STSC-Net: Leveraging Spatial, Temporal, and Semantic Context for Road Network Trajectory Prediction

Qingjie Liu, Meng Chen✉, and Li Pan✉

School of Software, Shandong University, Jinan, Shandong, China
qingjieliu@mail.sdu.edu.cn, {mchen, panli}@sdu.edu.cn

Abstract. This paper addresses the challenge of road network trajectory prediction in complex urban environments, where capturing both spatial and temporal dependencies is essential for accurate movement prediction. We propose a hybrid model that integrates Graph Neural Networks (GNNs) and Transformers to jointly model spatial relationships and temporal dynamics in road networks. The GNN component includes a fine-grained spatial and semantic encoder that captures road attributes such as connectivity, segment characteristics, and localized topological details, providing a deeper understanding of road network structure. The Transformer component models long-range temporal dependencies, complementing the spatial insights provided by the GNN. Furthermore, we introduce a Contextual Candidate Attention Module that leverages an attention mechanism to dynamically evaluate and prioritize candidate road segments, ensuring that the model's predictions are contextually informed and semantically rich. Experiments on real-world datasets demonstrate that our model outperforms state-of-the-art methods.

Keywords: Trajectory Prediction · Urban Mobility · Spatiotemporal Modeling.

1 Introduction

The rapid proliferation of GPS-enabled devices and advancements in data collection technologies have resulted in substantial spatiotemporal trajectory data. These advancements present significant opportunities to enhance the understanding of human mobility patterns and advance intelligent transportation systems [1, 5, 18, 11, 20, 24]. A valuable problem in human mobility mining is predicting the next road segment that a moving object will traverse based on its historical trajectory. As illustrated in Figure 1(a), a trajectory is composed of a sequence of road segments, each associated with a timestamp indicating traversal time. The objective of next visited segment prediction is to determine the most probable subsequent segment based on previously traversed segments and their timestamps.

✉ Corresponding author.

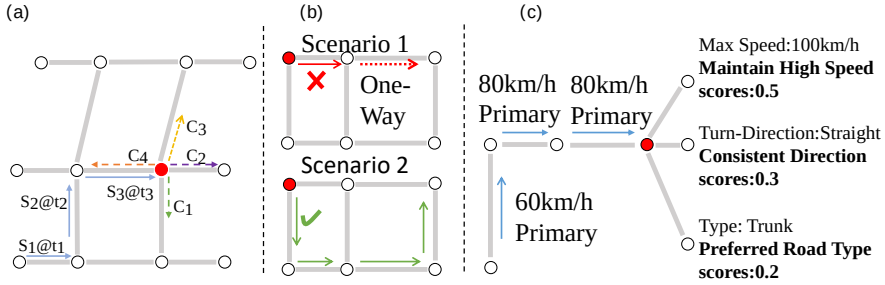

Fig. 1. Illustration of trajectory prediction and the rationale behind our proposed model. (a) Next road segment prediction: the moving object has passed through segments S_1, S_2, and S_3, with the goal of predicting the most likely next segment from candidates C_1, C_2, C_3, and C_4. (b) An example demonstrating how multi-hop attributes influence trajectory prediction. (c) An example showing how contextually relevant candidates improve prediction accuracy.

Traditional trajectory prediction methods, such as Markov Chains and Hidden Markov Models (HMMs) [2, 12], rely on historical data to forecast future movements, assuming that recent history heavily influences future behavior. While effective for short-term dependencies, these methods struggle with long-range dependencies and complex mobility patterns, thus limiting their performance in dynamic settings. In contrast, modern deep learning techniques, including Long Short-Term Memory (LSTM) networks [8] and Transformers [16], excel at modeling long-term dependencies and capture extended temporal relationships more effectively. Consequently, studies [10, 17, 22, 9] have shown that these advanced approaches significantly outperform traditional methods in capturing both temporal and spatial contexts.

While the deep models mentioned above effectively analyze temporal sequences, they often fail to capture the complex spatial-semantic relationships present in urban road networks. Predicting trajectories in these networks depends on dynamic attributes of road segments, including road type, length, direction, speed limit, and one-way restrictions, which are crucial for determining future paths. The road network's topology also influences trajectory predictions, as attributes from multiple segments can impact current routing decisions. Thus, both immediate and multi-hop attributes are essential for accurate trajectory prediction in complex urban settings. For example, as illustrated in Figure 1(b), a vehicle entering a complex area might initially choose a seemingly optimal route based on immediate segment information, such as suitable type (e.g., residential street), length, and direction. However, this path could lead to a dead-end or restricted zone due to one-way rules or speed limit changes that are not evident from nearby segments but occur further along. Therefore, capturing multi-hop attributes is critical, as relying solely on immediate road segment information may not suffice for accurate predictions.

Furthermore, the operational context of a vehicle significantly influences its choice among multiple road segments, as shown in Figure 1(c). For instance, when approaching an intersection with three options—one with a higher speed limit, another aligned with the vehicle's current heading for smoother travel, and a third that matches the vehicle's preferred road type—the vehicle evaluates these segments based on its recent speed, direction, and road type preferences. Each candidate is assigned a relevance score, making it more likely for the vehicle to select the segment that best fits its context. Thus, a model that dynamically prioritizes candidate road segments based on these factors will yield more accurate predictions than one that treats all candidate segments equally.

To this end, we propose a novel model that integrates the spatial-semantic relationships of road networks with the temporal dynamics of trajectories, while dynamically focusing on contextually relevant candidate segments. Our approach consists of three key components: First, we utilize a Graph Attention Network (GAT) to capture the complex spatial-semantic relationships in urban road networks, encompassing both immediate and multi-hop attributes. The GAT processes the road network graph to generate enriched embeddings for each segment, incorporating information such as type, length, direction, speed limits, and connectivity. Second, we introduce a Spatial-Semantic Enhanced Temporal Sequence Encoder, which captures temporal dependencies in the sequence of road segments traversed by moving objects. This encoder combines the GAT's spatial-semantic embeddings with temporal factors like time of day and day of the week, enhancing the Transformer's input to better represent movement patterns and the object's spatiotemporal features. Finally, we design a Contextual Candidate Attention Module that dynamically prioritizes candidate road segments based on their relevance to the object's current context, assigning scores to focus on the most appropriate options for predicting the next segment. By jointly modeling spatial-semantic information, temporal dynamics, and contextual relevance, our model effectively addresses the limitations of existing methods and improves prediction accuracy in complex urban environments.

Our contributions are summarized as follows:

- We tackle the challenge of modeling complex spatial-temporal dependencies in urban trajectory prediction by creating a unified framework that integrates the structural and semantic features of road networks with temporal movement dynamics. This comprehensive approach improves prediction accuracy and provides valuable insights into urban mobility patterns.
- We advance the modeling of candidate road segments by enriching their representations with comprehensive semantic attributes. Unlike existing methods that treat candidate segments uniformly or consider limited features, our approach dynamically prioritizes candidates based on their contextual alignment with the moving object's current context, thereby improving prediction accuracy in complex urban environments.
- We validate the effectiveness and robustness of our model through extensive experiments on various real-world datasets. The results consistently show

that our approach outperforms state-of-the-art methods, confirming its superiority and generalizability.

2 Related Work

Our research is situated within the broad area of human mobility prediction, a field focused on forecasting future movements using historical trajectory data. Within this domain, two principal approaches have emerged: next location prediction and next road segment prediction. While both approaches aim to predict subsequent movements, they differ in terms of the trajectory representation and the granularity of their prediction targets. In this section, we review recent advancements in both areas to contextualize our work within the existing body of literature.

2.1 Next Location Prediction

Recent approaches to next location prediction have largely adopted sequence-based methods and graph-based methods. Sequence-based methods rely on deep sequential models, such as convolutional neural networks (CNNs)[3, 25], recurrent neural networks (RNNs)[6, 10, 22], and Transformers[15], to capture the spatiotemporal information in location visit sequences. For example, CEM[3] employs one-dimensional CNNs to model the relative ordering of locations to predict the next location. DeepMove[6] uses attentional recurrent networks to predict movement patterns based on current and historical trajectories. ST-RNN [10]integrates spatiotemporal context into the hidden states of RNNs, while Flashback[22] incorporates spatiotemporal context to weight hidden states for next location prediction using RNNs. MCLP[15] employs the Transformer architecture to mine sequential patterns and incorporate multiple contextual factors for predicting the next location.

Graph-based models[13, 23, 19], on the other hand, enhance location semantic context and capture moving trend information by constructing transition graphs for users and locations. For instance, Graph-Flashback[13] builds a point of interest (POI) transition graph to improve POI representation, which is then input into an RNN for next POI prediction. GETNext[23] constructs a trajectory flow map to encode location transition information, and subsequently integrates a Transformer to predict the next location. MobGT[19] integrates spatial and temporal graph encoders to effectively capture both spatial and temporal features in users' mobility patterns.

2.2 Next Road Segment Prediction

In contrast to next location prediction, which focuses on user interests and preferences (e.g., predicting the next point of interest), next road segment prediction must account for the structure of the road network and the physical constraints of road segments. CSSRNN [17] extends RNNs by incorporating spatial context

to better capture the road network topology, while LPIRNN [17] treats each road segment prediction as an individual task, improving the modeling of topological influences. RA-LSTM [4] leverages road-aware features to assist LSTM in modeling temporal dependencies, enhancing its ability to capture both temporal and spatial characteristics of trajectory data. Although these approaches share some similarities with our proposed STSC-Net, they are limited in their ability to effectively model road segment semantics and complex spatial dependencies, thereby constraining their adaptability and overall performance.

To our knowledge, we are the first to explicitly model semantic dependencies in road networks, more effectively capturing their complexity and thereby improving the accuracy and adaptability of trajectory predictions.

3 Preliminaries

In this section, we present fundamental definitions and formally introduce the problem addressed in this paper.

Definition 1 (Road Network). *A road network is modeled as a directed graph $G = (V, E, \mathbf{F})$, where V represents the set of vertices corresponding to road segments, and $E \subseteq V \times V$ denotes the directed connections between these segments. An edge exists between two vertices if their road segments are directly connected, indicating possible movement. Each segment is defined by attributes such as road type, direction, length, speed limit, and geographic coordinates, which are organized in a feature matrix $\mathbf{F} \in \mathbb{R}^{|V| \times f}$, with $|V|$ being the number of segments and f the number of features for each segment.*

Definition 2 (Road Network Constrained Trajectory). *A trajectory is defined as an ordered sequence of road segments $\tau = (v_1, t_1), (v_2, t_2), \ldots, (v_n, t_n)$, with $t_1 \leq t_2 \leq \cdots \leq t_n$. Each $v_i \in V$ represents a traversed road segment, which is linked to a timestamp t_i, indicating when it was traversed.*

Definition 3 (Candidate Road Segments). *For a current road segment v_n, the set of candidate road segments $C(v_n) \subseteq V$ includes all segments directly reachable from v_n based on the adjacency relationships in E.*

Definition 4 (Next Road Segment Prediction). *Given a road network $G = (V, E, \mathbf{F})$ and a trajectory $\tau = (v_1, t_1), (v_2, t_2), \ldots, (v_n, t_n)$, the objective is to predict the next road segment $v_{n+1} \in V$ that the trajectory will reach at the subsequent time step t_{n+1}.*

4 Method

4.1 Model Overview

Figure 2 illustrates the architecture of our proposed model, which consists of three main modules: (1) Road Network Feature Encoding, (2) Spatial-Semantic

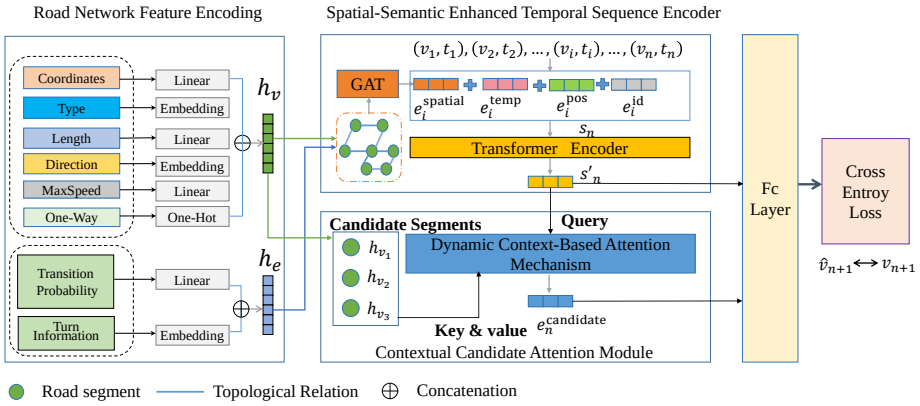

Fig. 2. Overall architecture of the proposed STSC-Net model.

Enhanced Temporal Sequence Encoder, and (3) Contextual Candidate Attention Module. First, we encode road segments and their connections as feature vectors that capture key spatial and semantic attributes of the road network. Second, we model spatial-semantic relationships and temporal dependencies by using a Graph Attention Network (GAT) to learn spatial-semantic embeddings and a Transformer-based sequence encoder for temporal modeling. Third, we employ an attention mechanism to dynamically weight the features of candidate road segments based on the current spatiotemporal context. Finally, we predict the next road segment by concatenating the context embedding and the weighted candidate embedding, then passing them through a fully connected layer to generate predicted probabilities for the candidate segments.

4.2 Road Network Feature Encoding

We begin by detailing how road network properties are encoded into feature vectors. Each node $v \in V$ is represented by a feature vector $\mathbf{h}_v = [\tilde{\mathbf{x}}_{\text{coord}}, \mathbf{e}_{\text{type}}, \tilde{x}_{\text{length}}, \mathbf{e}_{\text{direction}}, \tilde{x}_{\text{maxSpeed}}, \mathbf{e}_{\text{one\_way}}]$, which captures important road-specific attributes including geographic coordinates, road type, length, direction, speed limit, and one-way status.

- **Coordinates**, **Length**, and **MaxSpeed** are numerical features normalized using z-score normalization to ensure comparability across scales. After normalization, these features are encoded through a linear transformation: $\tilde{x} = W_{\text{num}} x' + b_{\text{num}}$, where x' represents the normalized numerical feature and W_{num}, b_{num} are learnable parameters for linear encoding.
- **Type** and **Direction** are categorical features embedded using embedding matrices: $\mathbf{e}_{\text{type}} = \mathbf{E}_{\text{type}}(\text{type}), \mathbf{e}_{\text{direction}} = \mathbf{E}_{\text{direction}}(\text{direction})$, where $\mathbf{E}_{\text{type}}$ and $\mathbf{E}_{\text{direction}}$ are embedding matrices mapping categorical values to embedding vectors.

– **One-Way Status** is a categorical feature indicating whether the road segment is one-way or two-way. This feature is encoded using one-hot encoding: $\mathbf{e}_{\text{one\_way}} = \text{one\_hot}(\text{one-way status})$, where $\text{one\_hot}(\cdot)$ is the one-hot encoding function mapping the categorical value to a binary vector.

Similarly, the properties of the edges E in the road network are encoded to capture important attributes that influence movement. Each edge $e \in E$ is represented by a feature vector $\mathbf{h}_e = [\tilde{p}_e, \mathbf{e}_{\text{turn}}]$, which includes

– **Transition Probability** is estimated from historical trajectory data, representing the likelihood of transitioning between road segments. This feature is normalized and encoded using a linear transformation: $\tilde{p}_e = W_{\text{edge}} p'_e + b_{\text{edge}}$, where p'_e is the normalized transition probability, and W_{edge} and b_{edge} are learnable parameters for the encoding.
– **Turn Information** represents the type of turn—straight, U-turn, left, or right—encoded as a categorical feature using an embedding matrix, $\mathbf{e}_{\text{turn}} = \mathbf{E}_{\text{turn}}(\text{turn type})$, where $\mathbf{E}_{\text{turn}}$ maps the categorical turn type to an embedding vector.

4.3 Spatial-Semantic Enhanced Temporal Sequence Encoder

To accurately predict the next road segment given a sequence of traversed segments, it is essential to understand spatial-semantic relationships and temporal dependencies in movement patterns. We propose a Spatial-Semantic Enhanced Temporal Sequence Encoder that combines a Graph Attention Network (GAT) with a Transformer-based sequence encoder. This architecture enables the model to learn detailed spatial-semantic embeddings of road segments, enhancing the Transformer's input and improving prediction accuracy.

Road Segment Spatial-Semantic Embedding via GAT We use a Graph Attention Network (GAT) to model spatial-semantic interactions among road segments in the network. It takes as input the features of road segments $\mathbf{h}_v$ and edge features $\mathbf{h}_e$. The goal is to create a spatial-semantic embedding $\mathbf{e}_i^{\text{spatial}}$ for each road segment v_i by aggregating information from neighboring segments through an edge-aware attention mechanism, which evaluates the importance of each neighboring segment $v_j \in \mathcal{N}(v_i)$ based on their respective node features $\mathbf{h}_{v_j}$ and edge features $\mathbf{h}_{e_{ij}}$:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_{v_i} \parallel \mathbf{W}\mathbf{h}_{v_j} \parallel \mathbf{W}e\mathbf{h}_{e_{ij}}]\right)\right)}{\sum_{v_k \in \mathcal{N}(v)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_{v_i} \parallel \mathbf{W}\mathbf{h}_{v_k} \parallel \mathbf{W}e\mathbf{h}_{e_{ik}}]\right)\right)}, \tag{1}$$

where $\mathbf{W}$ and $\mathbf{W}_e$ are learnable weight matrices for node and edge features, $\mathbf{a}$ is the attention vector, and $\parallel$ indicates concatenation.

The spatial-semantic embedding for node v_i is computed by aggregating information from its neighbors, weighted by the learned attention coefficients:

$$\mathbf{e}_i^{\text{spatial}} = \sigma\Big(\sum_{v_j \in \mathcal{N}(v_i)} \alpha_{ij} \mathbf{W}\mathbf{h}_{v_j}\Big), \tag{2}$$

where σ is a non-linear activation function, such as ReLU. This process allows the model to adaptively focus on the most relevant spatial-semantic information, enhancing its understanding of intricate dependencies in urban road networks.

Temporal Sequence Encoding with Enhanced Input The Temporal Sequence Encoder captures sequential dependencies among road segments traversed by moving objects by integrating spatial-semantic embeddings from the GAT, temporal factors like time of day and day of the week, and unique road segment identifiers. This integration enhances the Transformer's input, allowing for improved modeling of movement patterns. The input to the Transformer encoder includes:

- **Spatial-Semantic Embedding** ($\mathbf{e}_i^{\text{spatial}}$): Derived from the GAT for road segment v_i.
- **Temporal Embedding** ($\mathbf{e}_i^{\text{temp}}$): Mapped from temporal features, such as hour of day and day of the week at v_i, to embeddings.
- **Positional Embedding** ($\mathbf{e}_i^{\text{pos}}$): Sinusoidal positional encoding representing position i in the sequence.
- **Segment ID Embedding** ($\mathbf{e}_i^{\text{id}}$): Embedding for the road segment v_i.

The combined input embedding for v_i in the sequence is $\mathbf{s}_i = \mathbf{e}_i^{\text{spatial}} + \mathbf{e}_i^{\text{temp}} + \mathbf{e}_i^{\text{pos}} + \mathbf{e}_i^{\text{id}}$. Then, this enhanced input sequence $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ is passed through the Transformer encoder, which employs a multi-head self-attention mechanism to capture dependencies across the sequence,

The combined input embedding for v_i is $\mathbf{s}_i = \mathbf{e}_i^{\text{spatial}} + \mathbf{e}_i^{\text{temp}} + \mathbf{e}_i^{\text{pos}} + \mathbf{e}_i^{\text{id}}$. This enhanced input sequence $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ is then processed by the Transformer encoder, which uses a multi-head self-attention mechanism to capture dependencies throughout the sequence for generating the output embeddings,

$$\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_n = \text{Transformer}(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n), \tag{3}$$

where each embedding $\mathbf{s}'_i$ captures both the spatial-semantic and temporal patterns of the corresponding road segment.

4.4 Contextual Candidate Attention Module

To improve the prediction of the next road segment, we introduce the Contextual Candidate Attention Module, which focuses on adjacent candidate segments. While the Temporal Sequence Encoder captures overall movement patterns, this module refines prediction by utilizing features from neighboring segments. By dynamically weighting these candidates according to the current context, the model significantly enhances its predictive performance.

Dynamic Context-Based Attention Mechanism In this module, the attention mechanism utilizes the output $\mathbf{s}'_n$ from the Temporal Sequence Encoder as the query, and the features of candidate road segments as the keys and values.

- **Query** ($\mathbf{s}'_n$): The output embedding $\mathbf{s}'_n$ from the Temporal Sequence Encoder corresponding to the current road segment v_n.
- **Keys and Values** ($\mathbf{h}_v$): The features of candidate road segments $v \in \mathcal{C}(v_n)$, where $\mathcal{C}(v_n)$ denotes the candidate set of road segment v_n. These features are the node representations $\mathbf{h}_v$ encoded in the Road Network Feature Encoding.

The attention weights are computed by comparing the query with the keys:

$$\alpha_{nj} = \frac{\exp\left(\frac{\left(\mathbf{W}_k \mathbf{h}_{v_j}\right)^\top \left(\mathbf{W}_q \mathbf{s}'_n\right)}{\sqrt{d_k}}\right)}{\sum_{v_k \in \mathcal{C}(v_n)} \exp\left(\frac{\left(\mathbf{W}_k \mathbf{h}_{v_k}\right)^\top \left(\mathbf{W}_q \mathbf{s}'_n\right)}{\sqrt{d_k}}\right)}, \tag{4}$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable weight matrices for the queries and keys, d_k is the dimensionality of the key vectors, and α_{nj} represents the attention weight between candidate road segment v_j and the current context. The output is a weighted sum of the candidate features:

$$\mathbf{e}_n^{\text{candidate}} = \sum_{v_j \in \mathcal{C}(v_n)} \alpha_{nj} \mathbf{W}_v \mathbf{h}_{v_j}, \tag{5}$$

where $\mathbf{W}_v$ is a learnable weight matrix.

4.5 Prediction Layer

We combine the context embedding $\mathbf{s}'_n$ and the weighted candidate embedding $\mathbf{e}_n^{\text{candidate}}$, and pass through a fully connected layer with a softmax activation to produce the probability distribution over candidate road segments:

$$\hat{\mathbf{y}}_{n+1} = \text{softmax}\left(\mathbf{W}_o \left[\mathbf{s}'_n || \mathbf{e}_n^{\text{candidate}}\right] + \mathbf{b}_o\right), \tag{6}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are learnable parameters of the output layer, $\left[\mathbf{s}'_n || \mathbf{e}_n^{\text{candidate}}\right]$ denotes the concatenation of the context embedding and the weighted candidate embedding, and $\hat{\mathbf{y}}_{n+1}$ is the predicted probability distribution over the candidate road segments at time step $n+1$. The model is trained by minimizing the cross-entropy loss between the predicted probability distribution $\hat{\mathbf{y}}_{n+1}$ and the ground truth label.

5 Experiments

5.1 Experimental Setup

Datasets We conducted an experimental evaluation using two real-world taxi datasets, namely the Chengdu Dataset and the Jinan Dataset. The Chengdu dataset[1] contains taxi location data collected from a district in Chengdu, China,

[1] https://outreach.didichuxing.com

Table 1. Dataset statistics

	Chengdu Data	Jinan Data
Number of road segments	6,088	6,230
Number of trajectories	700,889	802,034
Average length of each trajectory	29.34	32.04

over the period from November 1 to November 30, 2016. The Jinan dataset contains comparable data collected from Jinan, China, from April 1 to May 1, 2023. Each record includes GPS coordinates and a timestamp. For both datasets, the location data from each taxi order was assembled into vehicle trajectories by sorting the locations chronologically. Subsequently, we used OpenStreetMap[2] to obtain road network data for the corresponding areas and applied a map-matching algorithm [21] to align the taxi locations with road segments. Trajectories with fewer than ten road segments were excluded. Finally, we partitioned the datasets based on timestamps: the first 20 days were used as the training set, the next 5 days as the validation set, and the final 5 days as the test set. Key statistical information for the datasets is provided in Table 1.

Evaluation Metrics Following similar works in the field [17, 9], we used Prediction Accuracy (ACC) and Negative Log-Likelihood (NLL) as evaluation metrics:

- **Prediction Accuracy (ACC):** This metric measures the percentage of correct predictions for the next road segment. Specifically, ACC determines whether the road segment with the highest predicted probability matches the true next segment at each step. For a test set containing N trajectories, ACC is defined as:

$$\mathrm{ACC} = \frac{1}{\sum_{i=1}^{N} len(i)} \sum_{i=1}^{N} \sum_{n=1}^{len(i)-1} \mathbb{I}\left(\arg\max_{v \in C(v_n)} P(v \mid v_{1:n}) = v_{n+1} \right), \quad (7)$$

  where $C(v_n)$ represents the set of candidate road segments that can be reached from v_n, $len(i)$ is the length of the ith trajectory, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the predicted segment matches the true segment, and 0 otherwise.
- **Negative Log-Likelihood (NLL):** This metric measures the model's ability to assign high probabilities to the correct next road segments over an entire test trajectory. For a test set containing N trajectories, NLL is defined as:

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{n=1}^{len(i)-1} \log P(v_{n+1} \mid v_{1:n}), \quad (8)$$

  where $P(v_{n+1} \mid v_{1:n})$ represents the probability assigned by the model to the next road segment v_{n+1} given the previous sequence of segments $v_{1:n}$.

[2] http://www.openstreetmap.org

Baselines To assess the effectiveness of STSC-Net, we evaluated its performance against several state-of-the-art baselines.

- **N-gram [14]:** N-gram models capture transition probabilities between road segments based on historical data. Second-order and third-order Markov models use the last two or three road segments, respectively, to predict the next segment.
- **RNN [7]:** The RNN baseline uses a recurrent neural network (RNN) to predict the next road segment from historical trajectory data.
- **CSSRNN [17]:** CSSRNN extends the standard RNN by incorporating spatial context to address the complexity of road network topology.
- **LPIRNN [17]:** LPIRNN integrates road network structure by treating each road segment prediction as an independent task, using topological information as an external input to the RNN.
- **MMTraj [9]:** MMTraj is a Transformer-based model that employs a multi-task learning framework to predict the next road segment and the final trajectory destination. It also uses Graph Convolutional Networks (GCNs) to model the road network topology.
- **Flashback [22]:** Flashback is a location-based trajectory prediction model that uses RNNs to capture sequential dependencies while incorporating spatiotemporal context to weight the hidden states for next location prediction. Spatiotemporal context is defined as the temporal and spatial distance from the current location. In our setting, locations are represented as road segments.
- **Graph-Flashback [13]:** Graph-Flashback uses GCNs to model a Point-of-Interest (POI) transition graph, enhancing POI representations before inputting them into an RNN model for next POI prediction. In our setting, POIs are represented as road segments, and the POI transition graph corresponds to the road network transition graph, with transition weights defined by estimated transition probabilities between road segments.
- **GETNext [23]:** GTET-Next constructs a trajectory flow map to encode location transition information into location embeddings, which are then combined with time embeddings and input into a Transformer model to capture sequential dependencies and predict the next location. Similar to Flashback, locations are represented as road segments.

Experiment Settings The model was implemented in PyTorch, and experiments were conducted on an Intel i9-10900K CPU and NVIDIA RTX 3090 GPU. Key hyperparameters are summarized below. The embedding dimensions for road segment identifiers and attributes were set to 64, and the temporal feature encoding dimension was set to 64. The Graph Attention Network (GAT) module consists of three layers with two attention heads per layer, each having an embedding dimension of 128. The GAT output layer has an embedding dimension of 64. The Transformer module has two layers, each with four attention heads. The Dynamic Context-Based Attention Mechanism uses four attention heads. A dropout rate of 0.2 was applied to the GAT and Transformer modules.

Table 2. Overall Performance of STSC-Net and Baseline Models

Methods	Chengdu		Jinan	
	ACC (%)	NLL	ACC (%)	NLL
N-gram (2nd-order)	85.23	8.31	87.45	7.82
N-gram (3rd-order)	86.31	7.78	88.59	7.36
RNN	87.08	7.29	88.98	7.08
CSSRNN	87.45	6.35	89.24	6.39
LPIRNN	87.49	6.31	89.19	6.42
MMTraj	87.77	5.92	89.38	6.03
Flashback	87.81	5.87	89.56	6.00
Graph-Flashback	88.05	5.73	89.77	5.96
GETNext	88.16	5.58	89.80	5.72
STSC-Net (Ours)	**88.82**	**5.21**	**90.33**	**5.19**

The model was optimized using the Adam optimizer, with an initial learning rate of 1×10^{-3} and a weight decay of 5×10^{-4}. The model was trained on both datasets using a batch size of 256 for 40 epochs. Optimal hyperparameters were determined via grid search.

5.2 Results and Discussion

Overall Performance Table 2 presents the performance of STSC-Net compared to baseline models on both datasets, evaluated in terms of ACC and NLL. We observer that:

- The N-gram models provide a useful baseline for understanding the benefits of advanced sequence modeling. Despite their simplicity, the 2nd-order and 3rd-order N-gram models achieve moderate prediction accuracies (e.g., 86.31% for Chengdu with 3rd-order) but fall short compared to deep learning approaches. This performance gap highlights the importance of capturing non-linear and long-range dependencies prevalent in urban mobility data, which simple probabilistic models like N-grams are unable to capture fully.
- Deep learning models (RNN, CSSRNN, LPIRNN and Flashback) exhibit significant performance improvements over traditional N-gram models. This result highlights the effectiveness of deep learning approaches in capturing complex temporal dependencies and non-linear relationships inherent in urban trajectory data.
- Incorporating graph neural networks into models (e.g., MMTraj, Graph-Flashback, and GETNext) further enhances prediction accuracy, highlighting the critical importance of modeling spatial topological structures.
- Models such as Graph-Flashback and GETNext underscore the critical importance of modeling transition patterns. By explicitly incorporating transition probabilities, they effectively capture inherent sequential dependencies and spatial relationships, thereby enhancing prediction accuracy.

Table 3. Ablation Study Results for STSC-Net

Variant	Chengdu		Jinan	
	ACC (%)	NLL	ACC (%)	NLL
STSC-Net	**88.82**	**5.21**	**90.33**	**5.19**
w/o RNFE	88.45	5.58	89.92	5.64
w/o TE	88.68	5.37	90.16	5.29
w/o CCAM	88.34	5.71	89.84	5.82

– Our proposed STSC-Net achieves the best performance by integrating spatial-semantic relationships, temporal dynamics, and contextual relevance. Compared to GETNext, STSC-Net increases accuracy by 0.66% and 0.53% on the two datasets and reduces NLL by 0.37 and 0.53. Since we focus on next road segment prediction, even a 0.5% improvement in accuracy can lead to significant cumulative benefits in large-scale applications.

Ablation Study To evaluate the contribution of each component in STSC-Net, we conducted an ablation study by systematically removing key modules and assessing their impact on performance. STSC-Net comprises three essential components: (1) Road Network Feature Encoding, (2) Temporal Sequence Encoder, and (3) Contextual Candidate Attention Module. We designed three variants to evaluate the effects of these components:

– **STSC-Net w/o RNFE**: We remove the Road Segment Spatial-Semantic Encoder ($\mathbf{e}^{spatial}$) and retain only the Temporal Sequence Encoder and Contextual Candidate Attention Module for trajectory prediction.
– **STSC-Net w/o TE**: We remove the temporal embedding $\mathbf{e}^{temp}$ in the Spatial-Semantic Enhanced Temporal Sequence Encoder.
– **STSC-Net w/o CCAM**: We remove Contextual Candidate Attention Module and use the Road Segment Spatial-Semantic Encoder and the Contextual Candidate Attention Module for trajectory prediction.

The results of the ablation study, presented in Table 3, illustrate the impact of each component on the model's accuracy (ACC) and Negative Log Likelihood (NLL) across the Chengdu and Jinan datasets.

Impact of Removing the Road Network Feature Encoding. Removing the Road Segment Spatial-Semantic Encoder led to a decrease in accuracy from 88.82% to 88.45% on the Chengdu dataset and from 90.33% to 89.92% on the Jinan dataset. Concurrently, the NLL increased from 5.21 to 5.58 for Chengdu dataset and from 5.19 to 5.64 for Jinan dataset. These declines indicate that the Road Segment Spatial-Semantic Encoder plays a pivotal role in capturing spatial and semantic relationships within the road network, which is critical for accurate trajectory prediction.

Impact of Removing the Temporal Embedding in the Temporal Sequence Encoder. Removing the Temporal Embedding resulted in a slight reduction in accuracy to 88.68% for Chengdu dataset and 90.16% for Jinan dataset, with NLL values increasing to 5.37 and 5.29, respectively. This suggests that temporal information is essential for modeling time-dependent mobility patterns, enhancing the model's ability to predict the next road segment based on temporal dynamics.

Impact of Removing the Contextual Candidate Attention Module. The most significant performance degradation was observed when the Contextual Candidate Attention Module was omitted. Accuracy dropped to 88.34% for Chengdu dataset and 89.84% for Jinan dataset, while NLL increased to 5.71 and 5.82, respectively. This substantial decline underscores the critical importance of the Contextual Candidate Attention Module in integrating context-sensitive semantics of candidate road segments, which significantly contributes to making informed and accurate predictions.

Summary of Ablation Study Results. The ablation study confirms that each component of STSC-Net is integral to its superior performance. The Road Network Feature Encoding is crucial for accurately capturing spatial and semantic relationships within the road network. Temporal Embedding effectively models time-dependent patterns, enhancing the model's predictive capabilities. The Contextual Candidate Attention Module plays a vital role in integrating contextual relevance, leading to more precise and confident predictions. The combined effect of these components validates the comprehensive design of STSC-Net, demonstrating that the integration of spatial, temporal, and contextual semantic information significantly enhances trajectory prediction performance in complex urban environments.

6 Conclusion and Future Work

In this paper, we addressed the challenge of road network trajectory prediction in complex urban environments by introducing STSC-Net, a context-aware model that jointly captures spatial, temporal, and semantic features through a combination of the Road Network Feature Encoding, the Spatial-Semantic Enhanced Temporal Sequence Encoder, and the Contextual Candidate Attention Module. By employing a GNN-based Road Network Feature Encoder, our model effectively captures the detailed structure of road networks, while the Transformer-based Sequence Encoder models temporal dynamics, allowing for a comprehensive understanding of movement patterns. The Contextual Candidate Attention Module further enhances prediction accuracy by focusing on contextually relevant candidate road segments using an attention mechanism. Our extensive experiments on real-world datasets demonstrated that STSC-Net outperforms state-of-the-art methods, validating the effectiveness and robustness of our approach in real urban environments. Future work will aim to improve the model's adaptability to various urban settings, incorporate real-time data

to enhance responsiveness in trajectory prediction, and integrate user-specific information to further personalize and refine prediction accuracy.

Acknowledgments. The authors would like to acknowledge the support provided by the National Key R&D Program of China under Grant 2023YFB4004503, and project ZR2023LZH016 supported by the Shandong Provincial Natural Science Foundation.

References

1. Chen, M., Liu, Q., Huang, W., Zhang, T., Zuo, Y., Yu, X.: Origin-aware location prediction based on historical vehicle trajectories. ACM Transactions on Intelligent Systems and Technology **13**(1), 1–18 (2021)
2. Chen, M., Liu, Y., Yu, X.: Nlpmm: A next location predictor with markov modeling. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 186–197. Springer (2014)
3. Chen, M., Zuo, Y., Jia, X., Liu, Y., Yu, X., Zheng, K.: Cem: A convolutional embedding model for predicting next locations. IEEE Transactions on Intelligent Transportation Systems **22**(6), 3349–3358 (2020)
4. Cui, J., Zhou, X., Zhu, Y., Shen, Y.: A road-aware neural network for multi-step vehicle trajectory prediction. In: Proceedings of International Conference on Database Systems for Advanced Applications. pp. 701–716. Springer (2018)
5. Deng, L., Zhao, Y., Sun, H., Yang, C., Xie, J., Zheng, K.: Fusing local and global mobility patterns for trajectory recovery. In: International Conference on Database Systems for Advanced Applications. pp. 448–463. Springer (2023)
6. Feng, J., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., Jin, D.: Deepmove: Predicting human mobility with attentional recurrent networks. In: Proceedings of the 2018 world wide web conference. pp. 1459–1468 (2018)
7. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
8. Hochreiter, S.: Long short-term memory. Neural Computation MIT-Press (1997)
9. Liu, K., Ruan, S., Xu, Q., Long, C., Xiao, N., Hu, N., Yu, L., Pan, S.J.: Modeling trajectories with multi-task learning. In: 2022 23rd IEEE International Conference on Mobile Data Management (MDM). pp. 208–213. IEEE (2022)
10. Liu, Q., Wu, S., Wang, L., Tan, T.: Predicting the next location: A recurrent model with spatial and temporal contexts. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)
11. Ni, J., Zhang, X., Jin, B., Zhang, F., Li, X., Huang, Q., Wang, P., Li, X., Xiao, N., Wang, Y., et al.: A localization system for gps-free navigation scenarios. In: International Conference on Database Systems for Advanced Applications. pp. 268–273. Springer (2022)
12. Qiao, Y., Si, Z., Zhang, Y., Abdesslem, F.B., Zhang, X., Yang, J.: A hybrid markov-based model for human mobility prediction. Neurocomputing **278**, 99–109 (2018)
13. Rao, X., Chen, L., Liu, Y., Shang, S., Yao, B., Han, P.: Graph-flashback network for next location recommendation. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 1463–1471 (2022)
14. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)

15. Sun, T., Fu, K., Huang, W., Zhao, K., Gong, Y., Chen, M.: Going where, by whom, and at what time: Next location prediction considering user preference and temporal regularity. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2784–2793 (2024)
16. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
17. Wu, H., Chen, Z., Sun, W., Zheng, B., Wang, W.: Modeling trajectories with recurrent neural networks. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. pp. 3083–3090 (2017)
18. Xu, S., Xu, J., Li, B., Fu, X.: Predicting where you visit in a surrounding city: A mobility knowledge transfer framework based on cross-city travelers. In: International Conference on Database Systems for Advanced Applications. pp. 334–350. Springer (2023)
19. Xu, X., Suzumura, T., Yong, J., Hanai, M., Yang, C., Kanezashi, H., Jiang, R., Fukushima, S.: Revisiting mobility modeling with graph: A graph transformer model for next point-of-interest recommendation. In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. pp. 1–10 (2023)
20. Xu, Y., Xu, J., Fang, J., Liu, A., Zhao, L.: When multitask learning make a difference: Spatio-temporal joint prediction for cellular trajectories. In: International Conference on Database Systems for Advanced Applications. pp. 207–223. Springer (2022)
21. Yang, C., Gidofalvi, G.: Fast map matching, an algorithm integrating hidden markov model with precomputation. International Journal of Geographical Information Science **32**(3), 547–570 (2018)
22. Yang, D., Fankhauser, B., Rosso, P., Cudre-Mauroux, P.: Location prediction over sparse user mobility traces using rnns. In: Proceedings of the twenty-ninth international joint conference on artificial intelligence. pp. 2184–2190 (2020)
23. Yang, S., Liu, J., Zhao, K.: Getnext: trajectory flow map enhanced transformer for next poi recommendation. In: Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval. pp. 1144–1153 (2022)
24. Yao, D., Guo, F., Wen, Z., Guo, Y., Cheng, P., He, Y., Bi, J.: Trajectory completion via context-guided neural filtering and encoding. In: International Conference on Database Systems for Advanced Applications. pp. 3–19. Springer (2024)
25. Zhong, T., Zhang, S., Zhou, F., Zhang, K., Trajcevski, G., Wu, J.: Hybrid graph convolutional networks with multi-head attention for location recommendation. World Wide Web **23**, 3125–3151 (2020)