From Counts to Choice: Choice-Driven Spatial-Temporal Counting Process Models

Chao Yang *
School of Data Science
The Chinese University of Hong Kong (Shenzhen)
Shenzhen, China
222043011@link.cuhk.edu.cn

Yiling Kuang *
Department of Statistics and Data Science
The Chinese University of Hong Kong
Hong Kong SAR, China
yilingkuang@link.cuhk.edu.hk

Shuang Li †
School of Data Science
The Chinese University of Hong Kong (Shenzhen)
Shenzhen, China
lishuang@cuhk.edu.cn

Abstract

Spatio-temporal event data—such as crime incidents or shared-mobility usage—are generated by human decisions in urban environment. Yet most existing models focus on statistical dependencies in time and space, overlooking cognitive and social factors that shape behavior. We argue that uncovering underlying *preferences* is essential, as they provide a structured link between observed event data and decision processes. We introduce a **preference-driven framework** that models event distributions through a two-stage "consider–then–choose" process: *sparse gating* captures limited attention, and *utility functions* guide selection within the consideration set. To capture heterogeneity, we employ a *mixture-of-experts* design that reveals distinct preference patterns across groups and contexts. The framework incorporates *sparse structural design*, and we analyze its theoretical properties by establishing approximation and generalization guarantees. Empirical studies on crime and bike-sharing datasets demonstrate competitive predictive accuracy while providing interpretable insights into behavioral drivers. By shifting focus *from counts to preferences*, our approach offers a behaviorally grounded and socially meaningful perspective for modeling event data, especially useful in urban life.

1 Introduction

Many real-world counting processes in urban environment, such as criminal incidents or bike-sharing usage, represent aggregate macro-level patterns that emerge from micro-level human decision-making. Traditional spatial-temporal models, such as those using Gaussian processes in the Log-Cox Gaussian Process model [19, 10] or incorporating triggering kernel functions in spatial-temporal point processes [21], focus primarily on capturing spatial-temporal dependencies. However, these models often fall short in addressing the underlying human decision-making processes and social influences that shape urban events [25, 11]. To truly understand these processes, we must model the human mechanisms behind the counts—not just their spatial-temporal correlations.

To bridge this gap, we propose a novel approach that *models these urban counting processes through the lens of human choice behavior*. Our key insight is twofold: *i)* Human choices in urban settings inherently involve a "consider-then-choose" process—individuals first narrow down options to a manageable consideration set (e.g., "Which areas are feasible for biking?") and then make refined

*Equal contribution
†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025 Workshop UrbanAI).

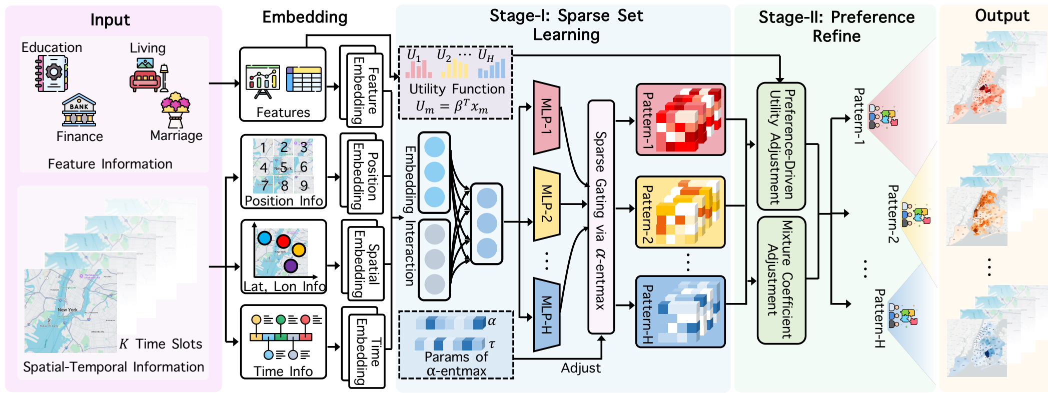

Figure 1: Model framework of **GLANCE**, from left to right: Input, Embedding Module, **Stage I**: sparse gates g_m^h filter options; **Stage II**: utilities U_m^h determine final choice probabilities, and Output.

selections within this subset. *ii)* Population-level counts in cities aggregate diverse preference patterns, where distinct subgroups (e.g., commuters vs. recreational cyclists) exhibit systematically different utilities for time-location pairs.

2 Preference-Driven Model for Spatio-Temporal Events

We view spatio-temporal event counts as *macro-level aggregates of many micro-level human choices*. Each decision corresponds to selecting a time–location pair from a vast universe of possibilities, shaped by cognitive limits and contextual cues. Inspired by the classical *consider–then–choose* paradigm [13], we model this process in two stages: first, a sparse attention mechanism filters and ranks feasible options; second, a utility function refines the final choice. At the population level, heterogeneity is captured by a mixture of latent decision-making patterns.

Shown in Fig. 1, we introduce the **G**ated **La**te**n**t Class **C**hoi**E** model, **GLANCE**, which integrates sparse attention, preference refinement, and population heterogeneity into a coherent framework.

2.1 Consider–Then–Choose Framework

Let $\mathcal{U} = \{(t_m, s_m)\}_{m=1}^{M}$ denote the universe of all discretized time–location pairs. For any individual, the effective choice set $\mathcal{C} \subseteq \mathcal{U}$ is *unknown*: people do not evaluate all M possibilities, but instead attend to a sparse subset shaped by context and cognitive limits. Our goal is to learn these latent *consideration sets* and the utilities guiding the final selection.

Stage I: Consideration via Sparse Attention. We introduce a gating vector $g \in \mathbb{R}^M$, where $g_m \in [0, 1]$ is the probability that option m enters the consideration set. Gates are generated from contextual features and learned end-to-end.

A key component is the α-*entmax* mapping [7], defined as

$$g = \alpha\text{-entmax}(\boldsymbol{z}) = \Big[(\alpha - 1)\boldsymbol{z} - \tau(\boldsymbol{z})\mathbf{1}\Big]_+^{\frac{1}{\alpha-1}},$$

where $\tau(\boldsymbol{z})$ ensures normalization. At $\alpha = 1$, this reduces to softmax (dense attention), while $\alpha > 1$ induces sparsity. Since it is convex and differentiable, the model can *learn sparse attention patterns directly from data*. Nonzero entries of g correspond to options predicted to belong to the consideration set, giving an interpretable representation of limited attention.

To generate scores $\boldsymbol{z}$, we embed each time–location pair into a d-dimensional vector. Let $X \in \mathbb{R}^{M\times d}$ be the shared embedding matrix. User- or event-specific features can be concatenated with these embeddings. We then apply two projection matrices $W_q, W_k \in \mathbb{R}^{d\times d'}$ and form

$$E = XW_q(XW_k)^\top \in \mathbb{R}^{M\times M}.$$

Because $W_q \neq W_k$, the interaction matrix E is generally asymmetric. Diagonal entries encode *intrinsic salience*, while off-diagonal terms capture how the presence of one option influences another (e.g., nearby stations competing for attention). Aggregating across rows,

$$\boldsymbol{z} = \sigma(E)\mathbf{1}, \quad \boldsymbol{z} \in \mathbb{R}^M,$$

where $\sigma(\cdot)$ is a nonlinear activation (e.g., ReLU, $\tanh$) and $\mathbf{1} \in \mathbb{R}^M$ is the all-ones vector. This produces a score vector summarizing both intrinsic and contextual influences before sparsification. Passing $\boldsymbol{z}$ through α-entmax yields g, a *data-driven estimate of the consideration set*.

Stage II: Choosing via Utility. Within the sparse set, selection is refined by a utility function. The utility of option m may be *feature-free* (a learnable scalar U_m) or *feature-dependent*, e.g.,

$$U_m = \beta^\top x_m,$$

where x_m may reuse or extend embeddings from X to encode socio-economic, temporal, or environmental attributes. The final choice probability is

$$f_m(\boldsymbol{z}, U) = \frac{g_m \exp(U_m)}{\sum_{m'=1}^{M} g_{m'} \exp(U_{m'})}.$$

This mirrors human decision-making: people first prune the vast universe into a manageable *consideration set*, then carefully choose among the survivors.

2.2 Capturing Population Heterogeneity

Human populations are rarely homogeneous. To model diverse decision rules, we introduce a mixture of H latent classes, each with its own sparse attention and utility functions.

Formally, class $h \in [H]$ defines a gating distribution $g^h = \alpha^h$-entmax$(\boldsymbol{z}^h)$ and utility U^h. The probability of selecting option m within class h is

$$f_m(\boldsymbol{z}^h, U^h) = \frac{g_m^h \exp(U_m^h)}{\sum_{m'=1}^{M} g_{m'}^h \exp(U_{m'}^h)}.$$

At the population level,

$$\mathbb{P}(m) = \sum_{h=1}^{H} \pi^h f_m(\boldsymbol{z}^h, U^h), \tag{1}$$

where π^h are nonnegative mixture weights summing to one. This structure uncovers interpretable subgroups—e.g., *commuters* who prioritize proximity versus *recreational users* who prefer socially vibrant options.

2.3 Likelihood and Training Objective

Suppose we observe N events. Each event is a realized time–location pair (t_i, s_i), encoded as a one-hot vector $y_i \in \mathbb{R}^M$ with $y_{im} = 1$ if the i-th event occurred at option m and $y_{im} = 0$ otherwise. Let $P_{im} = \mathbb{P}(m)$ denote the predicted probability of option m for event i. The log-likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_{im} \log P_{im}.$$

The model parameters are

$$\boldsymbol{\theta} = \Big\{ X, \{\pi^h, \alpha^h, W_q^h, W_k^h, \beta_h\}_{h=1}^{H} \Big\},$$

where X is a learnable embedding matrix for time–location alternatives. User or event-level covariates can be concatenated with X, so that both attention and utility adapt to context. Class-specific parameters (W_q^h, W_k^h, β_h) govern sparse attention and preferences, while π^h are mixture weights.

Maximizing $\mathcal{L}(\boldsymbol{\theta})$ aligns the model with observed event data, and the differentiability of α-entmax enables efficient, end-to-end gradient-based training.

3 Theoretical Analysis

In Theorem 1, we analyze the approximation error of our proposed model, which states that our finite latent class model can approximate any distribution of human preference parameters with arbitrary accuracy by increasing the number of latent classes H, and the maximum number of latent classes required is inversely proportional to the desired accuracy, in terms of the expected squared error. This bound holds for any underlying distribution of human preference parameters and event occurrences, and does not depend on the feature dimensions. This result resembles the universal approximation theorem for neural networks in a Barron space [5]. Details and proofs can be found in Appendix B.3.

In Theorem 2, we analyze our proposed model's generalization capability when trained on a finite data set. It demonstrates that the generalization bound is of order $\mathcal{O}(1/\sqrt{N})$, which implies stable performance improvements as the sample size N grows. Notably, this bound is independent of the number of latent classes H, regardless of the mixture distribution π, which is a desirable feature of the proposed model. The factor (M) that depends on the number of time-location pairs is included due to fact that the utility for each pair is treated separately. In practice, these pairs often have a lower-dimensional parameterization. In this case, it is easy to obtain a more favorable constant using such dimension reduction. Details and proofs can be found in Appendix B.4.

4 Experiments

4.1 Experimental Setup

Datasets We considered three real-world spatial-temporal datasets: *i) New York Crime*[3]. *ii) Chicago Crime*[4]. *iii) Shanghai Mobike*[5].

Baselines To evaluate the capability of our proposed models, we compare against commonly used baselines and state-of-the-art models. *i) ARMA* [4], *ii) CSI* [8], *iii) LGCP* [10, 18], *iv) NSTPP* [6], *v) DSTPP* [24], *vi) ST-HSL* [15], *vii) HintNet* [3], *viii) STNSCM* [9], *ix) UniST* [23], and *x) MNL (Multinomial Logic Choice Model)* [17, 12].

Evaluation Metrics. For evaluation, we group events into aggregate units i (e.g., one day or one week). For each unit, we form the empirical distribution $\boldsymbol{P}_i = (P_{i1}, \ldots, P_{iM}) \in \Delta_M$ by normalizing the observed counts across the M time–location options. Our model produces a corresponding predicted probability vector $\hat{\boldsymbol{P}}_i = (\hat{P}_{i1}, \ldots, \hat{P}_{iM}) \in \Delta_M$. We use two metrics: *(i) KL divergence*, which measures the discrepancy between predicted and empirical distributions, and *(ii) RMSE*, which captures numerical prediction error across options [25]:

$$\textit{(i)}\ \mathrm{KL}_i = D_{\mathrm{KL}}\Big(\hat{\boldsymbol{P}}_i \,\Big\|\, \boldsymbol{P}_i\Big) = \sum_{m=1}^{M} \hat{P}_{im} \log \frac{\hat{P}_{im}}{P_{im}}, \text{ and } \textit{(ii)}\ \mathrm{RMSE}_i = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \Big(\hat{P}_{im} - P_{im}\Big)^2}.$$

Model	NYC Crime		Chicago Crime		Shanghai Mobike	
	KL ↓	RMSE ↓	KL ↓	RMSE ↓	KL ↓	RMSE ↓
AMAR	0.65 +/- 0.06	0.62 +/- 0.08	0.70 +/- 0.10	0.68 +/- 0.06	0.46 +/- 0.08	0.42 +/- 0.04
CSI	0.67 +/- 0.08	0.66 +/- 0.03	0.68 +/- 0.12	0.65 +/- 0.09	0.47 +/- 0.04	0.43 +/- 0.05
LGCP	0.67 +/- 0.11	0.67 +/- 0.09	0.69 +/- 0.09	0.68 +/- 0.08	0.45 +/- 0.10	0.43 +/- 0.09
NSTPP	0.51 +/- 0.06	0.49 +/- 0.05	0.42 +/- 0.07	0.44 +/- 0.10	0.32 +/- 0.02	0.33 +/- 0.05
DSTPP	0.47 +/- 0.04	0.45 +/- 0.05	0.47 +/- 0.04	0.46 +/- 0.08	0.37 +/- 0.03	0.40 +/- 0.02
ST-HSL	0.56 +/- 0.06	0.52 +/- 0.05	0.49 +/- 0.04	0.52 +/- 0.06	0.38 +/- 0.05	0.43 +/- 0.03
HintNet	0.38 +/- 0.03	0.37 +/- 0.03	0.26 +/- 0.04	0.28 +/- 0.03	0.19 +/- 0.01	0.17 +/- 0.02
STNSCM	0.38 +/- 0.02	0.38 +/- 0.04	0.27 +/- 0.01	0.31 +/- 0.02	0.11 +/- 0.00	0.15 +/- 0.01
UniST	0.37 +/- 0.03	0.36 +/- 0.02	0.27 +/- 0.05	0.30 +/- 0.04	0.23 +/- 0.04	0.25 +/- 0.06
MNL	0.38 +/- 0.01	0.38 +/- 0.01	0.25 +/- 0.03	0.29 +/- 0.02	0.13 +/- 0.01	0.17 +/- 0.01
GLANCE	0.36 +/- 0.02	0.35 +/- 0.01	0.24 +/- 0.02	0.27 +/- 0.02	0.12 +/- 0.01	0.16 +/- 0.01

Table 1: Comparison of our model with baselines for prediction tasks, conducting using training data comprising 16,847 samples for NYC, 23,545 samples for Chicago, and 20,883 samples for Shanghai. Purple signifies the best result, while orange text indicates the second-best result. Performance metrics are averaged across three different runs, which reported as (Mean $+/-$ SD).

4.2 Results and Analysis

Analysis 1: Prediction Performance For prediction tasks, we utilize data from the final day as testing data, reserving remaining data as training data for all three datasets. The results in Tab. 1 demonstrate that GLANCE consistently surpasses the majority baseline methods or at least achieves competitive prediction accuracy.

[3]https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

[4]https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[5]https://github.com/Andrehinh/Interesting-python/tree/master/Mobike

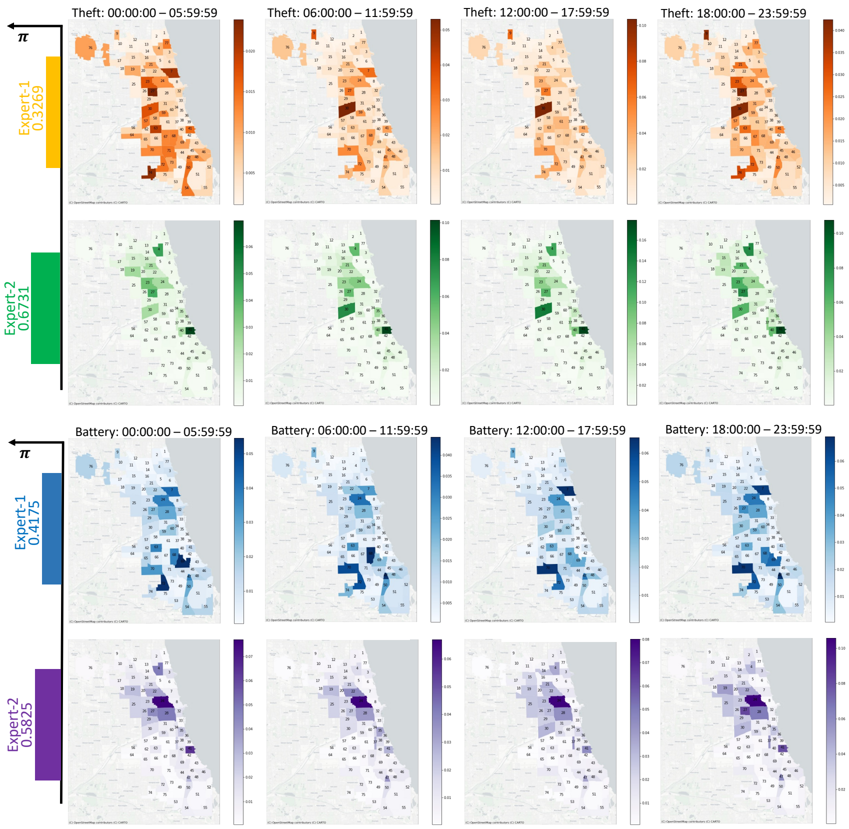

Figure 2: Mixing coefficient π^h (Left bar plots) and mixture pattern adjusted by utility score $(g^h \exp(U^h))$ for different latent class-h and different crime types, including theft and battery (Right heatmaps) from July 1 to July 31, 2024, in Chicago City. The selection of the number of experts is based on empirical experiments.

Analysis 2: Explain Human Decision Process Beyond raw accuracy, GLANCE uncovers latent behavioral structures that shed light on heterogeneous decision-making. Model selection via negative log-likelihood, training cost, and efficiency consistently suggests two latent classes for Chicago (Tab. 2, Appendix. C.3), striking a balance between parsimony and expressiveness. These latent classes reveal distinct crime decision patterns (Fig. 2):

- **Theft.** *Class 1 (33%)*: Smaller subgroup operating in the West and Far Southwest (Communities-30, -72, -74), with notable evening surges in Community-27. *Class 2 (67%)*: Larger subgroup centered on the South and West (Communities-41, -30), avoiding North/Central areas, but showing strong morning activity in Community-30.
- **Battery.** *Class 1 (42%)*: Spread across North and Far Southwest (Communities-7, -70, -72), generally avoiding Far North and South. *Class 2 (58%)*: Dominated by West Side activity, especially Community-24, with consistent high risk and nighttime surges in Community-41.

These findings highlight a key discovery: *different crime types are not only clustered in space and time but are also driven by distinct offender subgroups with different choice logics*. Unlike hotspot maps that aggregate over populations, GLANCE disentangles these heterogeneous behavioral strategies, enabling more precise and actionable interventions (e.g., tailoring patrols to theft vs. battery patterns).

5 Conclusion

We introduced GLANCE, a preference-driven framework for modeling spatio-temporal events in urban environments, which interprets aggregate counts as the macro-level outcome of numerous micro-level consider–then–choose decisions. By integrating sparse attention, flexible utility representations, and a mixture-of-experts architecture, our model effectively captures cognitive constraints and heterogeneous preferences across urban populations, all while maintaining interpretability. This approach not only enhances behavioral realism in urban event modeling but also provides a structured understanding of how human decisions shape complex urban phenomena.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021.

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[3] Bang An, Amin Vahedian, Xun Zhou, W Nick Street, and Yanhua Li. Hintnet: Hierarchical knowledge transfer networks for traffic accident forecasting on heterogeneous spatio-temporal data. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 334–342. SIAM, 2022.

[4] Shahab Araghinejad and Shahab Araghinejad. *Time Series Modeling*. Springer, 2014.

[5] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[6] Ricky TQ Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. *arXiv preprint arXiv:2011.04583*, 2020.

[7] Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

[8] Boor de. A practical guide to splines. 1978.

[9] Pan Deng, Yu Zhao, Junting Liu, Xiaofeng Jia, and Mulan Wang. Spatio-temporal neural structural causal models for bike flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4242–4249, 2023.

[10] Peter J Diggle, Paula Moraga, Barry Rowlingson, Benjamin M Taylor, et al. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

[11] Pu He, Fanyin Zheng, Elena Belavina, and Karan Girotra. Customer preference and station network in the london bike-share system. *Management Science*, 67(3):1392–1412, 2021.

[12] Yiqun Hu, David Simchi-Levi, and Zhenzhen Yan. Learning mixed multinomial logits with provable guarantees. *Advances in Neural Information Processing Systems*, 35:9447–9459, 2022.

[13] Mert Kimya. Choice, consideration sets, and attribute filters. *American Economic Journal: Microeconomics*, 10(4):223–247, 2018.

[14] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

[15] Zhonghang Li, Chao Huang, Lianghao Xia, Yong Xu, and Jian Pei. Spatial-temporal hypergraph self-supervised learning for crime prediction. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 2984–2996. IEEE, 2022.

[16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

[17] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

[18] Andrew Miller, Luke Bornn, Ryan Adams, and Kirk Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *International conference on machine learning*, pages 235–243. PMLR, 2014.

[19] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

[20] Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019.

[21] Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[22] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

[23] Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4095–4106, 2024.

[24] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3173–3184, 2023.

[25] Xiangyu Zhao and Jiliang Tang. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 497–506, 2017.

[26] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Sec. 4, the experimental results reflect the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix. E, we discuss the limitation of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: In Sec. 3 and Appendix. B, we provide the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Appendix. D, we report the reproducibility analysis, including computing infrastructure and hyper-parameter selection.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the access to the data in Sec. 4. And we commit to open-sourcing the implementation code immediately after camera-ready.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Sec. 4, Appendix D, and Appendix C, we specify all the training and test details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In Sec. 4 and Appendix C, we report error bars in figures and tables.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix. D, we report the needed information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Sec. E, we discuss both potential positive societal impacts and negative societal impacts of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We pose no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Sec. 4 and Appendix C, we provide the licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix Overview

In the following, we will provide supplementary materials to better illustrate our methods and experiments.

- Section. A provides the algorithm details of our proposed choice-driven spatial-temporal counting process model.
- Section. B derives theoretical guarantees for approximation error and generalization error.
- Section. C provides more detailed analysis about experiments on real-world datasets.
- Section. D reports the reproducibility analysis, including computing infrastructure and hyper-parameter selection.
- Section. E states the limitation, future research direction, and broader impacts of this work.

A Algorithm Details

A.1 Spatial-temporal embedding

We adopt a spatial-temporal embedding method akin to that described in [2, 1]. Initially, the region is segmented into distinct blocks. To encode block order, we introduce sinusoidal positional encoding for area position embedding. Subsequently, linear embedding is utilized for spatial information, typically latitude and longitude. Temporal information is encoded using sinusoidal positional encoding as described in [26]. In the context of decision-making, relevant static features can be encoded using one-hot semantic encoding, while dynamic features are encoded linearly. The embeddings for spatial, temporal, and relevant feature information are then combined via element-wise addition to generate the comprehensive embedding.

First, we divide the area into disjoint blocks. To inject a notation of block ordering we add sinusoidal positional encoding for these area position embedding. Then considering the spatial information which are usually latitude and longitude, we apply linear embedding. For encoding temporal information, we also adopt sinusoidal positional encoding a [26]. Considering other relevant features related to the decision-making process, we can apply one-hot semantic encoding for static feature and linear encoding for dynamic features. Finally, the embedding for spatial information, temporal information, and relevant feature information then directly element-wise addition together to obtain the overall embedding.

This approach is similar to the use of positional embeddings and feature embeddings in attention mechanisms, where initial embeddings are transformed through linear projections to capture more nuanced information. By combining the base embeddings A and B with these flexible projections, our model can more accurately represent and adapt to diverse preference patterns and social influences, enriching the overall decision-making framework.

A.2 Overall algorithm

The overall algorithm is shown in Alg. 1, which illustrates the learning process of all the model parameters for our proposed model in detail.

B Theoretical Details

B.1 Details of α-entmax

$$\alpha\text{-entmax}\,(\boldsymbol{z}) := \underset{\boldsymbol{p}\in\Delta^M}{\operatorname{argmax}} \quad \boldsymbol{p}^\top \boldsymbol{z} + \mathrm{H}_\alpha^{\mathrm{T}}(\boldsymbol{p}), \tag{2}$$

where $\Delta^M := \left\{\boldsymbol{p} \in \mathbb{R}^M : \sum_i p_i = 1\right\}$ is the probability simplex, and, for $\alpha \geq 1, \mathrm{H}_\alpha^{\mathrm{T}}$ is the Tsallis continuous family of entropies [22]:

$$\mathrm{H}_\alpha^{\mathrm{T}}(\boldsymbol{p}) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \sum_j \left(p_j - p_j^\alpha\right), & \alpha \neq 1 \\ -\sum_j p_j \log p_j, & \alpha = 1 \end{cases}$$

This family contains the well-known Shannon and Gini entropies, corresponding to the cases $\alpha = 1$ and $\alpha = 2$, respectively.

Algorithm 1 Learning the Model Parameters for the Mixture-of-Experts Model

Input: Observed data $\{y_{i,m}\}_{i=1}^N$, initial parameters

$$\boldsymbol{\theta} = \left\{\pi, A, B, \{[\alpha^h, W_A^h, W_B^h, U^h]\}_{h\in[H]}\right\}$$

Output: Optimized model parameters $\boldsymbol{\theta}^*$
Initialization: Initialize $\boldsymbol{\theta}$ randomly or heuristically.
Description: A and B serve as shared feature embeddings that encode the positional and contextual information necessary for understanding the preference distribution in generating the events.
repeat
  for each expert $h \in [H]$ **do**
    Compute gating function:

$$g^h = \alpha\text{-entmax}\left(\sigma(AW_A^h(BW_B^h)^T)\mathbf{1}\right).$$

  end for
  for each expert $h \in [H]$ **do**
    Compute probability:

$$P_{i,m} = \sum_{h=1}^{H} \pi^h \frac{g_m^h \exp\left(U_m^h\right)}{\sum_{m'=1}^{M} g_{m'}^h \exp\left(U_{m'}^h\right)}.$$

  end for
  Optimize: Maximize the likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{m=1}^{M}(P_{i,m})^{y_{i,m}}$$

  to update $\boldsymbol{\theta}$ using gradient descent or a similar optimization method.
until The likelihood is converged

B.2 Preliminaries and Assumptions for Theorems

Model recap (single class). For an alternative $m \in [M]$, let $x_m^\top$ denote the m-th row of the shared embedding matrix $X \in \mathbb{R}^{M\times d}$. Given class-specific projections $W_q, W_k \in \mathbb{R}^{d\times d'}$, define the interaction matrix

$$E = XW_q(XW_k)^\top \in \mathbb{R}^{M\times M}.$$

Let $\sigma(\cdot)$ be an elementwise nonlinearity (assumed 1-Lipschitz, e.g., ReLU or $\tanh$), and let $\mathbf{1} \in \mathbb{R}^M$ be the all-ones vector. Define scores $z = \sigma(E)\mathbf{1} \in \mathbb{R}^M$ and a sparse gating distribution $g = \mathrm{Entmax}_\alpha(z)$ for some $\alpha \in [1+\delta, 2]$ with $\delta > 0$. Utilities can be feature-free ($U_m \in \mathbb{R}$) or feature-dependent ($U_m = \beta^\top x_m$). The class-wise choice probability is

$$f_m(z, U) = \frac{g_m \exp(U_m)}{\sum_{m'} g_{m'} \exp(U_{m'})}, \qquad m \in [M].$$

With H classes, mixture weights π^h, and parameters $\{W_q^h, W_k^h, \beta_h, \alpha^h\}_{h=1}^H$, the population probability is

$$P(m) = \sum_{h=1}^{H} \pi^h f_m\left(z^h, U^h\right).$$

Assumptions used in proofs. Throughout the proofs we assume:

1. **Bounded embeddings:** $\|X\|_F \le C_X$. (If X is learned, the optimization is regularized so that this holds at the solution; if X is fixed features, this is immediate.)
2. **Bounded projections:** $\|W_q^h(W_k^h)^\top\|_F \le C_W$ for all h.
3. **Bounded utilities:** $\|\beta_h\|_2 \le C_U$ (feature-dependent case) or $|U_m^h| \le C_U$ (feature-free case).

4. **Sparse gating parameter:** $\alpha^h \in [1+\delta, 2]$ for some fixed $\delta > 0$ to avoid the softmax limit and ensure Lipschitz constants below remain finite.
5. **Lipschitz nonlinearity:** σ is 1-Lipschitz and monotone (true for ReLU, $\tanh$).

A useful Lipschitz property of Entmax_α. We use that for $\alpha \in [1+\delta, 2]$ the mapping $z \mapsto \mathrm{Entmax}_\alpha(z)$ is globally Lipschitz on $\mathbb{R}^M$ with a constant $L_{\mathrm{ent}}(\delta) = \mathcal{O}(1/\delta)$ (see, e.g., properties derived via strong convexity of the Tsallis-entropy regularizer; cf. [20, 7]). Formally, there exists $L_{\mathrm{ent}}(\delta) > 0$ such that

$$\left\|\mathrm{Entmax}_\alpha(z) - \mathrm{Entmax}_\alpha(z')\right\|_2 \;\le\; L_{\mathrm{ent}}(\delta)\,\|z - z'\|_2\,, \qquad \forall z, z' \in \mathbb{R}^M. \tag{3}$$

Mixture reduction. For any function class $\mathcal{F}$ that is convex in parameters, the empirical Rademacher complexity of mixtures satisfies

$$\sup_{\{\pi^h, f^h\}} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i \sum_{h=1}^{H} \pi^h f^h(x_i) \;\le\; \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i f(x_i), \tag{4}$$

because $\sum_h \pi^h f^h$ lies in the convex hull of $\mathcal{F}$ and the supremum over a convex hull is attained at an extreme point.

We now prove the two theorems.

B.3 Theorem 1

B.3.1 Details and Corresponding Analysis

Let $q_i^* \in \Delta_M$ denote the true choice probability distribution over M time–location pairs for event i, induced by an unknown distribution over latent preference parameters. Our H-class GLANCE model produces an approximation q_H^i.

Theorem 1 (Universal Approximation)**.** *For any $\epsilon > 0$, there exists a finite mixture with $H \le 2/\epsilon$ classes such that*

$$\frac{1}{N}\sum_{i=1}^{N} \|q_H^i - q_i^*\|^2 \le \epsilon.$$

This shows that enlarging H increases model capacity, and a sufficiently rich mixture can approximate *any* distribution of human preference parameters with arbitrary precision. The dependence $H = \mathcal{O}(1/\epsilon)$ resembles classical universal approximation results [5].

B.3.2 Proofs

Setup. Let μ_* be the (unknown) distribution over latent parameters $\phi = (\pi, \alpha, W_q, W_k, \beta)$. For a fixed context (here suppressed in notation; if contexts vary across events i, interpret all maps below pointwise in i), define the measurable map

$$\Phi : \phi \mapsto q(\phi) \in \Delta_M, \qquad q_m(\phi) \;=\; \sum_{h=1}^{H} \pi^h f_m\big(z^h(\phi), U^h(\phi)\big).$$

The *true* choice distribution is the pushforward mean

$$q^* \;=\; \mathbb{E}_{\phi\sim\mu_*}\big[\,q(\phi)\,\big] \in \Delta_M.$$

(If the context changes with i, define $q_i(\phi)$ and $q_i^* = \mathbb{E}[q_i(\phi)]$; the argument below applies to each i separately and then we average over i.)

Finite-support approximation via sampling (probabilistic method). Draw i.i.d. parameters $\phi^{(1)}, \dots, \phi^{(H)} \sim \mu_*$ and form the empirical mixture

$$q_H \;=\; \frac{1}{H}\sum_{h=1}^{H} q\big(\phi^{(h)}\big)\,.$$

Because $q(\phi) \in \Delta_M$ for all ϕ, we have $\|q(\phi)\|_2^2 \le \|q(\phi)\|_1 = 1$ (since all coordinates are nonnegative and sum to 1). Hence

$$\begin{aligned}\mathbb{E}\,\|\,q_H - q^*\,\|_2^2 &= \mathbb{E}\,\Big\|\,\tfrac{1}{H}\sum_{h=1}^{H}\big(q(\phi^{(h)}) - \mathbb{E}[q(\phi)]\big)\,\Big\|_2^2 \\ &= \frac{1}{H}\,\mathrm{Tr}\big(\mathrm{Cov}(q(\phi))\big) \;\le\; \frac{1}{H}\,\mathbb{E}\|q(\phi)\|_2^2 \;\le\; \frac{1}{H}.\end{aligned}$$

Therefore $\mathbb{E}\|q_H - q^*\|_2^2 \leq 1/H$. By the probabilistic method, there exists a realization of $\{\phi^{(h)}\}_{h=1}^H$ such that $\|q_H - q^*\|_2^2 \leq 1/H$.

High-probability and multi-i averaging. A standard vector Bernstein (or Hoeffding) inequality yields that, with probability at least $1-\eta$,

$$\|q_H - q^*\|_2^2 \;\leq\; \frac{c_1 + c_2 \log(1/\eta)}{H}$$

for universal constants c_1, c_2. When contexts vary across $i = 1, \ldots, N$, repeat the argument pointwise to get, with the same H,

$$\frac{1}{N}\sum_{i=1}^{N} \|q_H^i - q_i^*\|_2^2 \;\leq\; \frac{c_1 + c_2 \log(1/\eta)}{H}.$$

Choosing $H \geq 2(c_1 + c_2 \log(1/\eta))/\epsilon$ yields the stated $\mathcal{O}(1/\epsilon)$ rate. Absorbing constants gives the main-text statement "$H \leq 2/\epsilon$" up to universal multiplicative factors.

Conclusion. Thus a finite mixture with $H = \mathcal{O}(1/\epsilon)$ atoms suffices to approximate the population distribution arbitrarily well in average squared ℓ_2 error.

□

B.4 Theorem 2

B.4.1 Details and Corresponding Analysis

Suppose we observe N events $\{(t_i, s_i)\}_{i=1}^N$, each encoded as a one-hot $y_i \in \mathbb{R}^M$. We fit GLANCE parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood $\mathcal{L}_{\hat{\mathcal{D}}_N}(\boldsymbol{\theta})$. Let $\mathcal{L}_{\mathcal{D}_*}(\boldsymbol{\theta})$ denote the population loss under the true distribution $\mathcal{D}_*$. We analyze the generalization gap

$$\mathcal{L}_{\mathcal{D}_*}(\boldsymbol{\theta}) - \mathcal{L}_{\hat{\mathcal{D}}_N}(\boldsymbol{\theta}).$$

Assumptions. We assume: (i) the shared embedding X of time–location pairs has bounded Frobenius norm; (ii) projections satisfy $\|W_q^h (W_k^h)^\top\|_F \leq C_W$; (iii) utilities are bounded $\|\beta_h\|_2 \leq C_U$; (iv) $\alpha^h \in [1+\delta, 2]$ for some $\delta > 0$ to avoid degeneracy.

Theorem 2 (Generalization Bound)**.** *Under the above assumptions, the empirical Rademacher complexity of the GLANCE model class satisfies*

$$\mathfrak{R}_{\hat{\mathcal{D}}_N}(\mathcal{W}) = \tilde{\mathcal{O}}\left(\frac{M\, e^{C_U} C_W}{\delta\sqrt{N}}\right).$$

Consequently, with high probability,

$$\mathcal{L}_{\mathcal{D}_*}(\boldsymbol{\theta}) - \mathcal{L}_{\hat{\mathcal{D}}_N}(\boldsymbol{\theta}) \leq \tilde{\mathcal{O}}\left(\frac{M\, e^{C_U} C_W}{\delta\sqrt{N}}\right).$$

The bound decays at the standard $\mathcal{O}(1/\sqrt{N})$ rate, showing stable improvement with more data. Notably, it is *independent of the number of latent classes* H, so adding mixture components to capture heterogeneity does not compromise generalization. The dependence on M reflects the size of the time–location universe, though in practice low-dimensional embeddings in X yield smaller constants.

B.4.2 Proofs

We bound the generalization gap via the empirical Rademacher complexity of the probability outputs. Let $\mathcal{F}$ be the class of vector-valued functions mapping an input index i to $P_i(\boldsymbol{\theta}) = (P_{i1}, \ldots, P_{iM}) \in \Delta_M$:

$$\mathcal{F} \;=\; \Big\{ i \mapsto P_i(\boldsymbol{\theta}) = \sum_{h=1}^{H} \pi^h f\big(\boldsymbol{z}_i^h(\boldsymbol{\theta}), U^h(\boldsymbol{\theta})\big) \;:\; \boldsymbol{\theta} \in \mathcal{W} \Big\},$$

where $\mathcal{W}$ is the constrained parameter set described below.

Step 1: Symmetrization and vector contraction. Let $\ell(y, P) = -\sum_{m=1}^{M} y_m \log P_m$ be the log-loss. By standard symmetrization and the vector contraction inequality [16], if ℓ is L_ℓ-Lipschitz in P on the domain of interest, then with high probability

$$\mathcal{L}_{\mathcal{D}_*}(\boldsymbol{\theta}) - \mathcal{L}_{\hat{\mathcal{D}}_N}(\boldsymbol{\theta}) \;\lesssim\; L_\ell \cdot \mathfrak{R}_N(\mathcal{F}) \;+\; \tilde{\mathcal{O}}\big(\sqrt{\tfrac{1}{N}}\big).$$

Since the probabilities are bounded away from 0 due to bounded utilities and the gating normalization (see below), L_ℓ is finite and can be absorbed into the final constant. It remains to bound $\mathfrak{R}_N(\mathcal{F})$.

Step 2: Rademacher complexity of mixture reduces to single class. Let ϵ_{im} be i.i.d. Rademacher variables. Then

$$\begin{aligned}\mathfrak{R}_N(\mathcal{F}) &= \mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\theta}\in\mathcal{W}} \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\epsilon_{im}\sum_{h=1}^{H}\pi^h f_m\big(\boldsymbol{z}_i^h, U^h\big)\right] \\ &\le \mathbb{E}_\epsilon\left[\sup_{\{(\pi^h,\theta^h)\}}\sum_{h=1}^{H}\pi^h\cdot\frac{1}{N}\sum_{i,m}\epsilon_{im}\, f_m\big(\boldsymbol{z}_i^h, U^h\big)\right] \\ &\le \mathbb{E}_\epsilon\left[\sup_{\theta}\ \frac{1}{N}\sum_{i,m}\epsilon_{im}\, f_m\big(\boldsymbol{z}_i(\theta), U(\theta)\big)\right] \ = \ \mathfrak{R}_N(\mathcal{F}_1),\end{aligned} \tag{5}$$

where $\mathcal{F}_1$ is the single-class function class and we used the convexity bound (4). Thus *the mixture does not increase the complexity beyond that of one class*, explaining the independence of H in the final bound.

Step 3: Lipschitzness of $f(\cdot)$ in $(\boldsymbol{z}, U)$. Fix an index i and suppress it in notation. Consider two parameter settings inducing $(\boldsymbol{z}, U)$ and $(\boldsymbol{z}', U')$, and their corresponding gates $g = \mathrm{Entmax}_\alpha(\boldsymbol{z})$, $g' = \mathrm{Entmax}_\alpha(\boldsymbol{z}')$ with the same $\alpha \in [1+\delta, 2]$. Write the class-wise probability as

$$f_m(\boldsymbol{z}, U) \ = \ \frac{g_m e^{U_m}}{\sum_k g_k e^{U_k}} \ := \ \frac{a_m}{A}, \quad a_m = g_m e^{U_m},\ A = \sum_k a_k.$$

A direct Jacobian calculation (softmax-like) yields that, on the domain where $\|U\|_\infty \le C_U$ and $g \in \Delta_M$, the mapping $(g, U) \mapsto f$ is L_f-Lipschitz in the norm $\|(g,U)\| := \|g\|_2 + \|U\|_2$ with

$$L_f \ \le \ c_0\, e^{C_U}, \tag{6}$$

for an absolute constant $c_0 > 0$.[6] By the chain rule and (3), we obtain

$$\|f(\boldsymbol{z},U) - f(\boldsymbol{z}',U')\|_2 \ \le \ L_f\Big(L_{\mathrm{ent}}(\delta)\,\|\boldsymbol{z}-\boldsymbol{z}'\|_2 + \|U-U'\|_2\Big) \ \le \ c_1\,\frac{e^{C_U}}{\delta}\Big(\|\boldsymbol{z}-\boldsymbol{z}'\|_2 + \|U-U'\|_2\Big), \tag{7}$$

for a constant c_1 absorbing c_0 and the entmax Lipschitz factor.

Step 4: Bounding changes in $\boldsymbol{z}$ by parameter norms. Recall $E = XW_q(XW_k)^\top$. Using sub-multiplicativity and $\|\sigma(A)\|_F \le \|A\|_F$ for 1-Lipschitz σ, we have

$$\|E\|_F \ \le \ \|XW_q\|_F\,\|XW_k\|_F \ \le \ \|X\|_F^2\,\|W_q\|_F\,\|W_k\|_F.$$

Bounding the product with $\|W_q(W_k)^\top\|_F \le C_W$ and $\|X\|_F \le C_X$, we get $\|E\|_F \le C_X^2\, C_W$. Then

$$\|\boldsymbol{z}\|_2 \ = \ \|\sigma(E)\mathbf{1}\|_2 \ \le \ \|\sigma(E)\|_F\,\|\mathbf{1}\|_2 \ \le \ \|E\|_F\sqrt{M} \ \le \ C_X^2\, C_W\sqrt{M}. \tag{8}$$

Similarly, for two parameter settings,

$$\begin{aligned}\|\boldsymbol{z}-\boldsymbol{z}'\|_2 = \|\sigma(E)\mathbf{1} - \sigma(E')\mathbf{1}\|_2 \ &\le \ \|\sigma(E)-\sigma(E')\|_F\,\|\mathbf{1}\|_2 \ \le \ \|E-E'\|_F\sqrt{M} \\ &\le \ \sqrt{M}\Big(\|X\|_F^2\big\|W_q - W_q'\big\|_F\,\|W_k\|_F + \|X\|_F^2\|W_q'\|_F\,\big\|W_k - W_k'\big\|_F\Big) \\ &\le \ c_2\, C_X^2\sqrt{M}\,\big\|W_q(W_k)^\top - W_q'(W_k')^\top\big\|_F \ \le \ c_2\, C_X^2\sqrt{M}\ \cdot 2C_W,\end{aligned} \tag{9}$$

where in the last step we used a standard bilinear difference bound and the norm constraints (the constant c_2 absorbs the bilinear inequality constants). This shows z is Lipschitz in the projected interaction with constant $\mathcal{O}(C_X^2\sqrt{M})$.

[6]Sketch: $\partial f_m/\partial U_k$ is bounded by e^{C_U} times a probability-difference term; similarly $\partial f_m/\partial g_k$ is bounded by e^{C_U}. Summing over m and using Cauchy–Schwarz gives the stated Lipschitz bound in ℓ_2.

Step 5: Putting it together for $\mathcal{F}_1$. Using (7) and (9), the single-class mapping $\boldsymbol{\theta} \mapsto f(\boldsymbol{z}(\boldsymbol{\theta}), U(\boldsymbol{\theta}))$ is Lipschitz in the parameter block

$$\omega \;:=\; \big(W_q(W_k)^\top,\; \beta\big)$$

with constant

$$L_{\mathcal{F}_1} \;\lesssim\; \frac{e^{C_U}}{\delta}\Big(C_X^2\sqrt{M} \;+\; 1\Big).$$

Therefore, applying the *vector* Rademacher contraction inequality to the coordinate-wise linear forms $\sum_{i,m} \epsilon_{im} f_m(\cdot)$ yields

$$\mathfrak{R}_N(\mathcal{F}_1) \;\lesssim\; \frac{L_{\mathcal{F}_1}}{\sqrt{N}} \;\cdot\; \underbrace{\Big(\sup_{\theta\in\mathcal{W}} \|\omega\|_F\Big)}_{\le C_W + C_U} \;\lesssim\; \frac{e^{C_U}}{\delta\sqrt{N}}\Big(C_X^2\sqrt{M}+1\Big)\,(C_W + C_U).$$

Absorbing additive constants and C_X into $\tilde{\mathcal{O}}(\cdot)$ and recalling (5), we obtain the advertised form

$$\mathfrak{R}_N(\mathcal{F}) \;=\; \tilde{\mathcal{O}}\bigg(\frac{M\, e^{C_U}\, C_W}{\delta\sqrt{N}}\bigg).$$

Step 6: From complexity to generalization. Combining the symmetrization step with the above complexity bound, and absorbing the Lipschitz constant of the log-loss into the polylog factors, gives with high probability

$$\mathcal{L}_{\mathcal{D}_*}(\boldsymbol{\theta}) - \mathcal{L}_{\hat{\mathcal{D}}_N}(\boldsymbol{\theta}) \;\le\; \tilde{\mathcal{O}}\bigg(\frac{M\, e^{C_U}\, C_W}{\delta\sqrt{N}}\bigg),$$

which matches Theorem 2 in the main text (up to polylogarithmic factors in M and confidence $1-\eta$). $\square$

B.5 Remarks on Constants and Independence of H

Independence of H. The mixture-to-single-class reduction (5) explains why H does not appear in the bound: Rademacher complexity is convex, and the convex combination of classes does not expand the extremal envelope.

On the M factor. The M dependence enters through (8)–(9) (aggregating across M alternatives) and the vector contraction. In practice, alternatives are encoded in low-dimensional embeddings ($d \ll M$), and spatial/temporal structure further reduces effective capacity, improving constants.

On $\alpha \in [1+\delta, 2]$. The lower margin $\delta > 0$ ensures the entmax mapping remains Lipschitz with constant $L_{\mathrm{ent}}(\delta) = \mathcal{O}(1/\delta)$; taking $\alpha \downarrow 1$ (softmax) makes this constant blow up. Our bound explicitly reflects this via the $1/\delta$ factor.

C Experimental Details

C.1 Baseline Descriptions

We consider following commonly-used baselines and state- of-the-art models: *i) ARMA* [4]: Auto-Regression-Moving-Average is well known for predicting time series data. ARMA predicts the event number of a region solely based on the historical event records of the region, considering the recent time slots for a moving average. *ii) CSI* [8]: Cubic Spline Interpolation trains piecewise third-order polynomials which pass through event points of recent time slots, and then predicts the event number in the near future by the trained polynomials. *iii) LGCP* [10, 18]: Log-Gaussian Cox Process is a kind of Poisson process with varying intensity, where the log-intensity is assumed to be drawn from a Gaussian process. *iv) NSTPP* [6]: It applies neural ODEs as the backbone, which parameterized the temporal intensity with neural jump SDEs and the spatial PDF with continuous-time normalizing flows. *v) DSTPP* [24]: it leverages diffusion models to learn complex spatial-temporal joint distributions. *vi) ST-HSL* [15]: It proposes a Spatial-Temporal Self-Supervised Hypergraph Learning framework for crime prediction. *vii) HintNet* [3]: It performs a multi-level spatial partitioning to separate sub-regions with different risks and learns a deep network model for each level using spatial-temporal and graph convolutions *viii) STNSCM* [9]: A causality-based interpretation model for the bike flow prediction. *ix) UniST* [23]: A universal model designed for general urban spatial-temporal prediction across a wide range of scenarios. *x) MNL (Multinomial Logic Choice Model)* [12]: Degenerate the feature embedding of our method to time-location index embedding, while maintaining the consistent choice model framework.

C.2 How to explain results of other spatial-temporal models (e.g., LGCP)

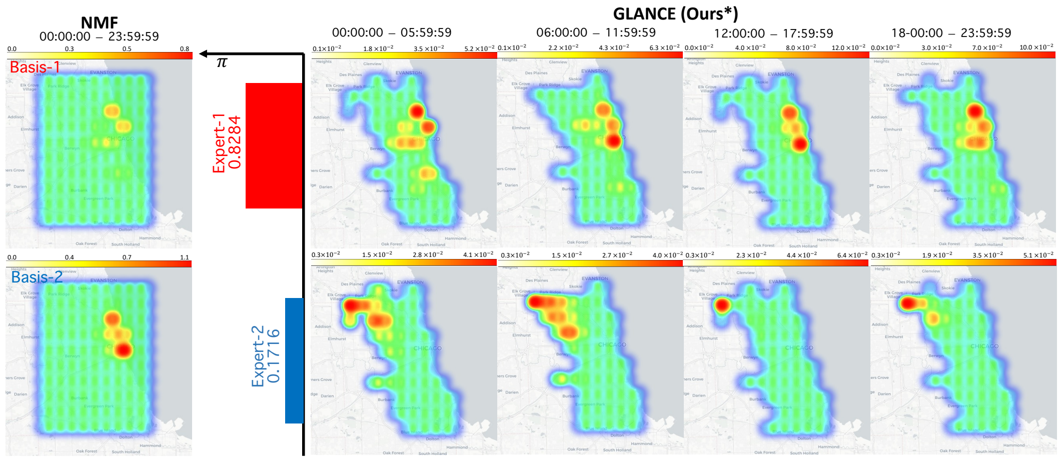

Figure 3: Comparison of the learned expert pattern of our choice model and the basis of non-negative matrix factorization (NMF) on Chicago City crime dataset. To align with the setting of LGCP-NMF, we partition the Chicago area into 10×10 area blocks. Left: NMF basis, Right: expert patterns learned by our model.

Consider a Log-Gaussian Cox Process (LGCP) [10], which is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponentiated Gaussian Process

$$Z(\cdot) \sim GP(0, k(\cdot, \cdot)), \tag{10}$$

$$\lambda(\cdot) \sim \exp(Z(\cdot)), \quad x_1, ..., x_N \sim PP(\lambda(\cdot)) \tag{11}$$

where $GP(\cdot)$ refer to a Gaussian Process, $PP(\cdot)$ refer to a Poisson process. $k(\cdot, \cdot)$ represents the squared exponential covariance function and x_i represents a countable collection of independent Poisson process with measure λ_i. It can be used to estimate the intensity surface of a spatial point process and therefore capture spatial patterns of data. LGCP-NMF [18] was proposed to use non-negative matrix decomposition [14] of Poisson process intensity surfaces to provide an interpretable feature space that parsimoniously describes the learned intensity matrix $\Lambda \in \mathbb{R}^{T \times S}$ from LGCP.

$$\Lambda \approx WB \tag{12}$$

where $W \in \mathbb{R}^{T \times H}$ is the weight matrix, and $B \in \mathbb{H}^{T \times S}$ is the basis matrix. S is the number of spatial grids, and T is the number of temporal intervals. H is the number of mixtures, which is set to be the same as our model.

Our model provides a refined alternative perspective to explain existing spatial-temporal models. Unlike LGCP-NMF, alter LGCP model being well-trained, we fit our model using a new objective function based on the least squared error between estimated probability of our model and the probability from the LGCP. This approach allows us to interpret the expert patterns learned by our model to explain the already fitted LGCP model and encompass more spatial-temporal details. Fig. 3 for Chicago exhibit two ways to explain the results from LGCP. LGCP-NMF captures few information in different bases while our model offers a more granular explanation at the same level of time-location pairs, thus better interpreting the results learned by LGCP.

C.3 More Experiments – Chicago Dataset

Selection of Latent Classes We first elucidate the number of latent classes selection process. Selecting an excessive number of classes can decrease training efficiency and lead to similar patterns across classes, reducing interpretability. Conversely, too few classes may fail to capture all event patterns. Therefore, during training, we empirically utilize the converged negative log-likelihood, time efficiency, and number of learnable parameters as selection criteria. Based on the experimental findings shown in Tab. 2, we choose two latent classes for Chicago datasets.

In Fig. 2, we have examined different crime patterns in Chicago based on the crime types, that align with perpetrators' anticipatory decision-making patterns. More analysis can be found in the main text.

# Expert	Neg. LL ↓	Time Cost (h) ↓	# Params
$H = 1$	5.64 +/- 0.03	0.58 +/- 0.02	1.369K
$H = 2$	5.38 +/- 0.01	0.65 +/- 0.01	2.732K
$H = 3$	5.46 +/- 0.02	0.68 +/- 0.01	4.063K
$H = 4$	5.40 +/- 0.02	0.71 +/- 0.01	5.396K

Table 2: Selection of the number of latent classes for Chicago crime dataset. Current selection of modules are highlighted in blue. Performance metrics are averaged across three different runs, which reported as (Mean $+/-$ SD).

Embedding		Expert	Utility	Metric				
(w/ prod.)	(w/ feat.)	(w/ multi.)	(w/ feat.)	Neg. LL ↓	KL ↓	RMSE ↓	Time (h) ↓	# Params
✗	✗	✗	✗	5.94	0.41	0.38	0.46	0.879K
✗	✓	✓	✓	5.67	0.32	0.34	0.75	4.019K
✓	✗	✗	✓	5.63	0.35	0.35	0.58	1.703K
✓	✗	✓	✓	5.38	0.24	0.27	0.65	2.732K
✓	✓	✗	✓	5.54	0.33	0.34	0.62	2.328K
✓	✓	✓	✗	5.60	0.30	0.29	0.70	4.734K
✓	✓	✓	✓	5.42	0.26	0.24	1.06	5.636K

Table 3: Ablation study using Chicago crime dataset with 23545 samples for different modules in embedding approach, number of experts, and construction of utility function. We use converged negative log-likelihood, prediction KL, prediction RMSE, and training time cost as metrics. "(w/ prod)": Use the product of two embeddings as the overall embedding. If "(w/o prod)", we only use a single embedding as the overall embedding. "(w/ feat)": Use the individual features in the construction of embeddings or utility function. "(w/ multi)": Indicate we use multiple experts or single expert.

Efficiency, Scalability, and Ablation Study For Chicago dataset, the results depicted in Fig. 4 affirm the high efficiency and good adaptability of our model for handling large-scale datasets and outperformance compared with deep neural network models. The ablation study in Tab. 3 further demonstrates that under our current modules combination, our model strikes a balance between model performance and efficiency. More analysis can be found in the main text.

D Reproducibility Analysis

D.1 Computing Infrastructure

All the real-world data experiments, including the comparison experiments with baselines, are performed on Ubuntu 20.04.3 LTS system with Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, 227 Gigabyte memory.

D.2 Hyper-Parameter Selection

We present the selected hyper-parameters on three real-world datasets in Tab. 4. The hyper-parameter selection metric is a trade-off between training converged log-likelihood, prediction performance, and time efficiency.

E Limitation & Broader Impacts

Limitation While the current methodological framework effectively incorporates spatial-temporal dynamics and individual attributes, it may inadequately account for critical external or unobservable confounders that could systematically bias model performance, especifically degrade the interpretability advantage. In future research, we can consider a deep consideration set choice model, attempting to focus on integrating attention mechanisms into the gating function of choice model. It has the potential to enhance the model's flexibility and enables the model to capture a broader range of information through neural networks.

Broader Impacts By explicitly modeling human decision-making in spatial-temporal events (e.g., crime, bike-sharing), our model provides actionable insights for policymakers to optimize resource allocation, improve public safety, and design human-centric urban infrastructure. The integration of choice theory with interpretable neural architectures advances transparent AI systems that align with human reasoning, benefiting domains like transportation (e.g., ride-sharing demand prediction)

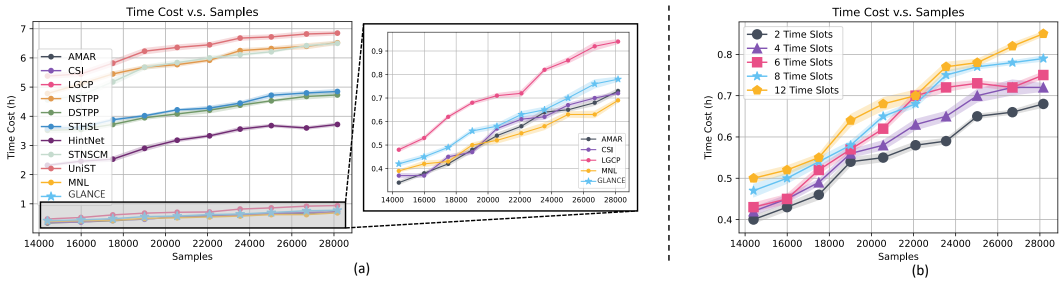

Figure 4: Scalability experiments for Chicago crime datasets with varying training samples and time slots. **(a)** Time cost v.s. training samples for all methods with fixed 4 time slots, and **(b)** Time cost v.s. training samples for our proposed method with varying time slots. All the experiments are conducted over 5 random runs and the standard error is reflected in the shaded areas.

Hyper-Parameters	Value Used		
	NYC Crime	Chicago Crime	Shanghai Mobike
Maximum Epochs	1000	1000	800
Batch Size	64	128	64
# Time Slot	4	4	6
# Area Block	77	77	100
# Latent Class	2	2	3
Embedding Dimension	32	32	32
Initial α	1.5	1.5	1.5
Learning Rate	1e-3	1e-3	5e-4
Optimizer	Adam	Adam	Adam

Table 4: Descriptions and values of hyper-parameters used for models trained on the three real-world datasets.

and public health (e.g., disease spread modeling). Moreover, the two-stage "consider-then-choose" paradigm offers a computational tool to test behavioral theories at scale, enabling new interdisciplinary collaborations between machine learning and social sciences. It is also should be noted that the theoretical guarantees (approximation/generalization) ensure robust performance across diverse populations, reducing biases in event prediction compared to traditional models.

In contrast, modeling individual choice behavior at high fidelity may inadvertently expose sensitive patterns in human mobility or preferences, requiring strict data anonymization protocols. And policymakers might prioritize model outputs over community engagement, marginalizing local knowledge in urban decision-making.