Efficient Temporal Tokenization for Mobility Prediction with Large Language Models

Haoyu He [*1] **Haozheng Luo** [*2] **Yan Chen** [2] **Qi R. Wang** [1]

Abstract

We introduce **RHYTHM** (Reasoning with Hierarchical Temporal Tokenization for Human Mobility), a framework that leverages large language models (LLMs) as spatio-temporal predictors and trajectory reasoners. RHYTHM partitions trajectories into daily segments encoded as discrete tokens with hierarchical attention, capturing both daily and weekly dependencies while substantially reducing sequence length. Token representations are enriched with pre-computed prompt embeddings via a frozen LLM, enhancing the model's ability to capture interdependencies without extensive computational overhead. By freezing the LLM backbone, RHYTHM achieves significant computational efficiency. Evaluation on three real-world datasets demonstrates a **2.4%** improvement in accuracy, **5.0%** increase on weekends, and **24.6%** reduction in training time compared to state-of-the-art methods.

1. Introduction

We propose **RHYTHM** (Reasoning with Hierarchical Temporal Tokenization for Human Mobility), a novel foundation model for human mobility prediction. Our approach reconceptualizes mobility modeling through structured temporal abstraction, combining efficient multi-scale temporal tokenization with the complex reasoning capabilities of pre-trained LLMs. This integration yields a computationally efficient yet powerful framework for trajectory prediction, offering superior accuracy while maintaining efficiency and scalability across diverse mobility contexts.

Our framework is grounded in the inherent patterns of human mobility. Research shows that human movement follows predictable daily and weekly rhythms (Cho et al., 2011; Gonzalez et al., 2008; Eagle & Pentland, 2006), with Song et al. (2010) demonstrating that 93% of daily trajectories are predictable. Yet modeling the complex interdependence between locations and temporal cycles remains challenging. Existing approaches either neglect long-term periodicity (Yang et al., 2020; Feng et al., 2018; Gambs et al., 2012) or fail to capture multi-scale temporal patterns (Hong et al., 2023; Wu et al., 2019). We address these limitations by decomposing trajectories into meaningful segments and tokenizing them into discrete representations. Our approach employs intra-segment attention for local patterns and inter-segment attention for long-range dependencies (see Figure 1), significantly reducing computational complexity. Each token is further enriched with pre-computed prompt embeddings derived from a frozen LLM, integrating trajectory context and task descriptions to enhance semantic understanding.

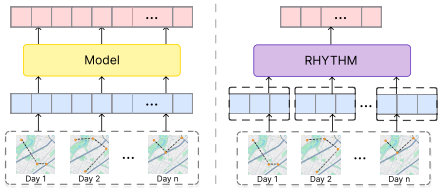

Figure 1. **Motivation for RHYTHM.** By partitioning trajectories into discrete tokens instead of a continuous stream, RHYTHM more effectively captures recurring mobility patterns.

Recent research highlights LLMs' dual capabilities as both representation extractors for spatio-temporal patterns and sophisticated reasoning engines (Chowdhery et al., 2023; Brown et al., 2020). These models excel through few-shot prompting (Brown et al., 2020), chain-of-thought reasoning (Pan et al., 2025; 2024; Wei et al., 2022), and in-context learning (Dong et al., 2024). While mobility-specific models like PMT (Wu et al., 2024b) and ST-MoE-BERT (He et al., 2024) lack LLM integration for modeling complex human movement correlations, RHYTHM addresses this limitation through an integrated LLM-based reasoning module. Our parameter-efficient strategy freezes the pre-trained LLM backbone, avoiding extensive fine-tuning while preserving reasoning capabilities. This approach effectively balances fine-grained spatio-temporal modeling with minimal computational overhead, making RHYTHM particularly suitable for resource-constrained settings.

[*]Equal contribution [1]Northeastern University, Boston, MA [2]Northwestern University, Evanston, IL. Correspondence to: Qi R. Wang <q.wang@northeastern.edu>.

Proceedings of the 3^{rd} Efficient Systems for Foundation Models Workshop at the International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

In summary, our key contributions are:

- We introduce a novel temporal tokenization scheme that represents daily mobility routines as discrete tokens, capturing multi-scale cyclical dependencies while markedly reducing sequence length.
- We propose an efficient prompt-guided integration of semantic trajectory cues and task descriptions via segment embeddings, enhancing interpretability of complex mobility behaviors.
- We design parameter-efficient LLM adaptation strategy leveraging frozen backbones, requiring only **12.37%** of model parameters and cutting computational cost by **24.6%**.
- Evaluation on three real-world datasets, yielding a **2.4%** overall accuracy gain and a **5.0%** improvement on weekend predictions over state-of-the-art baselines.

2. Method

In this section, we present **RHYTHM**, an LLM-driven architecture for prompt-guided representation learning of periodic spatio-temporal patterns (see Figure 2). paragraph

2.1. Problem definition

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ be a user's observed trajectory, where each observation $x_i = (t_i, l_i)$ pairs a timestamp t_i with a location $l_i \in \mathcal{L}$ drawn from a finite set $\mathcal{L}$. For a prediction horizon H, define the future time points $\mathcal{T} = \{t_{T+1}, \ldots, t_{T+H}\}$. Our goal is to predict the corresponding locations $\mathcal{Y} = \{l_{T+1}, \ldots, l_{T+H}\}$. Formally, we learn a mapping $f : (\mathcal{X}, \mathcal{T}) \mapsto \mathcal{Y}$ that takes the historical trajectory and future timestamps to the future location sequence.

2.2. Model structure

Spatio-Temporal Feature Encoding. For each observation x_i, we design temporal representations to encode the periodic nature of human mobility:

$$\mathbf{E}_i^{\text{temporal}} = \mathbf{E}^{\text{ToD}}(t_i) \| \mathbf{E}^{\text{DoW}}(t_i),$$

where $\cdot\|\cdot$ denotes the concatenation operation, $\mathbf{E}^{\text{ToD}}$ encodes time-of-day information, and $\mathbf{E}^{\text{DoW}}$ captures day-of-week patterns. These trainable embeddings transform discrete time indices into dense vector representations, yielding $\mathbf{E}_i^{\text{temporal}} \in \mathbb{R}^D$ where D corresponds to the input dimensionality of the underlying LLM.

The spatial representation $\mathbf{E}_i^{\text{spatial}} \in \mathbb{R}^D$ for location l_i is formulated as:

$$\mathbf{E}_i^{\text{spatial}} = \mathbf{E}^{\text{Loc}}(l_i) \| (W_{\text{coord}}[\text{lat}_i, \text{lon}_i]^T + b_{\text{coord}}),$$

where $\mathbf{E}^{\text{Loc}}$ represents the location-specific categorical embedding, while the latter term transforms geographic coordinates $(\text{lat}_i, \text{lon}_i)$ to the embedding domain through the transformation matrix $W_{\text{coord}} \in \mathbb{R}^{d_{\text{coord}} \times 2}$ and d_{coord} indicates the output dimension of the projection.

The unified spatio-temporal representation $\mathbf{E}_i \in \mathbb{R}^D$ is computed through element-wise summation:

$$\mathbf{E}_i = \mathbf{E}_i^{\text{temporal}} + \mathbf{E}_i^{\text{spatial}}.$$

When handling incomplete historical observations, we assign zero values to the spatial components, enabling the model to function solely with temporal features while maintaining consistent dimensionality.

Temporal Tokenization. Human movement behaviors demonstrate intrinsic hierarchical temporal patterns encompassing both local behavioral routines and cyclical regularities (Song et al., 2010; Gonzalez et al., 2008). To address these multi-scale dynamics, RHYTHM introduces a temporal segmentation strategy that separates fine-grained local behaviors from overarching temporal relationships, drawing inspiration from Liu et al. (2024c). Specifically, we divide the embedded representation $\mathcal{X}$ into N distinct, consecutive segments $\{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$, where each segment encodes semantically coherent time periods (such as daily sequences):

$$\mathbf{s}_i = \{E_{(i-1)L+1}, \ldots, E_{iL}\} \quad \text{for } i = 1, \ldots, N,$$

with L denoting the temporal span within each segment $\mathbf{s_i}$.

For capturing local sequential patterns within segments, we utilize intra-segment attention:

$$\tilde{\mathbf{E}}^{(i)} = \text{Attention}(\mathbf{s}_i),$$

We implement a pre-normalized transformer design incorporating gated feed-forward modules following Dubey et al. (2024), facilitating stable gradient propagation and enhancing representational capacity. Detailed architectural specifications can be found in Appendix C.

To efficiently encode relationships across segments, we introduce a trainable aggregation mechanism that compresses each segment into a compact token:

$$\mathbf{SE}_i = \text{Pool}\big(\tilde{\mathbf{E}}^{(i)}\big).$$

These compressed segment representations $\{\tilde{\mathbf{SE}}_1, \ldots, \tilde{\mathbf{SE}}_N\}$ are then processed through inter-segment attention to encode extended temporal relationships and distant dependencies:

$$\tilde{\mathbf{SE}}_{1:N} = \text{Attention}(\mathbf{SE}_{1:N}),$$

producing enhanced segment embeddings $\tilde{\mathbf{SE}}_\mathbf{i} \in \mathbb{R}^D$ that incorporate contextual signals from various temporal granularities. This design compresses the operational sequence length from T to N while maintaining both detailed temporal features and long-range correlations, thereby mitigating the computational burden associated with processing lengthy mobility sequences.

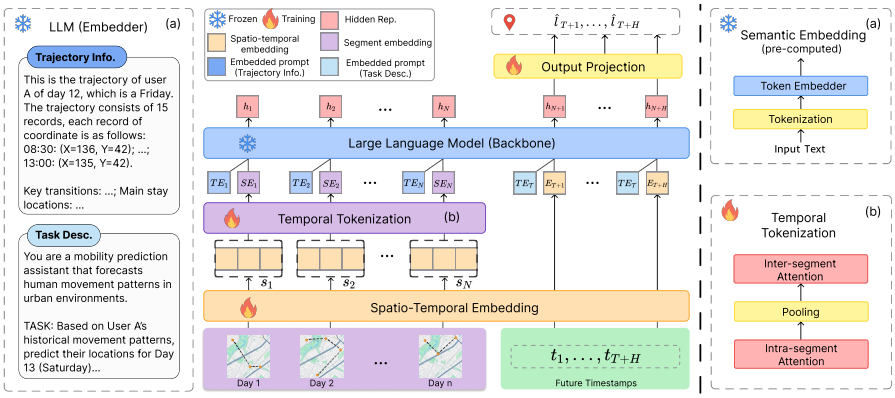

Figure 2. **The proposed architecture of RHYTHM.** Historical trajectories are first converted into spatio–temporal embeddings and discretized via temporal tokenization (**b**), enabling hierarchical attention to capture both local and global dynamics. Each segment token is enriched with semantic trajectory embeddings, while future time-step tokens integrate task-context descriptors (**a**). The resulting token sequence is fed into a frozen LLM backbone, and an output projection layer produces the final location predictions.

Semantic Context Integration. The integration of LLMs into mobility prediction presents a fundamental tension: maximizing semantic expressiveness while maintaining computational tractability. Existing approaches like LLM-Mob (Wang et al., 2023) adopt monolithic prompting strategies that encode entire trajectories as single, lengthy prompts—an approach that incurs prohibitive computational costs and dilutes fine-grained temporal signals. We propose a hierarchical prompting mechanism that decomposes trajectories into semantically coherent segments, each associated with a focused prompt that captures local movement patterns, transition behaviors, and activity semantics. This decomposition achieves an optimal balance: preserving the semantic richness necessary for accurate prediction while dramatically reducing the computational footprint.

Our semantic encoding pipeline operates at two granularities. For historical segments, we generate structured descriptions that capture movement patterns, stay durations, and transition dynamics. For future timestamps, we construct task-specific prompts that encode prediction objectives and temporal context, as detailed in Appendix E. These prompts are processed through frozen pre-trained LLMs to extract high-dimensional semantic representations.

To efficiently integrate these representations, we extract the final hidden state corresponding to the end-of-sequence (<EOS>) token, which naturally aggregates contextual information across the entire prompt. Formally, the semantic embedding for segment i is computed as:

$$\mathbf{TE}_i = \text{SelectLast}(\text{LLM}(\text{Prompt}(x_{(i-1)L+1:iL}))).$$

Task-oriented embeddings follow a similar formulation: $\mathbf{TE}_{\mathcal{T}} = \text{SelectLast}(\text{LLM}(\text{Prompt}(\mathcal{T})))$. Notably, all semantic embeddings are pre-computed offline, transforming what would be computationally intensive online inference into a one-time preprocessing step. This design enables RHYTHM to leverage the semantic understanding of LLMs without incurring runtime computational penalties.

Cross-representational Mobility Prediction. Given that the LLM's representation space naturally accommodates both temporal and semantic modalities, we integrate semantic information directly into the temporal representations without increasing sequence length. The unified representation $\mathbf{CE}_i$ for segment i is constructed through element-wise summation of segment representation $\tilde{\mathbf{SE}}_i$ and its corresponding semantic encoding $\mathbf{TE}_i$:

$$\mathbf{CE}_i = \tilde{\mathbf{SE}}_i + \mathbf{TE}_i.$$

For future time points, we combine temporal and task-specific representations: $\mathbf{CE}_{N+j} = \tilde{\mathbf{E}}_{T+j} + \mathbf{TE}_{\mathcal{T}}$.

These combined representations $\mathbf{CE}_i$ are subsequently processed by a frozen pre-trained LLM. The LLM transforms these inputs through its multi-layer architecture, leveraging its pre-trained knowledge to perform contextual reasoning over the fused spatio-temporal and semantic signals, producing final-layer hidden states h_i:

$$h_i = \text{LLM}(\mathbf{CE}_i).$$

A learnable projection head then transforms these representations into location-specific scores:

$$P(l_{T+j}|\mathcal{X}, \mathcal{T}) = \text{softmax}(W_o \tilde{\mathbf{h}}_{N+j} + \mathbf{b}_o),$$

where $W_o \in \mathbb{R}^{|\mathcal{L}| \times D}$ maps from the hidden dimension to the location vocabulary. This probability distribution over candidate locations enables the model to generate mobility predictions.

2.3. Computational Efficiency

RHYTHM's architectural design incorporates multiple strategies to optimize both computational efficiency and parameter utilization. Pre-computation of semantic representations using the frozen LLM occurs as a one-time preprocessing step, completely removing language model inference from the training and inference pipeline. Concurrently, the temporal segmentation strategy compresses the

Table 1. **Performance evaluation on the Kumamoto, Sapporo, and Hiroshima datasets.** We report Accuracy@k across multiple k thresholds (variance $\leq$ 2%). The highest values are **bolded**, and the second-best values are underlined. RHYTHM consistently outperforms all baselines in the majority of configurations.

	Kumamoto			Sapporo			Hiroshima		
Model	Acc@1	Acc@3	Acc@5	Acc@1	Acc@3	Acc@5	Acc@1	Acc@3	Acc@5
LSTM	0.2652	0.4799	0.5472	0.2310	0.3940	0.4526	0.2129	0.3775	0.4415
DeepMove	0.2779	0.4986	0.5683	0.2825	0.4672	0.5264	0.2804	0.4810	0.5477
PatchTST	0.2751	0.5018	0.5716	0.2703	0.4582	0.5168	0.2752	0.4839	0.5522
iTransformer	0.2609	0.4724	0.5412	0.2696	0.4500	0.5070	0.2804	0.4857	0.5523
TimeLLM	0.2712	0.4848	0.5535	0.2792	0.4746	0.5352	0.2698	0.4753	0.5426
CMHSA	0.2862	0.5182	0.5887	0.2890	**0.4901**	0.5525	0.2874	0.5001	0.5684
PMT	0.2697	0.4475	0.5187	0.2878	0.4896	0.5522	0.2850	0.4982	0.5668
COLA	0.2864	0.5186	0.5896	0.2847	0.4865	0.5497	0.2874	0.5013	0.5708
ST-MoE-BERT	0.2862	0.5155	0.5871	0.2869	0.4856	0.5480	0.2839	0.4925	0.5601
Mobility-LLM	0.2666	0.4793	0.5448	0.2838	0.4703	0.5288	0.2826	0.4856	0.5525
RHYTHM-LLaMA-1B	0.2929	0.5200	0.5835	0.2931	0.4876	0.5502	0.2913	0.5027	0.5753
RHYTHM-Gemma-2B	0.2923	0.5191	0.5932	**0.2943**	0.4896	**0.5545**	**0.2953**	**0.5074**	**0.5798**
RHYTHM-LLaMA-3B	**0.2941**	**0.5205**	**0.5947**	0.2938	0.4875	0.5523	0.2929	0.5032	0.5756

operational sequence length from $T+H$ to $N+H$, yielding a reduction in attention complexity from $\mathcal{O}((T+H)^2)$ to $\mathcal{O}((N+H)^2)$—a critical optimization for handling lengthy mobility traces. Additionally, by maintaining frozen LLM parameters throughout training, we achieve accelerated convergence while minimizing memory footprint. This confluence of design decisions empowers RHYTHM to handle extensive trajectory sequences without compromising prediction accuracy (demonstrated in Figure 4), rendering it particularly well-suited for deployment in resource-constrained environments and large-scale mobility applications.

3. Experiment

Models. We assess RHYTHM's mobility prediction capabilities using various pre-trained LLMs as backbones, sourced from Hugging Face with their original weights. The specific LLM configurations are listed in Appendix G.6.

Evaluation Metrics. We measure ranking quality using Accuracy@k and Mean Reciprocal Rank (MRR), complemented by Dynamic Time Warping (DTW) (Müller, 2007) and BLEU (Papineni et al., 2002) for trajectory-level evaluation. Metric definitions are provided in Appendix G.1.

Datasets. Our experiments utilize three urban mobility datasets from Kumamoto, Sapporo, and Hiroshima, obtained from YJMob100K (Yabe et al., 2024). Days are discretized into 48 half-hour intervals, with sparse observations across time slots. We partition data chronologically into training (70%), validation (20%), and test (10%) splits. Dataset specifications are detailed in Appendix F.

Baselines. We benchmark RHYTHM against established baselines including LSTM (Kong & Wu, 2018), DeepMove (Feng et al., 2018), PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024a), TimeLLM (Jin et al., 2024), PMT (Wu et al., 2024b), ST-MoE-BERT (He et al., 2024), CMHSA (Hong et al., 2023), COLA (Wang et al., 2024), and Mobility-LLM (Gong et al., 2024). See Appendix G.3 for baseline details.

Results. Table 1 demonstrates that RHYTHM consistently surpasses baseline methods on most criteria across all datasets. For Sapporo and Hiroshima, RHYTHM attains superior performance across all metrics. These results highlight RHYTHM's robust capability for mobility forecasting. The competitive performance of CMHSA and PMT in Accuracy@3 for Kumamoto can be attributed to their tailored attention architectures, which are particularly adept at identifying intermediate-ranked location candidates within this specific geographical area. Although Mobility-LLM employs a LLM-based framework, its performance falls short of RHYTHM, primarily due to its original optimization for visit intention prediction that emphasize semantic understanding. By comparison, RHYTHM combines temporal segmentation with LLM capabilities to capture hierarchical spatio-temporal patterns, emphasizing accurate location probability estimation. This architectural emphasis positions RHYTHM to achieve superior performance in ranking metrics. In summary, RHYTHM delivers a 2.4% gain in Accuracy@1 and 1.0% improvement in Accuracy@5 relative to the strongest competing approach.

Further empirical analyses encompassing spatial accuracy assessments, temporal pattern evaluation across daily and weekly cycles, computational efficiency benchmarks, model scaling experiments, and component-wise ablation experiments are presented in Appendix H given space limitations.

4. Conclusion

We present RHYTHM, a computationally efficient architecture for mobility prediction that employs temporal segmentation to encode spatio-temporal relationships and utilizes semantic representations to model periodic behaviors. Through the incorporation of frozen pre-trained LLMs as contextual reasoning modules, RHYTHM captures the underlying decision dynamics—especially for non-routine trajectories—while maintaining computational tractability. Furthermore, the framework's modular design facilitates straightforward adaptation across different pre-trained language models without architectural modifications.

Broader Impact

This paper presents a novel foundation model architecture for human mobility analysis, seeking to enhance the robustness and transferability of foundation models within spatio-temporal contexts. Although direct societal impacts are not immediately apparent, this work establishes fundamental capabilities for subsequent developments in urban infrastructure design, epidemiological monitoring, and transportation systems. Nevertheless, the model carries the risk of perpetuating or intensifying inherent biases within the training datasets, which could result in disparate prediction quality across different demographic groups or geographic regions.

Acknowledgments

The authors would like to thank the anonymous reviewers and program chairs for constructive comments.

H.H. and Q.R.W.'s work is supported by the National Science Foundation (NSF) under Grant Nos. 2125326, 2114197, 2228533, and 2402438, as well as by the Northeastern University iSUPER Impact Engine. H.L. is partially supported by the OpenAI Researcher Access Program. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. Any opinions, findings, conclusions, or recommendations expressed in the paper are those of the author and do not necessarily reflect the views of the funding agencies.

References

Alabdulmohsin, I., Neyshabur, B., and Zhai, X. Revisiting neural scaling laws in language and vision. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cabanas-Tirapu, O., Danús, L., Moro, E., Sales-Pardo, M., and Guimerà, R. Human mobility is well described by closed-form gravity-like models learned automatically from data. *Nature Communications*, 16(1):1336, 2025.

Chang, C., Wang, W.-Y., Peng, W.-C., and Chen, T.-F. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3):1–20, 2025.

Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, 2011.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Dang, W., Wang, H., Pan, S., Zhang, P., Zhou, C., Chen, X., and Wang, J. Predicting human mobility via graph convolutional dual-attentive networks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 192–200, 2022.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. A survey on in-context learning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Eagle, N. and Pentland, A. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10: 255–268, 2006.

Feng, J., Li, Y., Zhang, C., Sun, F., Meng, F., Guo, A., and Jin, D. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*, pp. 1459–1468, 2018.

Feng, J., Du, Y., Zhao, J., and Li, Y. AgentMove: A large language model based agentic framework for zero-shot next location prediction. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1322–1338, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Gambs, S., Killijian, M.-O., and del Prado Cortez, M. N. Next place prediction using mobility markov chains. In *Proceedings of the first workshop on measurement, privacy, and mobility*, pp. 1–6, 2012.

Gong, L., Lin, Y., Zhang, X., Lu, Y., Han, X., Liu, Y., Guo, S., Lin, Y., and Wan, H. Mobility-LLM: Learning visiting intentions and travel preference from human mobility data with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

He, H., Wu, X., and Wang, Q. Forecasting urban mobility using sparse data: A gradient boosted fusion tree approach. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*, pp. 41–46, 2023.

He, H., Luo, H., and Wang, Q. R. St-moe-bert: A spatial-temporal mixture-of-experts framework for long-term cross-city mobility prediction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge*, pp. 10–15, 2024.

Hong, Y., Zhang, Y., Schindler, K., and Raubal, M. Context-aware multi-head self-attentional neural network model for next location prediction. *Transportation Research Part C: Emerging Technologies*, 156:104315, 2023.

Hsu, S.-L., Tung, E., Krumm, J., Shahabi, C., and Shafique, K. Trajgpt: Controlled synthetic trajectory generation using a multitask transformer-based spatiotemporal model. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pp. 362–371, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Hu, J. Y.-C., Chang, P.-H., Luo, H., Chen, H.-Y., Li, W., Wang, W.-P., and Liu, H. Outlier-efficient hopfield layers for large transformer-based models. In *Forty-first International Conference on Machine Learning*, 2024.

Hu, J. Y.-C., Su, M., jui kuo, E., Song, Z., and Liu, H. Computational limits of low-rank adaptation (loRA) fine-tuning for transformer models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Huang, A. H., Wang, H., and Yang, Y. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kong, D. and Wu, F. Hst-lstm: A hierarchical spatial-temporal long-short term memory network for location prediction. In *Ijcai*, volume 18, pp. 2341–2347, 2018.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Liu, Q., Wu, S., Wang, L., and Tan, T. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, Y., Liao, X., Ma, H., He, B. Y., Stanford, C., and Ma, J. Human mobility modeling with limited information via large language models. *arXiv preprint arXiv:2409.17495*, 2024b.

Liu, Y., Qin, G., Huang, X., Wang, J., and Long, M. Autotimes: Autoregressive time series forecasters via large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024c.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Luo, H., Qiu, C., Su, M., Zhou, Z., Mehta, Z., Ye, G., Hu, J. Y.-C., and Liu, H. Fast and low-cost genomic foundation models via outlier removal. *arXiv preprint arXiv:2505.00598*, 2025.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

Luo, Y., Liu, Q., and Liu, Z. Stan: Spatio-temporal attention network for next location recommendation. In *Proceedings of the web conference 2021*, pp. 2177–2185, 2021.

Müller, M. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84, 2007.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Pan, Z., Luo, H., Li, M., and Liu, H. Conv-coa: Improving open-domain question answering in large language models via conversational chain-of-action. *arXiv preprint arXiv:2405.17822*, 2024.

Pan, Z., Luo, H., Li, M., and Liu, H. Chain-of-action: Faithful and multimodal question answering through large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Rao, X., Chen, L., Liu, Y., Shang, S., Yao, B., and 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.

Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science*, 327(5968): 1018–1021, 2010.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, X., Fang, M., Zeng, Z., and Cheng, T. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*, 2023.

Wang, Y., Zheng, T., Liang, Y., Liu, S., and Song, M. Cola: Cross-city mobility transformer for human trajectory simulation. In *Proceedings of the ACM on Web Conference 2024*, pp. 3509–3520, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, D., Hu, J. Y.-C., Li, W., Chen, B.-Y., and Liu, H. STanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. In *The Twelfth International Conference on Learning Representations*, 2024a.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Wu, X., He, H., Wang, Y., and Wang, Q. Pretrained mobility transformer: A foundation model for human mobility. *arXiv preprint arXiv:2406.02578*, 2024b.

Wu, Y., Lian, D., Jin, S., and Chen, E. Graph convolutional networks on user mobility heterogeneous graphs for social relationship inference. In *IJCAI*, pp. 3898–3904, 2019.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pp. 38087–38099. PMLR, 2023.

Yabe, T., Tsubouchi, K., Shimizu, T., Sekimoto, Y., Sezaki, K., Moro, E., and Pentland, A. Yjmob100k: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data*, 11(1):397, 2024.

Yang, D., Zhang, D., Zheng, V. W., and Yu, Z. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2014.

Yang, D., Fankhauser, B., Rosso, P., and 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.

Yang, S., Liu, J., and 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.

Zhang, Y. and Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2023.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. One fits all: Power general time series analysis by pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Supplementary Material

A. Limitations

Several important limitations warrant consideration regarding RHYTHM. First, the model's effectiveness depends significantly on the underlying pretrained LLMs, which were initially developed for natural language processing rather than spatial-temporal modeling. When pretrained models suffer from resource constraints, their ability to effectively encode human movement patterns may be compromised. Moreover, RHYTHM currently employs a non-autoregressive prediction paradigm for mobility forecasting. Although recent research has demonstrated promising results with autoregressive formulations for temporal sequences (Liu et al., 2024c), incorporating these techniques into RHYTHM represents a valuable direction for future development. Additionally, despite utilizing frozen LLMs to enhance computational efficiency, RHYTHM's training duration remains substantial, potentially limiting its deployment in time-sensitive scenarios. Nevertheless, RHYTHM establishes an innovative foundation model architecture specifically designed for mobility prediction, demonstrating improvements in both computational efficiency and forecast accuracy. This work provides a foundation for subsequent advances in model scalability and resource optimization. Future efforts will focus on developing sophisticated fine-tuning strategies to enhance performance while minimizing computational requirements, leveraging recent advances in parameter-efficient adaptation and model compression techniques (Luo et al., 2025; Hu et al., 2024; Dettmers et al., 2024; Xiao et al., 2023).

B. Related Work

Mobility Prediction. The field of human mobility prediction has progressed from classical statistical approaches to sophisticated deep learning architectures. Mechanistic models including the gravity framework (Cabanas-Tirapu et al., 2025) and radiation theory (Simini et al., 2012) forecast collective movement patterns through distance-decay and intervening opportunities, yet cannot capture individual-specific behaviors. Addressing this limitation, stochastic methods such as Markovian models (Gambs et al., 2012), decision tree techniques (He et al., 2023), and matrix decomposition (Yang

et al., 2014) have been developed to predict user-specific location sequences. Despite enabling personalized predictions, these approaches face challenges with data sparsity and complex temporal interdependencies characteristic of mobility data. Neural architectures introduced sequential modeling through LSTMs (Liu et al., 2016), enabling temporal context understanding, with attention-augmented extensions (Feng et al., 2018) mitigating gradient degradation. Nonetheless, such models frequently miss periodic behavioral patterns. Integrated frameworks including Graph-Flashback (Rao et al., 2022) and GCDAN (Dang et al., 2022) incorporate network topology for spatial modeling, though fixed-window constraints hinder their long-horizon prediction capabilities. The Transformer paradigm (Vaswani et al., 2017) transformed sequence modeling through self-attention for capturing distant dependencies. Notable extensions include STAN (Luo et al., 2021), which integrates spatial-temporal attention for location recommendation, COLA (Wang et al., 2024), which generalizes across urban environments, and GETNext (Yang et al., 2022), which separates personal patterns from collective trends. However, these Transformer variants maintain timestamp-based representations, resulting in quadratic computational growth for extended trajectories and lacking explicit modeling of nested temporal cycles (daily within weekly patterns). Contemporary research investigates large language models (LLMs) for mobility applications, exploiting their robust transfer learning properties (Gong et al., 2024; Liu et al., 2024b). Implementations such as LLM-Mob (Wang et al., 2023) and AgentMove (Feng et al., 2025) utilize strategic prompting for location forecasting and user identification, while TrajGPT (Hsu et al., 2024) employs generative modeling for trajectory synthesis. Yet these methods process mobility data as undifferentiated token sequences, disregarding inherent temporal structures and the representational disconnect between linguistic and spatio-temporal modalities. In contrast to current LLM approaches that handle trajectories as generic sequences, our method implements temporal segmentation to directly encode periodic structures (daily/weekly patterns), addressing representational gaps and modeling hierarchical temporal relationships for enhanced long-range mobility forecasting.

Time Series Foundation Models. Current time series foundation models fall into two primary paradigms: transformer-centric architectures and language model adaptations. Transformer-centric approaches (Wu et al., 2024a; Liu et al., 2024a; Nie et al., 2023) emphasize architectural innovations and attention mechanisms for temporal pattern extraction. Notably, PatchTST (Nie et al., 2023) employs segment-wise attention to model extended temporal relationships, while STanHop (Wu et al., 2024a) and Crossformer (Zhang & Yan, 2023) utilize multi-level attention architectures to encode both sequential patterns and hierarchical temporal organization. Language model adaptations (Liu et al., 2024c; Jin et al., 2024) repurpose pre-trained LLMs for time series analysis, demonstrating competitive performance in forecasting benchmarks. Specifically, AutoTime (Liu et al., 2024c) develops an autoregressive framework tailored for sequential dependency modeling, while TimeLLM (Jin et al., 2024) harnesses LLMs to learn complex temporal state transitions. Nevertheless, these approaches cannot adequately handle the unique characteristics of human mobility—particularly sudden spatial transitions and irregular temporal patterns. In contrast, RHYTHM combines specialized spatio-temporal representations with hierarchical modeling to capture these complex mobility dynamics effectively.

Cross-domain Adaptation of LLMs. LLMs transform from domain-specific language processors into general-purpose foundation models with advanced reasoning abilities spanning multiple disciplines (Alabdulmohsin et al., 2022; Brown et al., 2020). The combination of transformer architectures and large-scale pretraining enables exceptional knowledge transfer beyond textual domains. For visual understanding, frameworks like CLIP (Radford et al., 2021) create unified vision-language representations enabling zero-shot classification, while temporal modeling approaches including One-Fits-All (Zhou et al., 2023) and LLM4TS (Chang et al., 2025) achieve strong forecasting results by representing numerical data as discrete tokens. Within biomedical applications, specialized models such as BioBERT (Lee et al., 2020) and BioGPT (Luo et al., 2022) show substantial improvements on medical text processing, with instruction-aligned systems like Med-PaLM reaching near-expert performance on clinical questions (Singhal et al., 2023). Financial domain adaptations including FinBERT (Huang et al., 2023) and BloombergGPT (Wu et al., 2023) demonstrate marked advantages over generic models for market sentiment and entity extraction tasks.

To avoid resource-intensive complete retraining, efficient adaptation methods have emerged as preferred strategies. Low-Rank Adaptation (LoRA) (Hu et al., 2021) augments attention mechanisms with decomposed weight updates, while alternative strategies maintain frozen model parameters through modality-specific input transformations. Visual adaptation techniques (Alayrac et al., 2022; Tsimpoukelli et al., 2021) learn compact encoders that generate conditioning signals for static LLMs, whereas temporal sequence methods (Liu et al., 2024c; Jin et al., 2024) utilize learned projections for numerical-to-embedding conversion.

LLM applications in mobility analysis remain nascent, with current methods requiring substantial parameter updates. Mobility-LLM (Gong et al., 2024) updates model subsets during training, while LLM-Mob (Wang et al., 2023) employs

contextual prompting without explicit temporal structure. Conversely, RHYTHM preserves all LLM parameters unchanged, retaining pretrained capabilities while implementing a custom spatio-temporal encoding architecture optimized for mobility sequence processing.

C. Attention Implementation Details

We employ a pre-normalized transformer design that promotes training stability, augmented with gated feed-forward modules for enhanced representational capacity. The architectural formulation follows:

$$Z = \text{LayerNorm}(X) + \text{Multi-Head Attention}(\text{LayerNorm}(X)),$$
$$\tilde{Z} = Z + \text{GatedFFN}(\text{LayerNorm}(Z)),$$

with X denoting the input representations. The multi-head attention mechanism is defined as:

$$\text{Multi-Head}(X) = [\text{head}_1 \| \text{head}_2 \| \dots \| \text{head}_h] W_{\text{out}},$$
$$\text{head}_i = \text{Softmax}\left(\frac{XW_{q,i}(XW_{k,i})^\top}{\sqrt{d_k}}\right) XW_{v,i},$$

where we utilize h parallel attention heads, with $W_{q,i}, W_{k,i}, W_{v,i} \in \mathbb{R}^{d \times d_k}$ representing query, key, and value transformations for head i, and $W_{\text{out}} \in \mathbb{R}^{d \times d}$ serving as the final projection. The gated feed-forward component employs dynamic modulation:

$$\text{GatedFFN}(Z) = \text{FFN}(Z) \odot \sigma(W_{\text{gate}} Z),$$
$$\text{FFN}(Z) = W_2 \, \text{GELU}(W_1 Z),$$

where σ represents sigmoid activation, $\odot$ indicates Hadamard product, and $W_{\text{gate}} \in \mathbb{R}^{d \times d}$ controls the gating behavior. The feed-forward transformation implements a 4× dimension expansion through $W_1 \in \mathbb{R}^{4d \times d}$ and compression via $W_2 \in \mathbb{R}^{d \times 4d}$. Regularization through dropout follows both attention and feed-forward computations to mitigate overfitting.

D. Theoretical Guarantee

Our architectural decisions are grounded in robust theoretical foundations. Utilizing an LLM for universal sequence representation extraction offers two theoretical benefits: (1) convergence guarantees for model outputs, established by Zhou et al. (2023, Theorem E.2), and (2) uniform feature distribution properties within the LLM's final hidden representations, proven in Zhou et al. (2023, Theorem E.3). These characteristics collectively strengthen the downstream MLP classifier's learning capacity. Furthermore, given our transformer-based architecture, Ramsauer et al. (2021) establishes that transformers constitute a specific instantiation of contemporary Hopfield networks. This connection provides bounded memory retrieval guarantees for our LLM-based approach, as formalized in Hu et al. (2025, Lemma 3.2). Such theoretical underpinnings substantiate our design choices, with empirical results confirming these theoretical predictions.

E. Example Prompt

Trajectory Information

```
This is the trajectory of user <User_ID> of day <Day_ID> which is a <Day_of_Week>.
The trajectory consists of <N> records, each record of coordinate is as follows:
08:30:  (X=136, Y=42); 09:00:  (X=136, Y=42); 09:30:  (X=137, Y=41); 10:00:  (X=146,
Y=37); 10:30:  (X=145, Y=38); 11:00:  (X=144, Y=38); 11:30:  (X=135, Y=41); 12:00:
(X=135, Y=42); 12:30:  (X=135, Y=42); 13:00:  (X=135, Y=42).

Key transitions:  At 10:00:  (X=137, Y=41) → (X=146, Y=37); At 11:30:  (X=144, Y=38)
→ (X=135, Y=41).

Main stay locations:  (X=136, Y=42) from 08:30 to 09:30 (0.5 hours); (X=145, Y=38)
from 10:00 to 11:00 (0.5 hours); (X=135, Y=42) from 11:30 to 13:00 (1.5 hours).
```

Task Description

```
You are a mobility prediction assistant that forecasts human movement patterns in
urban environments.  The city is represented as a 200 x 200 grid of cells, where
each cell is identified by coordinates (X,Y). The X coordinate increases from left
(0) to right (199), and the Y coordinate increases from top (0) to bottom (199).

TASK: Based on User <User_ID>'s historical movement patterns, predict their
locations for Day <Day_ID> (<Day_of_Week>).  The predictions should capture expected
locations at 30-minute intervals throughout the day (48 time slots).  The model
should analyze patterns like frequent locations, typical daily routines, and
time-dependent behaviors to generate accurate predictions of where this user is
likely to be throughout the next day.

The previous days' trajectory data contains information about the user's typical
movement patterns, regular visited locations, transition times, and duration of
stays.  Key patterns to consider include:  home and work locations, morning and
evening routines, lunch-time behaviors, weekend vs.  weekday differences, and
recurring visit patterns.
```

F. Dataset

Table 2 presents comprehensive statistics for the three datasets utilized in our experiments.

Table 2. Dataset Statistics

City	Users	Duration	Spatial Resolution	Places
Kumamoto	3k	75 days	500m × 500m	40k
Sapporo	17k	75 days	500m × 500m	40k
Hiroshima	22k	75 days	500m × 500m	40k

G. Experiment Setting

G.1. Evaluation Metrics

Accuracy@k quantifies the fraction of instances where true locations appear among the top-k predictions:

$$\text{Accuracy@}k = \frac{1}{H}\sum_{i=1}^{H} \mathbb{1}(l_{T+i} \in \text{top-}k(\hat{p}_{T+i})),$$

with $\mathbb{1}(\cdot)$ denoting the indicator function and $\hat{p}_{T+i}$ representing the output probability vector.

Mean Reciprocal Rank (MRR) assesses ranking quality through reciprocal positions:

$$\text{MRR} = \frac{1}{H}\sum_{i=1}^{H} \frac{1}{\text{rank}(l_{T+i})},$$

where $\text{rank}(l_{T+i})$ indicates the ordinal position of the actual location.

Dynamic Time Warping (DTW) computes trajectory alignment cost:

$$\text{DTW}(\mathcal{Y}, \hat{\mathcal{Y}}) = \min_{\pi} \sum_{(i,j)\in\pi} d(l_{T+i}, \hat{l}_{T+j}),$$

with π defining an optimal alignment path and $d(\cdot,\cdot)$ measuring Euclidean separation.

BLEU evaluates sequence similarity via n-gram correspondences:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where p_n captures n-gram matching rates, w_n assigns importance weights, and BP adjusts for length discrepancies.

G.2. Settings

Our experiments employ 30-minute temporal granularity. We configure a historical context of 7 days comprising 336 time intervals, with predictions extending over 48 intervals (equivalent to 24 hours). The segment size is fixed at 48 time intervals throughout all experiments.

G.3. Baselines

We benchmark RHYTHM against three categories of baseline methods: LSTM-based architectures, transformer-based approaches, and LLM-based frameworks. Among transformer architectures, we evaluate PatchTST (Nie et al., 2023), PMT (Wu et al., 2024b), ST-MoE-BERT (He et al., 2024), CMHSA (Hong et al., 2023), iTransformer (Liu et al., 2024a), and COLA (Wang et al., 2024). Within this group, ST-MoE-BERT, PMT, and COLA represent current best-performing methods in mobility forecasting, while PatchTST and iTransformer are leading general-purpose temporal sequence models. To ensure equitable evaluation, we augment these general time series models with spatio-temporal encodings. For LLM-based comparisons, we include TimeLLM (Jin et al., 2024) and Mobility-LLM (Gong et al., 2024). TimeLLM achieves state-of-the-art results in LLM-driven time series prediction, which we enhance with spatio-temporal representations for consistent comparison. Mobility-LLM provides a comprehensive LLM framework supporting diverse mobility-related tasks. Our LSTM baselines comprise the classical LSTM (Kong & Wu, 2018) and DeepMove (Feng et al., 2018) architectures.

G.4. Computational Resource

All experiments are conducted on a single NVIDIA A100 GPU equipped with 40GB memory, paired with a 24-core Intel(R) Xeon(R) Gold 6338 processor running at 2.00GHz. Our implementation leverages PyTorch (Paszke et al., 2019) and integrates the Hugging Face Transformers library for model deployment.

G.5. Hyperparameters

We detail the training configurations employed across all models. The embedding dimensions are configured as follows: time-of-day and day-of-week embeddings utilize 128 dimensions each, categorical location representations use 256 dimensions, and coordinate projections employ 128 dimensions. Training optimization is performed using **AdamW** (Loshchilov & Hutter, 2019). We perform comprehensive hyperparameter optimization by evaluating learning rates within $\{1 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}\}$ and weight decay parameters from $\{0, 0.001, 0.01\}$. Optimal hyperparameters for each dataset are identified through rigorous validation experiments. To ensure fair evaluation, all models employ a uniform batch size of 64. Final parameter selections are determined by validation set performance.

G.6. LLM variants

We incorporated various pre-trained language models as text embedders and fixed backbone architectures in RHYTHM to assess performance across different model scales. Table 3 lists the foundation models utilized via the Hugging Face Transformers library, spanning parameter counts from 125M to 3B.

Table 3. List of LLMs used in RHYTHM.

Model	Parameters	HuggingFace Repository
OPT-125M	125M	facebook/opt-125m
OPT-350M	350M	facebook/opt-350m
LLaMA-3.2-1B	1.24B	meta-llama/Llama-3.2-1B
Qwen-2.5-1.5B	1.54B	Qwen/Qwen2.5-1.5B
DeepSeek-R1-1.5B	1.78B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
Gemma-2-2B	2.61B	google/gemma-2-2b-it
Phi-2	2.78B	microsoft/phi-2
LLaMA-3.2-3B	3.21B	meta-llama/Llama-3.2-3B

H. Extended Experiments

H.1. Geographical Evaluation

We assess RHYTHM with baselines using BLEU and DTW metrics, measuring sequence correspondence and trajectory alignment accuracy, respectively. Table 4 reveals that RHYTHM achieves optimal DTW scores on Sapporo, indicating superior spatial trajectory matching. Although COLA attains highest BLEU values across all datasets, RHYTHM secures second place for Kumamoto. These results reveal an inherent balance between precise sequence replication and spatial accuracy optimization. This performance difference may stem from COLA's posterior correction mechanism, which adjusts predictions toward observed location frequency distributions, potentially improving intermediate-rank predictions by reducing bias toward high-frequency locations. RHYTHM substantially surpasses LSTM architectures and transformer models through its temporal segmentation and semantic embedding strategies, achieving improved sequence consistency and location precision. This combination yields well-balanced performance for practical mobility applications. Regarding MRR evaluation, RHYTHM uniformly exceeds all competing methods with a 1.44% gain over the strongest alternative, confirming its robust ranking performance across varied movement patterns.

Table 4. **Geographical assessment of RHYTHM against baseline methods.** We report DTW (↓), BLEU (↑), and MRR (↑) metrics with variance below 2%. Best performance is shown in **bold**, with second-best results underlined.

	Kumamoto			Sapporo			Hiroshima		
Model	DTW	BLEU	MRR	DTW	BLEU	MRR	DTW	BLEU	MRR
LSTM	5014	0.1564	0.3860	4507	0.1716	0.3270	5908	0.1544	0.3113
DeepMove	4630	0.1746	0.4021	3818	0.1959	0.3887	4981	0.1933	0.3959
PatchTST	5251	0.1315	0.4021	4099	0.1784	0.3773	5021	0.1884	0.3945
iTransformer	6178	0.1275	0.3796	4074	0.1780	0.3730	5094	0.1789	0.3977
TimeLLM	5984	0.1285	0.3912	3915	0.2145	0.3902	5126	0.1988	0.3872
CMHSA	4490	0.1810	0.4158	3786	0.2299	0.4034	4841	0.2289	0.4086
PMT	4536	0.1524	0.3720	3799	0.2017	0.4026	4851	0.2009	0.4065
COLA	4446	**0.2064**	0.4164	3793	**0.2496**	0.3996	**4840**	**0.2445**	0.4095
ST-MoE-BERT	4691	0.1557	0.4151	3796	0.2102	0.4001	4889	0.2117	0.4031
Mobility-LLM	5603	0.1649	0.3858	3911	0.1917	0.3902	4985	0.2056	0.3990
RHYTHM-LLaMA-1B	4478	0.1793	0.4216	**3745**	**0.2496**	0.4045	5059	0.2083	0.4069
RHYTHM-Gemma-2B	**4416**	0.1928	0.4205	3995	0.2019	**0.4065**	4857	0.2109	**0.4173**
RHYTHM-LLaMA-3B	4470	0.1814	**0.4220**	4035	0.1917	0.4048	4935	0.2093	0.4140

H.2. Daily and Weekly Trend Analysis

We examine temporal performance variations of RHYTHM and competing methods on the Sapporo dataset, analyzing accuracy patterns across daily and weekly cycles (Figure 3). RHYTHM generally surpasses baseline approaches across temporal dimensions, particularly during evening commute periods and weekends, with performance gains of 3.4% and 5.0%, respectively. These observations align with Barbosa et al. (2018), who documented increased variability and irregular patterns in weekend mobility behaviors. Notably, RHYTHM's advantages diminish during highly predictable movement periods, including overnight hours and typical weekday routines, while demonstrating substantial improvements during less structured timeframes such as weekends and evening transitions when mobility patterns become more stochastic. This enhanced performance during irregular periods stems from RHYTHM's LLM-powered reasoning mechanisms, which effectively model complex behavioral factors underlying mobility decisions during non-routine scenarios. Conversely, baseline methods depend on rigid temporal heuristics, constraining their adaptability to fluctuating movement dynamics.

H.3. Transferability

To validate RHYTHM's generalizability across different pretrained architectures, we experiment with various backbone model scales and evaluate their mobility prediction performance (detailed in Table 1). We specifically investigate RHYTHM's behavior when equipped with LLaMA-3.2-1B, LLaMA-3.2-3B, and Gemma-2-2B as backbone models. Our findings reveal consistent performance improvements with increasing model capacity. Both LLaMA-3.2-3B and Gemma-2-2B configurations surpass LLaMA-3.2-1B across most evaluation metrics. These results confirm that RHYTHM's effectiveness scales proportionally with backbone model size, suggesting potential for further gains with larger architectures on expanded

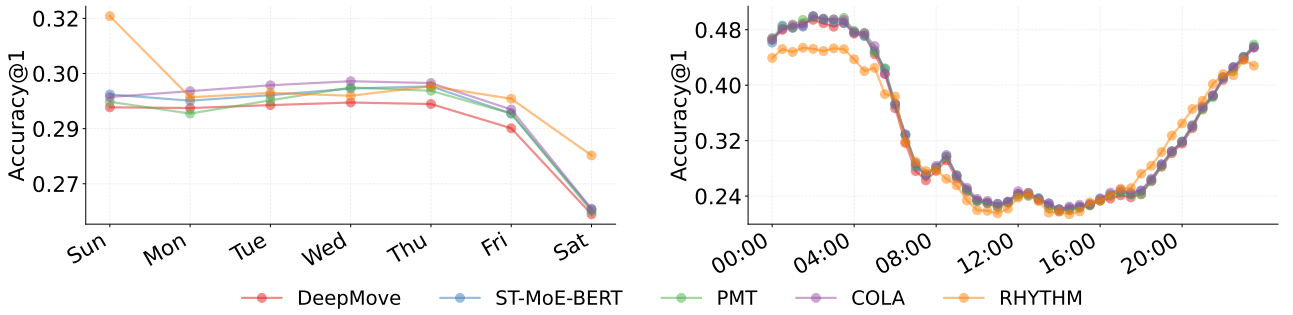

Figure 3. **Temporal performance patterns of RHYTHM and baselines on Sapporo data showing weekly (left) and daily (right) variations.** The results demonstrate systematic performance fluctuations across both diurnal and weekly cycles.

datasets. All models undergo 30 training epochs in our experiments. The larger LLaMA-3.2-3B variant likely benefits from extended training to reach optimal performance relative to the more efficient LLaMA-3.2-1B. Despite this constraint, LLaMA-3.2-3B maintains strong competitive performance against its smaller counterpart. Specifically, LLaMA-3.2-3B yields a 0.40% Acc@1 improvement over LLaMA-3.2-1B, demonstrating RHYTHM's effective scaling properties.

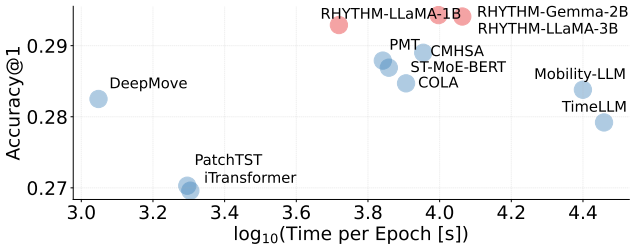

Figure 4. Computational efficiency versus predictive accuracy trade-offs for RHYTHM and baseline approaches on the Sapporo dataset.

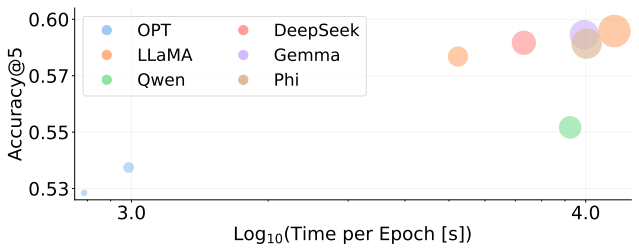

Figure 5. Computational performance across different LLM backbones using identical experimental settings from Table 5.

H.4. Training Speed

We benchmark RHYTHM's computational efficiency on the Sapporo dataset under consistent training settings. Experiments utilize a single NVIDIA A100 GPU with 40GB memory capacity. Figure 4 presents the comparative results. RHYTHM demonstrates superior training efficiency relative to most baseline approaches. While LSTM, DeepMove, PatchTST, and iTransformer exhibit faster per-epoch training, RHYTHM's accuracy gains justify the modest computational overhead. Remarkably, RHYTHM achieves training speeds on par with PMT, COLA, and ST-MoE-BERT while managing substantially larger parameter budgets, validating its efficient architectural design. Training time exhibits predictable scaling characteristics: the LLaMA-3B configuration requires 2.2× the computation of LLaMA-1B, while Gemma-2-2B increases training time by 1.9×.

H.5. Scaling Behavior

Model scalability represents a fundamental consideration for practical deployment. We investigate RHYTHM's scaling characteristics across diverse model capacities using pretrained LLMs spanning OPT, LLaMA-3.2, DeepSeek-R1, Gemma-2, Phi-2, and Qwen 2.5 architectures (see Table 3). Table 5 demonstrates consistent performance gains correlating with increased parameter counts. This scaling relationship necessitates balancing prediction quality against computational requirements. We examine three critical dimensions—predictive accuracy, model size, and per-epoch training duration—as in Figure 5. While LLaMA-3.2-3B delivers peak mobility prediction performance, LLaMA-3.2-1B emerges as the pragmatic choice for RHYTHM, optimally balancing accuracy improvements with resource efficiency.

Table 5. **Scaling analysis for RHYTHM.** We assess RHYTHM's scaling behavior across pre-trained models with different parameter counts. Evaluation encompasses Accuracy@k, MRR, and per-epoch training duration. Best performance is shown in **bold**, with second-best results underlined. Performance generally scales positively with increased model capacity across configurations.

Backbone	Time(s)	Acc@1	Acc@3	Acc@5	MRR
OPT-125M	787	0.2798	0.4726	0.5231	0.3819
OPT-350M	986	0.2837	0.4789	0.5343	0.3923
LLaMA-3.2-1B	5235	0.2929	0.5200	0.5835	0.4216
Qwen-2.5-1.5B	9241	0.2897	0.4873	0.5521	0.4049
DeepSeek-R1-1.5B	7308	0.2921	0.5164	0.5896	0.4188
Gemma-2-2B	9928	0.2923	0.5191	0.5932	0.4205
Phi-2	10047	0.2915	0.5166	0.5892	0.4183
LLaMA-3.2-3B	11566	**0.2941**	**0.5205**	**0.5948**	**0.4220**

H.6. Ablation study

We employ LLaMA-3.2-1B as the standard backbone across all ablation experiments. Component-wise analysis across three datasets (Table 6) reveals the relative importance of each architectural element. Excluding temporal tokenization causes the most severe degradation at 5.39%, while removing hierarchical attention (HA) reduces performance by 0.90%, establishing structured temporal encoding as RHYTHM's fundamental component. For semantic components, both trajectory embeddings and task prompts prove essential, with their joint removal decreasing performance by 1.82%. Task descriptions demonstrate slightly greater influence, contributing an extra 0.10% performance loss beyond trajectory information alone.

Table 6. **Ablation study of RHYTHM.** We examine the individual impact of each architectural component on model performance. Results report Accuracy@k metrics with variance below 2%. Best performances are marked in **bold**. All components demonstrate substantial contributions to RHYTHM's effectiveness across datasets.

	Kumamoto			**Sapporo**			**Hiroshima**		
Model	Acc@1	Acc@3	Acc@5	Acc@1	Acc@3	Acc@5	Acc@1	Acc@3	Acc@5
RHYTHM	**0.2929**	**0.5200**	**0.5835**	**0.2938**	**0.4866**	**0.5502**	**0.2913**	**0.5027**	**0.5753**
w/o HA	0.2917	0.5163	0.5881	0.2901	0.4856	0.5481	0.2895	0.4946	0.5657
w/o token	0.2801	0.5049	0.5764	0.2768	0.4775	0.5409	0.2749	0.4812	0.5535
w/o Traj info.	0.2914	0.5176	0.5891	0.2879	0.4842	0.5472	0.2858	0.4916	0.5633
w/o Task desc.	0.2895	0.5166	0.5889	0.2883	0.4839	0.5463	0.2882	0.4934	0.5648