

Entropy-Driven Curriculum for Multi-Task Training in Human Mobility Prediction

Tianye Fang
Technical University of Munich
Munich, Germany
tianye.fang@tum.de

Xuanshu Luo
Technical University of Munich
Munich, Germany
xuanshu.luo@tum.de

Martin Werner
Technical University of Munich
Munich, Germany
martin.werner@tum.de

Abstract—The increasing availability of big mobility data from ubiquitous portable devices enables human mobility prediction through deep learning approaches. However, the diverse complexity of human mobility data impedes model training, leading to inefficient gradient updates and potential underfitting. Meanwhile, exclusively predicting next locations neglects implicit determinants, including distances and directions, thereby yielding suboptimal prediction results. This paper presents a unified training framework that integrates entropy-driven curriculum and multi-task learning to address these challenges. The proposed entropy-driven curriculum learning strategy quantifies trajectory predictability based on Lempel-Ziv compression and organizes training from simple to complex for faster convergence and enhanced performance. The multi-task training simultaneously optimizes the primary location prediction alongside auxiliary estimation of movement distance and direction for learning realistic mobility patterns, and improve prediction accuracy through complementary supervision signals. Extensive experiments conducted in accordance with the *HuMob Challenge* demonstrate that our approach achieves state-of-the-art performance on GEO-BLEU (0.354) and DTW (26.15) metrics with up to 2.92-fold convergence speed compared to training without curriculum learning.

Index Terms—human mobility, entropy, curriculum learning, multi-task learning, Transformer, geospatial artificial intelligence

I. INTRODUCTION

The inherent regularity of human mobility data, which exhibits predictability of individual mobility patterns across diverse populations and travel distances [1], provides the foundation for numerous location-based applications, including urban planning and management, transportation optimization, epidemic modeling, and recommendation systems [2]–[7]. With the proliferation of pervasive user devices with passive location acquisition capabilities, unprecedented volumes of human mobility data have been collected, enabling data-driven approaches, particularly sequential deep learning models, to effectively extract human mobility patterns [8]–[11]. In comparison to handcrafted pattern matching [12]–[14] and Markov models [15]–[17], deep learning methods generally achieve superior long-term prediction performance.

Training these deep learning models conventionally involves random dataset shuffling and uniform mini-batch sampling, thereby implicitly treating all mobility data as equally difficult

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16DHBKI020.

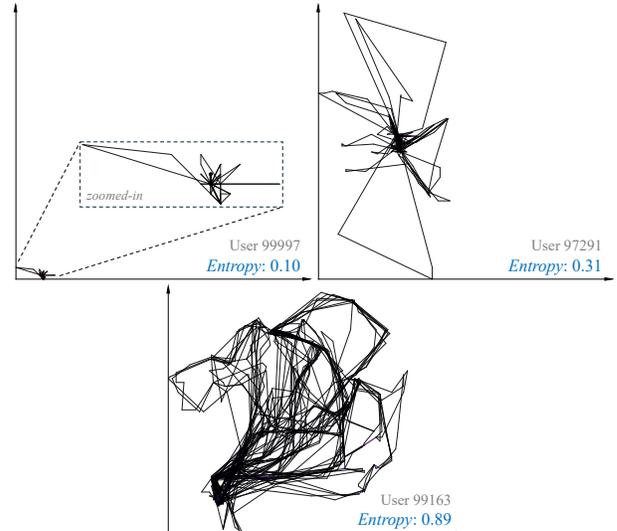


Fig. 1. Three trajectories with heterogeneous complexity in the YJMob100K dataset [18], presented at the same spatial scale. The entropy reflects trajectory predictability, which will be explained in Section III-A and III-B.

to learn, which conflicts with the intrinsic heterogeneous complexity of human mobility data, as illustrated in Fig. 1. Specifically, simple, repetitive routines (e.g., daily commutes) are fundamentally easier to model than irregular, sparse trajectories and complex trips with multiple stops. Consequently, such training procedures force models to learn from intricate examples early when the foundational understanding of mobility patterns remains undeveloped, leading to inefficient learning, training instability, premature convergence, and degraded prediction performance. Therefore, presenting human mobility data in order of increasing difficulty constitutes a rational training strategy, aligning with the concept of curriculum learning [19]–[21]. However, to the best of our knowledge, curriculum learning has not been extensively applied to human mobility prediction. Moreover, there is a lack of practical and validated methods to quantitatively evaluate trajectory complexity for data reordering in curriculum organization.

Furthermore, most human mobility prediction approaches predominantly focus on next-location prediction as a single training objective while neglecting other inherent mobility characteristics that could provide valuable supervision signals.

Concretely, human mobility inherently involves multifaceted decision-making processes where one may simultaneously consider destinations, means of transportation, and directional preferences based on urban structure or personal routines. By optimizing solely for location accuracy, models fail to capture these complementary aspects of mobility behaviors. Therefore, integrating multiple training objectives through multi-task learning (MTL) presents a promising approach to leverage these auxiliary mobility characteristics and enhance the primary next-location prediction [22], enabling models to learn more comprehensive and realistic representations of human mobility. Although several studies have utilized MTL and obtained superior prediction performance [10], [23], [24], their auxiliary task semantics are context-specific, depending heavily on underlying datasets. For example, the next activity [10], [23] and transportation modes [24] are not universally available in human mobility datasets, leading to limited generalization capabilities to other datasets and tasks.

To address these issues, this paper presents a unified training framework that integrates entropy-driven curriculum and MTL for human mobility prediction. Concretely,

- The proposed entropy-driven curriculum learning strategy quantifies trajectory predictability by estimating mobility entropy based on the Lempel-Ziv (LZ) compression principle [25], [26]. Training is then organized using data with increasing mobility entropy and prediction horizon, incorporating trajectory augmentation through mirroring and rotation, followed by fine-tuning exclusively on real trajectories for adaptation to original datasets.
- The primary location prediction task is jointly optimized alongside two universally applicable auxiliary estimations on moving distances and directions, which are inherently present in any mobility dataset and provide complementary supervision signals for more comprehensive mobility representations without additional data annotations.

Regarding the underlying model in the proposed framework, we present MoBERT, an encoder-only Transformer designed for human mobility prediction based on BERT [27], [28], which utilizes multi-feature embeddings, including temporal, spatial, and potential semantic information, coupled with feature interaction mechanisms to capture complex dependencies. We evaluate our method on YJMOb100K [18], an open-source urban-scale human mobility dataset utilized in the *HuMob Challenge* [29], [30], with standardized data splits and evaluation metrics: GEO-BLEU [31] and dynamic time warping (DTW) [32]. We adhere precisely to the same settings as the *HuMob Challenge* and demonstrate that our approach achieves state-of-the-art GEO-BLEU score of 0.354 and DTW distance of 26.15, with up to 2.92-fold convergence speed compared to training without the entropy-driven curriculum. For previously unseen cities, our method also exhibits superior zero-shot generalization capability. These results validate the effectiveness of the proposed framework for human mobility prediction. The code of this work is available here¹.

¹<https://github.com/tum-bgd/2025-IEEEBigData-HuMob>

II. BACKGROUND AND RELATED WORK

A. Human mobility prediction

A human mobility trajectory is a temporally ordered sequence of spatiotemporal points that captures an individual’s movement over time. Specifically, let a j -step trajectory be represented as $\mathbf{T}_{1:j} = \{p_1, \dots, p_j\}$, where each point p_i contains at minimum a spatial coordinate $l_i = (x_i, y_i)$ and a corresponding timestamp t_i . Human mobility prediction aims to forecast individuals’ future locations and movement patterns based on their historical trajectory data. Given an observed trajectory prefix of length n , denoted as $\mathbf{T}_{1:n}$, the objective is to predict the subsequent trajectory segment $\mathbf{T}_{n+1:n+k}$, where k is the prediction horizon, by maximizing the conditional probability $P(\mathbf{T}_{n+1:n+k} \mid \mathbf{T}_{1:n})$. The prediction accuracy depends on capturing the underlying spatiotemporal dependencies and behavioral patterns encoded in the observed trajectory. Spatial coordinates may be continuous geographical coordinates such as longitude and latitude, or discrete grid indices in rasterized spaces, making the prediction a regression or classification task, respectively. Meanwhile, each point p_i may also comprise semantic information s_i , e.g., the functional category of l_i , providing additional contribution to prediction.

Early approaches to human mobility prediction relied on either handcrafted rules [12]–[14] or Markov models [15]–[17], where the latter treats location transitions as probabilistic state changes. However, these methods are generally insufficient to characterize the complexity of human mobility and capture the nonlinear dependencies and long-range patterns inherent in human movement [33]. On one hand, human movement exhibits complex periodic patterns, and actual trajectory transitions do not conform to the assumptions of simple Markov chains. On the other hand, individuals often display different travel preferences, resulting in group-based rules performing poorly at the individual level. Moreover, these methods typically exploit only a limited history of past states, making it difficult to capture long-term dependencies and demonstrating limited scalability to increasingly large datasets.

The emergence of deep learning and the availability of big mobility data have revolutionized human mobility prediction. Recurrent neural networks and their variants have become the dominant approaches for modeling sequential trajectory data [8]–[11]. These models naturally handle variable-length sequences and capture temporal dependencies through their recurrent structure. However, they suffer from gradient vanishing problems in long sequences and process data sequentially, thereby limiting parallelization efficiency. There are also methods utilizing convolutional neural networks to extract local spatial patterns [34], [35], but they struggle with global spatial and temporal dependencies and lack generalization capability across different spatial areas. While graph neural networks excel at learning local spatial relationships, these models often lack explicit temporal modeling, and their global spatial understanding remains limited [36], [37]. Transformers [27] represent a more advanced approach by handling long-range spatiotemporal dependencies simultaneously. Although

Transformers typically require considerably large datasets for proper training [38], this is not problematic given the progressively expanding volume of human mobility data [18]. The analysis above motivates us to design a Transformer-like model in our proposed framework.

B. Curriculum learning

Curriculum learning is a training strategy mimicking human education, where learning progresses from simple to complex examples rather than through random sampling [20]. The core methodology of curriculum learning is organizing training data by difficulty and gradually introducing more challenging examples as model competence develops. This approach contrasts with conventional training that randomly samples mini-batches, treating all examples as equally difficult throughout the training process [19]. Curriculum learning typically outperforms random sampling for several reasons. First, starting with simple examples helps establish robust feature representations and prevents the model from struggling with complex data early in training. Second, the gradual increase in difficulty creates a smoother optimization landscape, reducing the likelihood of becoming trapped in poor local minima. Third, this approach naturally implements a form of regularization, as the model must maintain performance on simple examples while adapting to complex ones, thereby preventing overfitting to edge cases [21].

Unfortunately, to the best of our knowledge, curriculum learning has rarely been applied to human mobility prediction despite the prevalent heterogeneous complexity of human mobility data, as illustrated in Figure 1. One marginally relevant work is *ST-Curriculum Dropout* for spatial-temporal graph modeling to forecast traffic speed and taxi demand [39], which measures vertex difficulty according to its latent representation. This approach relies heavily on rational model initialization and is susceptible to variations during training, potentially inducing misleading curricula that deteriorate training performance. Therefore, this paper focuses on data itself, formulating mobility entropy as a feasible measure of trajectory predictability independent of underlying models, thereby providing a reliable foundation for curriculum organization.

C. Multi-task learning

Multi-task learning (MTL) is a paradigm where a model simultaneously learns multiple related tasks, sharing representations to improve generalization across all objectives [22]. Unlike single-task learning (STL), which optimizes for one specific target, MTL leverages task relationships through parameter sharing, typically implemented via shared encoder layers with task-specific heads. This architecture enables the model to learn more robust features by considering multiple aspects of the data simultaneously. MTL generally outperforms STL for following reasons. First, auxiliary tasks act as regularizers, preventing overfitting to the main task by introducing additional training signals. Second, shared representations must capture features relevant to all tasks, leading to more generalizable learned patterns. Third, MTL effectively

increases the amount of training data per shared parameter, as gradients from multiple tasks contribute to updating common layers. MTL is also beneficial when labeled data for the main task is limited, as related tasks can provide extra supervision.

In the context of human mobility prediction, the sparse and noisy nature of real-world mobility data exacerbate the challenge of learning robust representations from a single supervisory signal. Consequently, numerous methods have integrated MTL to optimize the primary mobility prediction task alongside activity type classification [10], [23] or transportation mode recognition [24]. However, these auxiliary tasks require additional annotations that may not be universally available. In contrast, our approach introduces distance and direction prediction as auxiliary tasks, as elaborated in Section III-D, which can be derived from any trajectory dataset without additional labeling, thereby ensuring broad applicability while maintaining the benefits of MTL for enhanced accuracy.

III. METHODOLOGY

This section aims to elucidate the proposed unified training framework. We first employ Fano’s inequality [40] to demonstrate that low-entropy trajectories are fundamentally more learnable, thereby establishing the theoretical foundation for entropy-driven curriculum organization. Subsequently, the normalized Lempel–Ziv mobility entropy estimator is proposed to quantify the predictability of mobility data. The entropy-driven curriculum learning procedure is then introduced along with trajectory augmentation strategies, followed by the design details of MoBERT, a BERT-based encoder-only Transformer for human mobility prediction that supports MTL.

A. The relationship between entropy and predictability

The relationship between entropy and predictability in human mobility is grounded in information theory. Entropy measures the uncertainty or randomness in a sequence of observations, where higher entropy indicates greater unpredictability. For human mobility, trajectories with low entropy correspond to highly regular movement patterns, such as repetitive commuting routes between home and workplace. Conversely, high-entropy trajectories reflect more exploratory or irregular behaviors, such as tourist activities spanning diverse urban areas. The connection between mobility entropy and predictability limits can be established by Fano’s inequality [40]. For a discrete random variable X with alphabet size Q and entropy $H(X)$, Fano’s inequality provides a lower bound on the probability of prediction error P_e :

$$H(X) \leq H(P_e) + P_e \log_2(Q - 1),$$

where $H(P_e) = -P_e \log_2 P_e - (1 - P_e) \log_2(1 - P_e)$ is the binary entropy of the error probability. Alternatively, let $\Phi = 1 - P_e$ be the maximum achievable prediction accuracy, we can rearrange the above inequality, yielding:

$$H(X) \leq -\Phi \log_2 \Phi - (1 - \Phi) \log_2(1 - \Phi) + (1 - \Phi) \log_2(Q - 1).$$

This relationship shows an upper bound on entropy given a certain level of predictability. The equality holds only when the

prediction is optimal. In the context of human mobility prediction, X represents the next location choice, Q corresponds to the total number of possible locations in a rasterized area, and $H(X)$ measures the uncertainty in location selection. When entropy approaches its maximum value $\log_2 Q$, the prediction accuracy cannot exceed random guessing ($\Phi \approx 1/Q$). This occurs when movement patterns are completely random, with each location equally possible for the next step regardless of history. Conversely, when entropy approaches zero, the bound allows for near-perfect prediction ($\Phi \approx 1$). This scenario corresponds to highly deterministic movement patterns where future locations are totally determined by past behavior. Therefore, trajectories with low entropy are not only more regular but also fundamentally more learnable, rendering them suitable for early training phases and vice versa.

B. The normalized Lempel–Ziv mobility entropy

The above theoretical foundation distinguishes the proposed entropy-driven method from heuristic difficulty measures, providing a principled approach grounded in information theory. In practice, we propose a normalized Lempel–Ziv mobility entropy estimator derived from Lempel–Ziv (LZ) compression [25], [26]. Our approach approximates mobility entropy by analyzing the rate at which new subsequences appear in the trajectory, specifically encompassing the following steps:

- 1) *Trajectory symbolization*: Two-dimensional (2D) coordinates $l = (x, y)$ are flattened into 1D through the transformation $x \cdot (y_{max} + 1) + y$, where y_{max} represents the maximum column index (counted from 0) in the spatial grid. This transformation preserves spatial relationships while facilitating sequential analysis.
- 2) *LZ parsing*: Symbolized trajectories are processed from left to right, maintaining a dictionary \mathcal{D} of observed subsequences. At each position m , the shortest subsequence of length q that has not previously appeared in \mathcal{D} is recorded as a phrase, added to \mathcal{D} , and parsing continues from position $m+q$. The process continues until the end of the sequence.
- 3) *LZ entropy estimation*: The entropy is computed as:

$$H_{LZ} = \frac{\ln N}{\bar{Q} \ln 2},$$

where $N = \sum_{u=1}^{|\mathcal{D}|} Q_u$, $\bar{Q} = N/|\mathcal{D}|$ is the average phrase length, and $|\mathcal{D}|$ denotes the number of phrases in \mathcal{D} . Longer \bar{Q} indicates higher compressibility and lower entropy, reflecting more predictable movement patterns, whereas shorter \bar{Q} suggests the converse [26].

- 4) *Normalization*: H_{LZ} is normalized for comparability across trajectories of different lengths by

$$H_{norm-LZ} = \frac{H_{LZ}}{\log_2 N} \in [0, 1].$$

H_{LZ} is normalized by dividing by $\log_2 N$, which is the maximum possible entropy for a fully random sequence, bounding the entropy measure between zero and one, where values approaching one indicate near-random mobility patterns, and values approaching zero suggest highly predictable behaviors.

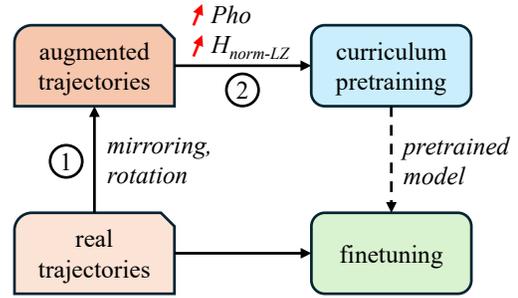


Fig. 2. The entropy-driven curriculum learning pipeline. ① Real trajectories are augmented by mirroring and rotation. ② Augmented trajectories are ordered with increasing $H_{norm-LZ}$ and subjected to progressively longer prediction horizon (Pho) to form a curriculum for pretraining. The pretrained model is subsequently finetuned using the raw dataset.

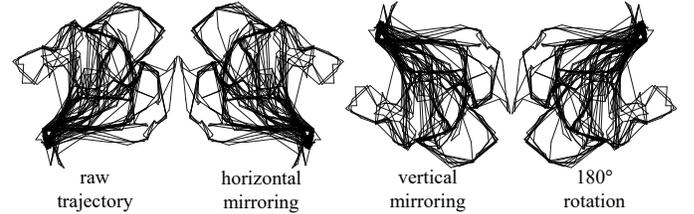


Fig. 3. Trajectory augmentation by mirroring and rotation.

C. Entropy-driven Curriculum Learning

Building upon the normalized Lempel–Ziv mobility entropy $H_{norm-LZ}$, the entropy-driven curriculum learning framework organizes training progression according to intrinsic trajectory complexity, with the complete pipeline illustrated in Fig. 2.

Real trajectories are first augmented through horizontal and vertical mirroring, as well as 180° rotation, quadrupling data volume while preserving underlying movement logic. Associated semantic information should also be transformed correspondingly to maintain consistency. Following augmentation, we perform entropy estimation of all trajectories by $H_{norm-LZ}$, with trajectories sorted by increasing entropy, creating a natural progression from highly predictable to complex behaviors. To further enrich training diversity, sorted trajectories are assigned increasing prediction horizon (Pho). For example, given a task requiring forecasting of the next five locations of a trajectory ($Pho = 5$), variant samples can be created with $Pho = 1, 2, \dots, 5$. The curriculum is then constructed by organizing augmented trajectories with increasing $H_{norm-LZ}$ and Pho , establishing progressive training difficulties.

This curriculum enables incremental learning, beginning with easily learnable patterns and gradually introducing more challenging scenarios. The entropy-based organization aligns training difficulty with information-theoretic predictability constraints, preventing early-phase overwhelm from complex patterns. Following curriculum pretraining, the model undergoes finetuning exclusively on real trajectories for optimal adaptation to authentic mobility characteristics. This two-stage pipeline balances large-scale pretraining on diverse augmented data with real-world adaptation, providing a principled ap-

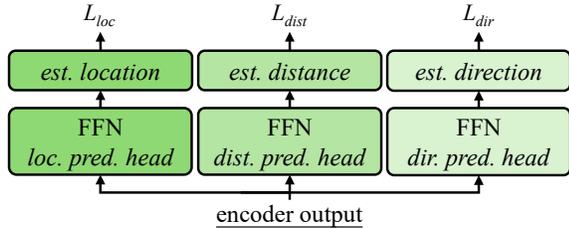


Fig. 4. The multi-task prediction head. Losses are computed by cross-entropy.

proach that leverages entropy-based complexity measures for structured learning progression and improved generalization compared to conventional random sampling.

D. Model design

To support MTL, this paper presents MoBERT, a BERT-like encoder-only model for human mobility prediction [28]. We adopt a BERT-style architecture for the following advantages. Unlike autoregressive models that generate predictions sequentially, BERT’s bidirectional encoding and parallel prediction capability prevent error accumulation in long-term forecasting. The self-attention mechanism naturally captures both periodic patterns (e.g., daily commutes) and irregular dependencies without the computational constraints of recurrent architectures. Moreover, the masked language modeling paradigm of BERT aligns well with the sparsity of mobility data, where trajectories are often missing or irregularly sampled [10].

The encoder part of MoBERT follows the architecture of the original BERT, where the input is adapted for human mobility data. Concretely, the encoder processes 3D input with dimensions $[B, M, E]$, where B denotes the batch size, M represents the length of historical trajectories, and E indicates the embedding dimension. The specific composition of E depends on the semantics provided in the dataset and minimally includes locations and corresponding timestamps.

For auxiliary tasks, we introduce distance and direction prediction, which are naturally available in any mobility dataset, avoiding dependence on specific datasets and ensuring the universality of MoBERT. These two straightforward tasks provide intrinsic spatial constraints: distance bounds the search space while direction provides orientation priors that complement the location prediction. In practice, both tasks can be formulated as classification or regression, depending on the dataset complexity and specific requirements. The prediction head of MoBERT is designed accordingly, as illustrated in Fig. 4. They share the same latent representations through the encoder while employing task-specific feed-forward networks (FFNs) [27]. We have the total loss

$$L = L_{loc} + \lambda_1 L_{dist} + \lambda_2 L_{dir},$$

where L_{loc} , L_{dist} , and L_{dir} represents the losses for location, distance, and direction estimation, respectively, and λ_1 , λ_2 denote the loss weights for auxiliary tasks. The values of λ_1 and λ_2 are determined via grid search within the range $[0, 1]$.

IV. DATASET AND ADAPTATIONS

To validate the effectiveness of the proposed method, this paper utilizes the YJMob100K dataset [18], a large-scale, real-world human mobility dataset jointly developed by multiple research institutions, serving as one of the most comprehensive open-source trajectory datasets available for academic research. This dataset was employed in the *HuMob Challenge* in 2023/24 [29], [30], with standardized train/test splits and evaluation metrics including GEO-BLEU [31] and DTW [32], facilitating transparent and fair comparison with existing studies that participated in the challenge. This section first briefly introduces the dataset, followed by the adaptations made to this dataset for the subsequent evaluations.

A. The YJMob100K dataset

The YJMob100K dataset captures over 110 million trajectory records through smartphone-based high-frequency GPS signals collected from approximately 100,000 active users within a major urban area of Japan during a continuous 75-day observation period, where the last 15 days are used for prediction. All raw geographic information was anonymized and discretized into a uniform 200×200 grid, where each cell measures 500 meters per side. Temporal data are similarly discretized into 48 uniform time slots of 30 minutes each, encompassing complete daily cycles. Each trajectory record consists of five key fields: an anonymous user identifier uId , day index dId ranging from 0 to 74, time slot index tId from 0 to 47, and spatial coordinates (x, y) indicating grid cell position. The dataset exhibits significant sparsity, reflecting realistic mobile phone usage patterns where users are not continuously tracked. The dataset includes comprehensive point-of-interest (POI) information for enhanced semantic understanding. Each grid cell contains counts for 85 anonymized POI categories covering dining, retail, education, healthcare, etc., enabling models to incorporate spatial semantics and urban context into mobility prediction tasks.

B. Dataset-specific adaptations

The YJMob100K dataset provides rich semantic information from which the following features can be derived. The day index dId can be further formulated into the day of the week ($dayOfWeek$) to reflect weekly patterns. The time slot index tId can also be derived to obtain time intervals ($timedelta$) between consecutive trajectory points, which capture temporal continuity, and day/night features ($dayNight$) to distinguish between diurnal and nocturnal activities. For POI, only the top- k most frequent POI categories, namely $topKPOI$, are considered, as most POI categories are rarely represented in the dataset. This approach emphasizes significant POIs while minimizing redundancy and noise, thereby enhancing the semantic representation of the environment. Consequently, the complete semantics s of each sample in this dataset is represented as $s = \{dayOfWeek, timedelta, dayNight, topKPOI\}$, while the timestamp t is defined as $t = \{dId, tId\}$. Combined with coordinates (x, y) , there are 8 features in total as input.

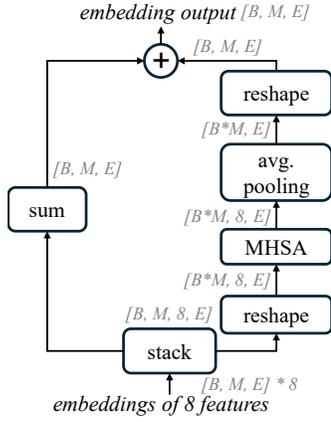


Fig. 5. Feature interaction pipeline. The output shapes are denoted in gray.

1) *Feature embedding interaction*: To effectively utilize all 8 features, we design a feature interaction module based on multi-head self-attention (MHSA), effectively fusing features and providing input to the MoBERT encoder. All feature embeddings are first stacked and undergo basic fusion via summation over the feature dimension, as shown in the left branch in Fig. 5. Subsequently, MHSA operates along the feature dimension to capture dynamic inter-dependencies, computing attention weights that adaptively emphasize different feature combinations based on context, for instance, prioritizing spatiotemporal interactions during commuting hours or *POI*-temporal relationships during leisure periods. The final representation combines both the basic fusion term (left) and the attention-based interaction term (right) via element-wise summation, preserving individual feature semantics while incorporating learned relational patterns that enhance the model’s ability to understand complex mobility behaviors. The results in Section V-B validate the effectiveness of this module.

2) *Curriculum stages*: Fig. 6 presents the distribution of $H_{norm-LZ}$ for trajectories in the YJMOb100K dataset, exhibiting a near-Gaussian distribution across the user population. Approximately 70% of users exhibit entropy values between 0.4 and 0.65, indicating moderate levels of predictability that balance routine patterns with exploratory behaviors. Fewer than 2% of users demonstrate extremely low entropy (below 0.2) or extremely high entropy (above 0.8), corresponding to highly repetitive or highly random movement patterns, respectively. Based on these statistics, we artificially divide the curriculum pretraining into three stages using augmented trajectories with increasing $H_{norm-LZ}$ and Pho , as given below:

- Stage 1: $H_{norm-LZ} < 0.4$, Pho is 3 days.
- Stage 2: $H_{norm-LZ} < 0.65$, Pho is 7 days.
- Stage 3: all data with the complete Pho of 15 days.

3) *MTL formulation*: The spatial area of the YJMOb100K dataset is discretized into a grid, rendering the primary location estimation a classification task. For auxiliary distance and direction estimation, we also formulate these as classification tasks. Specifically, for distance estimation, the Euclidean distance $dist$ between consecutive trajectory points

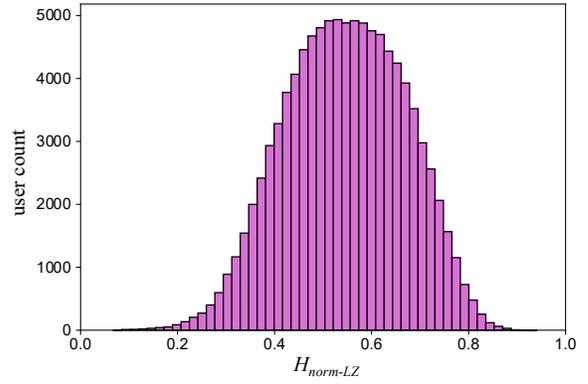


Fig. 6. The distribution of $H_{norm-LZ}$ for trajectories in YJMOb100K.

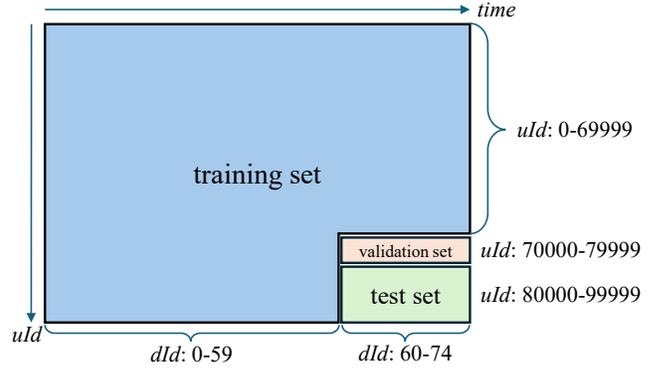


Fig. 7. Training, validation, and test set split of the YJMOb100K dataset.

is discretized into four classes: stationary ($dist < 500 m$), short-range ($500 m \leq dist < 1000 m$), medium-range ($1000 m \leq dist < 3500 m$), and long-range ($dist \geq 3500 m$), corresponding to typical movement scales such as stationary, walking, running/cycling, and driving behaviors. Movement direction prediction discretizes the geographic direction into nine classes, including four cardinal directions (N, E, S, W), four ordinal directions (NE, SE, SW, NW), and stationary.

V. EVALUATION

We evaluate our methods elaborated in Section III using the YJMOb100K dataset introduced in Section IV-A with corresponding adaptation strategies described in Section IV-B.

A. Experimental settings

The YJMOb100K dataset is split by uld into training, validation, and test sets with a 7:1:2 ratio, as illustrated in Fig. 7. We keep the test set identical to the *HuMob challenge* for fair comparison. This split prevents data leakage and enables evaluation of the generalization capabilities of models for unseen users. The training set contains 70,000 users, the validation set 10,000 users, and the test set 20,000 users. On the training set, the model is trained on the first 60 days of user trajectories and tested on predicting the subsequent 15-day period (days 60-74). The same metrics used in the *HuMob challenge* are adopted, namely GEO-BLEU [31]

and DTW [32]. GEO-BLEU measures the spatial coverage similarity between predicted and ground truth trajectories by computing n -gram overlap with equal weighting and a brevity penalty parameter, providing insights into the model’s ability to capture key spatial regions and movement hotspots. DTW evaluates the global spatiotemporal similarity between trajectory sequences by computing the minimum cumulative distance across all possible trajectory alignments using dynamic programming, effectively capturing both spatial accuracy and temporal sequence consistency while accommodating variable-length trajectories and temporal misalignments. Please refer to the cited papers for further details regarding these two metrics.

For experiments following the proposed curriculum learning pipeline, the model is pretrained for 60 epochs (20 for each stage) with an initial learning rate of 5×10^{-5} , followed by finetuning on real trajectories with a reduced initial learning rate of 2×10^{-5} . For experiments not following curriculum learning, the initial learning rate is also set to 5×10^{-5} . The Adam optimizer [41] with a batch size of 128 is employed universally for all experiments. MoBERT employs 8 BERT encoder layers and 8 self-attention heads (also for the feature interaction module) with 256-dimensional latent spaces.

B. Results

Beginning with a baseline MoBERT architecture using only essential input features including $l = (x, y)$ and $t = \{dId, tId\}$, we gradually add

- 1) Semantics s with $k = 3$ or 7 for *topKPOI*, obtaining MoBERT_{s3} and MoBERT_{s7};
- 2) Feature embedding interaction module (Section IV-B1), obtaining MoBERT_{s3/F} (taking s3 as an example);
- 3) Multi-task learning (Section III-D and Fig. 4), obtaining MoBERT_{s3/F/M};
- 4) Entropy-driven curriculum learning (Section III-B and IV-B2), obtaining MoBERT_{s3/F/M/E};

to observe the contribution of each components and naturally form ablation studies. We also compare the performance of the top three in the *HuMob Challenge 2023*, namely LP-BERT (1st) [42], GeoFormer (2nd) [43], MOBB (3rd) [44], where the leaderboard is online here² (last accessed: Aug. 28, 2025).

Table I presents the performance of MoBERT variants and the top three methods in the *HuMob Challenge 2023*, as evaluated by GEO-BLEU and DTW. It can be observed that MoBERT_{s3/F/M/E} with all four optimizations achieves state-of-the-art performance on both metrics, demonstrating the effectiveness of the proposed approaches. Particularly for DTW, which quantifies the temporal and spatial alignment of entire trajectories, MoBERT_{s3/F/M/E} achieves 12.7% lower (better) DTW values than LP-BERT, demonstrating that MoBERT_{s3/F/M/E} can capture both global trends and local variations more accurately than competing methods.

We highlight the performance gains of current MoBERT variants in parentheses compared to their previous variants

TABLE I
PERFORMANCE COMPARISON OF MOBERT VARIANTS AND OTHER STUDIES ON GEO-BLEU (LARGER IS BETTER) AND DTW (SMALLER IS BETTER). NUMBERS IN PARENTHESES SHOW DIFFERENCES BETWEEN THE CURRENT MOBERT VARIANT AND THE PREVIOUS MOBERT VARIANT.

Model	GEO-BLEU (\uparrow)	DTW (\downarrow)
MoBERT	0.264	30.06
MoBERT _{s3}	0.287 (+0.023)	29.11 (-0.95)
MoBERT _{s7}	0.284 (+0.020)	28.98 (-1.08)
MoBERT _{s3/F}	0.307 (+0.020)	28.93 (-0.18)
MoBERT _{s3/F/M}	0.335 (+0.028)	28.16 (-0.77)
MoBERT _{s3/F/M/E}	0.354 (+0.019)	26.15 (-2.01)
MOBB (3 rd) [44]	0.327	38.65
GeoFormer (2 nd) [43]	0.316	26.22
LP-BERT (1 st) [42]	0.344	29.96

TABLE II
EPOCHS REQUIRED TO REACH TARGET LOSSES FOR MoBERT_{s3/F/M/E} AND MoBERT_{s3/F/M} (W/ OR W/O CURRICULUM LEARNING).

Target losses	MoBERT _{s3/F/M/E}	MoBERT _{s3/F/M}
training loss ≤ 2.6	37 (1.95-fold speed)	72
validation loss ≤ 2.1	38 (2.92-fold speed)	111

to demonstrate the contribution of newly applied optimization strategies. Note that both MoBERT_{s3} and MoBERT_{s7} are compared with the vanilla MoBERT. These results reveal that MTL contributes most significantly to GEO-BLEU, while entropy-driven curriculum learning optimizes DTW most effectively. MoBERT_{s3} and MoBERT_{s7} achieve superior GEO-BLEU and DTW performance, respectively. We select MoBERT_{s3} for the following experiments as the *HuMob Challenge* prioritizes GEO-BLEU over DTW. Furthermore, all four optimization strategies improve performance on both metrics compared to cases when they are absent, demonstrating that all four strategies are effective for human mobility prediction.

Table II compares the convergence speed by counting epochs required to reach target training/validation losses between MoBERT_{s3/F/M/E} and MoBERT_{s3/F/M} (with and without curriculum learning). Although the contribution of entropy-driven curriculum learning to GEO-BLEU is not the largest, it dramatically accelerates the training process and achieves up to 2.92-fold convergence speed, which supports our analysis of curriculum learning and validates that the proposed $H_{norm-LZ}$ is an appropriate difficulty measurement for human mobility data. Our final model corresponds to a validation loss close to 2.1, so the achieved acceleration ratio has considerable reference value.

C. Hyperparameter selection

We now adopt grid search for hyperparameter selection. Table III lists the MTL results with varying λ_1 and λ_2 . Starting with MoBERT_{s3/F}, auxiliary tasks are introduced with different weights. Introducing only the direction prediction task and gradually increasing its weight from 0.2 to 0.8 results in a 5.6% improvement in GEO-BLEU and

²<https://connection.mit.edu/humob-challenge-2023/>

TABLE III
GRID SEARCH RESULTS OF MTL WITH VARYING λ_1 AND λ_2 .

auxiliary task	weight	GEO-BLEU (\uparrow)	DTW (\downarrow)
none (MoBERT _{s3/F})	-	0.307	28.93
distance estimation	$\lambda_1 = 0.2$	0.323	29.15
distance estimation	$\lambda_1 = 0.5$	0.324	28.87
distance estimation	$\lambda_1 = 0.8$	0.322	28.77
distance estimation	$\lambda_1 = 1.0$	0.321	29.02
direction estimation	$\lambda_2 = 0.2$	0.320	28.78
direction estimation	$\lambda_2 = 0.5$	0.322	28.47
direction estimation	$\lambda_2 = 0.8$	0.325	28.32
direction estimation	$\lambda_2 = 1.0$	0.321	28.52
both (MoBERT _{s3/F/M})	$\lambda_1 = 0.5$	0.335	28.16
	$\lambda_2 = 0.8$		

TABLE IV
RESULTS OF MoBERT_{s3/F/M/E} WITH DIFFERENT MODEL HYPERPARAMETERS.

# encoder layers	# heads	GEO-BLEU (\uparrow)	DTW (\downarrow)
8	8	0.354	26.15
8	16	0.309	35.16
4	8	0.343	26.83
4	16	0.326	31.19

a 2.1% decrease in DTW. However, further increasing the weight to 1.0 leads to a decline in model performance. This suggests that moderately increasing the direction task weight enables the model to more accurately capture trajectory spatial trends, while excessively high weights may undermine the learning of the primary task. A similar pattern is observed for the movement distance prediction task: when the distance estimation task weight is set to 0.5, the model achieves the highest GEO-BLEU value. The largest improvement arises when both sub-tasks are active, where λ_1 and λ_2 are set to 0.5 and 0.8, respectively, indicating that two sub-tasks provide complementary constraints and lead to more stable and accurate estimation. We select the best weights for single tasks and identify the optimal pair of weights, where λ_1 and λ_2 are set to 0.5 and 0.8, respectively. These settings are maintained for MoBERT_{s3/F/M} and MoBERT_{s3/F/M/E}, with their performances presented in Table I and II.

We also determine the optimal model hyperparameters, including the number of encoder layers and the number of self-attention heads. Table IV illustrates the results of MoBERT_{s3/F/M/E} with varying model hyperparameters. A larger number of encoder layers yields improved Geo-BLEU and DTW scores. We do not evaluate additional encoder layers due to resource limitations. By contrast, increasing the number of self-attention heads produces lower accuracy. More heads allocate each head a smaller sub-dimension, thereby reducing its representation capacity. Therefore, we select 8 encoder layers and 8 self-attention heads for model construction.

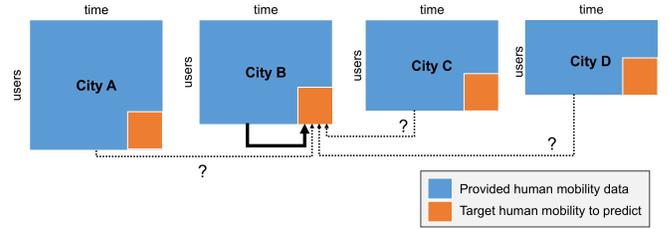


Fig. 8. Multi-City Human Mobility Prediction problem setting in the *HuMob Challenge 2024* (adapted from [30]).

TABLE V
CITY-WISE PERFORMANCE COMPARISON.

Model	Data	GEO-BLEU (\uparrow)	DTW (\downarrow)
Llama-3-8B-Mob [45]	City B	0.354	25.39
1st place in 2024	City C	0.296	20.57
(trained on City A+B)	City D	0.321	31.94
LP-BERT [42]	City B	0.309	23.30
1st place in 2023	City C	0.268	23.81
(trained on all cities)	City D	0.303	38.89
MoBERT _{s3/F/M/E}	City B	0.329	23.43
Ours	City C	0.314	18.33
(trained on City A only)	City D	0.328	38.03

D. Cross-city generalization

Human mobility models often face different geographic settings, transport infrastructure, and behavioral patterns across cities. A model trained in a single city may therefore transfer poorly. The *HuMob Challenge 2024* highlighted this question under the theme *Multi-City Human Mobility Prediction*, and asked whether data from additional cities can boost the target-city forecast, as depicted in Fig. 8. The dataset is enlarged to four cities, denoted as City A, B, C, and D, where City A is the dataset used for all previous evaluations (as shown in Fig. 7). For each city, 3,000 user trajectories formed the test set.

We directly use MoBERT_{s3/F/M/E} trained on City A to perform inference on the other three cities without any fine-tuning. To ensure fairness, the test sets are exactly the same as those used in the *HuMob Challenge 2024*. Table V presents a city-wise performance comparison with Llama-3-8B-Mob (1st in 2024) [45] and LP-BERT (1st in 2023) [42]. MoBERT_{s3/F/M/E} achieves an average Geo-BLEU of 0.324, matching the performance of Llama-3-8B-Mob while substantially outperforming LP-BERT. On City B, where Llama-3-8B-Mob benefits from training exposure, MoBERT_{s3/F/M/E} performs within 7% of the LLM-based model while significantly outperforming LP-BERT despite its access to City B. More notably, MoBERT_{s3/F/M/E} achieves the highest Geo-BLEU scores on both City C and D, surpassing both competitors. Similar patterns also emerge in DTW metrics, particularly for City C, indicating robust trajectory shape fitting capabilities for unseen urban environments.

These results reveal that architectural design and training strategies can be more impactful than simply expanding

training data coverage across multiple cities. The zero-shot generalization capability of MoBERT_{s3/F/M/E} with only 7.02 million parameters, approximately 1/6 the size of Llama-3-8B-Mob, demonstrates that compact models with well-designed entropy-driven curriculum learning and MTL can efficiently extract transferable mobility patterns. Conversely, the underwhelming performance of LP-BERT despite access to the most comprehensive multi-city training data suggests that naive data aggregation may introduce conflicting patterns that impede effective learning. The findings indicate that model capacity for extracting generalizable features, rather than geographic data diversity alone, determines cross-city transferability, with the proposed training approaches, potentially enabling adaptation to diverse urban contexts without finetuning.

VI. CONCLUSION

This paper presents a unified training framework that integrates entropy-driven curriculum learning with MTL for human mobility prediction. The entropy-driven curriculum learning strategy employs normalized LZ compression to quantify trajectory predictability, organizing training progression from simple to complex patterns based on information-theoretic principles grounded in Fano’s inequality. This approach achieves up to 2.92-fold convergence speed compared to conventional training by random sampling.

The MTL component simultaneously optimizes location prediction alongside auxiliary distance and direction estimation tasks, providing complementary supervision signals that are universally available in any mobility dataset. The proposed MoBERT architecture incorporates multi-feature embeddings and attention-based feature interaction mechanisms to capture complex spatiotemporal dependencies.

Experimental results on the YJMob100K dataset demonstrate state-of-the-art performance, achieving a GEO-BLEU score of 0.354 and a DTW distance of 26.15, outperforming existing methods including top performers from the *HuMob Challenge 2023*. Cross-city generalization experiments reveal robust transferability, with the method achieving competitive performance on unseen urban environments while using significantly fewer parameters than other approaches. The method trained exclusively on one city achieves comparable or superior results to models trained on multiple cities, indicating effective extraction of generalizable mobility patterns.

ACKNOWLEDGMENT

We would like to thank all members of the Professorship of Big Geospatial Data Management at Technical University of Munich for their insightful discussions and comments.

REFERENCES

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
 [2] K. Zhao, S. Tarkoma, S. Liu, and H. Vo, “Urban human mobility data mining: An overview,” in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1911–1920.

[3] M. Haraguchi, A. Nishino, A. Kodaka, M. Allaire, U. Lall, L. Kuei-Hsien, K. Onda, K. Tsubouchi, and N. Kohtake, “Human mobility data and analysis for urban resilience: A systematic review,” *Environment and Planning B: Urban Analytics and City Science*, vol. 49, no. 5, pp. 1507–1535, 2022. [Online]. Available: <https://doi.org/10.1177/23998083221075634>
 [4] X. Song, H. Kanasugi, and R. Shibusaki, “Deeptransport: prediction and simulation of human mobility and transportation mode at a citywide level,” in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI’16. AAAI Press, 2016, p. 2618–2624.
 [5] L. Pappalardo, E. Manley, V. Sekara, and L. Alessandretti, “Future directions in human mobility science,” *Nature Computational Science*, vol. 3, no. 7, pp. 588–600, Jul 2023. [Online]. Available: <https://doi.org/10.1038/s43588-023-00469-4>
 [6] D. Lian, Y. Ge, F. Zhang, N. J. Yuan, X. Xie, T. Zhou, and Y. Rui, “Content-aware collaborative filtering for location recommendation based on human mobility data,” in *2015 IEEE International Conference on Data Mining*, 2015, pp. 261–270.
 [7] G. Alix, N. Yanin, T. Pechlivanoglou, J. Li, F. Heidari, and M. Papagelis, “A mobility-based recommendation system for mitigating the risk of infection during epidemics,” in *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, 2022, pp. 292–295.
 [8] Q. Liu, S. Wu, L. Wang, and T. Tan, “Predicting the next location: A recurrent model with spatial and temporal contexts,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, Feb. 2016. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/9971>
 [9] R. Jiang, X. Song, Z. Fan, T. Xia, Q. Chen, S. Miyazawa, and R. Shibusaki, “DeepUrbanMomentum: An online deep-learning system for short-term urban mobility prediction,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/11338>
 [10] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, “DeepMove: Predicting human mobility with attentional recurrent networks,” in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, p. 1459–1468. [Online]. Available: <https://doi.org/10.1145/3178876.3186058>
 [11] D. Kong and F. Wu, “HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 2341–2347. [Online]. Available: <https://doi.org/10.24963/ijcai.2018/324>
 [12] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, “Semantic trajectories modeling and analysis,” *ACM Comput. Surv.*, vol. 45, no. 4, Aug. 2013. [Online]. Available: <https://doi.org/10.1145/2501654.2501656>
 [13] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, “Semantic trajectories: Mobility data computation and annotation,” *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 3, Jul. 2013. [Online]. Available: <https://doi.org/10.1145/2483669.2483682>
 [14] M. Suzuki, S. Furuta, and Y. Fukazawa, “Personalized human mobility prediction for humob challenge,” in *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, ser. HuMob-Challenge ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 22–25. [Online]. Available: <https://doi.org/10.1145/3615894.3628501>
 [15] Y. Qiao, Z. Si, Y. Zhang, F. B. Abdesslem, X. Zhang, and J. Yang, “A hybrid markov-based model for human mobility prediction,” *Neurocomputing*, vol. 278, pp. 99–109, 2018, recent Advances in Machine Learning for Non-Gaussian Data Processing. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231217314455>
 [16] H. Shi, Y. Li, H. Cao, X. Zhou, C. Zhang, and V. Kostakos, “Semantics-aware hidden markov model for human mobility,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 3, pp. 1183–1194, 2021.
 [17] H. Wang, Y. Li, D. Jin, and Z. Han, “Attentional markov model for human mobility prediction,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2213–2225, 2021.
 [18] T. Yabe, K. Tsubouchi, T. Shimizu, Y. Sekimoto, K. Sezaki, E. Moro, and A. Pentland, “Yjmob100k: City-scale and longitudinal

- dataset of anonymized human mobility trajectories,” *Scientific Data*, vol. 11, no. 1, p. 397, Apr 2024. [Online]. Available: <https://doi.org/10.1038/s41597-024-03237-9>
- [19] X. Wang, Y. Chen, and W. Zhu, “A survey on curriculum learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2022.
- [20] M. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010.
- [21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML ’09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48. [Online]. Available: <https://doi.org/10.1145/1553374.1553380>
- [22] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997. [Online]. Available: <https://doi.org/10.1023/A:1007379606734>
- [23] J. Feng, Y. Li, Z. Yang, Q. Qiu, and D. Jin, “Predicting human mobility with semantic motivation via multi-task attentional recurrent networks,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 5, pp. 2360–2374, 2022.
- [24] D. Lei, M. Xu, and S. Wang, “A deep multimodal network for multi-task trajectory prediction,” *Information Fusion*, vol. 113, p. 102597, 2025.
- [25] A. Lempel, M. Cohn, and W. Eastman, “A class of balanced binary sequences with optimal autocorrelation properties,” *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 38–42, 1977.
- [26] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/>
- [29] *HuMob-Challenge ’23: Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*. New York, NY, USA: Association for Computing Machinery, 2023.
- [30] *HuMob’24: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge*. New York, NY, USA: Association for Computing Machinery, 2024.
- [31] T. Shimizu, K. Tsubouchi, and T. Yabe, “GEO-BLEU: similarity measure for geospatial sequences,” in *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL ’22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: <https://doi.org/10.1145/3557915.3560951>
- [32] R. Bellman and R. Kalaba, “On adaptive control processes,” *IRE Transactions on Automatic Control*, vol. 4, no. 2, pp. 1–9, 1959.
- [33] S. Wang, J. Cao, and P. S. Yu, “Deep learning for spatio-temporal data mining: A survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3681–3700, 2022.
- [34] B. Shen, X. Liang, Y. Ouyang, M. Liu, W. Zheng, and K. M. Carley, “Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 724–733. [Online]. Available: <https://doi.org/10.1145/3219819.3219931>
- [35] S. Fiorini, G. Pilotti, M. Ciavotta, and A. Maurino, “3d-clost: A cnn-lstm approach for mobility dynamics prediction in smart cities,” in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3180–3189.
- [36] Y. Liu, Y. Rong, Z. Guo, N. Chen, T. Xu, F. Tsung, and J. Li, “Human mobility modeling during the COVID-19 pandemic via deep graph diffusion infomax,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14 347–14 355, Jun. 2023. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/26678>
- [37] K. Wang, Z. Peng, M. Cai, H. Wu, L. Liu, and Z. Sun, “Coupling graph neural networks and travel mode choice for human mobility prediction,” *Physica A: Statistical Mechanics and its Applications*, vol. 646, p. 129872, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378437124003819>
- [38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [39] H. Wang, J. Chen, T. Pan, Z. Fan, X. Song, R. Jiang, L. Zhang, Y. Xie, Z. Wang, and B. Zhang, “Easy begun is half done: Spatial-temporal graph modeling with st-curriculum dropout,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4668–4675, Jun. 2023. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/25590>
- [40] T. S. Han and S. Verdu, “Generalizing the fano inequality,” *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1247–1251, 1994.
- [41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [42] H. Terashima, N. Tamura, K. Shoji, S. Katayama, K. Urano, T. Yonezawa, and N. Kawaguchi, “Human mobility prediction challenge: Next location prediction using spatiotemporal bert,” in *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, ser. HuMob-Challenge ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1–6. [Online]. Available: <https://doi.org/10.1145/3615894.3628498>
- [43] A. V. Solatorio, “Geoformer: Predicting human mobility using generative pre-trained transformer (gpt),” in *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, ser. HuMob-Challenge ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 11–15. [Online]. Available: <https://doi.org/10.1145/3615894.3628499>
- [44] R. Koyama, M. Suzuki, Y. Nakamura, T. Mimura, and S. Ishiguro, “Estimating future human trajectories from sparse time series data,” in *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, ser. HuMob-Challenge ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 26–31. [Online]. Available: <https://doi.org/10.1145/3615894.3628500>
- [45] P. Tang, C. Yang, T. Xing, X. Xu, R. Jiang, and K. Sezaki, “Instruction-tuning llama-3-8b excels in city-scale mobility prediction,” in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge*, ser. HuMob’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1–4. [Online]. Available: <https://doi.org/10.1145/3681771.3699908>