Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Social Good

DART: Navigating Last-Mile Heterogeneity in Instant Delivery via Distribution-Adaptive Splines

Hao Xiong[1] , **Yang Gao**[2] , **Haiyong Luo**[2] , **Fang Zhao**[1] , **Dan Luo**[3]
[1]Beijing University of Posts and Telecommunications
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3]Beijing Forestry University
xmr2015211989@bupt.edu.cn, gaoyang23s@ict.ac.cn, yhluo@ict.ac.cn, zfsse@bupt.edu.cn, dluo_work@bjfu.edu.cn

Abstract

On-demand delivery platforms rely on Travel Time Estimation (TTE) to balance courier earnings and overdue risks. In collaboration with one of China's largest platforms, we address a critical "Fairness Gap" in TTE: current systems fail to capture complex delivery patterns in GNSS-denied environments, subjecting couriers handling high concurrent order volumes to disproportionate pressure due to overdue deliveries. Analyzing 1.27 million real-world trajectories, we attribute this bias to unique challenges in GNSS-denied scenarios: distributional heterogeneity, structural heterogeneity, and contextual uncertainty. To bridge this gap, we propose **DART** (Distribution-Adaptive Robust Timing). DART incorporates a Learnable Adaptive Spline (LAS) encoder with a gradient-driven knot migration mechanism to enhance non-linear expressiveness for outliers, significantly improving long-tail accuracy. Furthermore, a Spatio-Temporal Transition Graph (STTG) reconstructs the latent topology by integrating sequence semantics, such as Wi-Fi-sensed arrival merchant timestamps. At the same time, a Distribution Gating Mechanism characterizes delivery time distributions under distinct contexts. Through extensive experiments and large-scale online A/B testing, DART not only reduces MAE by 14.0% in complex environments but also decreases the Order Overdue Rate by 1.7% (saving $24,000 daily), demonstrating how AI effectively reconciles operational efficiency with labor fairness.

1 Introduction

The rapid expansion of the Online-to-Offline (O2O) economy has transformed on-demand delivery into a critical urban infrastructure, supporting the livelihoods of millions of gig workers [Dablanc *et al.*, 2017]. In practice, to maximize operational efficiency, platforms typically employ a "batch dispatching" strategy, where a courier receives multiple concurrent orders and must plan a complex sequence of pickups and drop-offs within strict time windows [Wang, 2019]. In this high-concurrency workflow, Delivery Time Estimation (DTE), a domain-specific specialization of TTE, acts as the central synchronization mechanism. It dictates not only the route planning but also the promised delivery time for each customer. Therefore, precise DTE is not merely an algorithmic metric; it is the fundamental guarantee for couriers to handle high workloads and secure their economic returns safely. This problem directly impacts labor fairness and the sustainability of the gig economy, highlighting its profound social importance.

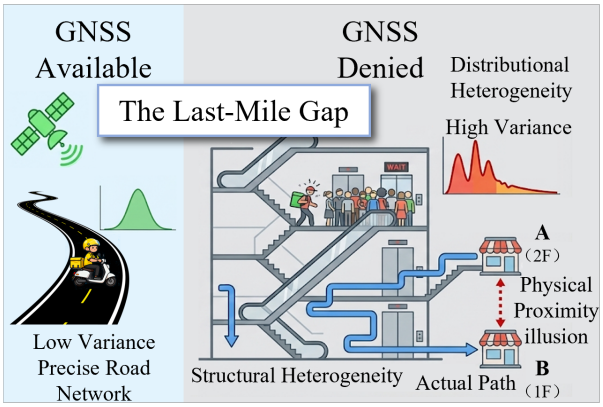

Figure 1: Conceptual illustration of the "Last-Mile Gap." Left (GNSS Available): Delivery relies on predictable road networks and clear GPS signals, characterized by low-variance distributions. Right (GNSS Denied): In GNSS-denied complex environments, models face both distributional and structural heterogeneity challenges.

Delivery Time Estimation is traditionally formulated as a regression problem on road networks, aiming to map trajectories to durations. While existing methods have achieved high accuracy in open outdoor environments by modeling traffic speeds and distances [Wu and Wu, 2019], the "last-mile" delivery in GNSS-denied environments—such as indoor shopping malls and dense commercial complexes—presents a fundamentally different challenge. As conceptually illustrated in Figure 1, different from outdoor scenarios where physical distance is a reliable proxy for time, delivery in these zones is governed by "invisible" stochastic events, such as waiting for elevators, security checks, or locating merchants in complex topologies [Yang *et al.*, 2020]. Current industrial models of-

ten ignore these fine-grained environmental heterogeneities, applying uniform estimation standards. This creates a systemic **"Fairness Gap"**: couriers who are willing to take on higher order counts suffer from disproportionate prediction errors. Consequently, they face severe, unmitigated overdue penalties for environmental complexities beyond their control, leading to algorithmic inequality [Ding *et al.*, 2022].

Bridging this fairness gap to restore fair and efficient dispatching is non-trivial due to three critical challenges: i) Distributional Heterogeneity: High-load delivery data exhibits a heavy-tailed distribution driven by stochastic delivery scenarios [Zhu *et al.*, 2020], making it difficult for standard regression models to fit outliers without sacrificing global accuracy [Liu *et al.*, 2024b]; ii) Structural Heterogeneity: Complex delivery zones act as topological "blind zones" lacking physical maps [Guo *et al.*, 2022], where existing models relying on Euclidean distance cannot perceive latent vertical traversal costs, despite recent advances in spatial pattern mining [Yuan *et al.*, 2025; Wang *et al.*, 2025]; iii) Contextual Uncertainty: Couriers engaged in high-concurrency delivery constantly switch between diverse contexts, leading to multi-peak delivery time distributions that are hard to disentangle without explicit context signals, similar to dynamic preference shifts in sequential modeling [Qi *et al.*, 2024].

In this paper, we propose a **Distribution-Adaptive Robust Timing** framework, called **DART**, to address the above challenges. DART explicitly models the heterogeneities in last-mile logistics through three innovations: i) A **Learnable Adaptive Spline (LAS) encoder** with gradient-driven knot migration to capture long-tail patterns; ii) A **Spatio-Temporal Transition Graph (STTG)** to reconstruct latent building topologies from crowd-sourced sequences; and iii) A **D-Gating Mechanism** to adaptively switch prediction strategies based on transit contexts.

In summary, the key contributions of the paper are as follows:

- To the best of our knowledge, we are the first to identify and formulate the **"Fairness Gap"** in last-mile TTE within GNSS-denied environments. We shift the research focus from pursuing average accuracy to achieving distributional robustness, ensuring equitable performance for high-workload couriers.
- We design a Distribution-Adaptive Robust Timing framework called **DART**. Specifically, we devise a Learnable Adaptive Spline (LAS) encoder to automatically detect and focus on feature subspaces that trigger long-tail effects by dynamically increasing the knot density. We also develop a Spatio-Temporal Transition Graph (STTG) to reconstruct latent topologies and a D-Gating Mechanism to disentangle contextual uncertainties.
- Deployed in partnership with one of the largest On-demand Food Delivery (OFD) platforms in China, which provided delivery data, partial delivery context features, and cloud infrastructure for large-scale A/B testing, we implement DART on a real-world on-demand platform and conduct extensive experiments based on 1.27 million trajectories and a large-scale Online A/B Test. The evaluation results show that DART outperforms state-of-the-art baselines by 14% in prediction accuracy. More importantly, it reduces the Order Overdue Rate by 1.7%, decreasing daily fines by approximately $24,000.

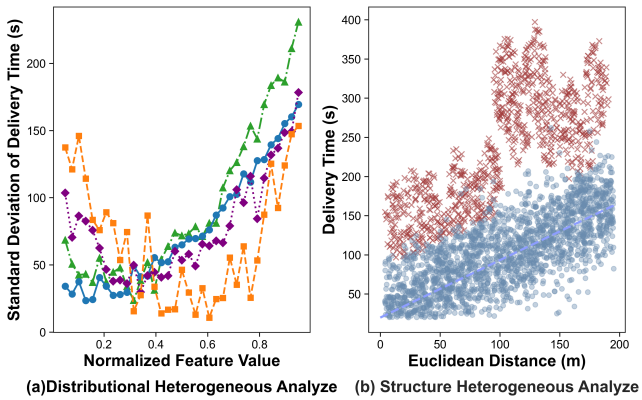

Figure 2: Heterogeneity analysis. (a) Non-linear Volatility: Standard deviation fluctuates with workload, indicating high uncertainty. (b) Inefficacy of Distance: Global correlation is low, but long-tail samples (red) form distinct topological clusters, suggesting recoverable patterns.

2 Motivation

2.1 Problem Formulation

We consider a set of historical delivery orders $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ in GNSS-denied urban environments, where $y_i \in \mathbb{R}^+$ is the ground-truth delivery time. The input $\mathbf{x}_i$ is a composite feature vector derived from multi-source sensed sequences, including crowd-sourced Wi-Fi sensed arrival merchant time sequence $\mathcal{S}_i$ and delivery specific attributes $\mathcal{A}_i$ [Ding *et al.*, 2022]:

$$\mathbf{x}_i = \Phi(\mathcal{S}_i, \mathcal{A}_i) \tag{1}$$

Given the heteroscedastic nature of urban delivery, we treat this as a conditional density estimation task. Our goal is to learn a probability density function $P(y|\mathbf{x}; \Theta)$ that maximizes the likelihood of observed delivery time, rather than fitting a deterministic regressor.

Observations

Opportunity. We analyzed real-world delivery trajectories that have ground truth labels by Bluetooth beacon sensing (deployed only at specific partner merchants) to identify the roots of estimation bias (Fig. 2), revealing two phenomena distinguishing GNSS-denied environments from outdoor logistics.

The first is **feature-dependent non-linear volatility**. Fig. 2a shows that the standard deviation of the delivery time fluctuates violently with real-time features such as workload. This implies that high workload scenarios introduce exponentially growing non-linearities. Existing models, assuming uniform distributions, tend to overfit dominant "head" samples while sacrificing accuracy on challenging tail cases.

The second is the **inefficacy of physical distance**. Unlike outdoor scenarios, Fig. 2b reveals a deceptive correlation between Euclidean distance and time. However, separating samples by walk speed reveals that long-tail instances are not random noise but form distinct topological clusters. This points to a promising avenue: anomalies likely stem from systematic latent constraints (e.g., elevators); thus, by effectively decoupling these latent context distributions, it is possible to rectify distance distortion and achieve high-fidelity estimation.

Challenges. However, realizing this opportunity is nontrivial due to three obstacles rooted in data heterogeneity. Primarily, the task is impeded by **distributional heterogeneity**, where long-tail variance is latently correlated with specific feature combinations, causing traditional discriminative models to yield unreliable predictions for outliers. Furthermore, the environment exhibits significant **structural heterogeneity**; in GNSS-denied settings, Euclidean distance fails to characterize spatial dependencies, causing performance degradation in geometric-based GNNs and necessitating adaptive graph construction. Finally, the model must contend with **contextual complexity**, where diverse events are implicitly entangled. Without explicit supervision, capturing this multimodal distribution requires generative modeling capabilities that surpass standard regression approaches.

3 Design

The overall framework of **DART** is illustrated in Figure 3. To tackle the intrinsic heterogeneity in GNSS-denied delivery, DART operates in three distinct stages: (1) A **LAS Encoder** that mitigates distributional heterogeneity via gradient-driven spline knots; (2) A **STTG** module that captures topological constraints from inaccurate physical spatial relations; and (3) A **D-Gating Mechanism** that manages contextual uncertainty through adaptive expert routing.

3.1 Learnable Adaptive Spline (LAS) Encoder

Standard MLPs suffer from global interference, where updates for long-tail samples distort the manifold for normal data. We propose the LAS Encoder to decouple these patterns using spline-based local activations.

Spline-based Neural Transformation

As depicted in Figure 4, unlike rigid linear weights, LAS computes the activation of the j-th neuron at layer l as an aggregation of learnable B-spline functions $\phi_{i,j}(\cdot)$. For an input vector $\mathbf{x}^{(l)} \in \mathbb{R}^{d_{in}}$, the transformation is defined as:

$$x_j^{(l+1)} = \sum_{i=1}^{d_{in}} \phi_{i,j}(x_i^{(l)}) = \sum_{i=1}^{d_{in}} \sum_{k=1}^{G} w_{i,j,k} \cdot B_{k,d}(x_i^{(l)}; \mathbf{t}_{i,j}) \tag{2}$$

where $B_{k,d}(\cdot)$ denotes the k-th B-spline basis function of degree d. The learnable coefficients $w_{i,j,k}$ control the function's amplitude, while the adaptive knot vector $\mathbf{t}_{i,j}$ dynamically adjusts the resolution density across the input domain.

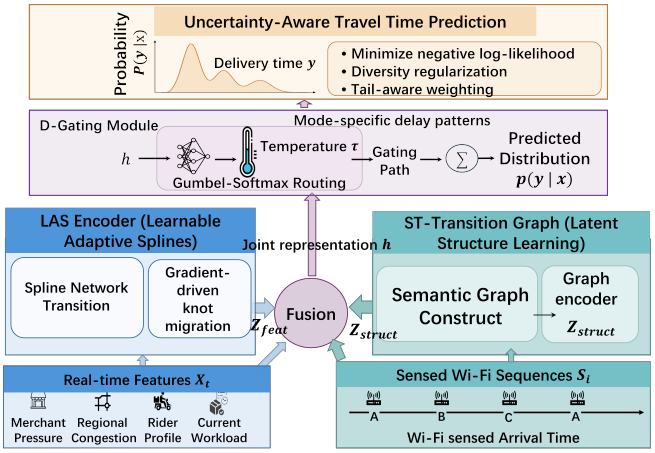

Figure 3: The overall framework of **DART**. (1) Distribution heterogeneity (top-left): Real-time features are processed by the LAS Encoder. (2) Structural heterogeneity (bottom-left): The STTG constructs a semantic graph based on merchant pairs. (3) Contextual uncertainty (right): Fused embeddings are routed by the D-Gating Mechanism.

Gradient-Driven Knot Migration

To robustly fit extreme outliers without sacrificing global stability, we employ a learnable knot mechanism. The gradients of the task loss $\mathcal{L}$ flow into knot positions via the chain rule:

$$\frac{\partial \mathcal{L}}{\partial t_m} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial y_j} \sum_k w_{i,j,k} \frac{\partial B_{k,d}(x_i; \mathbf{t})}{\partial t_m} \tag{3}$$

This allows knots to migrate adaptively: concentrating in high-error tail regions to increase local plasticity while remaining sparse in normal regions to ensure smoothness. Upon convergence, the activation function becomes a shape-optimized spline that provides dense support specifically for long-tail intervals identified by gradient descent.

Upon convergence, the knot vectors evolve from a uniform grid to a distribution-adapted constellation, denoted as $\mathbf{T}^* = \{\mathbf{t}_{i,j}^*\}$. Consequently, the final embedding $\mathbf{z}_{feat}$ captures the heterogeneous patterns through these spatially adaptive basis functions. The j-th dimension of the output is formalized as:

$$\mathbf{z}_{feat,j} = \sum_{i=1}^{d_{in}^{(L)}} \phi_{i,j}(x_i^{(L-1)}; \mathbf{t}_{i,j}^*) \tag{4}$$

Here, the activation function $\phi(\cdot; \mathbf{t}_{i,j}^*)$ is no longer a static mapping but a **shape-optimized spline** that provides dense support specifically for long-tail intervals identified by the gradient descent.

Theoretical Guarantee: Local Plasticity

The mathematical foundations of spline-based networks, particularly their universal approximation capabilities and numerical stability, have been rigorously established in classical approximation theory [De Boor, 2001] and recent advances in spline-based networks [Liu *et al.*, 2024a; Liu *et al.*, 2024c]. Building upon these theoretical guarantees, we focus on the Local Lipschitz Continuity to explicitly quantify the robustness of our LAS Encoder against outliers.

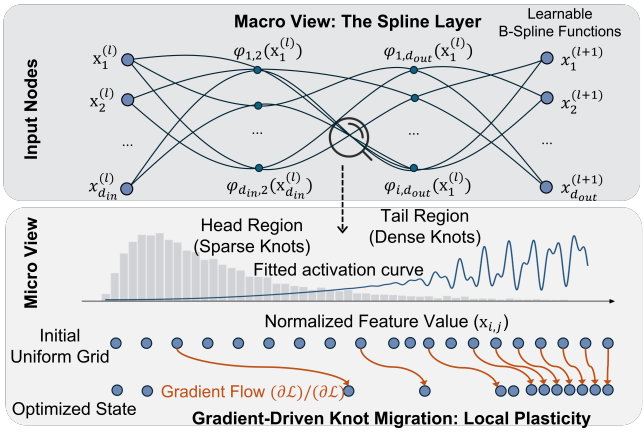

Figure 4: Detailed architecture of the LAS Encoder. (Top) Spline-based layer replacing rigid MLP weights. (Bottom) Gradient-driven knot migration: Knots (blue dots) dynamically move to high-error regions, enabling the model to fit specific long-tail patterns without affecting the global manifold.

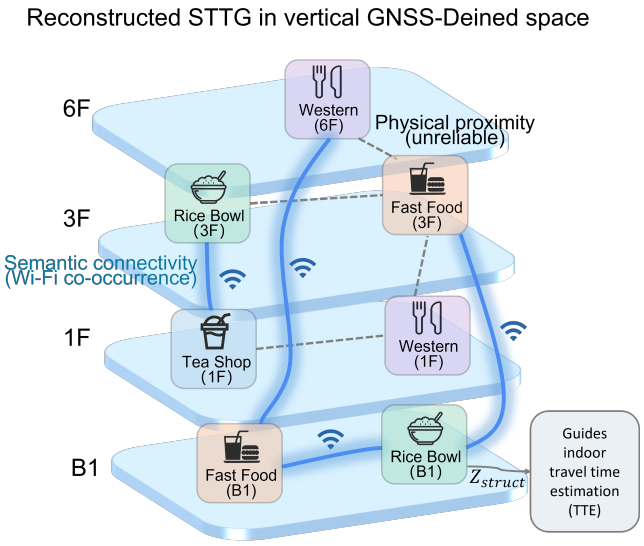

Figure 5: Dual-primal semantic graph based on merchant pairs.

Proposition 1 (Local Lipschitz Bound). *For a cubic B-spline function $\phi(x)$, the local Lipschitz constant L_ϕ over an interval $[t_j, t_{j+1})$ is bounded by:*

$$L_\phi^{[t_j,t_{j+1})} \leq \frac{k}{\min_{m\in\mathcal{I}_j}(t_{m+1}-t_m)} \max_{m\in\mathcal{I}_j} |w_{m+1}-w_m| \quad (5)$$

This bound demonstrates that gradient updates from outliers are strictly confined to the local interval $\mathcal{I}_j$, preventing catastrophic interference with the global representation.

3.2 Spatio-Temporal Transition Graph (STTG)

To capture topological constraints from inaccurate physical spatial relations, we design a dual-primal semantic graph $G_S = \{V_{OD}, E_S\}$ where nodes represent merchant pairs rather than static locations (see Figure 5).

We formulate the semantic structure of the dual-primal graph G_S by defining node attributes and edge connectivity. First, regarding node attributes, we adopt Wi-Fi arrival detection technology [Guo *et al.*, 2022] to accurately track the courier's status. Unlike static physical distance, we compute a dynamic semantic distance $d^{\text{Wi-Fi}}$ by accumulating the horizontal and vertical displacements derived from the sequence of Wi-Fi fingerprints as follows:

$$d_{o_i}^{\text{Wi-Fi}} = \sum_{t=1}^{T-1} \text{Haversine}(l_{t+1}, l_t) + \sum_{t=1}^{T-1} |f_{t+1}-f_t| \cdot d_{\text{floor}} \quad (6)$$

This metric serves as a robust proxy for the actual traversal cost (e.g., incorporating vertical detours), effectively compensating for the limitations of Euclidean distance. Second, for edge construction, we define the adjacency matrix $\mathbf{A}$ to capture the latent connectivity between tasks. Instead of relying on physical proximity, we establish edges between merchant pairs that exhibit **high temporal proximity**. Specifically, we link pairs whose historical delivery time $\bar{t}_{OD}$ are within a connectivity threshold derived from the global average t_m:

$$A_{ij} = \begin{cases} 1, & \text{if } \bar{t}_{OD} \leq t_m \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

This ensures that the subsequent GAT aggregates features from tasks that are temporally correlated and share similar transit patterns.

Structure Propagation. We employ a multi-head Graph Attention Network (GAT) to encode these dependencies. Initial embeddings e_i (comprising GPS and arrival time) are updated by aggregating semantic neighbors:

$$\mathbf{z}_{struct} = \frac{1}{K}\sum_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \mathbf{W}\mathbf{h}_j\right) \quad (8)$$

where α_{ij}^k denotes the attention coefficient for head k. This explicitly injects latent topological friction into the final representation $\mathbf{z}_{struct}$.

3.3 D-Gating Mechanism

The final stage explicitly addresses **Contextual Uncertainty** (e.g., waiting vs. transit contexts) by adaptively switching prediction strategies.

We first fuse the distributional features $\mathbf{z}_{feat}$ and structural context $\mathbf{z}_{struct}$ via Cross-Attention to obtain $\mathbf{h}$. We then employ a Mixture of Gaussians output, where mixing coefficients π_k are determined by a **Gumbel-Softmax** gating network:

$$\pi_k = \frac{\exp((g_k(\mathbf{h})+\epsilon_k)/\tau)}{\sum_{j=1}^{K}\exp((g_j(\mathbf{h})+\epsilon_j)/\tau)}, \quad \epsilon \sim \text{Gumbel}(0,1) \quad (9)$$

Here, τ anneals from 1.0 to 0.1, enabling a transition from soft mixture learning to hard, interpretable mode selection.

The model is trained to maximize the likelihood of the ground truth y. To prevent mode collapse and ensure stability, we augment the Negative Log-Likelihood (NLL) with orthogonality and dispersion constraints.

4 Experiments

4.1 Experimental Setup

Datasets and Heterogeneity Analysis

We utilized a large-scale industrial dataset from Meituan, one of the biggest OFD platforms in China, spanning four months (April–August 2023) across major Chinese cities (Beijing, Shanghai, Jiangsu). The dataset comprises 1.27 million orders, 20,000 active couriers, and ground truth collected via Bluetooth beacons. To rigorously evaluate DART under varying environmental heterogeneities, we partitioned the test data into four subsets of distinct complexity:

- **D1 & D2 (High Complexity):** Large-scale multi-story shopping malls with complex indoor topologies and high order density, representing the most challenging "long-tail" scenarios.
- **D3 (Data Scarce):** Small-scale buildings with extreme data sparsity (0.7% of total volume), testing few-shot generalization.
- **D4 (Medium Complexity):** Semi-indoor pedestrian streets representing balanced complexity.

Implementation

To ensure data reliability, ground truth arrival times were captured via Bluetooth beacons at partner merchants, mitigating GPS drift and manual reporting errors. The dataset is partitioned chronologically: the first two months for training, followed by one month each for validation and testing. All baselines and U-STAR utilize identical splits for rigorous comparison. Offline training was conducted on an NVIDIA A100 GPU (PyTorch 1.13), while real-time performance was evaluated on a standard Meituan Cloud production node (Intel Xeon Gold 6240, single-core) to assess operational viability. Key hyperparameters include $d_{model} = 64$, cubic B-splines ($d = 3$) with initial grid $G = 15$ for the DA-Spline Encoder, and $K = 5$ mixture components for ADisAM (τ annealed 1.0 to 0.1). Optimization employs Adam ($LR = 1e - 4$, Batch=64) with early stopping.

Baselines

We benchmark DART against 12 baselines across three categories. Note that foundation models (e.g., [Li *et al.*, 2024; Wang *et al.*, 2024]) are excluded as their high latency is incompatible with the millisecond-level constraints of real-time dispatching.

Industrial Heuristics & Statistical Models. We first establish performance lower bounds using fundamental non-learning approaches: **AVG** (historical averages) and **GPS/Wi-Fi** (physical sensing without semantic understanding). To evaluate signal processing limits, we include **TransLoc** [Yang *et al.*, 2020], which utilizes HMM and Kalman Filtering for trajectory smoothing. Additionally, **KDE-Reg** combines kernel density estimation with regression to test whether DART's complex mixture density assumption yields significant gains over simple statistical fitting.

Dataset	KDE-Reg *(Stat.)*	P2Loc- *(DL)*	IGT *(Graph)*	DART (Ours)
D1 (High)	78.3/120.6	62.7/90.2	65.1/95.3	**52.2/89.1**
D2 (High)	55.4/88.3	37.9/**61.4**	39.8/72.5	**33.7**/63.3
D3 (Sparse)	62.2/108.2	58.1/105.3	49.3/95.4	**45.7/78.7**
D4 (Medium)	42.2/75.2	35.2/65.0	33.2/62.9	**30.4/60.1**

Table 1: Performance comparison with best-in-class baselines. Metrics are reported as MAE / RMSE.

Spatio-Temporal Deep Learning. To assess the capability of capturing non-linear correlations and noise, we compare against **EDCN** [Chen *et al.*, 2021] (temporal gating for sequence dependencies), **P2Loc-** [Ding *et al.*, 2022] (modeling courier-merchant interactions), and **MWSL-TTE** [Wang *et al.*, 2023] (weakly supervised learning for label noise). We also include the recent SOTA **NAO** [Guo *et al.*, 2024], which optimizes dynamic noise distributions via MLE, serving as a direct competitor to our D-Gating mechanism in modeling aleatoric uncertainty.

State-of-the-Art Graph Models. Finally, we benchmark against frontier graph learning methods to validate our structural modeling. **GSL-QR** [Zhang *et al.*, 2023] employs implicit graph structure learning with quantile regression, providing a contrast to our explicit STTG topology reconstruction. Furthermore, we compare DART with **IGT** [Zhou *et al.*, 2023] (Inductive Graph Transformer), the current SOTA in structural modeling, to verify the superiority of our spline-based LAS encoder over Transformers in handling long-tail distribution heterogeneity.

Metric

In terms of metrics, beyond standard MAE and RMSE, we explicitly report Tail-MAE and Tail-RMSE calculated on the top 15% of samples with the longest travel times among identical Origin-Destination pairs. These metrics serve as a rigorous robustness check for structural anomalies and unexpected delays, distinguishing our approach from scenarios where conventional global thresholds are sufficient.

4.2 Overall Performance Comparison

Table 1 presents the comprehensive results across datasets with varying complexities, where DART consistently achieves superior performance. In high-complexity scenarios (D1 & D2), DART reduces MAE by **19.8%** (in D1) compared to the state-of-the-art Graph Transformer (IGT). This validates that our **LAS Encoder** effectively captures high-frequency fluctuations in long-tail samples, overcoming the "oversmoothing" limitation of global attention mechanisms used in IGT. Furthermore, DART demonstrates robust few-shot generalization in data-scarce environments (D3). Compared to the Deep Learning baseline P2Loc-, DART reduces MAE by **21.3%**, confirming that the **STTG** module successfully extracts transferable semantic topology rather than overfitting structural priors like standard sequence models. Finally, regarding multimodal uncertainty (D4), the **D-Gating Mechanism** enables DART to outperform deterministic baselines (e.g., P2Loc-) with a **7.5%** reduction in RMSE. By explicitly disentangling latent delivery modes, DART prevents

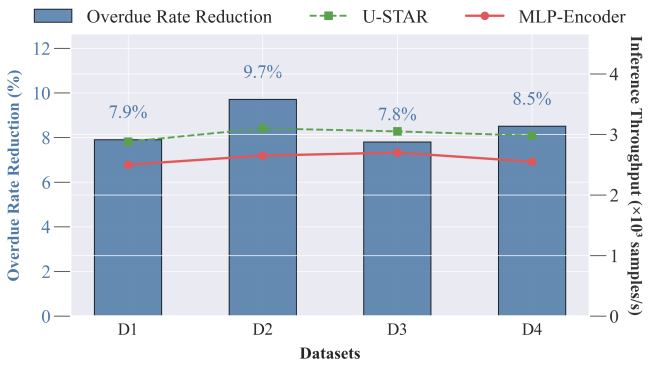

Figure 6: Efficiency-performance trade-off. DART significantly reduces Overdue Rate (Left Axis) while maintaining high inference throughput comparable to MLP (Right Axis), verifying its feasibility for real-time industrial deployment.

large deviations caused by unimodal assumptions. These significant improvements in accuracy, particularly in long-tail and data-scarce areas, provide a solid foundation for fair and efficient dispatching in real-world deployment, as detailed in the following section.

4.3 Online Deployment

To validate industrial feasibility, we first analyzed the computational cost (Figure 6). DART achieves a substantial reduction in the Overdue Rate across all datasets while maintaining an inference throughput ($\times 10^3$ samples/sec) comparable to a lightweight MLP baseline. This confirms that the local support property of B-splines ensures the high efficiency required for real-time dispatching.

We subsequently deployed DART in a large-scale Online A/B Test on the Meituan OFD platform, covering 16 major cities with a 50/50 traffic split. Over the course of one month, the model demonstrated a significant operational impact. The Overdue Rate decreased by **1.7%**, effectively mitigating customer complaints and reducing penalties for couriers, thereby directly addressing labor fairness challenges. Furthermore, the improved accuracy optimized the dispatching "buffer time," yielding estimated operational savings of **$24,000 per day**. These metrics definitively validate the real-world business impact of our proposed method.

Finally, we examined the model's robustness against increasing consecutive order counts (Figure 7). While the baseline suffers from exponential growth in cumulative delivery MAE error (CMAE) under high concurrency, DART maintains a stable error curve even as the courier workload increases to 7 consecutive orders. This demonstrates DART's ability to capture the non-linear physics of congestion, ensuring reliability and stability when the system is under maximum stress.

4.4 In-depth Analysis

Unlike opaque deep learning black boxes, DART demonstrates clear interpretability in handling data heterogeneity.

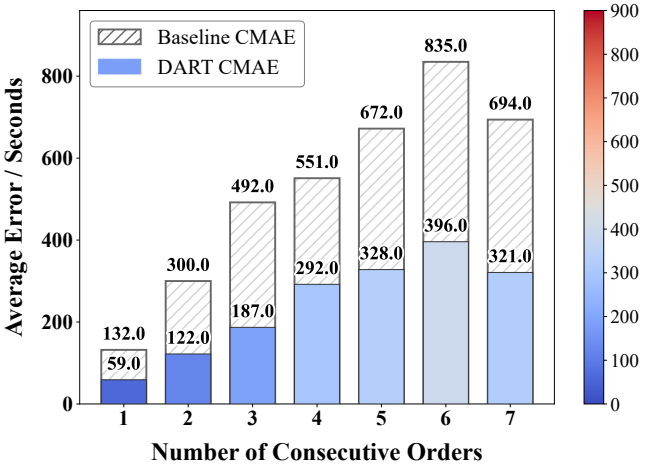

Figure 7: Robustness under high-pressure workload. Unlike the Baseline, which suffers exponential error growth with the increase in concurrent orders, DART effectively suppresses error accumulation.

Feature Level: Gradient-Driven Adaptation

We visualized the knot density distribution in Figure 8 to reveal the model's resource allocation strategy. Unlike the Fixed Grid baseline, which maintains a rigid uniform density, DART evolves into a distinct "U-shaped" distribution. It strategically reduces knot density in the data-rich central region, effectively identifying frequent patterns as "easy samples" that require less model capacity. These freed resources are then migrated to the sparse tail regions, where density is significantly amplified. This confirms that DART's topology adaptation is driven by task difficulty (gradients) rather than mere data quantity, allowing it to suppress error divergence in outlier regions without sacrificing performance in the dominant central distribution.

Mode Level: Hierarchical Functional Specialization

We further elucidated the latent structure learned by the D-Gating mechanism by analyzing Expert Selection Frequency against sample difficulty (Figure 9). The empirical distribution manifests an explicit hierarchical specialization. In the deterministic regime (low deviation, $0-3$ min), Experts 2, 3, and 4 exhibit sequential dominance, indicating that the model has successfully disentangled the latent factors of predictable transit into orthogonal functional subspaces. Conversely, in the stochastic tail (>5 min), the routing shifts towards Expert 5 with significantly higher entropy. This suggests that for extreme outliers, DART automatically adopts an ensemble-like approach to handle irreducible uncertainty, rather than relying on a single deterministic mapping.

Ablation Study

To quantify the contribution of each module, we evaluated variants of DART (Table 2).

- **w/o LAS Encoder:** Reverting to a standard MLP causes Tail RMSE to deteriorate by 29.4%, confirming that global activation functions fail to model heavy-tailed distributions.

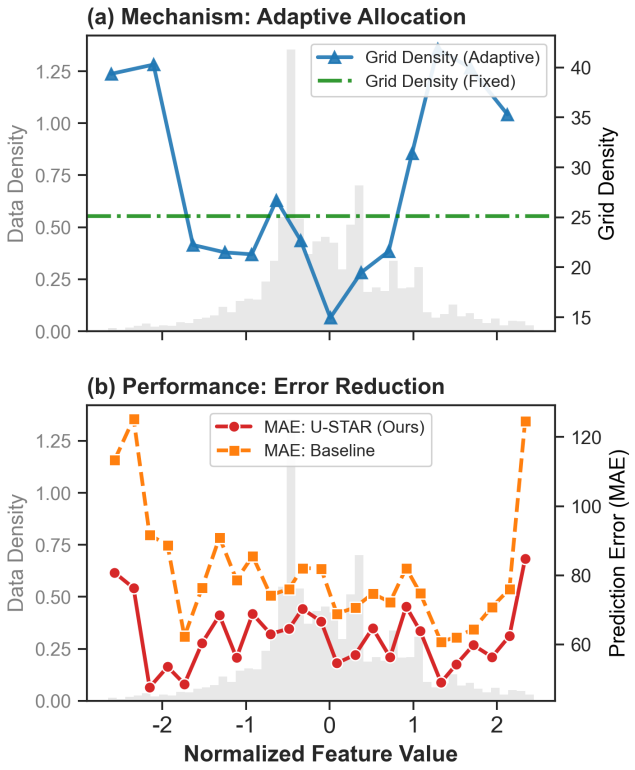

Figure 8: Gradient-driven resource reallocation. (a) Mechanism: DART (Blue) forms a "U-shaped" density, reducing resources in the easy center and reallocating them to the hard tails compared to the uniform baseline (Green). (b) Performance: This strategy (Red) significantly suppresses error divergence in tail regions compared to the baseline (Orange).

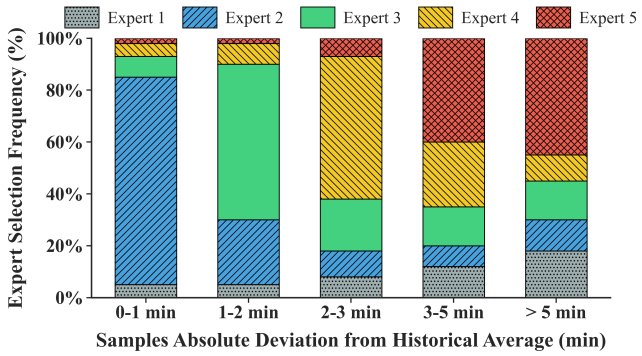

Figure 9: Expert selection frequency vs. sample difficulty. Experts 2–4 show sequential dominance in low-deviation regimes (predictable), while the high-variance tail ($>$ 5 min) triggers Expert 5 with high routing entropy, indicating an ensemble response to uncertainty.

Model Variant	MAE	RMSE	Tail MAE	Tail RMSE
w/o STTG	59.39	95.12	178.33	198.95
w/o D-Gating	73.10	112.14	211.34	228.57
w/o LAS (MLP)	59.08	97.28	185.00	204.59
DART (Full)	**45.43**	**79.19**	**128.98**	**162.30**

Table 2: Ablation study.

- **w/o STTG:** Removing the graph module degrades Overall MAE by 29.1%, highlighting the necessity of modeling non-Euclidean semantic topology.
- **w/o D-Gating:** Using a single Gaussian head increases MAE by 58.9%, proving that capturing multimodal uncertainty is critical for accuracy.

5 Related Work

Spatio-Temporal Prediction and Uncertainty. Travel Time Estimation is fundamental to logistics efficiency [Lyu *et al.*, 2023]. Modern approaches have evolved from simple sequence modeling to capturing complex latent patterns: STAMImputer [Wang *et al.*, 2025] and STPro [Yuan *et al.*, 2025] utilize attention mechanisms and hierarchical prototypes to address data sparsity and spatial heterogeneity, while Cuss-Ctpm [Qi *et al.*, 2024] models continuous temporal dynamics. Despite these advances, real-world delivery faces significant stochasticity. Recent works, such as ST-nFBST [Liu *et al.*, 2024b], introduce Bayesian significance testing to quantify uncertainty. However, existing methods suffer from two limitations in our context: 1) They predominantly rely on explicit road network graphs, which are unavailable in GNSS-denied environments; 2) They often assume unimodal distributions, failing to disentangle the multimodal conditional heteroskedasticity inherent in heavy-tailed delivery data.

Spline-based Neural Networks. Kolmogorov-Arnold Networks (KANs) [Liu *et al.*, 2024c] have emerged as interpretable alternatives to MLPs by parameterizing edges with learnable splines. While promising, current implementations rely on fixed grid points. This rigid discretization induces spectral bias [Rahaman *et al.*, 2019], where the model captures low-frequency trends but underfits the high-frequency fluctuations of tail outliers. To address this, we propose a distribution-adaptive mechanism that dynamically allocates resolution to capture high-frequency details in the long tail.

6 Conclusion

In this work, we address the "Last-Mile Gap" in GNSS-denied logistics, where conventional models fail due to environmental complexity. We propose **DART**, a framework targeting distributional and structural heterogeneity. By synergizing the **LAS Encoder** and **STTG**, DART effectively captures long-tail outliers and latent topological constraints. Experiments on massive industrial datasets show that DART outperforms SOTA baselines by reducing MAE by **14.0%**. Furthermore, online A/B testing confirms significant reductions in overdue rates and operational costs. Crucially, DART promotes algorithmic fairness by accurately identifying long-tail delivery trip samples to mitigate unfair penalties, fostering a more equitable gig economy. Future work will focus on cross-region generalization.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62261042, the Beijing Natural Science Foundation under Grant 4232035 and 4254084, the Yibin high-level Introduction talents project under Grant 2024YG01 and the China Postdoctoral Science Foundation under Grant 2024M750200.

References

[Chen *et al.*, 2021] Bo Chen, Yichao Wang, Zhirong Liu, Ruiming Tang, Wei Guo, Hongkun Zheng, Weiwei Yao, Muyu Zhang, and Xiuqiang He. Enhancing explicit and implicit feature interactions via information sharing for parallel deep ctr models. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3757–3766. ACM, October 2021.

[Dablanc *et al.*, 2017] Laetitia Dablanc, Eleonora Morganti, Niklas Arvidsson, Johan Woxenius, Michael Browne, and Neila Saidi. The rise of on-demand 'instant deliveries' in european cities. In *Supply Chain Forum: An International Journal*, volume 18, pages 203–217. Taylor & Francis, 2017.

[De Boor, 2001] Carl De Boor. *A Practical Guide to Splines*. Springer-Verlag New York, revised edition, 2001.

[Ding *et al.*, 2022] Yi Ding, Dongzhe Jiang, Yu Yang, Yunhuai Liu, Tian He, and Desheng Zhang. P2-loc: A person-2-person indoor localization system in on-demand delivery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–24, 2022.

[Guo *et al.*, 2022] Baoshen Guo, Weijian Zuo, Shuai Wang, Wenjun Lyu, Zhiqing Hong, Yi Ding, Tian He, and Desheng Zhang. Wepos: Weak-supervised indoor positioning with unlabeled wifi for on-demand delivery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–25, 2022.

[Guo *et al.*, 2024] Xiaoyu Guo, Weiwei Xing, Jun Fang, Jia Chen, Xi Chen, and Ruipeng Gao. Noise-aware optimization for mobile crowdsensing-based travel time estimation. *IEEE Transactions on Vehicular Technology*, 73(3):4067–4080, 2024.

[Li *et al.*, 2024] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5351–5362, New York, NY, USA, 2024. Association for Computing Machinery.

[Liu *et al.*, 2024a] Ming Liu, Song Bian, Bo Zhou, and Paul Lukowicz. ikan: Global incremental learning with kan for human activity recognition across heterogeneous datasets. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pages 89–95, 2024.

[Liu *et al.*, 2024b] Zehua Liu, Jingyuan Wang, Zimeng Li, and Yue He. Full bayesian significance testing for neural networks in traffic forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[Liu *et al.*, 2024c] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[Lyu *et al.*, 2023] Wenjun Lyu, Haotian Wang, Yiwei Song, Yunhuai Liu, Tian He, and Desheng Zhang. A prediction-and-scheduling framework for efficient order transfer in logistics. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Qi *et al.*, 2024] Lianyong Qi, Yuwen Liu, Weiming Liu, Shichao Pei, Xiaolong Xu, Xuyun Zhang, Yingjie Wang, and Wanchun Dou. Counterfactual user sequence synthesis augmented with continuous time dynamic preference modeling for sequential poi recommendation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[Rahaman *et al.*, 2019] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[Wang *et al.*, 2023] Hongjun Wang, Zhiwen Zhang, Zipei Fan, Jiyuan Chen, Lingyu Zhang, Ryosuke Shibasaki, and Xuan Song. Multi-task weakly supervised learning for origin–destination travel time estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11628–11641, 2023.

[Wang *et al.*, 2024] Shuliang Wang, Xinyu Pan, Sijie Ruan, Haoyu Han, Ziyu Wang, Hanning Yuan, Jiabao Zhu, and Qi Li. Diffcrime: A multimodal conditional diffusion model for crime risk map inference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3212–3221, New York, NY, USA, 2024. Association for Computing Machinery.

[Wang *et al.*, 2025] Yiming Wang, Hao Peng, Senzhang Wang, Haohua Du, Chunyang Liu, Jia Wu, and Guanlin Wu. Stamimputer: spatio-temporal attention moe for traffic data imputation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025.

[Wang, 2019] Zi Wang. Time predictions in uber eats. https://www.infoq.com/articles/uber-eats-time-predictions, 2019. Accessed: 2019-12-15.

[Wu and Wu, 2019] Fan Wu and Lixia Wu. Deepeta: a spatial-temporal sequential neural network model for estimating time of arrival in package delivery system. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 774–781, 2019.

[Yang *et al.*, 2020] Yu Yang, Yi Ding, Dengpan Yuan, Guang Wang, Xiaoyang Xie, Yunhuai Liu, Tian He, and Desheng Zhang. Transloc: transparent indoor localization with uncertain human participation for instant delivery. In *Pro-*

ceedings of the 26th Annual International Conference on Mobile Computing and Networking, pages 1–14, 2020.

[Yuan *et al.*, 2025] Shilu Yuan, Xiaoyu Li, Wenqian Mu, Ji Zhong, Meng Chen, Haoliang Sun, and Yongshun Gong. Spatio-temporal prototype-based hierarchical learning for od demand prediction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025.

[Zhang *et al.*, 2023] Lei Zhang, Xin Zhou, Zhigang Zeng, Yang Cao, Yong Xu, Min Wang, Xindong Wu, Ye Liu, Laizhong Cui, and Zhiqi Shen. Delivery time prediction using large-scale graph structure learning based on quantile regression. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3403–3416. IEEE, 2023.

[Zhou *et al.*, 2023] Xin Zhou, Jinglong Wang, Yong Liu, Xingyu Wu, Zhiqi Shen, and Cyril Leung. Inductive graph transformer for delivery time estimation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 679–687, New York, NY, USA, 2023. Association for Computing Machinery.

[Zhu *et al.*, 2020] Lin Zhu, Wei Yu, Kairong Zhou, Xing Wang, Wenxing Feng, Pengyu Wang, Ning Chen, and Pei Lee. Order fulfillment cycle time estimation for on-demand food delivery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2571–2580, 2020.