A HST-GT ALGORITHM

Our proposed HST-GT algorithm for node representations learning at time t ($t \geq 1$) is given in Algorithm 1.

Algorithm 1 HST-GT for Node Representation Learning

```
Input: WC-graph $\mathcal{G} = \{V, E\}$, spatial representation $u^{S(t)}$, temporal data $X^{T(t)}$, background data $X_{BG}^{(t)}$
Output: Node representation $u^{(t)}$
$u^{S'(t-1)} \leftarrow$ Eq.(7), (8) with $u^{S(t-1)}$, $\mathcal{G}$ in spatial model.
$u^{T(t)} \leftarrow$ Eq.(10) with $X^{T(t)}$, $X_{BG}^{(t)}$ in temporal model.
$u^{ST(t)} \leftarrow$ Eq.(12), (14) with $u^{S'(t-1)}$, $u^{T(t)}$ in ST correlation mining model.
$u^{S(t)} \leftarrow$ Eq.(15), (16) with $u^{S'(t-1)}$, $u^{T(t)}$ in ST fusion model.
$u^{(t)} \leftarrow \{u^{S(t)}, u^{T(t)}, u^{ST(t)}\}$
Return $u^{(t)}$
```

B METRIC DEFINITIONS

Four widely used metrics, i.e., Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE), are adopted for the prediction accuracy evaluation. They are defined as:

$$\begin{aligned}
RMSE &= \sqrt{\frac{1}{N}\sum_{t}^{\mathcal{T}_{\text{test}}}\sum_{r_i \in R^{(t)}}\left\|\hat{y}_{r_i}^{(t)} - y_{r_i}^{(t)}\right\|^2} \\
MAE &= \frac{1}{N}\sum_{t}^{\mathcal{T}_{\text{test}}}\sum_{r_i \in R^{(t)}}\left|\hat{y}_{r_i}^{(t)} - y_{r_i}^{(t)}\right| \\
MAPE &= \frac{1}{N}\sum_{t}^{\mathcal{T}_{\text{test}}}\sum_{r_i \in R^{(t)}}\left|\frac{\hat{y}_{r_i}^{(t)} - y_{r_i}^{(t)}}{y_{r_i}^{(t)} + \varepsilon}\right| \\
SMAPE &= \frac{1}{N}\sum_{t}^{\mathcal{T}_{\text{test}}}\sum_{r_i \in R^{(t)}}\left|\frac{2(\hat{y}_{r_i}^{(t)} - y_{r_i}^{(t)})}{\hat{y}_{r_i}^{(t)} + y_{r_i}^{(t)}}\right|
\end{aligned} . \quad (21)$$

where $\hat{y}_{r_i}^{(t)}$ denotes the delivery time estimation for full-link route r_i at time t, and $y_{r_i}^{(t)}$ denotes the ground-truth. $\mathcal{T}_{\text{test}}$ is the test period, N is the number of orders in the test data, and $R^{(t)}$ denotes the set of full-link delivery routes at time t. For the MAPE, a small value close to zero (ϵ) is added to the denominator to avoid zero division.

C SUPPLEMENTARY MATERIAL FOR REPRODUCIBILITY

C.1 Dataset Details

In this section, we describe (i) the details of real-world dataset, (ii) the feature extraction method, and (iii) a synthetic dataset we proposed.

C.1.1 *Real-world Dataset Description.*

The detail information of the real-world dataset from one of the largest online e-commerce retailers in China in as below:

- ***Full-link delivery data***: The delivery network in e-commerce retailer (e.g., JD, Amazon) is in the warehouse-distribution integration mode, and the spatial network data includes (i) delivery unit information such as warehouses ID and sorting centers ID; (ii) distribution connection information such as warehouse-sorting center connections and sorting center-sorting center connections. To save costs and improve efficiency, the company has designed a set of delivery routes in advance. When an order is created, the company will assign a delivery route for it. A full-link delivery route is a delivery unit sequence including a warehouse and multiple-level sorting centers. In total, there are 334 warehouses, 559 sorting centers, 1177 warehouse-sorting center connections, 2971 sorting center-sorting center connections, and 15040 routes in the delivery network.
  The historical delivery data includes 1.152 million orders of the delivery network from February 4, 2022 to March 6, 2022, which logs detailed (i) spatial information such as the route ID in the distribution network; and (ii) temporal information such as order creating time, order departure warehouse time, first sorting time, and receive time recorded at the second level. To protect privacy, the e-commerce removes all customer IDs and commodity types.
- ***Historical temporal data***: Under the warehouse-distribution integration mode in e-commerce retailer, the temporal data is aggregated statistics for some metrics (e.g., sales volume), and we combine the historical temporal data from online sales and offline logistics together. In online sales, the temporal data includes hourly cumulative sales volume and daily cumulative sales volume. In offline logistics, the temporal data includes hourly amount of packages processing in warehouse, and hourly amount of packages carrying in delivery routes. Further, we integrate the background temporal information including time of day (e.g., 15:00), day of week (e.g., Friday), and sales promotion period (e.g., Spring Festival) into the temporal data.

Moreover, we provide examples of the spatial data in Table 3 and temporal data in Table 4, 5, and 6.

Table 3: Spatial data: an example of full-link delivery route

Field	Value
RouteID	00001
Warehouse ID	110007124
Level 1 Sorting Center ID	800000809
Level 2 Sorting Center ID	800003403
Level 3 Sorting Center ID	800008415
Level 4 Sorting Center ID	800001617
Level 5 Sorting Center ID	800005373

Table 4: Temporal data: an example of histrical delivery order

Field	Value
OrderID	JDV005350750003
RouteID	00001
Order Creating Time	2022/2/19 15:01:46
Order Departure Warehouse Time	2022/2/19 17:22:59
First Sorting Time	2022/2/19 21:43:41
Receive Time	2022/2/21 09:55:39

Table 5: An example of temporal data of a warehouse

Field	Value
Warehouse ID	110007124
Time	2022/2/19 15:00:00
Hourly Amount of packages in Warehouse	6357
Hourly Cumulative Sales Volume	3477
Daily Cumulative Sales Volume	17582

Table 6: An example of temporal data of a sorting center

Field	Value
Sorting Center ID	800000809
Time	2022/2/19 15:00:00
Hourly Amount of packages in Delivery Routes	5710
Hourly Cumulative Sales Volume	1708
Daily Cumulative Sales Volume	7358

Further, **to clarify the full-link delivery route**, we detail the real-world logistic delivery route as below:

A full-link delivery route is a delivery unit sequence including a warehouse and multiple-level sorting centers. An example is shown in Table 3. The delivery route is different from path, as shown in Fig 19, the packages are delivered in **Route 00001** from *Warehouse 110007124* to *Level 1 Sorting Center 800000809* to *Level 2 Sorting Center 800003403* to *Level 3 Sorting Center 800008415* to *Level 4 Sorting Center800001617* to *Level 5 Sorting Center 800005373*. In **Route 00001**, the **backbone node sequence** is *deterministic*.

However, there are lots of paths in the road network when transporting packages between two backbone nodes (e.g., from warehouse 110007124 to level 1 sorting center 800000809) and these paths are *uncertain*.

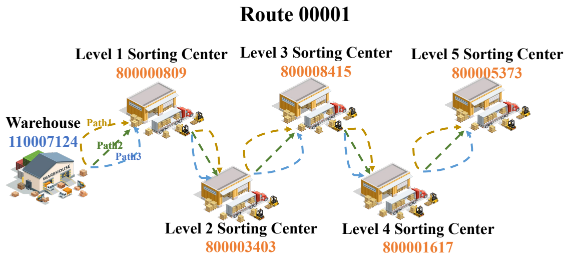

Figure 19: An example of full-link delivery route

For an example, as shown in Figure 20, when an order is created, the customer would fill in a **shipping address**, then the e-commerce platform will select a delivery route for the order according to the address, and send the order information to the warehouse in the pre-determined route. When packages accumulate to an amount or a time threshold is reached, **a truck of packages** will be delivered to level 1 sorting center in the delivery route. After disassembly and assembly processes, the packages are delivered to level 2 sorting center in the delivery route, and the process is repeated until the packages are received by customers.

In addition, when packages are delivered between warehouse and sorting center or between sorting centers in delivery route, truck drivers choose paths based on their preferences. Given the predetermined delivery routes and uncertain paths, it is still a challenge to estimate accurate time for each order since customer create the order under **dynamically changing temporal conditions**, and the dynamic temporal features could make **different impacts** on warehouses and sorting centers in delivery routes.

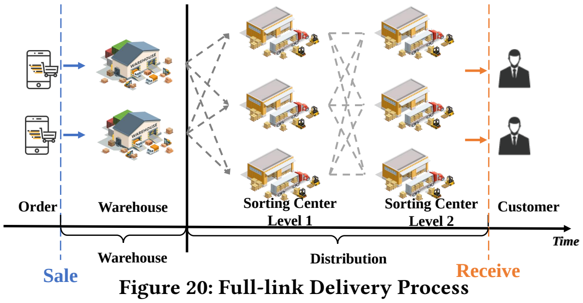

Figure 20: Full-Link Delivery Process

C.1.2 ***Feature Extraction***. The details of spatial features extraction and temporal features extraction are as below:

- ***Online Temporal Data.*** At time t, the online temporal data consists of two parts:
  (i) the online temporal data for warehouse W_i is denoted as $X_{W_i}^{T(t)}$, including the amount of packages processed in warehouse W_i as $X_{W_i}^{T(t)}[0]$, hourly cumulative sales volume related to warehouse W_i as $X_{W_i}^{T(t)}[1]$, daily cumulative sales volume elated to warehouse W_i as $X_{W_i}^{T(t)}[2]$, the total amount of packages being processed in downstream delivery routes of warehouse W_i as $X_{W_i}^{T(t)}[3]$.
  (ii) the online temporal data for sorting center S_j is denoted as $X_{S_j}^{T(t)}$, including the total amount of packages delivered in current routes as $X_{S_j}^{T(t)}[0]$, hourly cumulative sales volume related to sorting center S_j as $X_{S_j}^{T(t)}[1]$, daily cumulative sales volume related to sorting center S_j as $X_{S_j}^{T(t)}[2]$, the total amount of packages sent by upstream delivery units as as $X_{S_j}^{T(t)}[3]$.
  In addition, at time t, the background temporal data is denoted as $X_{BG}^{(t)}$, including time of day as $X_{BG}^{(t)}[0]$ (woking time:1, non-woking time:0), day of week as $X_{BG}^{(t)}[1]$ (woking day:1, non-woking day:0) and sales promotion period as $X_{BG}^{(t)}[2]$ (sales promotion period:1, normal period:0).
- ***Feature Extraction.*** At time 0, we feed the restructured IDs of warehouse and sorting centers into two embedding layers respectively to initialize the trainable spatial node representations $u_W^{S(0)}$ and $u_S^{S(0)}$.
  At time t, we define the temporal features of warehouse W_i as $X_{W_i}^{T(t)} = \{X_{W_i}^{T(t)}[0], X_{W_i}^{T(t)}[1], X_{W_i}^{T(t)}[2], X_{W_i}^{T(t)}[3]\}$, the temporal features of sorting center S_j as $X_{S_j}^{T(t)} = \{X_{S_j}^{T(t)}[0],$

$X_{S_j}^{T(t)}[1], X_{S_j}^{T(t)}[2], X_{S_j}^{T(t)}[3]\}$, and the background temporal data $X_{BG}^{(t)} = \{X_{BG}^{(t)}[0], X_{BG}^{(t)}[1], X_{BG}^{(t)}[2]\}$

C.1.3 Synthetic Dataset. To benefit the SIGKDD community and improve the reproducibility of our work, we provide the codes of our proposed HST-GT framework and a synthetic dataset in github repository (**https://github.com/zxh991103/HST-GT**), which includes data of 30 warehouses, 44 sorting centers, 312 delivery routes and hourly temporal data in 10 days. The detail of the dataset is described in the **Dataset Readme**.

Besides, the codes, synthetic dataset, and other data are stored in **Google Drive**, with Google Colaboratory, our work could be reproduced quickly.

C.2 Baseline Details

We compare our model with four categories of baselines including segment-based spatial methods, segment-based spatial-temporal methods, end-to-end spatial methods, and end-to-end spatial-temporal methods.

Segment-based Spatial Methods:

- **MLP/Xgboost**[4, 18]: We utilize MLP and Xgboost to represent traditional machine learning algorithms, and learn spatial information from the homogeneous warehouse-center graph.
- **Multi-MLP/Multi-Xgboost**: We set double MLP/ Xgboost as Multi-MLP/Multi-Xgboost to learn spatial information from the heterogeneous warehouse-center graph.

Segment-based Spatial-Temporal Methods:

- **ST-LSTM/ST-GRU** [29]: ST-LSTM is a deep learning approach that combines spatial-temporal features for short-term forecast, and we utilize it to predict delivery time. Then, we develop a variant of GRU, i.e., ST-GRU for prediction.
- **Multi-ST-LSTM/Multi-ST-GRU** : Double ST-LSTM/ ST-GRU are set as Multi-ST-LSTM/ Multi-ST-GRU to learn spatial information for heterogeneous graph.

End-to-End Spatial Methods:

- **GCN/GAT/GT**: We utilize homogeneous graph neural networks including graph convolution network (GCN)[17], graph attention network (GAT)[33] and graph transformer (GT)[45] to mine graph knowledge by considering the warehouse-center graph as a homogeneous graph.
- **RGCN/HAN/HS-GT** : We also utilize heterogeneous graph neural networks including relational graph convolution network (RGCN)[27], heterogeneous graph attention network (HAN)[36] and heterogeneous spatial graph transformer (HS-GT)[13] to learn nodes representations for the warehouse-center heterogeneous graph.

End-to-End Spatial-Temporal Methods:

- **ConstGAT/ST-GCN**: ConstGAT[8] is a model for travel time estimation, and ST-GCN[44] is a framework for traffic forecasting. We reform them on the delivery time estimation task by changing the warehouse-center graph into a homogeneous graph.
- **HTGT** [7]: Heterogeneous temporal graph transformer (HTGT) is an intelligent system for evolving android malware detection, and we transform its task to delivery time estimation.
- **ST-GT/ST-RGCN/ST-HAN**: We reform the spatial model in HTGT with GT/RGCN/HAN as ST-GT/ST-RGCN/ST-HAN.

C.3 Ablation Model Details

We develop the variants of our model to verify the effectiveness of each component including:

- **HST-GT without Spatial-Temporal Correlation Mining Model (w/o STCM)**: In order to verify the impact of spatial-temporal related information, the spatial-temporal correlation mining model is removed.
- **HST-GT without Spatial-Temporal Fusion Model (w/o STFM)**: we remove the spatial-temporal fusion model to verify the role of the temporal information in enhancing the spatial representations.
- **HST-GT with Spatial-Temporal Correlation Mining Model Reverse QKV of Attention (w/ CMRA)**: we reverse the query (Q), key (K) and value (V) of attention mechanism to verify the effect of specific attention in spatial-temporal related models.
- **HST-GT with Spatial-Temporal Fusion Model Reverse QKV of Attention (w/ FMRA)**: we reverse the query (Q), key (K) and value (V) of attention mechanism to verify the effect of specific attention in spatial-temporal fusion model.
- **HST-GT without Downstream Information Model (w/o DIM)**: The downstream information in prediction module is removed to verify its effect.
- **HST-GT without Contextual Information Model (w/o CIM)**: The contextual information in prediction module is removed to verify its effect.

C.4 Further Results

C.4.1 Parameter Sensitivity.

- **Effect of Different Hidden Sizes**: Fig. 21 shows the performance comparison *w.r.t.* the hidden size in our model. Overall, the performance is relatively stable under different sizes. It indicates that a smaller hidden size has insufficient representation while a larger embedding size may increase complexity and cause over-fitting, and the best hidden size is 64. We present the result of MAE since there are similar observations on other metrics.

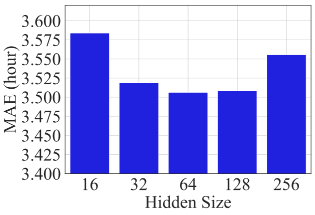

Figure 21: Effect of different hidden sizes

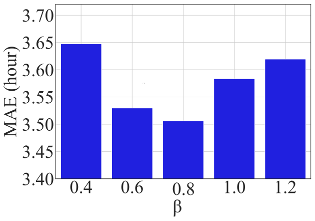

Figure 22: Performance under different values of β

Table 7: Performance comparison of different spatial-temporal datasets

Dataset	Metric	Area 1 (SG1)	Area 2 (SG2)	Area 3 (SG3)	Overall (FG)
Time Slice 1 (TS1)	MAE	3.4889	3.4098	3.5124	3.4471
	MAPE	0.0931	0.8994	0.0979	0.0910
	SMAPE	0.0822	0.0810	0.0853	0.0811
	RMSE	4.7897	4.7105	4.8130	4.7522
Time Slice 2 (TS2)	MAE	3.5129	3.4801	3.7988	3.5561
	MAPE	0.0997	0.0909	0.1102	0.0982
	SMAPE	0.0903	0.0816	0.1017	0.0870
	RMSE	4.8730	4.7863	5.3222	4.9505
Time Slice 3 (TS3)	MAE	3.8246	3.5132	3.9742	3.7982
	MAPE	0.1109	0.0972	0.1120	0.1097
	SMAPE	0.1027	0.0891	0.1092	0.1071
	RMSE	5.5907	4.8078	5.6880	5.3128
Overall (FT)	MAE	3.5217	3.4494	3.7191	**3.5063**
	MAPE	0.0985	0.0917	0.1079	**0.0954**
	SMAPE	0.0891	0.0825	0.1041	**0.0842**
	RMSE	4.9637	4.7701	5.2252	**4.8009**

Besides, we compared our model with the state-of-the-art method (HTGT[7] with Adaption) on seven spatial-temporal datasets including TS1-FG, TS2-FG, TS3-FG, FT-FG, FT-SG1, FT-SG2, and FT-SG3. As shown in Fig 23, 24, 25, and 26, we find that our model outperforms the HTGT on several spatial-temporal datasets, which indicates the robustness of HST-GT.

- **Performance with different** β: We further study the sensitivity of our model to the parameter β. We present the result based on MAE in Fig. 22. We found the overall performance is stable with different values of β. In our work, we select 0.8 as the value of β, which produces the best result for our model.

C.4.2 Robustness Experiments.

To fully evaluate our model, we choose three areas in real-world dataset, each area, each includes about 100 warehouses and more than 200 relevant sorting centers. We select three time slices (***Time Slice*** *i*). Then, 16 spatial-temporal datasets are combined by different time slices and areas. Each dataset is the same as the original dataset in terms of the spatial structure and temporal structure. By utilizing spatial-temporal dataset in different sizes, we validate the spatial-temporal robustness of our model.

Same as the experiment setting applied on the origin dataset, we select the first 80% of data as the training data and the remaining as the test data. As shown in Table 7, we train and test our model on these datasets, and the model performs stably under different spatial-temporal conditions. Our proposed **HST-GT** is effective for logistics companies with different sizes as long as it is the full-link delivery in the warehouse-distribution integration mode.

C.4.3 Feature-level Ablation Study.

In order to verify the validity of the temporal features, we design extra feature-level ablation experiments by utilizing the synthetic dataset. There are 11 feature-level variants of our model:

HST-GT without $X_{\underline{W}}^{T(t)}$ [0](w/o W0),
HST-GT without $X_{\underline{W}}^{T(t)}$ [1](w/o W1),
HST-GT without $X_{\underline{W}}^{T(t)}$ [2](w/o W2),
HST-GT without $X_{\underline{W}}^{T(t)}$ [3](w/o W3),
HST-GT without $X_{\underline{S}}^{T(t)}$ [0](w/o S0),
HST-GT without $X_{\underline{S}}^{T(t)}$ [1](w/o S1),
HST-GT without $X_{\underline{S}}^{T(t)}$ [2](w/o S2),
HST-GT without $X_{\underline{S}}^{T(t)}$ [3](w/o S3),
HST-GT without $X_{\underline{BG}}^{T(t)}$ [0](w/o BG0),
HST-GT without $X_{\underline{BG}}^{T(t)}$ [1](w/o BG1),
HST-GT without $X_{\underline{BG}}^{T(t)}$ [2](w/o BG2).

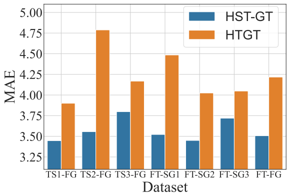

Figure 23: MAE Comparision

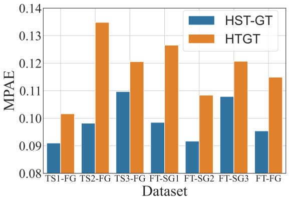

Figure 24: MPAE Comparision

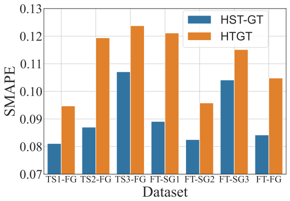

Figure 25: SMPAE Comparision

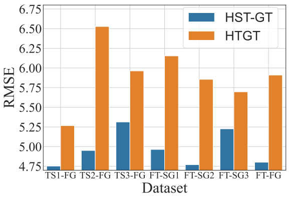

Figure 26: RMSE Comparision

The result of feature-level ablation experiments is shown in Table 8.

We find that all temporal features are usefull to predict the full-link delivery time, and the feature that make the greatest impact on predicting time in warehouse is the daily cumulative sales volume

Table 8: Results of feature-level ablation experiments

Variant	Metric	Warehouse Time	Sorting Time	Full-link Delivery Time
w/o W0	MPAE	0.0625	0.0448	0.0914
	SMPAE	0.0575	0.0363	0.0834
w/o W1	MPAE	0.0565	0.0431	0.0824
	SMPAE	0.0524	0.0349	0.0759
w/o W2	MPAE	0.1049	0.0658	0.1592
	SMPAE	0.0907	0.0541	0.1387
w/o W3	MPAE	0.0843	0.0530	0.1419
	SMPAE	0.0859	0.0440	0.1302
w/o S0	MPAE	0.0711	0.0768	0.1186
	SMPAE	0.0598	0.0695	0.1070
w/o S1	MPAE	0.0453	0.0665	0.0850
	SMPAE	0.0352	0.0617	0.0815
w/o S2	MPAE	0.0413	0.0559	0.0705
	SMPAE	0.0365	0.0525	0.0651
w/o S3	MPAE	0.0625	0.0833	0.1089
	SMPAE	0.0571	0.0710	0.0912
w/o BG0	MPAE	0.0484	0.0866	0.0908
	SMPAE	0.0401	0.0679	0.0814
w/o BG1	MPAE	0.0414	0.0547	0.0784
	SMPAE	0.0358	0.0466	0.0701
w/o BG2	MPAE	0.0538	0.0714	0.0977
	SMPAE	0.0492	0.0655	0.0847
HST-GT	MPAE	**0.0395**	**0.0338**	**0.0488**
	SMPAE	**0.0385**	**0.0330**	**0.0471**

($X_W^{T(t)}[2]$), which the feature that make The most significant impact on predicting time in downstream sorting centers is the total amount of packages delivered in current routes ($X_S^{T(t)}[0]$).

C.4.4 Other Results.

- As shown in Fig27, we compare the overall full-link delivery time prediction results. It can be seen that (i) the curve of HST-GT is the closest to the ground truth, indicating great performance on most of the test data; (ii) the curve of HST-GT is more smooth than the state-of-the-art method HTGT, indicating the appropriate training settings of our model to avoid overfitting.
- We show the sorting rank scores of sorting centers in Fig.28, which are sorted by the decrease of the score. In real-world scenarios, the sorting centers in distribution network are divided into three levels, and we also found two inflection points in the curve, which indicates great practical performance of our sorting rank algorithm.

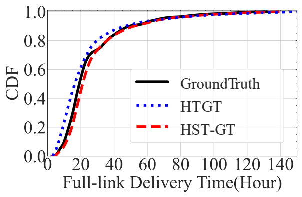

Figure 27: Delivery Time

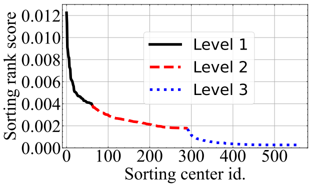

Figure 28: Sorting Rank Score

D DETIALED DISCUSSION

Lessons learned: Based on the results from our paper, we summarize the following lessons learned:

- *Full-link time estimation in the warehouse-distribution integration e-commerce should consider the heterogeneous contextual information*, which makes it different from the traditional warehouse-distribution separation mode. As shown in Fig. 3 and Fig. 2, the time in warehouse is significantly influenced by downstream sorting centers and the sorting time is also affected by the neighbor sorting centers. Similar results are obtained in the comparison of our model with baselines in Table 1, where end-to-end methods outperform the segment-based methods.
- *Dynamic temporal data has different influences on heterogeneous delivery units*. As shown in Fig. 4, due to the heterogeneity of delivery units, the sales volume has different impacts on warehouse and sorting centers, and the difference is captured by the specific attention mechanisms designed in our model. The performance of the model when reversing the attention parameters is shown in Fig. 14 and Fig. 16, and the results validate that the critical attention modules of our model efficiently exploit the specific effect of heterogeneous nodes.

Limitation: We evaluate our HST-GT model on a real-world dataset from one of the largest online e-commerce retailers in China. However, due to the privacy and sensitivity of data, the period of the dataset is one month and the detailed geographic location is removed. We further expect to analyze our model with a dataset with long periods and richer geographical information.