Adaptive Quantile Guidance in Diffusion Models: Multi-Dataset Learning for Pandemic Time Series

Jane Odum
School of Computing
University of Georgia
Athens, USA
jane.odum@uga.edu

John Mller
School of Computing
University of Georgia
Athens, USA
jamill@uga.edu

***Abstract*—Probabilistic forecasting of epidemiological time series demands accurate uncertainty quantification, especially during volatile outbreak periods. Existing diffusion-based forecasting approaches often use static quantile parameters, which limit their responsiveness to real-time shifts in uncertainty and spikes. We introduce Adaptive Quantile Guidance (AQG), a novel approach that dynamically recalibrates diffusion model quantiles by using recent residual forecast errors. AQG adaptively adjusts prediction intervals and significantly enhances forecast accuracy and uncertainty calibration in rapidly changing epidemiological contexts. Evaluations on multiple real-world pandemic datasets (COVID-19, Influenza-like Illness, and Respiratory Syncytial Virus and others) demonstrate AQG's ability to reduce forecasting errors by up to 73.3% compared to state-of-the-art baselines.**

***Index Terms*—Time Series forecasting, Generative models, Diffusion Model, Quantile guidance, Transfer Learning, Pretraining.**

I. Introduction

The COVID-19 pandemic, alongside seasonal epidemics such as influenza, stresses the importance of accurate and timely forecasting of disease metrics (e.g., incidence, hospital admissions, mortality). These predictive models guide public health interventions and resource allocation, yet building reliable forecasts remains challenging due to evolving patterns, sparse data, and non-stationary dynamics [1]. Current statistical approaches, such as ARIMA or state space models, often struggle to flexibly capture complex temporal and cross-series relationships without substantial domain-specific tuning [2]. More recently, machine learning-based forecasting methods, including deep recurrent or transformer models, have shown promise in achieving finer predictive distributions that can adapt to rapidly changing conditions [3], [4].

Diffusion models, initially popularised in image generation tasks [5], [6], have begun to gain traction in the time series domain. By defining a forward noise and reverse denoising process over continuous timesteps, diffusion-based forecasters can represent complex, multimodal predictive distributions [7]. Early works indicate that diffusion models can capture intricate temporal dependencies and produce coherent probabilistic forecasts. However, two key challenges remain: (1) *Initialization and Transferability*; Training diffusion models from scratch on a single target time series domain can be data-inefficient and slow, and (2) *Uncertainty Calibration*; While diffusion models naturally produce stochastic samples, controlling and calibrating the quantile levels of their predictive distribution is non-trivial.

In this paper, we address these challenges by introducing a *multi-task pre-training* strategy and an *adaptive quantile guidance* mechanism for diffusion-based time series forecasting. First, inspired by the recent shifts in language modeling and vision tasks where large-scale pre-training has led to substantial gains in downstream performance [8], [9], we pre-train a diffusion model on a broad collection of epidemiological series. This pre-training allows the model to internalise general temporal dynamics before fine-tuning on a specific pandemic-related dataset, such as COVID-19 incidence or Influenza-Like Illness (ILI) time series. In doing so, we leverage knowledge transfer to improve both accuracy and data efficiency in the downstream task. A key intuition behind our Adaptive Quantile Guidance (AQG) is that real-world epidemiological time series are highly non-stationary: the level and nature of uncertainty can change abruptly during outbreaks, seasonal transitions, or shifts in reporting. Traditional diffusion-based forecasters with fixed quantile parameters are unable to respond to these changes, often producing prediction intervals that are either too narrow, missing critical surges, or too wide, diluting actionable insights.

Secondly, we propose an *adaptive quantile guidance* technique, extending recent work on guidance mechanisms in diffusion models [10]. Instead of relying on fixed quantiles, our approach adaptively adjusts quantile levels during the reverse diffusion process based on the model's internal representations and the observed data patterns. This dynamic adjustment yields more reliable quantile estimates, providing well-calibrated predictive intervals that are essential for epidemiological decision-making. Miscalibrated forecasts can lead to either over- or under-preparation in healthcare and policy responses, making improved uncertainty representation a vital contribution.

We evaluate our approach on multiple pandemic forecasting tasks, comparing multi-dataset pre-trained diffusion models and adaptive quantile guidance to baseline models and established forecasting approaches. The results show improved accuracy, calibration, and adaptability.

II. Background

Recent progress in deep learning for time series forecasting has been driven by three key innovations: *(i) pre-training strategies* that learn universal temporal representations from large and diverse datasets, *(ii) diffusion models* for robust probabilistic generation, and *(iii) guidance mechanisms* for controllable sample synthesis. In this work, we focus on a multi-dataset pre-training approach, wherein we train a diffusion model on multiple source datasets to learn general temporal features, and subsequently fine-tune the model on a target pandemic dataset that contains relatively limited observations.

A. Pre-Training for Temporal Representations

Pre-training has proven to be an effective for time series analysis, inspired by its success in natural language processing (NLP) [8], GPT-3 [12], and computer vision (CV) [13], [14]. By exposing a model to multiple datasets, for instance, different seasonal or epidemiological signals, we can capture shared temporal structures (e.g., seasonality, recurrent events, anomalies) that transfer to downstream tasks. This is particularly advantageous in pandemic forecasting, where the available data in any *single* disease domain (e.g., influenza or COVID-19) may be sparse or highly nonstationary.

Commonly, *masked reconstruction* is used to learn generalized features by forcing a model to infer randomly masked subsequences from their surrounding temporal context. This idea is central to the masked language modeling objective introduced in BERT [8], which compels the network to understand the underlying structure of the input in order to reconstruct the missing data. When applied to multiple datasets, the model gains exposure to a broader variety of temporal patterns, thus becoming more robust upon fine-tuning.

Formally, let

$$\mathbf{y} = (y_1, \ldots, y_L)^\top \in \mathbb{R}^L$$

denote a univariate time series of length L. In addition, let

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,P} \\ \vdots & \ddots & \vdots \\ x_{L,1} & \cdots & x_{L,P} \end{bmatrix} \in \mathbb{R}^{L \times P} \qquad \mathbf{x}_{t,\cdot} = [\sin(2\pi t/52)],$$

be a matrix of P auxiliary covariates (in our case, sinusoidal components). In some cases, x is not always present in all experiments. The sinusoidal components encode annual seasonality and immunological factors relevant to disease transmission. No other covariates are used, each row $\mathbf{x}_{t,\cdot}$ aligns with time step t.

We select a masking objective $\mathcal{M} \subseteq \{1, \ldots, L\}$ indicating which entries of $\mathbf{y}$ to reconstruct, and train a predictor f_θ to minimize the expected reconstruction error over the combined pre-training corpus $\mathcal{D}_{\text{multi-pre}}$:

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{\mathbf{y} \sim \mathcal{D}_{\text{multi-pre}},\, \mathcal{M}} \left[\sum_{t \in \mathcal{M}} \left\| y_t \; - \; f_\theta\big(\mathbf{y}_{\setminus \mathcal{M}},\, \mathbf{x}\big) \right\|^2 \right]. \quad (1)$$

We do not include an explicit time trend feature (such as t or polynomial terms t^2) in $\mathbf{x}$, because the masked-reconstruction objective in (1) forces f_θ to infer long-term trends, seasonality, and abrupt shifts directly from the surrounding values of $\mathbf{y}$. In practice, modern sequence architectures, using sinusoidal positional encodings and self-attention or recurrent layers, are able to model smooth drift and periodic patterns without requiring hand-crafted time-trend regressors [19], [57]. Studies have shown that adding raw time-trend covariates often yields negligible gains when such inductive biases are already present [3].

B. Denoising Diffusion Probabilistic Models (DDPMs)

Diffusion models [5], [6] provide a principled framework for probabilistic time series generation via iterative denoising. This approach differs from conventional autoregressive models by refining noisy latents into coherent forecasts step by step, allowing a flexible representation of uncertainty.

a) Forward Process.: An original sequence $\mathbf{x}_0$ is progressively corrupted over T steps:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\, \beta_t\,\mathbf{I}\big), \quad (1)$$

where $\{\beta_t\}_{t=1}^T$ is a chosen variance schedule and $\mathbf{x}_T$ approaches Gaussian noise after sufficient corruption steps.

b) Reverse Process.: A neural network ϵ_θ aims to *invert* the corruption, iteratively denoising $\mathbf{x}_T$ back to a valid sample. Concretely, at each reverse step,

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\Big(\mathbf{x}_{t-1};\, \tfrac{1}{\sqrt{\alpha_t}}\big(\mathbf{x}_t \; - \; \tfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\big), \sigma_t^2\mathbf{I}\Big), \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. A simplified training loss [6] is:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\, \mathbf{x}_0,\, \epsilon}\Big[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, t)\|^2\Big]. \quad (3)$$

For time series forecasting, ϵ_θ can condition on past observations or covariates, to allow the model to learn generative dynamics from multiple datasets during pre-training. Subsequent fine-tuning on a single target dataset adapts the learned representations to the domain-specific forecasting challenge (e.g., influenza or COVID-19).

C. Guidance Mechanisms in Diffusion Models

Diffusion models produce samples by iteratively removing noise from an initial Gaussian variable $x_T \sim \mathcal{N}(0, I)$ down to a clean sample x_0. To steer this process toward specific outcomes, such as generating trajectories that exceed a given quantile or satisfy a class label, researchers have developed two principal *guidance* techniques.

a) : **Classifier Guidance** An auxiliary classifier $p_\phi(c \mid x_t)$ is trained to predict a condition c (for example, a class label or a risk-threshold exceedance) from the noisy intermediate state x_t. At each denoising step, its log-probability gradient is added to the model's own score:

$$\tilde{s}(x_t, t \mid c) = \nabla_{x_t} \log p_\theta(x_t) \; + \; w\, \nabla_{x_t} \log p_\phi(c \mid x_t) \quad (4)$$

$\nabla_{x_t} \log p_\theta(x_t)$ is the unconditional score output by the diffusion model, and $w \geq 0$ controls the strength of the classifier's influence. A higher w pushes samples more strongly toward regions the classifier deems likely for c, but requires a separate network trained across all noise levels $\{x_t\}$ and can reduce sample diversity.

b) : **Classifier-Free Guidance** Instead of training two models, classifier-free guidance trains a single network ϵ_θ to predict noise both with and without conditioning. Concretely, the model learns:

$$\epsilon_\theta(x_t, t, c) \quad \text{and} \quad \epsilon_\theta(x_t, t, \varnothing), \tag{5}$$

where $\varnothing$ denotes the absence of any condition. At inference, one combines these two outputs:

$$\epsilon_\theta^{\text{guided}}(x_t, t, c) = (1 + w)\, \epsilon_\theta(x_t, t, c) \; - \; w\, \epsilon_\theta(x_t, t, \varnothing) \tag{6}$$

with w again modulating guidance strength. This approach achieves similar control over the sample distribution without an external classifier, trading off diversity for conditional fidelity, and incurs two model evaluations per timestep.

Both methods use a static guidance weight w chosen before sampling, which cannot adapt to actual forecast performance or changing data dynamics. In forecasting scenarios, where biases and volatility can shift rapidly, fixed guidance can under or overcorrect. This limitation motivates *Adaptive Quantile Guidance* (AQG), introduced in Section III-B, which dynamically updates the effective guidance weight based on recent residuals to maintain calibrated intervals in real time.

III. Methodology

We want to learn a diffusion-based forecaster that (1) leverages multi-dataset pretraining to internalize diverse temporal dynamics and (2) employs an online calibration loop, Adaptive Quantile Guidance (AQG), to adjust its target quantile κ_t at inference, to ensure well-calibrated predictive intervals under shifting data uncertainty.

A. Diffusion-Based Forecasting Backbone

Given historical observations $\mathbf{y}_{\text{obs}} \in \mathbb{R}^{L \times C}$, we form the input $\mathbf{x}_0 = \mathbf{y}_{\text{obs}}$. A forward noising process corrupts $\mathbf{x}_0$ over T steps:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\; \beta_t I\big), \quad \bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s). \tag{7}$$

A neural network $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to predict the injected noise via

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \big\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\; t) \big\|^2. \tag{8}$$

The learned reverse process

$$p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t) = \mathcal{N}\big(\mu_\theta(\mathbf{x}_t, t),\, \sigma_t^2 I\big), \tag{9}$$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \Big(\mathbf{x}_t - \tfrac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(\mathbf{x}_t, t) \Big) \tag{10}$$

recovers $\mathbf{x}_0 \approx \mathbf{y}_{\text{obs}}$ and, when conditioned on these observations, generates future forecasts.

B. Adaptive Quantile Guidance

Adaptive Quantile Guidance (AQG) embeds an online calibration loop into the diffusion sampling process to adaptively tune the forecast quantile parameter $\kappa_t \in [0, 1]$. In our case, we begin with initial $\kappa_t = 0.5$. κ_t represents the target quantile level at time t.

Let $y_t \in \mathbb{R}$ be the observed value at time t and $\hat{y}_t$ the corresponding median forecast using quantile κ_t. We maintain a buffer of the last w one-step residuals:

$$r_{t-i} = y_{t-i} - \hat{y}_{t-i}, \quad i = 1, \ldots, w, \tag{11}$$

r_{t-i} represents the forecast error at time $t - i$. We assign exponential recency weights to emphasize recent performance:

$$\alpha_i = \rho^{i-1}, \quad 0 < \rho < 1, \tag{12}$$

where ρ is the decay factor controlling the rate of weight decay.

The weighted exceedance ratio, which tells us how often our forecasts have been too low over the last w steps, is given by:

$$q_t = \frac{\sum_{i=1}^{w} \alpha_i\, \mathbf{I}(r_{t-i} > 0)}{\sum_{i=1}^{w} \alpha_i}, \tag{13}$$

$\mathbf{I}(\cdot)$ is the indicator function. For example, $q_t = 0.7$ implies that, after recency weighting, 70% of the last w forecasts underestimated the truth, signaling that the next quantile κ_{t+1} should be increased accordingly.

We then update the quantile parameter using a clipped gradient descent step:

$$\kappa_{t+1} = \text{clip}\left(\kappa_t + \eta \cdot (q_t - \kappa_t), \epsilon, 1 - \epsilon\right), \tag{14}$$

$\eta \in \mathbb{R}^+$ is the quantile learning rate and $\epsilon \in (0, 0.5)$ bounds κ_t away from the extremes to ensure numerical stability.

During reverse diffusion at timestep $u \in \{1, \ldots, T\}$, let $\mu_\theta(\mathbf{x}_t, t) \in \mathbb{R}^d$ and $\sigma_t \in \mathbb{R}^+$ be the standard DDPM mean and scale parameters, where $\mathbf{x}_t \in \mathbb{R}^d$ represents the noisy state at step t and d is the dimensionality of the time series. We define the quantile regression loss:

$$\mathcal{L}_q = \frac{1}{w} \sum_{i=1}^{w} \left[\kappa_t \cdot \max(r_{t-i}, 0) + (1 - \kappa_t) \cdot \max(-r_{t-i}, 0)\right]. \tag{15}$$

The quantile loss $\mathcal{L}_q$ penalizes asymmetric errors based on the current quantile target. We then sample from the guided distribution:

$$\mathbf{x}_{t-1} \sim \mathcal{N}\left(\mu_\theta(\mathbf{x}_t, t) + \sigma_t \cdot \kappa_t \cdot \nabla_{\mathbf{x}_t} \mathcal{L}_q, \sigma_t^2 \mathbf{I}\right), \tag{16}$$

$\nabla_{\mathbf{x}_t} \mathcal{L}_q \in \mathbb{R}^d$ is the gradient of the quantile loss with respect to the current noisy state, and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. This gradient term nudges samples toward the desired quantile level while preserving the stochastic nature of the diffusion process.

Algorithm 1 AQG for Multi-Horizon Forecasting

Require: Horizons H, buffer length w, decay ρ, learning rate η, clip bound ϵ, initial quantile κ_0

for $h = 1$ to H **do**
  $\kappa^{(h)} \leftarrow \kappa_0$
  Initialize residual buffer $r_{t-i}^{(h)} \leftarrow 0$ for $i = 1, \ldots, w$
end for
for $t = 1$ to T_{pred} **do**
  for $h = 1$ to H **do**
    Generate forecast $\hat{y}_t^{(h)}$ using quantile $\kappa^{(h)}$
    Compute residual: $r_t^{(h)} \leftarrow y_t^{(h)} - \hat{y}_t^{(h)}$
    Update buffer and compute weighted exceedance ratio:

$$q_t^{(h)} \leftarrow \frac{\sum_{i=1}^{w} \rho^{i-1} \cdot \mathbf{I}(r_{t-i}^{(h)} > 0)}{\sum_{i=1}^{w} \rho^{i-1}}$$

    Update quantile: $\kappa^{(h)} \leftarrow \text{clip}(\kappa^{(h)} + \eta \cdot (q_t^{(h)} - \kappa^{(h)}), \epsilon, 1 - \epsilon)$
  end for
end for

C. Multi-Dataset Pre-Training and Fine-Tuning

We now describe how we combine M heterogeneous source datasets to pre-train the diffusion denoiser, and how to carry those learned parameters forward into fine-tuning on a specific target domain. Let $\{D_m\}_{m=1}^{M}$ be M source domains (e.g. ILI, COVID-19, RSV). For each domain D_m, let $\mathbf{x}_0^{(m)} \in \mathbb{R}^{L \times C}$ be a time-series sample of length L; here L is the length of each input window and C the number of channels.

In our setting we build purely autoregressive inputs by stacking $C-1$ lagged copies of the univariate series alongside its current window. Concretely, for each domain D_m, a sample is

$$\mathbf{x}_0^{(m)} = \begin{bmatrix} y_{t-L+1} & y_{t-L+2} & \cdots & y_t \\ y_{t-L} & y_{t-L+1} & \cdots & y_{t-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{t-L-C+2} & y_{t-L-C+3} & \cdots & y_{t-C+1} \end{bmatrix} \in \mathbb{R}^{L \times C},$$

In addition to these autoregressive channels, we include a simple sinusoidal time-of-year encoding to capture annual seasonality:

$$\mathbf{s}_t = \left[\sin(2\pi t/52),\ \cos(2\pi t/52) \right]^\top \in \mathbb{R}^2,$$

which we append as two constant channels across the entire L-step window when seasonal effects are important.

We also learn a p-dimensional embedding for each domain:

$$e_m \in \mathbb{R}^p,$$

and condition the denoiser on both the noisy input and the domain embedding:

$$\epsilon_\theta(\mathbf{x}_t, t, e_m) : \mathbb{R}^{L \times C} \times \{1, \ldots, T\} \times \mathbb{R}^p \longrightarrow \mathbb{R}^{L \times C}. \quad (17)$$

Stage 1: Multi-Dataset Pre-Training. We optimize

$$\mathcal{L}_{\text{pre}} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\substack{\mathbf{x}_0^{(m)} \sim D_m \\ t \sim [1,T],\, \epsilon \sim \mathcal{N}(0,I)}} \Big\| \epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0^{(m)} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\ t,\ e_m\big) \Big\|^2 \quad (18)$$

This yields shared denoiser parameters θ_{pre} and domain embeddings $\{e_m\}$ that capture broad temporal patterns and noise structures across all M source datasets.

Stage 2: Fine-Tuning with Adaptive Quantile Guidance. We initialize the model parameters $\theta \leftarrow \theta_{\text{pre}}$ and retain only a single embedding e_{tar} for the target domain. We then optimize

$$\mathcal{L}_{\text{fine}} = \underbrace{\mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\ t,\ e_{\text{tar}}\right) \right\|^2}_{\mathcal{L}_{\text{diff}}} + \underbrace{\gamma\, \mathcal{L}_q}_{\substack{\text{AQG pinball} \\ \text{loss from Sec. III-B}}}, \quad (19)$$

where $\mathcal{L}_{\text{diff}}$ is the usual denoising-diffusion regression loss, now conditioned on e_{tar}. $\mathcal{L}_q$ is the quantile (pinball) loss over the residual buffer, enforcing calibration at the current κ_t. $\gamma > 0$ balances the strength of the AQG calibration term.

During fine-tuning: Every reverse-diffusion step is conditioned on e_{tar}. In parallel, κ_t is updated online via AQG (Sec. III-B) using the buffer of one-step residuals on the target series. Each sampling step is guided by the gradient of $\mathcal{L}_q$ at the current κ_t.

IV. Experiments

We evaluate our Adaptive Quantile Guidance (AQG) framework on a diverse collection of epidemiological and related time series, to demonstrate its generality across different domains and geographic contexts. We compare our *adaptive quantile guidance* versus *fixed quantile*, and other state-of-the-art baseline models for time series forecasting, and also assess the benefit of *multi-dataset pre-training*. Additionally, we analyze the *nonmonotonic* forecasting behavior across different horizons and perform few-shot training to validate the significance of our multi-dataset pre-training over a strong baseline.

A. Dataset Scope

To evaluate our approach across a spectrum of temporal dynamics, we first curate three core respiratory-disease series from the U.S. Centers for Disease Control and Prevention (CDC). The COVID-19 dataset (2020–2023) features rapid surges and declines; we use look-back windows of $\{1, 2, 3, 4, 5, 6\}$ weeks to capture its abrupt shifts. The Influenza-Like Illness (ILI) series (2002–2020) exhibits strong annual seasonality; we employ look-backs of $\{6, 12, 24, 36, 48, 60\}$ weeks to encompass both yearly and multi-year patterns. The Respiratory Syncytial Virus (RSV) data (2010–2020) follows a biennial cycle; we again use $\{6, 12, 24, 36, 48, 60\}$-week windows to resolve its two-year

periodicity. Each of these time series spans multiple years to ensure exposure to both baseline fluctuations and outbreak peaks.

To test geographic and pathogen-type generalization and transferability, we augment our evaluation with four additional datasets: national COVID-19 case counts from Japan, Canada, and Italy, and a Dengue hospitalization series in the United States. These extensions introduce a different temporal and seasonal profiles, challenging the model to transfer learned diffusion dynamics and adaptive calibration across locations and diseases.

- **COVID-19 Japan / Canada / Italy:** Weekly case counts from national health agencies.
- **Dengue (USA):** Weekly counts of Dengue hospitalizations, capturing a different pathogen and seasonal profile.

All series are min–max scaled to [0,1]. We split each dataset 80/20 chronologically and use rolling 4-week forecast windows to evaluate how well AQG adapts to shifting uncertainty. We report MAE for point accuracy and SMAPE for scale-independent error. Baselines include TSDiff, D-Linear, N-Linear, PatchTST, Informer, ARIMA(6,0,4), and MOMENT. To isolate effects, we compare our full Pretrain+AQG model against two ablations: one without adaptive quantiles and one without pre-training. A detailed calibration analysis (e.g. CRPS) is left for future work.

B. Few-Shot Performance

We assess full model performance when only 20%, 40%, or 60% of each target series is available for training (the remainder used for testing). Table I reports MAE at a 4-week horizon on three held-out test domains.

TABLE I: Few-Shot MAE (4-week horizon) under varying train splits

Method	20%	40%	60%	100%
COVID-19 (USA)				
Ours	2.458	1.583	1.263	0.134
MOMENT	2.589	2.076	1.842	0.224
COVID-19 (Japan)				
Ours	2.187	1.846	0.947	0.118
MOMENT	2.374	2.035	1.392	0.173
Dengue (USA)				
Ours	2.013	1.729	0.812	0.095
MOMENT	2.218	1.947	1.158	0.206

Across all splits, our pretraining+AQG model outperforms MOMENT, especially in the low-data (20% and 40%) regimes, illustrating the benefit of transfer learning combined with adaptive calibration.

C. Ablation: Pretraining vs. AQG

Table II isolates the contributions of multi-dataset pretraining and AQG on the 100% training split (4-week horizon MAE).

Removing pre-training or AQG degrades performance, confirming that both components are essential for best results.

TABLE II: Ablation of Pretraining and AQG (Average MAE)

Variant	USA	Canada	Dengue	Japan
No Pretraining	0.151	0.264	0.149	0.139
No Adaptive Quantile	0.146	0.122	0.125	0.124
Full Model	0.134	0.118	0.095	0.105

D. Nonmonotonic Horizon Behavior

Unlike most forecasting methods whose errors grow with lead time, our diffusion-based model can exhibit *nonmonotonic* error patterns. For instance, as Table III shows, the ILI SMAPE is 32.48 at a 48-week horizon but then falls to 30.71 at 60 weeks. This dip shows the model's capacity to leverage seasonality and lag effects, making some mid-range forecasts more accurate than shorter-term ones, a phenomenon also observed in other score-based generators [27]. In practice, this means we must assess performance across all horizons rather than assuming error always increases with forecast length.

V. Results

Our Pretrain+AQG model consistently outperforms all baselines on COVID-19, ILI and RSV series (Table III). For example, at a 6-week horizon on COVID-19 it achieves 0.102 MAE versus 0.171 for TSDiff and 0.141 for Pretrain-only—a 40% relative drop. On ILI, it cuts SMAPE to 26.75% compared to 36.72% for TSDiff and 29.07% for PatchTST. These gains hold across short and long horizons, especially during volatile periods, this confirms that both multi-dataset pre-training and adaptive quantile guidance are critical. Ablations show removing either component degrades performance, and we even observe nonmonotonic error trends (e.g. a dip in SMAPE at horizon 24 for ILI), showing how dynamic quantile steering lets the diffusion backbone capture complex seasonal and lagged patterns.

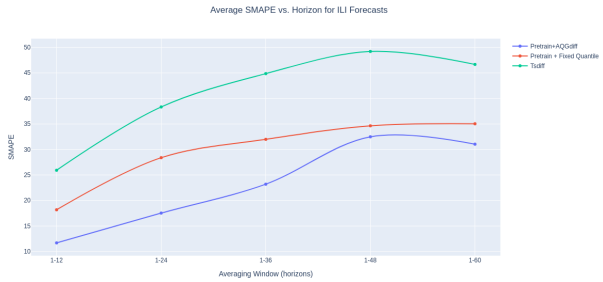
A. Additional Results on Geographic Extensions

Table IV reports 1–6-step MAE on COVID-19 (Japan, Canada) and Dengue (USA), plus the average MAE across all 18 horizons. The lowest average (best) is highlighted in **blue bold**, the second-lowest in **green bold**, and the highest (worst) in **red bold**.

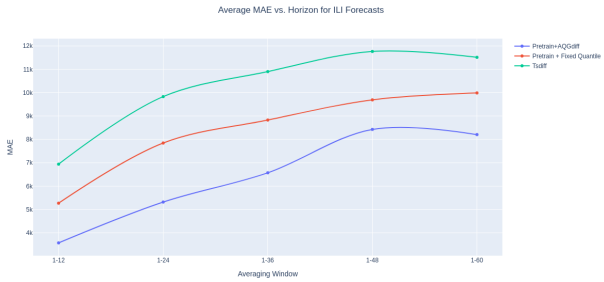
Our model consistently outperforms these strong baselines across horizons and datasets, showing it's ability to transfer knowledge to new geographies and diseases.

VI. Related Work

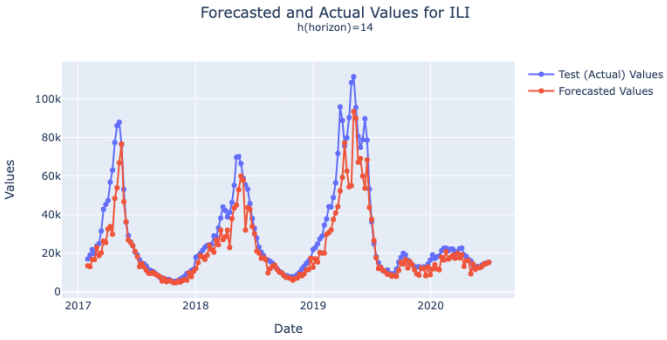
Modern time series forecasting research has advanced along three major models: (1) foundation model adaptation, (2) probabilistic generative methods, and (3) specialized neural architectures. In this section, we provide a comprehensive review of key contributions in each area to highlight their strengths, limitations, and the open challenges that motivate our work.



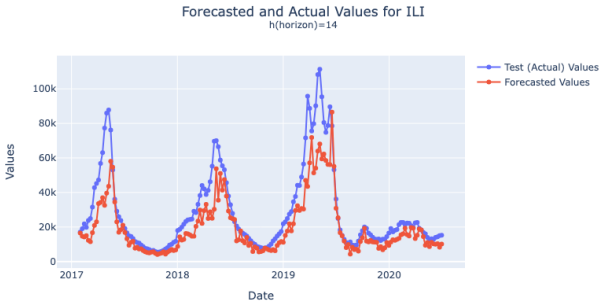
(a) Average SMAPE scores for diffusion baselines



(b) Average MAE scores for diffusion baselines



(c) forecast for 14 weeks ahead using Rolling Evaluation AQG + pretrain



(d) forecast for 14 weeks ahead using Rolling Evaluation (TSdiff

Dataset	Horizon	Tsdiff		D-Linear		N-Linear		PatchTST		Informer		ARIMA(6, 0, 4)		Pretrain Alone		Pretrain+AQGdiff	
		MAE	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
COVID-19	1	0.205	21.43	0.293	15.76	0.268	13.01	0.285	15.82	0.222	14.76	0.216	16.83	0.148	17.63	0.122	13.10
	2	0.155	17.78	0.443	21.42	0.374	18.88	0.398	18.98	0.348	24.24	0.380	24.73	0.119	13.05	0.097	10.74
	3	0.135	13.76	0.541	27.79	0.559	29.51	0.538	25.68	0.249	13.17	0.481	31.88	0.121	13.11	0.128	11.95
	4	0.269	23.67	0.652	36.06	0.725	39.69	0.716	36.95	0.442	25.59	0.603	39.78	0.151	15.42	0.129	12.58
	5	0.265	23.24	0.720	40.80	0.869	49.61	0.849	41.66	0.629	36.20	0.646	41.76	0.125	13.72	0.104	11.86
	6	0.171	15.08	0.805	48.05	1.012	58.55	1.017	55.50	0.623	35.84	0.766	49.66	0.141	11.26	0.102	11.08
	Avg	0.20	19.15	0.576	31.65	0.635	34.87	0.634	32.43	0.419	24.97	0.515	34.10	0.134	14.03	0.113	11.88
ILI	6	0.158	15.41	0.811	33.57	0.981	21.99	0.494	28.41	0.981	33.95	0.330	24.77	0.131	12.27	0.115	10.82
	12	0.247	25.93	0.916	36.06	0.744	32.57	0.613	27.99	1.201	42.57	0.548	42.53	0.187	18.19	0.127	11.70
	24	0.344	38.33	1.021	36.65	0.901	37.16	0.761	31.77	1.287	59.76	0.843	61.92	0.274	28.39	0.186	17.54
	36	0.389	44.86	1.024	35.33	0.902	34.79	0.731	29.67	1.379	61.10	0.962	67.81	0.315	31.99	0.235	23.20
	48	0.423	49.20	1.068	36.24	0.925	35.50	0.787	31.20	1.540	62.49	0.956	63.64	0.348	34.64	0.304	32.48
	60	0.411	46.66	1.097	37.87	0.978	37.80	0.856	31.82	1.495	61.52	0.866	55.32	0.356	35.04	0.294	30.71
	Avg	0.329	36.72	0.989	35.95	0.836	34.00	0.707	29.07	1.313	53.56	0.750	52.66	0.268	26.75	0.210	21.07
RSV	6	0.274	28.78	0.352	37.14	0.263	34.98	0.244	26.76	0.284	35.89	0.285	30.06	0.196	25.21	0.173	23.91
	12	0.483	51.57	0.658	67.21	0.806	76.67	0.471	57.44	0.573	69.77	0.461	47.56	0.433	47.92	0.359	46.86
	24	0.751	77.79	0.923	84.93	0.972	88.91	0.815	78.01	0.728	86.49	0.702	67.97	0.639	65.39	0.643	61.46
	36	0.716	72.99	0.864	73.35	0.922	77.44	0.711	77.98	0.847	89.48	0.890	74.43	0.827	79.88	0.711	65.30
	48	0.799	75.60	0.838	79.37	0.808	85.20	0.822	84.90	0.826	78.63	0.798	73.94	0.831	77.58	0.716	67.53
	60	0.669	69.60	0.936	88.47	0.908	85.89	0.804	80.10	0.990	88.55	0.780	73.72	0.931	79.36	0.818	78.83
	Avg	0.620	62.72	0.761	71.74	0.779	74.84	0.644	67.53	0.708	74.80	0.652	61.28	0.642	62.56	0.57	57.31

TABLE III: Forecasting Results (MAE and SMAPE) for COVID-19, ILI, and RSV datasets. In each average row, the lowest average is shown in bold blue and the second lowest is green, with the worst performance shown in red.

Method	Japan						Avg	Canada						Avg	Dengue (USA)						Avg
	H1	H2	H3	H4	H5	H6		H1	H2	H3	H4	H5	H6		H1	H2	H3	H4	H5	H6	
PatchTST	0.247	0.286	0.312	0.433	0.478	0.391	**0.358**	0.273	0.306	0.223	0.384	0.417	0.492	**0.349**	0.237	0.332	0.284	0.378	0.409	0.455	**0.349**
MOMENT	0.128	0.187	0.211	0.294	0.263	0.143	**0.204**	0.134	0.192	0.256	0.274	0.146	0.218	**0.203**	0.184	0.119	0.258	0.134	0.276	0.298	**0.212**
Ours (Full)	0.159	0.115	0.112	0.091	0.083	0.153	**0.119**	0.153	0.118	0.094	0.108	0.095	0.067	**0.106**	0.134	0.128	0.058	0.089	0.052	0.106	**0.095**

TABLE IV: Additional Results (MAE) on Japan, Canada, and Dengue (USA) at horizons 1–6, with average MAE per dataset.

A. Foundation Models for Temporal Reasoning

Recent work adapts large pre-trained models (LLMs and vision-language architectures) for temporal tasks. [28] shows frozen LLMs can be effectively repurposed for time series forecasting via learned tokenization strategies, leveraging pre-trained representations without extensive fine-tuning. [29] reveals LLMs exhibit strong few-shot forecasting capabilities without task-specific training. Vision-language models have also been adapted, with [30] introducing patch-based tokenization that treats time series segments as "words" to capture long-range dependencies.

Limitations: Despite their promise, foundation models face significant challenges. First, they often struggle with precise temporal alignment, as demonstrated by [31] which identified spectral gaps in financial data modelling. Second, hybrid approaches like [32] that combine parameter-efficient tuning with quantization typically lack robust mechanisms for uncertainty quantification. These limitations hinder their applicability in dynamic environments where uncertainty estimation is critical.

B. Probabilistic Forecasting via Generative Models

Diffusion models have emerged as a leading approach for probabilistic forecasting, modeling complex temporal distributions through iterative denoising. Recent innovations include transformer-based denoising for scalability ([33]), hierarchical noise schedules for multi-scale dependencies ([34]), and training stabilization techniques like spectral normalization ([35]). Additional advances include adaptive layer normalization for scaling ([56]) and iterative refinement for accuracy ([11]).

Applications span time series imputation ([58]), anomaly detection ([38]), spatio-temporal modeling ([39]), and non-stationary data handling through learnable noise schedules ([11]). Conditional forecasting methods integrate hierarchical mechanisms ([41], [42]), while efficiency improvements include latent space compression ([43]) and high-frequency preservation techniques ([44], [45]).

Limitations: Diffusion methods face challenges including reliance on rigid noise priors that limit robustness under distribution shifts, and high computational overhead that hinders real-time deployment ([36]).

Alternative generative approaches include GANs handling irregular sampling ([36]), VAEs combined with temporal normalizing flows ([47]), and emerging techniques like energy-based models and implicit quantile networks, though these remain less mature and computationally intensive.

C. Specialized Forecasting Architectures

Task-specific architectures have been designed to capture unique temporal dependencies and facilitate interpretability. The [48] pioneered attention-based feature selection for multistep prediction, allowing the model to focus on relevant historical data points. [49] introduced a modular architecture that supports cross-domain generalization, making it applicable to various forecasting tasks. Patch-based tokenization, as introduced in [30], treats time series segments as discrete tokens, improving long-range dependency capture. [50] utilizes optimal transport methods to align multivariate time series, enhancing the model's ability to capture cross-channel dependencies.

Limitations: Despite their strengths, many specialized architectures suffer from scalability issues and require full retraining to handle concept drift. Studies like [34] report accuracy degradations of 12–15% over time, highlighting the need for more adaptive and computationally efficient solutions.

Our work addresses these challenges by integrating adaptive quantile guidance with transfer learning through pre-training on multiple datasets. This approach enables the model to dynamically adjust prediction intervals based on local residual patterns using knowledge transferred across domains. When we combine these approaches, we provide a unified framework for dynamic, uncertainty-aware forecasting that avoids the computational burdens of full retraining and adapts efficiently to evolving data distributions.

VII. Conclusions and Future Work

We proposed a diffusion-based forecaster that leverages multi-dataset pre-training and an adaptive quantile guidance loop. Across COVID-19, Dengue, ILI and RSV series, our approach cuts MAE by 26–33% versus fixed quantiles and boosts accuracy by 28–45% compared to training from scratch, even yielding surprisingly lower errors at some mid-range horizons.

Looking ahead, we aim to slash inference costs, by exploring latent-space diffusion or distilled denoisers and to scale to ultra-long horizons with sparse attention. We also plan to weave in external covariates (e.g. vaccination rates, mobility metrics, interventions) to improve causal interpretability and better connect our forecasts to real-world policy actions.

References

[1] C. Viboud *et al.*, "Rapid assessment of pandemic potential of emerging influenza viruses," *Proc. Natl. Acad. Sci.*, 2018.

[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2018.

[3] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks," *Int. J. Forecasting*, 2020.

[4] H. Wen *et al.*, "Multi-Scale LSTM for Time Series Forecasting," in *Proc. AAAI*, 2017.

[5] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," in *Proc. ICML*, 2015.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *NeurIPS*, 2020.

[7] A. Rasul *et al.*, "Autoregressive Diffusion Models for High-Fidelity Image Generation," in *ICLR*, 2021.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT*, 2019.

[9] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners," OpenAI Tech. Rep., 2019.

[10] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NeurIPS*, 2021.

[11] A. Kollovieh *et al.*, "Predict, Refine, Synthesize, Self-Guiding: Diffusion Models for Time Series Forecasting," in *UAI*, 2023.

[12] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," arXiv:2005.14165, 2020.

[13] K. He *et al.*, "Masked Autoencoders Are Scalable Vision Learners," in *CVPR*, 2022.

[14] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021.

[15] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," arXiv:2207.12598, 2022.

[16] C. Saharia *et al.*, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," arXiv:2205.11487, 2022.

[17] W. Feller, *An Introduction to Probability Theory and Its Applications*. Wiley, 1968.

[18] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *NeurIPS*, 2014.

[19] C. Lu *et al.*, "Guided Diffusion for Fast Inverse Design," *Nat. Mach. Intell.*, 2023.

[20] A. Gu *et al.*, "Efficiently Modeling Long Sequences with Structured State Spaces," in *ICLR*, 2022.

[21] CDC, "FluView: Influenza-Like Illness Data," https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, accessed Month Day, Year.

[22] CDC, "COVID-19 National Data Tracker," https://covid.cdc.gov/covid-data-tracker, accessed Month Day, Year.

[23] CDC, "Respiratory Syncytial Virus National Trends," https://www.cdc.gov/rsv/research/rsv-net/dashboard.html, accessed Month Day, Year.

[24] Y. Tashiro, T. Matsubara, S. Wulff *et al.*, "CSDI: Conditional Score-based Diffusion Models for Probabilistic Time-Series Imputation," in *NeurIPS*, 2021.

[25] Y. Nie, Z. Liu, Y. Wang *et al.*, "A Patching Strategy for Time Series Transformer," in *ICLR*, 2023.

[26] H. Zhou, S. Zhang, J. Peng *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," in *AAAI*, 2021.

[27] G. Daras, Y. Dagan, A. Dimakis, and C. Daskalakis, "Consistent Diffusion Models: Mitigating Sampling Drift by Learning to be Consistent," in *NeurIPS*, 2023.

[28] M. Jin *et al.*, "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models," arXiv:2310.01728, 2024.

[29] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large Language Models Are Zero-Shot Time Series Forecasters," arXiv:2310.07820, 2024.

[30] Y. Nie, N. H. Nguyen, P. Sinthong *et al.*, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," arXiv:2211.14730, 2023.

[31] J. Yang *et al.*, "Generative Pre-Trained Diffusion Paradigm for Zero-Shot Time Series Forecasting," arXiv:2406.02212, 2024.

[32] A. Das, M. Faw, R. Sen, and Y. Zhou, "In-Context Fine-Tuning for Time-Series Foundation Models," arXiv:2410.24087, 2024.

[33] D. Cao, W. Ye, Y. Zhang, and Y. Liu, "General-purpose Diffusion Transformers for Time Series Foundation Model," arXiv:2409.02322, 2024.

[34] L. Shen, W. Chen, and J. Kwok, "Multi-Resolution Diffusion Models for Time Series Forecasting," in *ICLR*, 2024.

[35] T. B. Ifriqi *et al.*, "On Improved Conditioning Mechanisms and Pre-training Strategies for Diffusion Models," arXiv:2411.03177, 2025.

[36] H. Huang, M. Chen, and X. Qiao, "Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns," in *ICLR*, 2024.

[37] M. Letafati, S. Ali, and M. Latva-aho, "Conditional Denoising Diffusion Probabilistic Models for Data Reconstruction Enhancement in Wireless Communications," arXiv:2310.19460, 2024.

[38] A. Mousakhan, T. Brox, and J. Tayyub, "Anomaly Detection with Conditioned Denoising Diffusion Models," arXiv:2305.15956, 2023.

[39] H. Wen *et al.*, "DiffSTG: Probabilistic Spatio-Temporal Graph Forecasting with Denoising Diffusion Models," arXiv:2301.13629, 2024.

[40] Y.-J. Lu *et al.*, "Conditional Diffusion Probabilistic Model for Speech Enhancement," arXiv:2202.05256, 2022.

[41] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting," arXiv:2101.12072, 2021.

[42] X. Han, H. Zheng, and M. Zhou, "CARD: Classification and Regression Diffusion Models," arXiv:2206.07275, 2022.

[43] S. Feng, C. Miao, Z. Zhang, and P. Zhao, "Latent Diffusion Transformer for Probabilistic Time Series Forecasting," *Proc. AAAI*, vol. 38, no. 11, pp. 11979–11987, 2024.

[44] N. Chen *et al.*, "WaveGrad: Estimating Gradients for Waveform Generation," arXiv:2009.00713, 2020.

[45] T. Yan, H. Zhang, T. Zhou, Y. Zhan, and Y. Xia, "ScoreGrad: Multivariate Probabilistic Time Series Forecasting with Continuous Energy-based Generative Models," arXiv:2106.10121, 2021.

[46] S. R. Cachay, B. Zhao, H. Joren, and R. Yu, "DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting," arXiv:2306.01984, 2023.

[47] Y. Li, X. Lu, Y. Wang, and D. Dou, "Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement," arXiv:2301.03028, 2023.

[48] B. Lim, S. O. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," arXiv:1912.09363, 2020.

[49] S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik, "UniTS: A Unified Multi-Task Time Series Model," arXiv:2403.00131, 2024.

[50] J. Chen *et al.*, "From Similarity to Superiority: Channel Clustering for Time Series Forecasting," arXiv:2404.01340, 2024.

[51] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[52] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?" arXiv:2205.13504, 2022.

[53] H. Zhou, S. Zhang, J. Peng *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," arXiv:2012.07436, 2021.

[54] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A Family of Open Time-series Foundation Models," arXiv:2402.03885, 2024.

[55] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley, 2015.

[56] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," *arXiv preprint arXiv:2212.09748*, 2023. [Online]. Available: https://arxiv.org/abs/2212.09748

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[58] Y. Tashiro, D. Yoon, and Y. Kim, "CSDI: Conditional Score-Based Diffusion Models for Probabilistic Time Series Imputation," *arXiv preprint arXiv:2107.03502*, 2021. [Online]. Available: https://arxiv.org/abs/2107.03502