evaluates the loss at $\boldsymbol{\theta} + \Delta$ and $\boldsymbol{\theta} - \Delta$. The gradient is estimated as

$$\nabla L(\boldsymbol{\theta}) \;\approx\; \frac{L(\boldsymbol{\theta} + c_k\,\Delta) - L(\boldsymbol{\theta} - c_k\,\Delta)}{2\,c_k}\,\Delta \tag{14}$$

where Δ has i.i.d. components in $\{\pm 1\}$ and c_k is a perturbation scale at iteration k. Crucially, SPSA requires only two loss evaluations per iteration, regardless of the number of parameters.

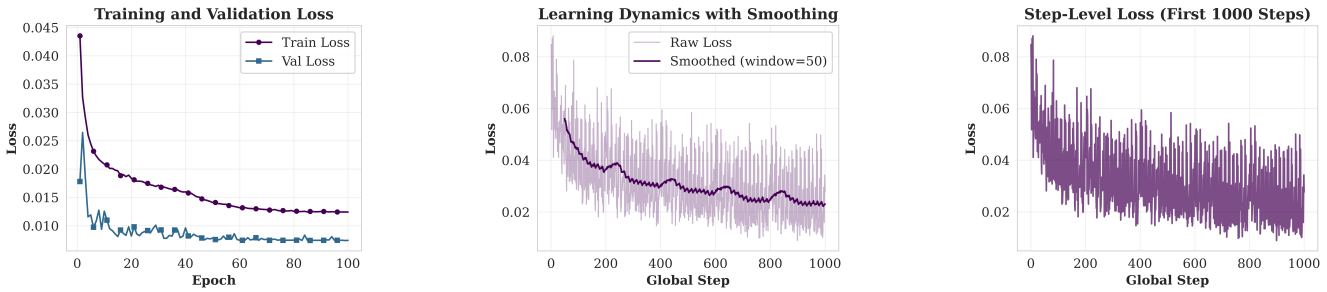
We use decaying step sizes a_k and perturbation scales c_k,

$$a_k = a\,(A + k)^{\alpha}, \qquad c_k = c\,k^{\gamma}, \tag{15}$$

with hyperparameters $A = 80$, $a = 0.05$, $c = 0.1$, $\alpha = 0.602$, and $\gamma = 0.101$, following standard SPSA practice. We further reduce variance by averaging the gradient estimate over two independent perturbation draws per iteration. Training runs for 100 epochs. Each epoch consists of 200 randomly sampled batches from the training set, with batch size 32. Within a batch, we apply SPSA updates per sample (an online training style) rather than aggregating gradients across the batch. We reset the SPSA iteration counter at the start of each epoch, effectively restarting the learning-rate schedule and preventing overly rapid decay. After each epoch, we evaluate the model on a held-out validation subset using the same preprocessing and decoder to obtain all M trajectories, which are then compared

Algorithm 5: SPSA training for quantum multi-modal trajectory model

Input: triplets $(\mathbf{h}, b_{1:T}, \Delta p_{1:T}^{\mathrm{gt}})$; number of epochs; batches per epoch; batch size; a, c, A, α, γ; λ
Output: parameters $\boldsymbol{\theta}$
initialize $\boldsymbol{\theta}$
for *epoch* $= 1$ **to** *number of epochs* **do**
  $k \leftarrow 0$
  partition and shuffle triplets into batches of given batch size
  foreach *batch* **do**
    foreach $(\mathbf{h}, b_{1:T}, \Delta p_{1:T}^{\mathrm{gt}})$ *in batch* **do**
      $k \leftarrow k + 1$
      $\Delta \leftarrow \textsc{Rand}\{\pm 1\}^{|\boldsymbol{\theta}|}$
      $a_k \leftarrow a(A{+}k)^{-\alpha}$
      $c_k \leftarrow c\,k^{-\gamma}$
      $\boldsymbol{\theta}^{+} \leftarrow \boldsymbol{\theta} + c_k\Delta$
      $\boldsymbol{\theta}^{-} \leftarrow \boldsymbol{\theta} - c_k\Delta$
      $\{p_{1:T}^{(m,+)}\}_{m=1}^{M} \leftarrow \textsc{Forward}(\mathbf{h}, b_{1:T}, \boldsymbol{\theta}^{+})$
      $L^{+} \leftarrow \textsc{Loss}(\{p_{1:T}^{(m,+)}\}_{m=1}^{M}, b_{1:T}, \Delta p_{1:T}^{\mathrm{gt}}, \lambda)$
      $\{p_{1:T}^{(m,-)}\}_{m=1}^{M} \leftarrow \textsc{Forward}(\mathbf{h}, b_{1:T}, \boldsymbol{\theta}^{-})$
      $L^{-} \leftarrow \textsc{Loss}(\{p_{1:T}^{(m,-)}\}_{m=1}^{M}, b_{1:T}, \Delta p_{1:T}^{\mathrm{gt}}, \lambda)$
      $g \leftarrow \dfrac{L^{+} - L^{-}}{2c_k}\,\Delta$
      $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - a_k g$
return $\boldsymbol{\theta}$



(a) Global loss behaviour over 100 epochs. (b) Loss (raw & smoothed). (c) Per-step loss trace.

Fig. 2: Global and step-level training signals. **(a)** Training and validation losses decrease smoothly and track closely, indicating good generalization under residual prediction and bounded-angle normalization. **(b)** Raw and smoothed loss show steady improvement with gentle undulations aligned with SPSA schedule resets. **(c)** Step-wise loss traces confirm the same trend at finer granularity.

to ground truth. All circuit parameters are initialized with small random values drawn from $\mathcal{N}(0, 0.05^2)$, and a fixed random seed is used for initialization and data shuffling to ensure reproducibility. We also save parameter checkpoints $\boldsymbol{\theta}$ periodically for potential early stopping and post-hoc analysis.

III. Experiments and Results

A. Evaluation Setup

a) Prediction Horizon and Output Format.: For each scenario the model produces $M{=}16$ trajectory hypotheses, each covering a 2.0 s horizon at 0.1 s resolution, together with confidence scores $\pi^{(m)}$. We interpret these as M plausible futures for the SDV: the highest-confidence trajectory can be treated as the primary prediction, while the remaining modes capture uncertainty and multi-modality in complex scenes.

b) Validation Procedure.: During validation, each scenario's 1.1 s history is fed through the model to obtain the $M{=}16$ predicted trajectories and confidences; displacement and ranking metrics are then computed per scenario and aggregated. Errors are evaluated in the local lane-aligned frame and converted to meters for reporting. We also evaluate a simple baseline (lane-following or CTRV extrapolation from preprocessing) by comparing its trajectory $\{b_t\}$ (zero residual) to the ground truth, yielding baseline ADE/FDE. In practice, the model's minADE and minFDE are significantly lower than the baseline's errors, indicating that the quantum model learns meaningful residual corrections to the nominal path. For all metrics we use a large, representative subset of the validation set, 50 batches of 32 scenarios (1,600 scenarios), due to simulation constraints, and report averages over these scenes. We repeat evaluation with different random seeds and observe minimal variation, suggesting stable performance and reliable estimates. The reported quantities (minADE, minFDE, Recall@K, mAP) can therefore be regarded as robust indicators of the model's accuracy on the validation split.

B. Training dynamics and convergence

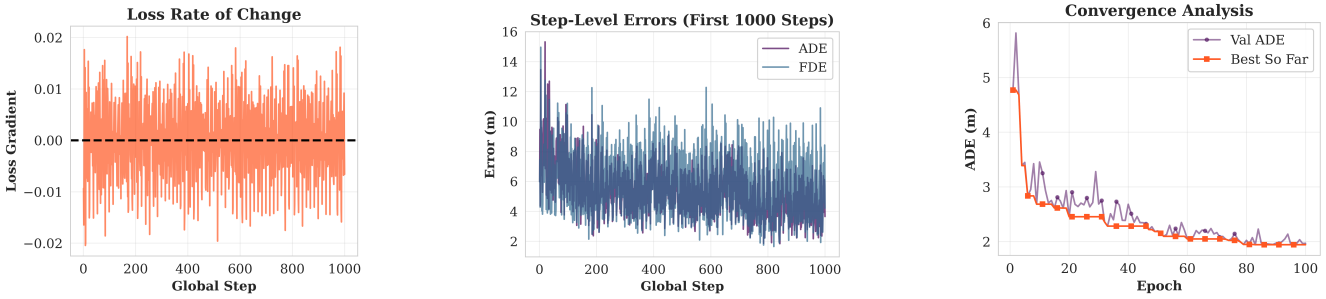
a) Global loss behaviour.: Training and validation losses decrease smoothly over 100 epochs as shown in Fig. 2a, with the validation curve closely tracking the training curve, evidence that the shallow, 9-qubit circuits generalize without overfitting. Two design choices make this possible: (i) predicting *residuals* in a lane-aligned frame compresses output range and removes most rigid-body motion; (ii) feature normalization bounds all rotation angles, keeping SPSA perturbations well-conditioned.

b) Fine-grained learning signals.: The per-step traces reveal how the hybrid system progresses. The raw step-level loss and its smoothed trajectory (Fig. 2b, Fig. 2c) show steady improvement with gentle undulations that coincide with the epoch-wise SPSA schedule resets. The finite-difference loss rate (Fig. 3a) oscillates around a negative mean, which is characteristic of objectives with a $\min$ over modes: when a new hypothesis becomes the best explainer for a scene, the local slope changes sign but the global trend remains downward. Step-level ADE/FDE (Fig. 3b) steadily contract in amplitude, indicating that volatility becomes localized to a shrinking subset of challenging scenes.

c) Convergence and stability.: A 'best-so-far' validation ADE forms a staircase that drops at epochs $\sim$6, 18, 35, 52, 74 and 83 (3c); those steps align with parameter regimes where the shallow attention and decoder phases re-align to the Fourier readout. The rolling standard deviation of ADE collapses by an order of magnitude from early to late training (4a), and the epoch-to-epoch improvement trend tapers to near-zero after $\sim$70 epochs (4b), both consistent with well-behaved convergence. In parallel, variance over training settles to a tight band (4c). Together these diagnostics show that the shallow circuits admit stable optimization and predictable convergence under SPSA.

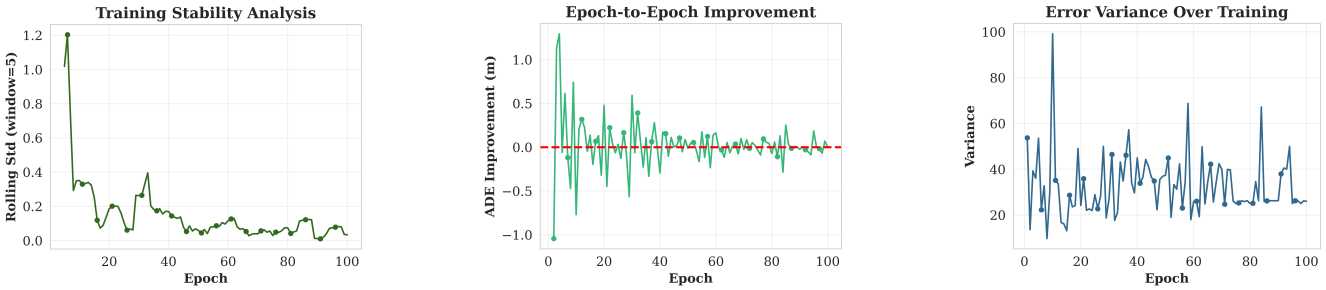
C. Accuracy over time and against a kinematic baseline

a) ADE/FDE across epochs and miss rates.: Average and terminal errors decline rapidly in the first 10–15 epochs and then plateau gracefully (Fig. 5a). Miss rates track this improvement: Miss@2 m and Miss@4 m drop steadily and stabilize (Fig. 5b). The pattern is mechanistically consistent with our design: the attention encoder quickly captures short-horizon kinematics, while the Fourier decoder refines terminal placement via phase adjustments.



(a) Finite-difference loss rate. (b) Step-level ADE/FDE. (c) Best-so-far validation ADE.

Fig. 3: Fine-grained learning signals and discrete convergence steps. **(a)** Finite-difference loss rate oscillates around a negative mean; sign flips reflect mode switches under the min-over-modes objective while the global trend decreases. **(b)** Step-level ADE/FDE amplitudes contract over training, confining volatility to a shrinking subset of scenes. **(c)** Best-so-far validation ADE forms a staircase with drops at epochs ∼6, 18, 35, 52, 74, and 83, consistent with phase realignments between the shallow attention/decoder and the Fourier readout.



(a) Rolling stdev of ADE collapses. (b) Epoch-to-epoch improvement rate tapers. (c) Variance settles to a tight band.

Fig. 4: Convergence and stability diagnostics under SPSA. **(a)** Rolling ADE standard deviation drops by roughly an order of magnitude from early to late training. **(b)** Epoch-to-epoch improvement rate tapers after ∼70 epochs, indicating approach to a stable optimum. **(c)** Error variance settles into a narrow band. Together these trends indicate stable optimization and predictable convergence for the shallow 9-qubit circuits.

b) Percentile structure of errors.: Error percentiles reveal *where* the gains accrue. The high-volume region improves markedly, P90 ADE falls to approximately $3.5\,\mathrm{m}$ to $4.0\,\mathrm{m}$ and P95 to approximately $9\,\mathrm{m}$ to $10\,\mathrm{m}$, whereas P99 remains approximately constant (Fig. 5c). An alternate percentile progression view (Fig. 5d) confirms the same trend with shaded P90-P100 and P95-P100 bands. This profile is exactly what a truncated Fourier basis ($B{=}8$) should produce: strong bias toward smooth, lane-conforming adjustments (head and mid-tail), with a small long-tail associated with abrupt late decisions.

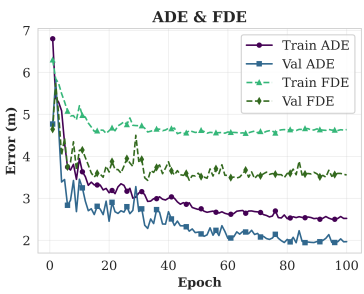
c) Comparison to a strong kinematic baseline.: Against a lane/CTRV extrapolation, the model's ADE curve closes the gap epoch by epoch (Fig. 5e), and a time-profiled error curve averaged across epochs shows a consistent advantage across the prediction horizon (Fig. 5f). Importantly, our approach delivers this accuracy *and* a multi-hypothesis distribution in a single quantum pass-an efficiency benefit inherent to the superposition-based decoder.

D. Multi-modal generation: coverage and efficiency

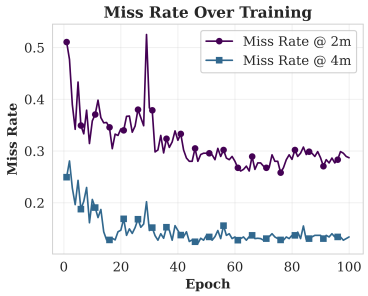
a) Minimum errors over the top-K ranked hypotheses.: The minimum ADE and FDE measured over the top-K confidence-ranked modes decrease monotonically with K (Fig. 6a, Fig. 6b); at the model-selection epoch (83), the values are $1.952\,\mathrm{m}/3.694\,\mathrm{m}$ for $K{=}5$ and $1.942\,\mathrm{m}/3.562\,\mathrm{m}$ for $K{=}16$ (Fig. 6c). The shallow slope indicates that even lower-ranked modes remain close to the best hypothesis, consistent with the decoder's coherent phase-family construction.

b) Oracle set quality.: Best-K (oracle) curves evidence the quality of the hypothesis *set*. Best-K ADE and FDE fall well below the corresponding Top-K values in later epochs (Fig. 7a, Fig. 7b; see also Fig. 7c, which compacts the ADE/FDE Best-K means across $K \in \{1, 5, 10\}$ and shows a brief warm-up (epochs ∼5–20) followed by a steady decline that tapers after ∼70 epochs). This is a key result: with only 9 qubits and shallow entanglement, a single forward pass routinely *contains* at least one trajectory that lies close to the realized future among its superposed modes.

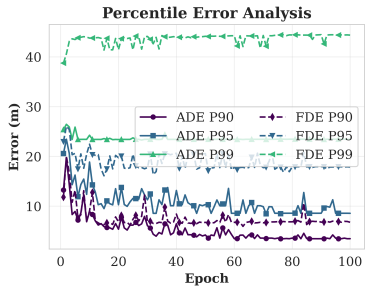
c) Top-K end-point dispersion.: Top-K FDE measured on the highest-confidence hypotheses exhibits stable behaviour across training and decreases with K (Fig. 8a). The level and smoothness of these curves reflect consistent long-horizon end-point dispersion under a stringent measure applied uniformly across heterogeneous scenes.



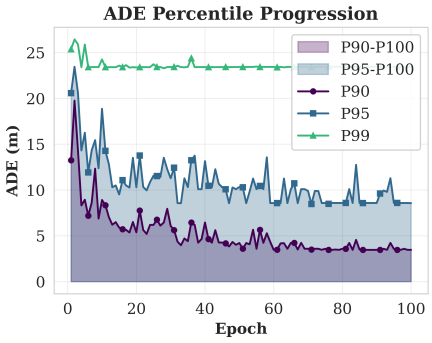
(a) ADE/FDE across epochs.



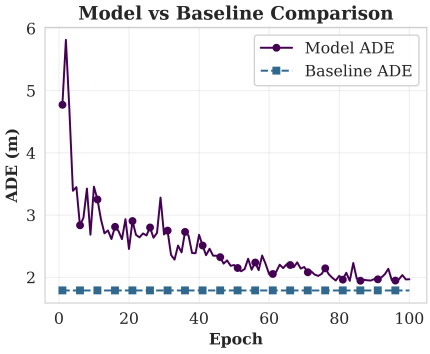
(b) Miss@2 m and Miss@4 m



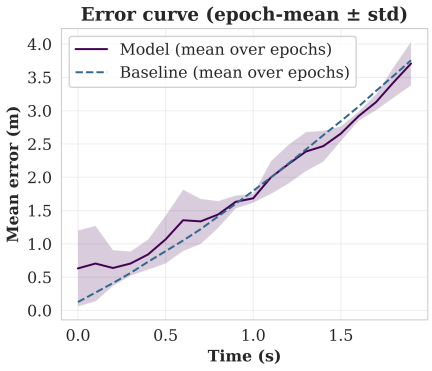
(c) Percentile ADE (P50–P99).



(d) Progression with shaded P90–P100 and P95–P100.



(e) ADE vs. lane/CTRV baseline.



(f) Error over horizon (epoch-mean ± std).

Fig. 5: Accuracy overview across training. **(a)** ADE/FDE drop quickly in the first 10–15 epochs then taper, indicating fast capture of short-horizon kinematics with later refinement. **(b)** Miss@2 m and Miss@4 m decrease and stabilize. **(c)** Percentile ADE (P50–P99) shows substantial shrinkage of the bulk and mid-tail, with little change in the extreme tail. **(d)** P90–P100 and P95–P100 bands narrow steadily, indicating improved error concentration. **(e)** ADE vs. lane/CTRV baseline: the model opens and sustains a gap over the kinematic prior. **(f)** Horizon-wise error curves over 0 s–2.0 s show a consistent advantage across time.

d) Top-K vs. Best-K at the selection epoch.: A compact bar plot contrasts Top-K (confidence-ranked) and Best-K (oracle) ADE at epoch 83 (Fig. 8b). The oracle bars quantify the *representational headroom* available from the same shallow circuits; the Top-K bars confirm that the chosen ranking still captures a large share of that headroom at modest K. This gap is useful: it indicates that further improvements can be achieved by purely algorithmic changes to scoring, without modifying circuit depth or qubit count.

E. Ranking quality, recall and calibration

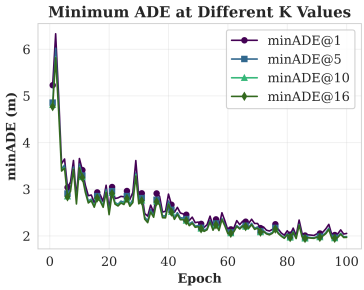
a) Hit rates and recall.: Hit@1 increases over training while the "oracle within top-K" rates remain high (Fig. 8c). Across horizons (1.0 s, 2.0 s), thresholds (2 m, 4 m), and $K \in \{1, 5, 10, 16\}$, recall stays consistently strong, and at the selection epoch coverage is near-uniform across these settings.

b) Calibration and precision–recall summaries.: Expected Calibration Error evolves gradually and remains stable across epochs, as seen in the ECE row of the temporal performance heatmap (Fig. 9), consistent with our deterministic, scene-level confidence mapping derived from the latent spectrum. Scene-wise mAP remains nearly constant, and Average Precision stays high across epochs (Fig. 8d); at the selection epoch, AP at the standard distance thresholds is effectively saturated.

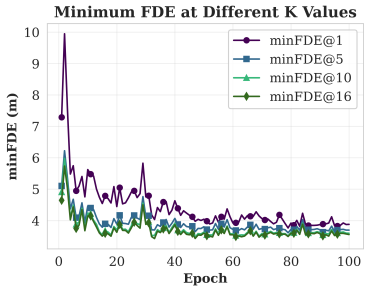
F. Aggregated views and metric relations

a) All metrics at the selection epoch.: At the selection epoch, the aggregate snapshot shows meter-scale ADE, multi-meter FDE, and further reductions for *minADE*/*minFDE* at modest K. Miss@2 m and Miss@4 m are substantially lower than for the kinematic baseline, Hit@1 is improved, and calibration/precision–recall measures remain stable, indicating strong accuracy and broad hypothesis coverage from a lightweight quantum model. Although we do not aim for state-of-the-art performance, it is useful to contextualize the numbers: recent large classical models on the WOMD typically report ADE/FDE in the sub-meter to ∼1–2 m range for comparable short-horizon settings, using orders of magnitude more parameters and richer scene encoders. Our 9-qubit, $\sim 10^3$-parameter model sits above these best-in-class error levels, but does so while improving over a strong kinematic baseline and providing single-pass multi-modality in a hardware-aligned quantum formulation.

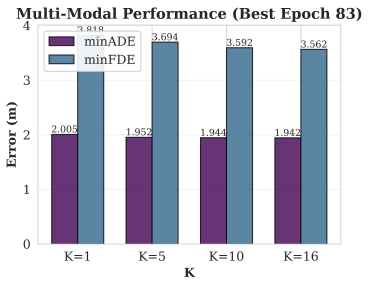
b) Temporal overview.: A normalized performance heatmap aligns the trajectories of all major metrics across the 100 epochs (Fig. 9): the ADE/FDE rows progressively darken



(a) Min ADE vs. top-K (confidence-ranked).

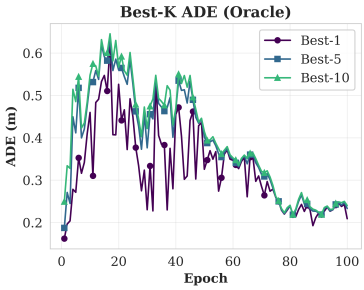


(b) Min FDE vs. top-K (confidence-ranked).

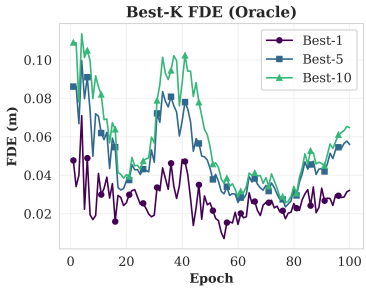


(c) Selection epoch (83): key values for K=5 and K=16.

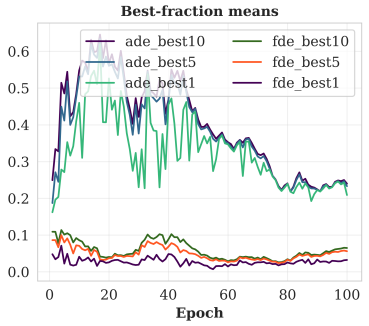
Fig. 6: Minimum ADE/FDE over the top-K confidence-ranked hypotheses decrease monotonically with K, with shallow slopes—indicating that even lower-ranked modes remain close to the best hypothesis. At the model-selection epoch (83), values reported in the text are read from the right panel.



(a) Best-K (oracle) ADE over epochs.



(b) Best-K (oracle) FDE over epochs.



(c) Best-K means (ADE upper cluster, FDE lower cluster).

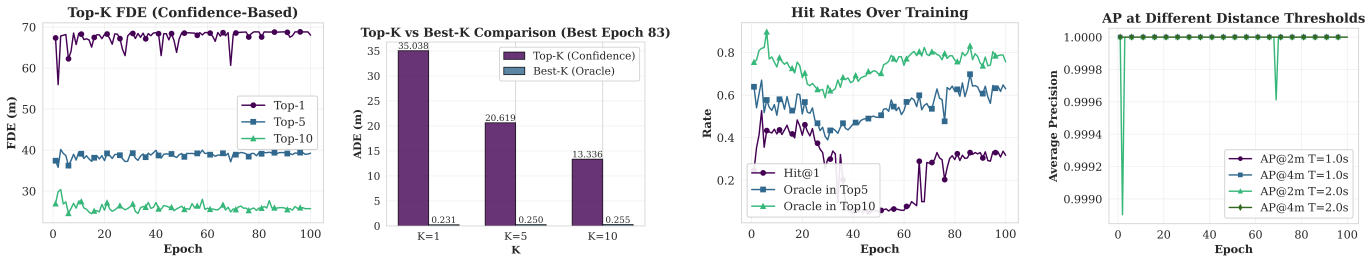
Fig. 7: Oracle set quality across training. Left/middle: Best-K ADE and FDE over epochs. Right: compact summary of ADE/FDE Best-K means for $K \in \{1, 5, 10\}$ showing a brief warm-up (epochs ∼5–20), steady decline, and taper after ∼70 epochs (meters after de-normalization).

(improve), miss rates cool in tandem, and Hit@1 brightens as training proceeds, showing coherent evolution across metrics from a single training run.

G. Why shallow circuits are effective here

The results above arise from pairing *task structure* with *quantum inductive bias*. Residual learning in an ego-centric lane frame ensures the quantum layers model only small, structured deviations, well within the expressivity of 9 qubits with linear/ladder entanglement. The Fourier decoder encodes a smoothness prior matched to short-horizon driving statistics: it captures the bulk of trajectories at very low circuit depth and, via global phase superposition, produces multiple plausible futures without additional passes. SPSA with bounded rotations then yields stable optimization despite the non-smooth "min over modes" objective, as reflected in the loss, variance and convergence diagnostics (Fig. 2b, Fig. 2c, Fig. 3a, Fig. 4a, Fig. 3c).

a) Resource-efficient design that drives accuracy.: Our performance gains stem from design choices that shrink the function class the quantum circuit must represent while preserving the behaviours that matter for short-horizon driving. Predicting *residuals* in an ego-lane frame atop a map/CTRV baseline means the circuit models only small, smooth corrections instead of full trajectories, concentrating probability mass near lane-consistent futures and explaining the rapid ADE/FDE decline and lower slope of the error-over-time curves relative to the baseline (Fig. 5e). The decoder's truncated Fourier basis with phase superposition simultaneously imposes a low-frequency smoothness prior and emits multiple coherent hypotheses in a *single* pass, yielding strong Recall@K and stable mAP without scaling inference cost with K (summarized above; see Fig. 6c). Shallow entanglement with per-qubit rotations keeps the parameter count tiny and the SPSA landscape well-conditioned, producing the clean convergence and shrinking variance we observe (Fig. 2b, Fig. 4a, Fig. 4c). Finally, confidences derived from the latent spectrum (FFT of the decoded state) mean ranking quality emerges from the same physics-shaped representation that generates trajectories, explaining the strong Top-K curves and high 'oracle in Top-5/10' rates without an extra learned head (Fig. 6c, Fig. 8c). Collectively, these design decisions allow a 9-qubit, shallow-depth model to deliver meter-level accuracy, broad hypothesis coverage and stable training, achieving useful performance while remaining computationally and parameter efficient.



(a) Top-K FDE across training. (b) Epoch 83: Top-K vs. Best-K ADE. (c) Hit@1 and oracle-in-Top-K. (d) Average Precision across epochs.

Fig. 8: Ranking, dispersion, and calibration overview. **(a)** Top-K FDE (confidence-ranked) decreases with K and remains stable across training, indicating controlled long-horizon dispersion. **(b)** At the selection epoch (83), the gap between Top-K (ranked) and Best-K (oracle) ADE measures representational headroom, with ranking already capturing much of it at modest K. **(c)** Hit@1 rises over training while 'oracle in Top-K' rates stay high, indicating strong ranking quality. **(d)** AP stays consistently high across epochs, summarizing precision–recall behaviour under the adopted evaluation protocol.

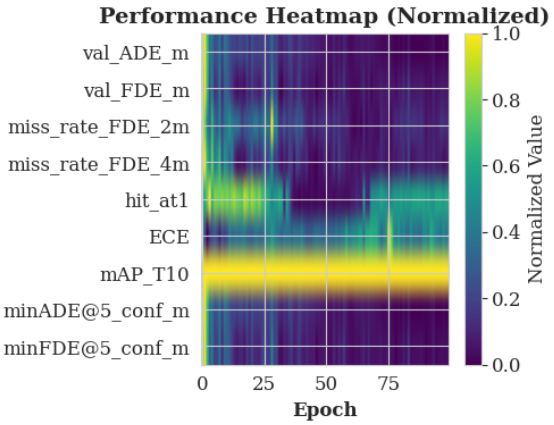


Fig. 9: Temporal overview: normalized performance heatmap over 100 epochs showing coherent evolution across metrics.

IV. Discussion

A. Complexity, expressivity and multi-modality

The decoder operates on n=9 qubits, i.e. a 2^n=512-dimensional Hilbert space, while the full model uses only $N_\theta \approx 1.2\times10^3$ trainable angles across attention, feedforward and decoder circuits. From a classical viewpoint, compositions of trigonometric rotations and sparse entanglement implement a rich non-linear feature map over the SDV history, with the truncated Fourier readout acting as a compact linear head in a low-frequency basis. Oracle metrics expose the effective capacity of this setup: Best-K (oracle) ADE/FDE curves lie well below their Top-K (confidence-ranked) counterparts for moderate K (e.g. $K \in \{5, 10, 16\}$), showing that in many scenes at least one of the M=16 hypotheses lies close to the realized future. The remaining gap is largely due to ranking rather than representation—the shallow circuits can already place some modes near ground truth within the truncated Fourier basis, but the spectrum-derived confidences do not always surface these modes at the very top. The multi-modal construction is deliberately frugal: all M hypotheses share a single decoder and differ only by global phase offsets applied before Fourier readout. This 'phase sweep' produces a correlated family of futures that share a smooth backbone but diverge along a few phase directions, without increasing depth or qubit count, yielding a hypothesis set that is expressive for short-horizon motion while keeping ranking simple and hardware-aligned.

B. Relationship to classical baselines and large models

Relative to the lane/CTRV baseline, the model opens and maintains an ADE/FDE gap across the prediction horizon, while also reducing miss rates and improving Hit@1 and related recall measures. This indicates that the quantum components learn useful residual structure beyond the deterministic prior, even under tight parameter and depth budgets, with the largest gains appearing in the bulk and mid-tail of the scene distribution where the lane-aligned residual and Fourier priors are most appropriate. Absolute error levels remain above those of large classical GNN and transformer architectures on WOMD, which is expected given that such models employ orders of magnitude more parameters, ingest full multi-agent and HD-map context, and are trained end-to-end with backpropagation. The goal here is not to match those systems, but to show that a small quantum sub-module, designed for NISQ constraints, can train stably, produce coherent multi-modal outputs and reliably outperform a strong kinematic baseline. From a deployment perspective, the single-pass multi-modality of the Fourier decoder is particularly attractive: all M trajectories are generated in one quantum forward pass and confidences are derived from the same latent spectrum, so decoding cost is effectively constant in M, in contrast to classical multi-head or iterative schemes whose inference time scales with the number of modes.

C. Limitations and scope

This study has important limitations. All experiments are run in classical simulation, so noise, decoherence and device-specific constraints are not captured; although the circuits are shallow and low-parameter to be NISQ-friendly, the reported metrics do not yet reflect real hardware behaviour. The formulation is SDV-centric with strong preprocessing: the

lane-aligned residual representation compresses the prediction problem but omits rich interaction patterns between agents and much of the HD-map structure beyond local lane direction. Consequently, highly interactive or rare manoeuvres such as merges, near-collisions and abrupt lane changes remain challenging for a low-frequency Fourier basis, which is visible in the persistent long tail of large ADE/FDE cases. The evaluation further targets a short (2 s) horizon on a large but still limited subset of WOMD due to simulation cost, and the work does not claim quantum advantage over strong classical forecasters. Instead, it should be viewed as a controlled feasibility study: under realistic resource constraints, shallow, few-qubit circuits can be integrated into a physics-shaped pipeline and trained to deliver non-trivial, multi-modal trajectory forecasts on a modern autonomous-driving benchmark.

V. Conclusion and Future Work

This paper presented a compact, physics-shaped hybrid quantum architecture for short-horizon trajectory forecasting in autonomous driving. By working in an ego-centric, lane-aligned frame and predicting residuals on top of a map/CTRV baseline, the model focuses on smooth, meter-scale corrections instead of full trajectories. A transformer-inspired quantum attention encoder, a deep but parameter-lean feedforward stack, and a Fourier-based decoder together form a 9-qubit, $\sim 10^3$-parameter pipeline that generates $M{=}16$ trajectory hypotheses in a single pass, with confidences derived deterministically from the latent spectrum.

On a representative subset of the WOMD, the model attains meter-scale ADE/FDE, improves consistently over a strong kinematic baseline, and shows stable miss rates, Hit@1, and precision–recall behaviour across training. Multi-modal metrics indicate that, for many scenes, at least one of the 16 hypotheses lies close to the realized future, and that the simple spectrum-based ranking captures a substantial fraction of this representational headroom. Training with SPSA remains stable despite non-differentiable components, demonstrating that shallow, few-qubit circuits can be optimized reliably in this setting.

The study remains limited to classical simulation, a single-SDV formulation with strong preprocessing, and a short prediction horizon. Future work includes deploying the architecture on near-term hardware, incorporating richer multi-agent and HD-map context, exploring more expressive or adaptive quantum decoders, and refining ranking and calibration mechanisms. Beyond autonomous driving, the residual, physics-shaped formulation and phase-based multi-modality explored here are applicable to other resource-constrained sequential prediction problems in robotics, logistics, and human motion forecasting.

References

[1] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.

[2] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15303–15312, 2021.

[3] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9710–9719, 2021.

[4] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14403–14412, 2021.

[5] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on intelligent vehicles*, vol. 1, no. 1, pp. 33–55, 2016.

[6] J. Wang, T. Ye, Z. Gu, and J. Chen, "Ltp: Lane-based trajectory prediction for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17134–17142, 2022.

[7] K. Long, Z. Sheng, H. Shi, X. Li, S. Chen, and S. Ahn, "Physical enhanced residual learning (perl) framework for vehicle trajectory prediction," *Communications in Transportation Research*, vol. 5, p. 100166, 2025.

[8] F. Chen, Q. Zhao, L. Feng, C. Chen, Y. Lin, and J. Lin, "Quantum mixed-state self-attention network," *Neural Networks*, vol. 185, p. 107123, 2025.

[9] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, P. Massey, R. Halavati, M. Y. Niu, A. Zlokapa, *et al.*, "Tensorflow quantum: A software framework for quantum machine learning," *arXiv preprint arXiv:2003.02989*, 2020.

[10] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

[11] L. Lin, S. Gong, S. Peeta, and X. Wu, "Long short-term memory-based human-driven vehicle longitudinal trajectory prediction in a connected and autonomous vehicle environment," *Transportation research record*, vol. 2675, no. 6, pp. 380–390, 2021.

[12] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7386–7400, 2021.

[13] R. Huang, G. Zhuo, L. Xiong, S. Lu, and W. Tian, "A review of deep learning-based vehicle motion prediction for autonomous driving.," *Sustainability (2071-1050)*, vol. 15, no. 20, 2023.

[14] J. Liu, X. Mao, Y. Fang, D. Zhu, and M. Q.-H. Meng, "A survey on deep-learning approaches for vehicle trajectory prediction in autonomous driving," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 978–985, IEEE, 2021.

[15] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2375–2384, 2019.

[16] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision*, pp. 683–700, Springer, 2020.

[17] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, *et al.*, "Tnt: Target-driven trajectory prediction," in *Conference on robot learning*, pp. 895–904, PMLR, 2021.

[18] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, *et al.*, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7814–7821, IEEE, 2022.

[19] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, *et al.*, "Scene transformer: A unified architecture for predicting multiple agent trajectories," *arXiv preprint arXiv:2106.08417*, 2021.

[20] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu, "Hivt: Hierarchical vector transformer for multi-agent motion prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8823–8833, 2022.

[21] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," *arXiv preprint arXiv:2207.05844*, 2022.

[22] A. Seff, B. Cera, D. Chen, M. Ng, A. Zhou, N. Nayakanti, K. S. Refaat, R. Al-Rfou, and B. Sapp, "Motionlm: Multi-agent motion forecasting as language modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.

[23] C. Bai, X. Gao, M. Chen, W. Guo, D. Rong, C. Yang, and S. Jin, "Hybrid data-model driven trajectory prediction on highways: Integrating anticipatory interaction awareness and personalized driving preferences,"

Transportation Research Part C: Emerging Technologies, vol. 180, p. 105351, 2025.

[24] N. Singh and S. R. Pokhrel, "Modeling feature maps for quantum machine learning," *arXiv preprint arXiv:2501.08205*, 2025.

[25] N. Singh and S. R. Pokhrel, "Modeling and benchmarking quantum machine learning for genomic data analysis," *IEEE Transactions on Artificial Intelligence*, pp. 1–13, 2025.

[26] S. Y.-C. Chen, S. Yoo, and Y.-L. L. Fang, "Quantum long short-term memory," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8622–8626, IEEE, 2022.

[27] Y. Takaki, K. Mitarai, M. Negoro, K. Fujii, and M. Kitagawa, "Learning temporal data with a variational quantum recurrent neural network," *Physical Review A*, vol. 103, no. 5, p. 052414, 2021.