Balancing Confidence and Precision: A Framework for Real-Time Bus Arrival Time Prediction with Uncertainty Quantification

Beiyu Song[a,*], Changlin Li[a], Edward Chung[a] and Hongbo Ye[a]

[a] <Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic Univerisity>, <Hong Kong SAR>, <China>
beiyu10.song@connect.polyu.hk, changlin.li@connect.polyu.hk, edward.cs.chung@polyu.edu.hk, hongbo.ye@polyu.edu.hk
* Corresponding author

January 19, 2025

Keywords: Bus arrival time prediction, Uncertainty-aware prediction, Distributional-free prediction interval generation

Abstract

Accurate and reliable real-time bus arrival time (BAT) predictions are crucial for improving passenger satisfaction and operational efficiency. Existing pointwise BAT prediction models have demonstrated their effectiveness in estimating single values close to the true arrival time. However, there is a lack of research on quantifying the uncertainties associated with these predictions, which is essential for better passenger planning and enhancing the credibility and reliability of bus operators. This paper introduces UncertBAT, a novel framework designed to address this gap. UncertBAT provides not only the predicted BAT but also an arrival time window with a high degree of confidence. The model incorporates conformalized quantile regression and a grouping calibration mechanism to address challenges posed by data skewness and variability, ensuring an optimal balance between prediction confidence and precision. Several experiments conducted in the study demonstrate the model's effectiveness in BAT prediction.

1 INTRODUCTION

Accurate information on bus arrival times is beneficial for both bus operators in implementing effective control strategies of bus systems and passengers in planning their journeys (Altinkaya & Zontul, 2013). For continuously operating bus services, it is essential to provide up-to-the-minute updates on the predicted arrival times at every bus stop.

One significant limitation of current online bus arrival time prediction methods (Li *et al.*, 2023, Petersen *et al.*, 2019) is that they only provide a single value, which brings significant inconvenience for passengers. For example, since these predictions are not always accurate, passengers are likely to miss the bus if the predicted time is later than the actual arrival. This will lead to a prolonged waiting time, especially on low-frequency routes. Furthermore, knowing the earliest and latest expected arrival times allows passengers to manage their schedules more flexibly. Despite extensive research on online prediction, few efforts have been dedicated to quantifying the uncertainty.

From the passengers' perspective, it is beneficial to know the bus's arrival time at each stop along a route, as well as quantify the uncertainty of these predictions, which can be represented as the probability of the bus arriving within a specific time range, defined as the arrival time window (ATW). Although uncertainty quantification has been widely researched in other transportation-related fields (Tang *et al.*, 2022, Zhang & Mahadevan, 2020, Liang *et al.*, 2023), its application in bus arrival time prediction is quite novel and exhibits two challenges: (i) Based on a bus's

current location, its arrival times at future bus stops show huge heterogeneity. For example, for bus stops further along the route, arrival times will exhibit greater variation and uncertainty. (ii) There is a trade-off between achieving high-confidence predictions (certainty) and maintaining a narrow arrival time window (precision) (Yaniv & Foster, 1995). A high confidence requires a broader arrival time window, which can reduce the precision of the prediction and result in lower practicality. Conversely, a more precise prediction requires a narrower arrival time window, which increases the risk of passengers missing the bus. It is challenging to adjust a flexible time window optimized for both certainty and practical usage, especially for arrival times that vary widely and share skewness along multi-step prediction horizons.

To this end, we propose an **uncert**ainty-aware **b**us **a**rrival **t**ime prediction framework (UncertBAT) to predict the BATs with quantified uncertainty metric. The main contributions are summarized as follows:

- We propose a framework for real-time BAT prediction with uncertainty quantification that accounts for varying prediction horizons. This framework can generate adjusted arrival time windows for different practical usages.
- We develop a quantile regression with monotonicity for sequential prediction of BATs to ensure the predicted values strictly follow the increasing order of quantiles. It provides initial values for either single-value predictions or arrival time windows for further adjustment.
- We introduce a grouping calibration mechanism to refine the windows by narrowing the lower and upper boundaries. It optimizes the balance between high confidence and narrow arrival time windows.

Table 1 – *Important notations.*

Notation	Description
$\hat{T}_\alpha$	αth quantile prediction for BAT
$\hat{\mathcal{T}}_k^{low}, \hat{\mathcal{T}}_k^{up}$	initial lower and upper boundaries for group k
$\mathcal{T}_k^{low}, \mathcal{T}_k^{up}$	adjusted lower and upper boundaries
S_k^{low}, S_k^{up}	conformal scores for group k
s_m^{low}, s_n^{up}	lower and upper adjustment scores

2 Preliminaries

Please refer to Table 1 for the notations used in this paper.

- Definition 1: Trips. A trip is defined as a journey of a bus that follows a designated path from one designated location to another designated location, covering K bus stops.
- Definition 2: Arrival Time Window (ATW). ATW is defined as a time interval with lower bound T^{low} and upper bound T^{up} for bus arrival times, i.e., $\mathcal{ATW} = \left[T^{\text{low}}, T^{\text{up}}\right]$.
- Definition 3: Coverage Rate (CR). CR is defined as the probability of an actual arrival time T falling within the ATW. CR measures how confident/reliable the window is.

2.1 Uncertainty-aware Quantification

At time t, given the same inputs as above, the objective is to predict the lower and upper boundaries of the arrival time $\left(\mathcal{T}^{\text{low}}, \mathcal{T}^{\text{up}}\right) \in \mathbb{R}^{k\times 2}$, defined as the ATW. The following are the details of the two uncertainty-aware tasks.

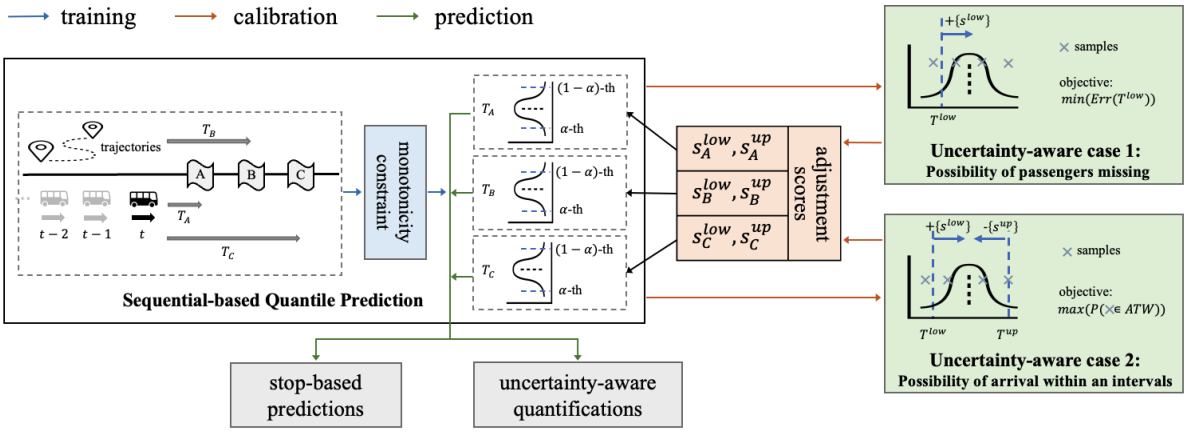


Figure 1 – *The architecture of UncertBAT.*

Uncertainty-aware Task 1: Possibility of Passengers Missing. At time t, given the trajectory of a bus trip between stop $K-k-1$ and stop $K-k$, and a predefined maximum possibility p_{miss} of missing the bus, the task is to predict the lower boundaries of the arrival time $\mathcal{T}^{low} = \{T^{low}_{K-k}, T^{low}_{K-k+1}, \ldots, T^{low}_{K}\} \in \mathbb{R}^k$. The objectives are: (i) minimizing the difference between coverage rate $\mathcal{CR}$ and $1-p_{miss}$, (ii) minimizing the gap between $\mathcal{T}^{low}$ and the true arrival time $\mathcal{T}$.

Uncertainty-aware Task 2: Possibility of Bus Arriving Within an Interval. Given a desired width of arrival time window i_{arrival}, the task is to predict both the lower and upper boundaries of the arrival time $\left(\mathcal{T}^{low}, \mathcal{T}^{up}\right) = \left\{\left(T^{low}_{K-k}, T^{up}_{K-k}\right), \left(T^{low}_{K-k+1}, T^{up}_{K-k+1}\right), \ldots, \left(T^{low}_{K}, T^{up}_{K}\right)\right\} \in \mathbb{R}^{k\times 2}$. The objectives are: (i) minimizing the difference between the with of the arrival time window $\mathcal{ATW}$ and i_{arrival}, (ii) maximizing the coverage rate $\mathcal{CR}$.

3 Model Framework

In this section, we introduce the technical details of the UncertBAT framework, as illustrated in Fig. 1.

3.1 Sequential-based Quantile Prediction for Bus Arrival Time

3.1.1 Sequential Embedding

Sequential patterns are crucial for generating rich information from a bus's trajectory. We use the Echo State Networks (ESNs) (Lukoševičius & Jaeger, 2009) to embed sequential features for each trajectory because the fixed-weight-based reservoir in ESNs is more computationally efficient than traditional Recurrent Neural Networks (RNNs). The state is updated recurrently as follows:

$$h(t) = \mathbf{tanh}\left(W_1 x(t) + W h(t-1) + b\right), \tag{1}$$

where $x(t) \in \mathbb{R}^D$ is the input denoting the bus's current location (we use the time elapsed from the previous stop $K-k-1$ along with the stop information); W_1 and b are learnable weight and bias parameters; W is a sparse and fixed matrix; and **tanh** is the activation function. Then, the high-dimensional embedding of a bus's trajectory at time t is generated from:

$$y(t) = W_2 h(t), \tag{2}$$

where W_2 is a weight matrix.

3.1.2 Quantile Prediction with Monotonicity

After generating the sequential embedding from ESNs, we incorporate contextual features that reveal trip-specific or stop-specific information. By concatenating the contextual features ($y_{\text{cxt}}(t)$) with the sequential embedding $y(t)$, the arrival time to k stops ahead can be predicted using:

$$\hat{T} = \mathbf{FFN}\left(y_{\text{cxt}}(t) \oplus y(t)\right) = \hat{\mathcal{Q}}\left(x(t)\right), \tag{3}$$

where **FFN** denotes a fully connected feed-forward network with two layers. Here, we define $\hat{\mathcal{Q}}(\cdot)$ as the quantile prediction function for an observation x.

In the second layer of the feed-forward network, instead of generating a pair of parameters to maximize the similarity between predicted arrival times and the ground truths, we aim to estimate the αth percentile $\hat{T}_\alpha$ for quantifying the possibility of having a point smaller than $\hat{T}_\alpha$ with $\mathbf{P}\left(\hat{T} < \hat{T}_\alpha\right) \approx \alpha$. The values of α and $1-\alpha$ help determine the initial lower and upper boundaries for a predicted arrival time sample. To address the possible crossing prediction for different quantiles, we incorporate a monotonicity constraint into the pinball loss for training:

$$\mathcal{L}\left(T, \left\{\hat{T}_{q_1}, \hat{T}_{q_2}, \ldots, \hat{T}_{q_M}\right\}\right) = \sum_{i=1}^{M} \mathcal{L}_{pin}\left(T, \hat{T}_{q_i}\right) + \lambda \sum_{j=2}^{M} \mathcal{L}_{mon}\left(\hat{T}_{q_{j-1}}, \hat{T}_{q_j}\right). \tag{4}$$

Here, we consider predicting M quantiles in an increasing order ($q_1 < q_2 < \cdots < q_M$). Specifically, $\mathcal{L}_{pin}$ is the pinball loss function (Koenker & Hallock, 2001) defined as:

$$\mathcal{L}_\alpha\left(T, \hat{T}_\alpha\right) = \max\left(\alpha\left(T - \hat{T}_\alpha\right), (1-\alpha)\left(\hat{T}_\alpha - T\right)\right), \tag{5}$$

which penalizes positive and negative residuals with parameters $1-\alpha$ and α, respectively. $\mathcal{L}_{mon}$ is the newly-added monotonicity constraint to penalize situations where the predicted quantiles are not in the correct order, defined as:

$$\mathcal{L}\left(\hat{T}_1, \hat{T}_2\right) = \max\left(0, \hat{T}_1 - \hat{T}_2\right). \tag{6}$$

The overall loss is balanced between the pinball loss and the monotonicity constraint with a parameter λ.

3.2 Grouping Calibration for Uncertainty Estimation

Quantile predictions can provide the lower and upper boundaries for bus arrival time uncertainty quantification. However, strict model specifications will lead to inadequate coverage in finite samples (Steinwart & Christmann, 2011). To achieve a stricter control of the miscoverage rate, conformalized quantile regression (CQR) (Romano *et al.*, 2019) is a way. However, CQR has two limitations when being implemented for our uncertainty quantification for bus arrival time:

- **Global Calibration Issues**. The empirical quantiles of positive and negative residuals are computed from the entire calibration dataset, serving as global constants to adjust initial lower and upper boundaries. However, bus arrival times vary with prediction horizons, which means the predictions for one stop ahead and several stops ahead share different distributions, which cannot be adjusted precisely by a global calibration value.

- **Precision Maintenance**. While selecting the conformal score can ensure the windows meet the desired coverage level, it cannot guarantee a small prediction range to maintain precision, which is crucial for practical application.

To address the above-mentioned two issues of CQR, we group arrival times at K stops into horizon-based categories $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K\}$. Specifically, for stop k, $\mathcal{T}_j$ refers to the arrival time when the bus is traveling between stops $k-j$ and $k-j-1$. The objective is to categorize arrival times into groups with similar variance. Within a specific group k, for a sample $\mathcal{T}_{k,i}$, the initial lower and upper boundaries can be predicted as $\hat{\mathcal{T}}_{k,i}^{\text{low}} = \hat{\mathcal{Q}}_{low}(x_i)$ and $\hat{\mathcal{T}}_{k,i}^{\text{up}} = \hat{\mathcal{Q}}_{up}(x_i)$. We compute the conformal scores for both sides as $S_{k,i}^{\text{low}} = \left(\mathcal{T}_{k,i} - \hat{\mathcal{T}}_{k,i}^{\text{low}}\right)$ and $S_{k,i}^{\text{up}} = \left(\hat{\mathcal{T}}_{k,i}^{\text{up}} - \mathcal{T}_{k,i}\right)$.

The lower and upper adjustment scores are chosen from the candidates as:

$$s^{\text{low}} = \left\{\mathbf{EQ}\left(S_k^{\text{low}}, m\right) \mid m \in \{1, 2, \ldots, M\}\right\}, \tag{7}$$

and

$$s^{\text{up}} = \left\{\mathbf{EQ}\left(S_k^{\text{up}}, n\right) \mid n \in \{1, 2, \ldots, N\}\right\}. \tag{8}$$

Here, $\mathbf{EQ}(\cdot)$ is the empirical quantile function that selects the mth M-quantile or nth N-quantile from the conformal score lists S_k^{low} and S_k^{up} grouped by k stops ahead. The lower and upper boundaries are adjusted as: $\mathcal{T}_{k,i}^{\text{low},m} = \hat{\mathcal{T}}_{k,i}^{\text{low}} + s_m^{\text{low}}$, and $\mathcal{T}_{k,i}^{\text{up},n} = \hat{\mathcal{T}}_{k,i}^{\text{up}} - s_n^{\text{up}}$. The values of the lower and upper adjustment scores for group k are decided to form a window shaped by $\left[\mathcal{T}_{k,i}^{\text{low},m}, \mathcal{T}_{k,i}^{\text{up},n}\right]$ with different optimization objectives in the following section.

3.3 Uncertainty-aware Quantification

After getting the predicted quantiles through Eq. (4) and the lower and upper adjustment scores, we can utilize the score candidates to adjust the boundaries for uncertainty-aware tasks.

Uncertainty-aware Task 1: Probability of Passengers Missing the Bus. It can be interpreted as predicting a lower boundary while making testing samples have a predefined p_{miss} probability to be larger than the boundary. The optimization intends to minimize the width between predicted early arrival and actual arrival times:

$$\begin{aligned} \underset{m}{\text{minimize}} \quad & \mathcal{ATW}\left(\mathcal{T}_{k,i}^{\text{low},m}, \mathcal{T}\right) && \text{(9a)} \\ \text{subject to:} \quad & \mathcal{CR}\left(\mathcal{T}_{k,i}^{\text{low},m}, +\infty\right) \approx p_{\text{miss}}, && \text{(9b)} \\ & m \in \{1, 2, \ldots, M\}, && \text{(9c)} \end{aligned}$$

with constraints set to meet the coverage requirement while keeping the search space aligned with the number of empirical quantiles. m is selected with a parameter γ to balance the high coverage rate and the narrow window with:

$$\begin{aligned} m_\gamma = \arg\min_m \sum^{i \in \mathcal{T}_k} \Big[& -\gamma \mathcal{CR}\left(\mathcal{T}_{k,i}^{\text{low},m}, +\infty\right) \\ & + (1-\gamma)\, \mathcal{ATW}\left(\mathcal{T}_{k,i}^{\text{low},m}, \mathcal{T}\right)\Big]. \end{aligned} \tag{10}$$

Then, the γ is chosen along with the best mth M-quantiles that decides the best lower adjustment scores:

$$m = \arg\min_{m_\gamma} \sum^{i \in \mathcal{T}_k} \left| \mathcal{CR}\left(\mathcal{T}_{k,i}^{\text{low},m_\gamma}, +\infty\right) - p_{\text{miss}} \right|. \tag{11}$$

Uncertainty-aware Task 2: Probability of Bus Arriving Within an Interval. In this case, both lower and upper boundaries are needed to make up a prediction window close

Table 2 – *The two routes.*

	20A	**340**
No. of stops	34	20
Scheduled time of a trip (min)	45	40
Headway (min)	30	25
No. of trips	658	1201
Mean/std of travel times (min)	20/13	19/12

to i_{arrival}. To restrict the width of the prediction window to be i_{arrival} while maximizing the probability of the true arrival time falling within it, the optimization becomes:

$$\underset{m,n}{\text{maximize}} \quad \mathcal{CR}\left(\mathcal{T}_{k,i}^{\text{low},m}, \mathcal{T}_{k,i}^{\text{up},n}\right) \tag{12a}$$

$$\text{subject to:} \quad \mathcal{ATW}\left(\mathcal{T}_{k,i}^{\text{low},m}, \mathcal{T}_{k,i}^{\text{up},n}\right) \approx i_{\text{arrival}}, \tag{12b}$$

$$m \in \{1, 2, \ldots, M\}, \tag{12c}$$

$$n \in \{1, 2, \ldots, N\}. \tag{12d}$$

After getting the mth and nth quantiles that best select the lower and upper adjustment scores s_m^{low} and s_n^{up} for K groups, the lower and upper boundaries can be achieved.

4 Experiments

4.1 Two Real-World Datasets

We select two bus routes with low-frequency services from two cities. The details of the two routes are listed in Table 2.

- **Bus 20A (inbound), Hong Kong (HK).** This dataset contains records of the 30-second-updated estimated time of arrival (ETA) of the next several buses at each stop, collected from DATA.GOV.HK[1]. The data covers 26 days from March to April 2024.
- **Bus 340 (outbound direction), Brisbane.** This dataset contains bus trajectories extracted from the GTFS-realtime data from DATA.QLD.GOV[2]. The data collection period was September 2024.

4.2 Implementation Details

The data is split into training, validation, calibration and test sets in a ratio of 4:1:3:2. The training and validation sets are used to train the quantile prediction function, the calibration set is used to find the best adjustment scores in the grouping calibration submodule, and the test set is used to make the pointwise predictions and generate lower and upper boundaries for quantification.

All hyperparameters are tuned based on the validation set through experiments. The model is implemented in Pytorch and executed on the CPU. We use the Adam optimizer with a learning rate of 0.001. The training spans 100 epochs, with an early stopping threshold of 30 epochs based on validation performance to avoid overfitting. The batch size is 48.

[1] https://data.gov.hk/sc-data/dataset/hk-ogcio-st_div_04-transport-bus-route-list-and-eta-spcific-bus-stop

[2] https://www.data.qld.gov.au/dataset/translink-real-time-data/resource/d92375b5-a291-49cc-aae8-4a7180d7984f?inner_span=True

4.3 Evaluation Metrics

We use Prediction Interval (PI) (the width of ATW) and Coverage Rate (CR) to evaluate performance. A smaller PI indicates a narrower ETA window and thus a more precise prediction, while a larger CR indicates a more confident prediction with a higher number of samples falling in the ETA window.

4.4 Baselines

we consider the following baselines: (1) **Historical Average (HA)** categorizes training samples grouped by trips and prediction horizons. The lower and upper boundaries are calculated based on the empirical quantile function of the training distribution. (2) **Distributional Forecasting (DF)** quantifies uncertainties by predicting the parameters of a probability distribution. We select the lognormal distribution as it fits the training samples better than other distributions (Li *et al.*, 2010). (3) **Bayesian Backpropagation (BBP)** (Blundell *et al.*, 2015) incorporates Bayesian inference principles to estimate prediction uncertainties and outputs a distribution over possible outcomes. We use a Gaussian distribution with tuned standard deviations around zero mean as the prior distribution. (4) **Monte Carlo Dropout (MC Dropout)** (Gal & Ghahramani, 2016) leverages dropout in the training phase to estimate uncertainties. In our model, we add two dropout layers between three fully connected layers, with a dropout rate of 0.5. (5) **Conformalized Quantile Regression (CQR)** combines conformal prediction with quantile regression to generate prediction intervals. The quantiles used for training and calibration are the same as those in our model.

Additionally, we include two variants of our model to test the effectiveness of the proposed mechanisms: (1) **UncertBAT-G** excludes grouping during calibration. The lower and upper adjustment scores are calculated and chosen from the whole calibration set $\mathcal{T}$. (2) **UncertBAT-M** removes the monotonicity constraint in Eq. (4) and relies solely on the pinball loss for quantile prediction.

4.5 Effectiveness Evaluation

4.5.1 Performance Comparison

We evaluate our model's performance against the baselines. The results of uncertainty quantification in Tasks 1 and 2 are shown in Table 3.

In Table 3, we report the prediction interval (PI) and the coverage rate (CR) for different p_{miss} and i_{arrival} values. For example, for 20A's test in UncertBAT under the case of $p_{\text{miss}} = 50\%$, the average gap between the lower boundaries and true arrival times is 1.58 minutes, and the test samples have a probability of 49.9% to be larger than the boundary with a 0.1%'s difference with the predefined p_{miss}.

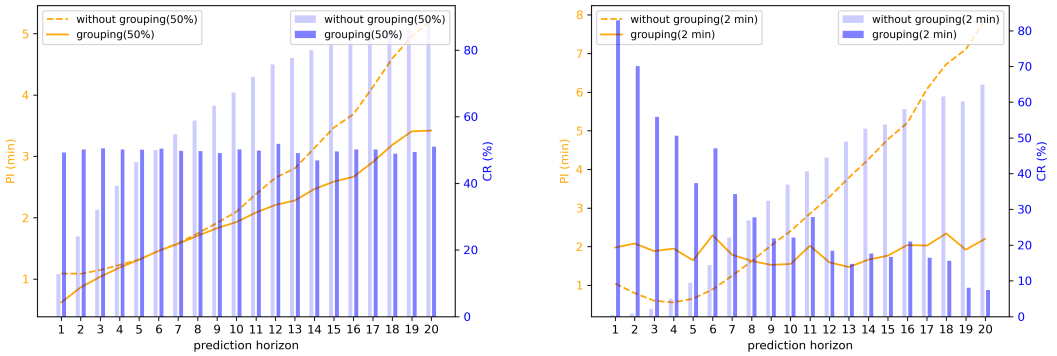
In Task 1, it is observed that only based on historical distribution, HA cannot handle unseen samples for 20A's test, resulting in coverage rates far from the target p_{miss}. DF and CQR tend to make wider predictions, especially when CR=90%, with DF providing much earlier predictions to meet the CR requirement for two datasets and CQR for 340's test. MC Dropout and BBP perform similarly when making pointwise-like quantification (CR=50% when the boundary is predicted between test samples); however, as p_{miss} increases, the prediction interval also enlarges. Overall, our model provides tighter prediction intervals, with an average improvement of 24.5%, and reduces underestimation errors compared to other baselines that can handle uncertainty estimation.

In Task 2, MC Dropout shows a significant decrease in test sample coverage as the prediction interval enlarges (9.5%, 4.5% and 23.4% decreases when PI = 2, 3, and 4min for 20A's test). CQR demonstrates the opposite pattern with 44.6%, 14.6%, and 5.3% decreases, indicating it is more likely to provide decentralized uncertainties. As a parametric model, the increase in CR in

Table 3 – *Performance of different models for Tasks 1 and 2 with different p_{miss} and $i_{arrival}$*

Route	**Models**	$p_{\text{miss}} = 50\%$		$p_{\text{miss}} = 75\%$		$p_{\text{miss}} = 90\%$		$i_{\text{arrival}} = 2\text{min}$		$i_{\text{arrival}} = 3\text{min}$		$i_{\text{arrival}} = 4\text{min}$	
		PI(min)	CR(%)	PI(min)	CR(%)	PI(min)	CR(%)	PI(min)	CR(%)	PI(min)	CR(%)	PI(min)	CR(%)
20A	HA	1.87	42.6	1.93	61.2	2.10	74.5	2.01	34.3	3.01	49.1	3.55	53.9
	DF	1.92	52.4	2.15	71.1	3.36	90.1	2.19	42.0	3.14	46.1	4.41	52.4
	BBP	1.63	51.0	1.89	74.0	3.28	93.1	2.00	38.0	3.00	49.4	4.17	68.5
	MC Dropout	1.57	48.9	2.02	76.7	2.56	85.3	1.99	39.2	3.33	59.9	4.10	56.6
	CQR	1.88	49.5	2.12	73.7	2.50	87.7	2.02	24.0	3.22	53.6	4.03	70.0
	UncertBAT	1.58	49.9	1.88	75.3	2.52	90.1	1.93	43.3	3.10	62.8	4.09	73.9
	UncertBAT-G	1.88	50.2	2.28	81.0	2.62	90.1	1.95	21.5	3.42	58.5	4.39	75.2
340	HA	3.51	49.8	3.92	75.0	5.19	90.0	2.19	22.1	2.88	27.9	3.92	34.5
	DF	2.00	53.6	2.37	74.6	3.40	89.4	1.88	28.2	2.83	35.8	4.09	41.4
	BBP	1.67	50.5	2.10	75.4	*	*	1.89	32.7	2.80	51.6	*	*
	MC Dropout	1.70	50.0	2.03	73.3	2.89	86.7	2.00	36.3	2.82	50.8	4.00	64.4
	CQR	2.06	52.8	2.73	72.6	3.83	82.5	2.21	20.6	3.08	43.3	3.81	57.4
	UncertBAT	1.67	53.7	1.96	76.9	2.51	90.6	1.87	34.0	2.96	53.8	3.77	65.0
	UncertBAT-G	2.04	56.9	2.43	80.8	2.85	92.0	2.20	18.5	2.51	30.9	3.89	60.0

* Result could not be obtained due to limited generated samples.



(a) *CR = 50%* (b) *PI = 2min*

Figure 2 – *Ablation study on Route 20A, UncertBAT v.s. UncertBAT-G. The x-axis refers to the prediction horizon.*

Table 4 – *Effectiveness of involving monotonicity.*

Models	$\mathrm{P}(\hat{\mathcal{Q}}_{0.1} < \hat{\mathcal{Q}}_{0.25})$	$\mathrm{P}(\hat{\mathcal{Q}}_{0.25} < \hat{\mathcal{Q}}_{0.5})$
UncertBAT	0.957	0.967
UncertBAT-M	0.797	0.927

DF does not align with the enlarged prediction intervals compared with others, which indicates more centrally distributed learned candidates with wider yet sparser tails, together with the results from Task 1. Overall, our model achieves an average 19.7% increase in coverage rate.

4.5.2 Ablation Study

We conduct experiments on two variants, UncertBAT-G and UncertBAT-M, to test the effectiveness of the proposed mechanisms.

Without grouping in calibration, UncertBAT-G shows a wide margin of performance loss, either with a wider PI with a similar CR or a limited CR with a comparable PI as shown in Table 3. Results across different cases in Tasks 1 and 2 along prediction horizons are demonstrated in Figure 2. With predefined CRs, Figure 2a shows that involving grouping in UncertBAT maintains relatively stable coverage rates across varying prediction horizons. From passengers' perspectives, it is crucial to ensure consistent arrival time windows regardless of the bus's proximity.

We also exclude the monotonicity constraint during the training phase. The constraint ensures that the independent parameters in the final prediction layer do not introduce uncertainties to quantile-based orders. We run UncertBAT and UncertBAT-G 5 times each and report the probability of the Ath quantile predictions being smaller than the Bth quantile ($A < B$) based on the validation dataset in Table 4. The results show significant improvement in maintaining the order consistency between predictions and required quantiles.

5 Conclusion

This paper fills the research gap in quantifying the uncertainties related to real-time bus arrival time prediction. We introduce UncertBAT, a novel framework designed to optimize arrival time windows by balancing high confidence and precision through the incorporation of conformalized quantile regression and a grouping mechanism based on sequential information. Extensive experiments demonstrate the model's effectiveness and utility in handling varying uncertainty-aware tasks relevant to real Bus Transit Systems. Our future work involves accounting for more

complex and dynamic traffic environment uncertainties while maintaining robust quantification.

References

Altinkaya, Mehmet, & Zontul, Metin. 2013. Urban bus arrival time prediction: A review of computational models. *International Journal of Recent Technology and Engineering (IJRTE)*, **2**(4), 164–169.

Blundell, Charles, Cornebise, Julien, Kavukcuoglu, Koray, & Wierstra, Daan. 2015. Weight uncertainty in neural network. *Pages 1613–1622 of: International conference on machine learning*. PMLR.

Gal, Yarin, & Ghahramani, Zoubin. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *Pages 1050–1059 of: international conference on machine learning*. PMLR.

Koenker, Roger, & Hallock, Kevin F. 2001. Quantile regression. *Journal of economic perspectives*, **15**(4), 143–156.

Li, Changlin, Lin, Shuai, Zhang, Honglei, Zhao, Hongke, Liu, Lishan, & Jia, Ning. 2023. A sequence and network embedding method for bus arrival time prediction using GPS trajectory data only. *IEEE Transactions on Intelligent Transportation Systems*, **24**(5), 5024–5038.

Li, Zheng, Hensher, David A, & Rose, John M. 2010. Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. *Transportation research part E: logistics and transportation review*, **46**(3), 384–403.

Liang, Jian, Ke, Jintao, Wang, Hai, Ye, Hongbo, & Tang, Jinjun. 2023. A Poisson-based distribution learning framework for short-term prediction of food delivery demand ranges. *IEEE Transactions on Intelligent Transportation Systems*.

Lukoševičius, Mantas, & Jaeger, Herbert. 2009. Reservoir computing approaches to recurrent neural network training. *Computer science review*, **3**(3), 127–149.

Petersen, Niklas Christoffer, Rodrigues, Filipe, & Pereira, Francisco Camara. 2019. Multi-output bus travel time prediction with convolutional LSTM neural network. *Expert Systems with Applications*, **120**, 426–435.

Romano, Yaniv, Patterson, Evan, & Candes, Emmanuel. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, **32**.

Steinwart, Ingo, & Christmann, Andreas. 2011. Estimating conditional quantiles with the help of the pinball loss.

Tang, Xiaolin, Yang, Kai, Wang, Hong, Wu, Jiahang, Qin, Yechen, Yu, Wenhao, & Cao, Dongpu. 2022. Prediction-Uncertainty-Aware Decision-Making for Autonomous Vehicles. *IEEE Transactions on Intelligent Vehicles*, **7**(4), 849–862.

Yaniv, Ilan, & Foster, Dean P. 1995. Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, **124**(4), 424.

Zhang, Xiaoge, & Mahadevan, Sankaran. 2020. Bayesian neural networks for flight trajectory prediction and safety assessment. *Decision Support Systems*, **131**, 113246.