From Counts to Preferences: Preference-Driven Models for Spatio-Temporal Event Data

Chao Yang[1,*] **Yiling Kuang**[2,*] **Shuang Li**[1,†]
[1]The Chinese University of Hong Kong, Shenzhen
School of Data Science
[2]The Chinese University of Hong Kong
Department of Statistics and Data Science

Abstract

Spatio-temporal event data—such as crime incidents or shared-mobility usage—are generated by human decisions. Yet most existing models focus on statistical dependencies in time and space, overlooking the 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 the focus *from counts to preferences*, our approach offers a behaviorally grounded and socially meaningful perspective for modeling event data.

1 INTRODUCTION

Many real-world counting processes—such as crime incidents or bike-sharing usage—reflect aggregate patterns that emerge from micro-level human decisions. Traditional spatio-temporal models, including Gaussian processes in Log-Gaussian Cox models (Møller et al., 1998; Diggle et al., 2013) and triggering kernels in spatio-temporal point processes (Reinhart, 2018), have focused on capturing statistical dependencies across space and time. While powerful for correlation modeling, these approaches often overlook the cognitive mechanisms and social factors that drive such events (Zhao and Tang, 2017; He et al., 2021). For instance, a crime is not merely the outcome of spatial hotspots, but a deliberate act shaped by perceived risks, norms, and situational opportunities. To better explain event data generated by human beings, we must move beyond correlations and incorporate the human preferences that give rise to counts.

We propose a **preference-driven framework** that models spatio-temporal event data through the lens of human decision-making. Our key insight is twofold: *(i)* human decisions typically follow a "consider–then–choose" process, where individuals first narrow down options to a manageable consideration set (e.g., "Which areas are feasible for biking?") and then make refined selections within this subset; and *(ii)* population-level counts reflect heterogeneous preferences, with distinct subgroups (e.g., commuters vs. recreational cyclists) exhibiting systematically different utilities for time-location pairs.

To capture these mechanisms, we integrate discrete choice theory (Levin and Milgrom, 2004; Bentz and Merunka, 2000) with interpretable neural architectures that explicitly model the two-stage decision-making process. In the first stage (consider), a sparse gating function (Correia et al., 2019) identifies salient time-location candidates, simulating how humans filter numerous options into a context-aware consideration set. This sparsity adaptively "zooms in" on critical regions, mirroring attention mechanisms in human cognition (Peters et al., 2019). In the second stage (choose), a refined selection process further evaluates the shortlisted options using learnable utility func-

*Equal Contribution. †Corresponding Author. Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

tions, ultimately determining the final decision.

To capture heterogeneity, we employ a *mixture-of-experts* framework (Jacobs et al., 1991; Shazeer et al., 2017), where each expert represents a distinct mode of "consider–then–choose" behavior. This design not only improves predictive accuracy but also reveals interpretable preference patterns across subpopulations and contexts. The framework incorporates *sparse structural design*, and we analyze its theoretical properties by establishing approximation and generalization guarantees.

Our contributions are threefold: *(i)* We reinterpret spatio-temporal event distributions as outcomes of human decision-making, explicitly modeling the "consider–then–choose" mechanism. *(ii)* We introduce sparse gating and utility-based ranking functions that link event counts to underlying preferences, improving interpretability. *(iii)* We demonstrate empirically, on crime and bike-sharing datasets, that our framework achieves competitive predictive performance while uncovering social and cognitive drivers of human behavior.

By shifting the focus *from counts to preferences*, we aim to provide a behaviorally grounded and socially meaningful perspective for modeling spatio-temporal event data.

2 RELATED WORK

Spatio-Temporal Modeling. Modeling spatio-temporal event data has traditionally focused on statistical dependencies in space and time. Deep learning methods such as recurrent and temporal convolutional networks capture sequential dynamics (Yu et al., 2017b; Wang et al., 2017; Wu et al., 2019), while convolutional and graph neural networks model spatial relations (Li et al., 2022; Yu et al., 2017a). Although effective for prediction, these approaches typically function as black boxes, offering limited interpretability.

Spatio-temporal point processes (STPPs) (Moller and Waagepetersen, 2003; Diggle, 2006; Reinhart, 2018) take a more principled view by modeling intensity functions over continuous space and time. Parametric versions such as Log-Gaussian Cox processes (LGCP) (Møller et al., 1998; Diggle et al., 2013) emphasize interpretability, while non-parametric methods (Chen et al., 2020) provide flexibility but struggle with scalability. Neural TPPs (Zhang et al., 2020; Zuo et al., 2020) extend these methods, but still treat events primarily as correlated signals, rather than as outcomes of human decision-making. Mixture-of-Experts (MoE) designs have been applied to applications such as ride-hailing (Rahman et al., 2024), event forecasting (Liu et al., 2023), and traffic prediction (Jiang et al., 2024), capturing heterogeneity but without modeling the cognitive processes that produce the events.

Choice Models. In contrast, choice models are designed to explain how individuals select from a set of alternatives. Classical models include the multinomial logit (McFadden, 1972), Markov chain choice (Blanchet et al., 2016), non-parametric (Farias et al., 2009), and mixed logit (McFadden and Train, 2000). Neural extensions (Arkoudi et al., 2023; Ko and Li, 2023; Wang et al., 2023) learn flexible feature-to-utility mappings and achieve strong predictive accuracy. However, most choice models assume that all alternatives are evaluated simultaneously, which is implausible in spatio-temporal event data: each decision corresponds to a time–location pair, and no human realistically considers every possible combination.

Behavioral research suggests instead that humans adopt a "Consider–Then–Choose" (CTC) strategy (Hauser, 2014; Manzini and Mariotti, 2014): first filtering a manageable set of options, then making a refined choice. Existing CTC models often impose strong inductive biases, such as threshold rules (Jagabathula and Rusmevichientong, 2017; Kimya, 2018; Wang et al., 2022; Aouad et al., 2019), randomization (Gallego and Li, 2024; Akchen and Mitrofanov, 2025), or graph-based heuristics (Jagabathula and Vulcano, 2018; Jagabathula et al., 2022). These designs are useful but rely on fixed heuristics for forming consideration sets, which makes them rigid and limits their adaptability across diverse spatio-temporal contexts.

Our Position. We view spatio-temporal event data as macro-level aggregates of human *preferences*, where each count reflects many micro-level decisions. Our contribution is to bring *choice theory* into this setting, modeling counts as outcomes of an implicit "consider–then–choose" process. A data-driven sparse gating mechanism mimics attention limitations, while utility functions capture selection within consideration sets. Combined with a mixture-of-experts design, this framework explains event distributions as preference-driven rather than correlation-driven, capturing heterogeneous decision modes with interpretability and theoretical grounding.

3 BACKGROUND

We review two classical perspectives that motivate our work: *spatio-temporal models*, which capture event dynamics through statistical dependencies in space and time, and *discrete choice models*, which explain how individuals select from competing alternatives based on

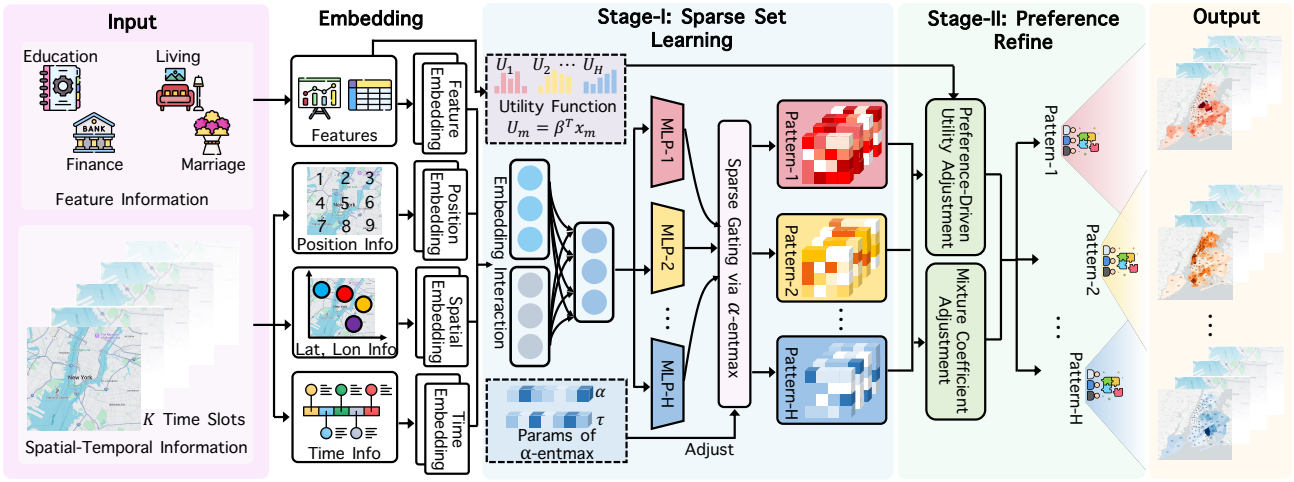

Figure 1: Model framework of **GLANCE**, from left to right: Input, Embedding Module, **Stage I**: sparse gates g_m^h filter options; **Stage II**: utilities U_m^h determine final choice probabilities, and Output.

utilities. Together, these form the foundation for our preference-driven view of spatio-temporal event data.

3.1 Traditional Spatio-Temporal Models: Event Dynamics

Spatio-temporal event data are typically represented as sequences of tuples (t_i, s_i), where t_i denotes the event time and $s_i = (x_i, y_i)$ the spatial location. A central tool is the *conditional intensity function*:

$$\lambda(t, s \mid \mathcal{H}_t) = \lim_{\Delta t, \Delta s \to 0} \frac{\mathbb{E}[N((t, t+\Delta t] \times (s, s+\Delta s]) \mid \mathcal{H}_t]}{\Delta t \cdot \Delta s},$$

which describes the instantaneous event rate at (t, s) given the history $\mathcal{H}_t$ (i.e., all events before time t).

Two canonical examples are:

- **Log-Gaussian Cox Process (LGCP):** $\lambda(t, s) = \exp(Z(t, s))$, where Z is drawn from a Gaussian process with mean $\mu(t, s)$ and covariance $k((t, s), (t', s'))$ (Diggle et al., 2013).
- **Self-exciting Point Process (Hawkes type):**

$$\lambda(t, s \mid \mathcal{H}_t) = \lambda_0(s) + \sum_{t_j < t} g(t - t_j, s - s_j),$$

  where $\lambda_0(s)$ is a baseline intensity and g a kernel capturing past-event influence.

These models are well-suited for natural phenomena (e.g., earthquakes, rainfall, neuronal spikes), but for *human-generated events* they fall short: treating events as stochastic signals misses the decision-making processes that drive them.

3.2 Discrete Choice Models: Modeling Human Decisions

Discrete choice models provide a framework for explaining how individuals select among alternatives. Let $\mathcal{U}$ denote the universe of all items (e.g., products, routes, or time–location options). At each decision point, an individual faces a subset $\mathcal{C} \subseteq \mathcal{U}$, called the *choice set*.

Each alternative $m \in \mathcal{C}$ is assigned a latent *utility* U_m. In the classical Plackett–Luce model (Luce, 1959; Plackett, 1975), the choice probability is

$$\mathbb{P}(m \mid \mathcal{C}) = \frac{\exp(U_m)}{\sum_{m' \in \mathcal{C}} \exp(U_{m'})}.$$

This highlights two ingredients: *(i)* the *choice set* $\mathcal{C}$ defines feasible options, and *(ii)* the *utility function* U, depending on item and context features, captures behavioral complexity.

Traditional choice models thus offer a utility-based view of decision-making but assume the *choice set is known and fixed*. In spatio-temporal event data, each decision corresponds to a time–location choice, making the set vast and implicit. To address this, we introduce sparse structural mechanisms that mimic human attention limitations, inspired by the "consider–then–choose" process.

4 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 (Kimya, 2018), 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.

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

4.1 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 (Correia et al., 2019), 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.

4.2 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 $g^h = \alpha^h\text{-entmax}(\boldsymbol{z}^h)$ 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.

4.3 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.

5 THEORETICAL ANALYSIS

Our framework interprets spatio-temporal events as the aggregate outcome of human choices. Beyond empirical performance, it is important to understand its *theoretical properties*. We focus on two fundamental questions:

1. *Approximation:* Can a finite mixture of latent "consider–then–choose" models approximate arbitrary preference distributions?
2. *Generalization:* With finitely many observed events, how close is the fitted model to the population truth?

5.1 Approximation Error

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 (Barron, 2002). The proof (Appendix B.2) adapts covering arguments to our mixture-of-sparse-utilities setting.

5.2 Generalization Error

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 $\|W_q^h (W_k^h)^\top\|_F \le C_W$; (iii) utilities are bounded $\|\beta_h\|_2 \le C_U$; (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,

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

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. Details are provided in Appendix B.3.

Summary. Together, Theorems 1 and 2 show that GLANCE is both *expressive*, approximating complex preference distributions, and *statistically reliable*, generalizing stably from finite samples.

6 EXPERIMENTS

In experiments, we evaluate GLANCE as a preference-driven model for spatio-temporal events. Our focus is on how well the model captures *event distributions*—that is, the probability of where and when events occur—rather than just correlations. Specifically, we ask:

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).

Model	NYC Crime		Chicago Crime		Shanghai Mobike	
	KL ↓	RMSE ↓	KL ↓	RMSE ↓	KL ↓	RMSE ↓
ARMA	$0.65_{\pm 0.06}$	$0.62_{\pm 0.08}$	$0.70_{\pm 0.10}$	$0.68_{\pm 0.06}$	$0.46_{\pm 0.08}$	$0.42_{\pm 0.04}$
CSI	$0.67_{\pm 0.08}$	$0.66_{\pm 0.03}$	$0.68_{\pm 0.12}$	$0.65_{\pm 0.09}$	$0.47_{\pm 0.04}$	$0.43_{\pm 0.05}$
LGCP	$0.67_{\pm 0.11}$	$0.67_{\pm 0.09}$	$0.69_{\pm 0.09}$	$0.68_{\pm 0.08}$	$0.45_{\pm 0.10}$	$0.43_{\pm 0.09}$
NSTPP	$0.51_{\pm 0.06}$	$0.49_{\pm 0.05}$	$0.42_{\pm 0.07}$	$0.44_{\pm 0.10}$	$0.32_{\pm 0.02}$	$0.33_{\pm 0.05}$
DSTPP	$0.47_{\pm 0.04}$	$0.45_{\pm 0.05}$	$0.47_{\pm 0.04}$	$0.46_{\pm 0.08}$	$0.37_{\pm 0.00}$	$0.40_{\pm 0.02}$
ST-HSL	$0.56_{\pm 0.06}$	$0.52_{\pm 0.05}$	$0.49_{\pm 0.04}$	$0.52_{\pm 0.06}$	$0.38_{\pm 0.05}$	$0.43_{\pm 0.03}$
HintNet	$0.38_{\pm 0.03}$	$0.37_{\pm 0.03}$	$0.26_{\pm 0.04}$	$0.28_{\pm 0.03}$	$0.19_{\pm 0.01}$	$0.17_{\pm 0.02}$
STNSCM	$0.38_{\pm 0.02}$	$0.38_{\pm 0.04}$	$0.27_{\pm 0.01}$	$0.31_{\pm 0.02}$	$0.11_{\pm 0.00}$	$0.15_{\pm 0.01}$
UniST	$0.37_{\pm 0.03}$	$0.36_{\pm 0.02}$	$0.27_{\pm 0.05}$	$0.30_{\pm 0.04}$	$0.23_{\pm 0.04}$	$0.25_{\pm 0.06}$
MNL	$0.38_{\pm 0.01}$	$0.38_{\pm 0.01}$	$0.25_{\pm 0.03}$	$0.29_{\pm 0.02}$	$0.13_{\pm 0.01}$	$0.17_{\pm 0.01}$
GLANCE	$0.33_{\pm 0.02}$	$0.34_{\pm 0.01}$	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.12_{\pm 0.01}$	$0.16_{\pm 0.01}$

Table 2: Ablation study using Chicago crime dataset 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. Our current choice is shaded in light purple.

Embedding		Expert	Utility	Metric				
(w/ prod.)	(w/ feat.)	(w/ multi.)	(w/ feat.)	Neg. LL ↓	KL ↓	RMSE ↓	Time (h) ↓	# Params
✗	✗	✗	✗	$5.94_{\pm 0.03}$	$0.41_{\pm 0.04}$	$0.38_{\pm 0.04}$	$0.46_{\pm 0.00}$	0.879K
✗	✓	✓	✓	$5.67_{\pm 0.05}$	$0.32_{\pm 0.01}$	$0.34_{\pm 0.03}$	$0.75_{\pm 0.02}$	4.019K
✓	✗	✗	✓	$5.63_{\pm 0.05}$	$0.35_{\pm 0.04}$	$0.35_{\pm 0.06}$	$0.58_{\pm 0.02}$	1.703K
✓	✗	✓	✓	$5.31_{\pm 0.01}$	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.65_{\pm 0.01}$	2.732K
✓	✓	✗	✓	$5.54_{\pm 0.08}$	$0.33_{\pm 0.02}$	$0.35_{\pm 0.03}$	$0.62_{\pm 0.02}$	2.328K
✓	✓	✓	✗	$5.60_{\pm 0.06}$	$0.30_{\pm 0.00}$	$0.29_{\pm 0.02}$	$0.70_{\pm 0.02}$	4.734K
✓	✓	✓	✓	$5.43_{\pm 0.02}$	$0.26_{\pm 0.01}$	$0.27_{\pm 0.03}$	$1.06_{\pm 0.08}$	5.636K

- **Q1 (Event Distribution Accuracy):** How well does GLANCE capture spatio-temporal event distributions compared with classical models?
- **Q2 (Interpretability):** Does GLANCE uncover meaningful latent preference patterns that shed light on human decision-making processes?
- **Q3 (Efficiency):** How well does GLANCE scale with data size and model complexity, and what is the contribution of each architectural component?
- **Q4 (Connections to Classical Models):** Can GLANCE recover or reinterpret traditional spatio-temporal models (e.g., Hawkes processes, LGCP) as special cases, thereby providing a unifying perspective?

We primarily study the Chicago crime dataset as the main benchmark, with additional validation on other datasets in Appendix C.

6.1 Experimental Setup

Datasets. We evaluate GLANCE on three large-scale spatio-temporal datasets that capture distinct aspects of human behavior: *(i) New York Crime*[1]: Records of felony, misdemeanor, and violation incidents reported to the NYPD. *(ii) Chicago Crime*[2]: Comprehensive incident reports from the Chicago Police Department,

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

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

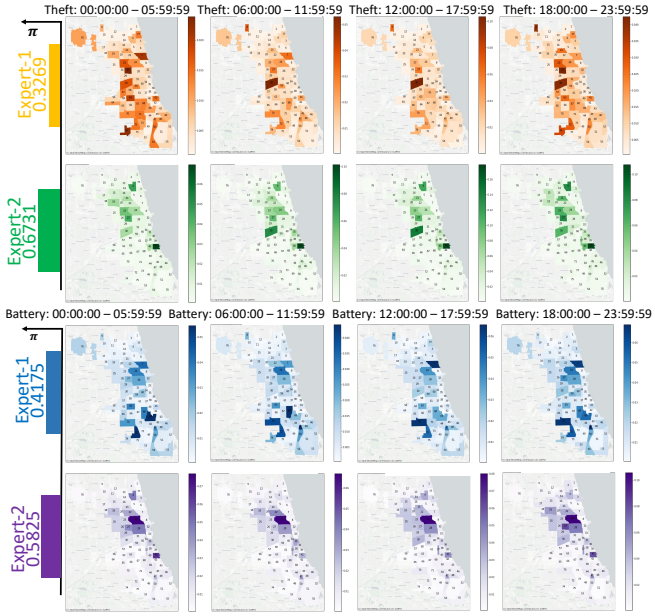

Figure 2: Mixing coefficient π^h (**Left bar plots**) and mixture pattern adjusted by utility score ($g^h \exp(U^h)$) for different latent class-h and different crime types, including theft and battery (**Right heatmaps**) from July 1 to July 31, 2024, in Chicago City. The selection of the number of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

used as our main benchmark. *(iii) Shanghai Mobike*[3]: Bicycle-sharing rental data in Shanghai. Since mobility behavior varies sharply across time, we divide the day into distinct slots and further separate workday morning and evening rush hours to avoid overlap with other periods.

Baselines. We compare GLANCE with a wide range of state-of-the-art approaches, spanning classical time-series models, spatio-temporal point processes, and recent deep learning methods: *ARMA* (Araghinejad, 2013) and *CSI* (De Boor and De Boor, 1978) capture purely temporal dependencies; *LGCP* (Diggle et al., 2013; Miller et al., 2014) and *NSTPP* (Chen et al., 2020) model spatial-temporal intensities via probabilistic processes; *DSTPP* (Yuan et al., 2023), *ST-HSL* (Li et al., 2022), *HintNet* (An et al., 2022), *STNSCM* (Deng et al., 2023), and *UniST* (Yuan et al., 2024) represent advanced neural spatio-temporal architectures. Finally, we include *MNL (Multinomial Logit)* (McFadden, 1972; Hu et al., 2022) as the canonical discrete choice model, which serves as a natural behavioral baseline for our preference-driven approach. (Details of each baseline are provided in Appendix C.1.)

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

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:

$$\mathrm{KL}_i = D_{\mathrm{KL}}\Big(\hat{\boldsymbol{P}}_i \,\Big\|\, \boldsymbol{P}_i\Big) = \sum_{m=1}^{M} \hat{P}_{im} \log \frac{\hat{P}_{im}}{P_{im}}.$$

(ii) RMSE, which captures numerical prediction error across options (Zhao and Tang, 2017):

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \Big(\hat{P}_{im} - P_{im}\Big)^2}.$$

We report mean ± standard deviation of each metric across all evaluation units and random seeds.

6.2 Results and Analysis

A1: Prediction Accuracy (Q1). GLANCE achieves highly accurate probability estimates that mirror observed spatio-temporal event distributions. For instance, in Chicago, Community-41 consistently registers the highest crime rates, while Communities-42 and -8 remain lowest; Community-30 spikes midday (12:00–18:00) and Community-27 rises strongly at night (18:00–24:00) (Fig. 11, Appendix C.3.2). These fine-grained temporal patterns are faithfully captured by GLANCE, not just at the aggregate level but across specific neighborhoods and hours. In one-day-ahead forecasts, GLANCE surpasses nearly all baselines across datasets (Tab. 1), and maintains stable accuracy in 1–3 day horizons (Tabs. 8 and Tab. 9, Appendix C.5). This shows that GLANCE is not only competitive as a short-term predictor but also robust in capturing structural regularities that generalize to longer horizons—an essential property for real-world deployment.

A2: Interpreting Human Decision Processes (Q2). 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. 3), striking a balance between parsimony and expressiveness. In our current setting, the gating network uses

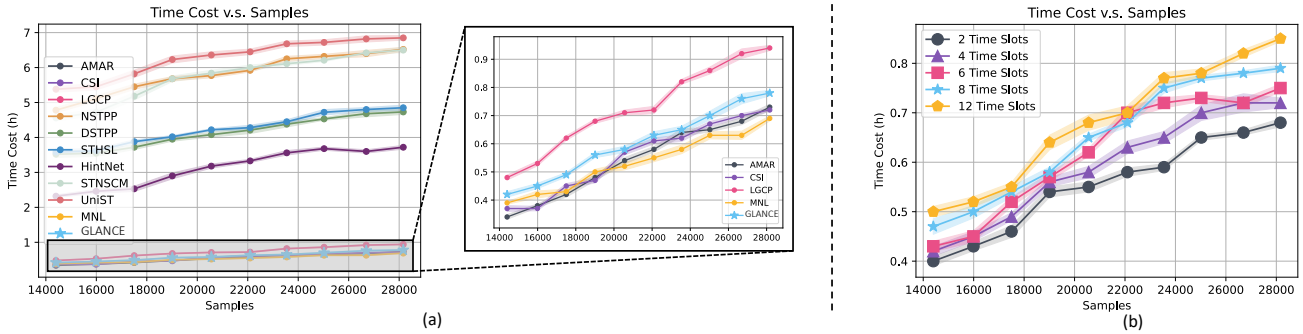

Figure 3: Scalability experiments for Chicago crime datasets with varying training samples and time slots. **Left**: Time cost v.s. training samples for all methods with fixed 4 time slots, and zoom in for methods with low time cost. **Right**: Time cost v.s. training samples for our proposed method with varying time slots. All the experiments are conducted over three random runs and the standard error is reflected in the shaded areas.

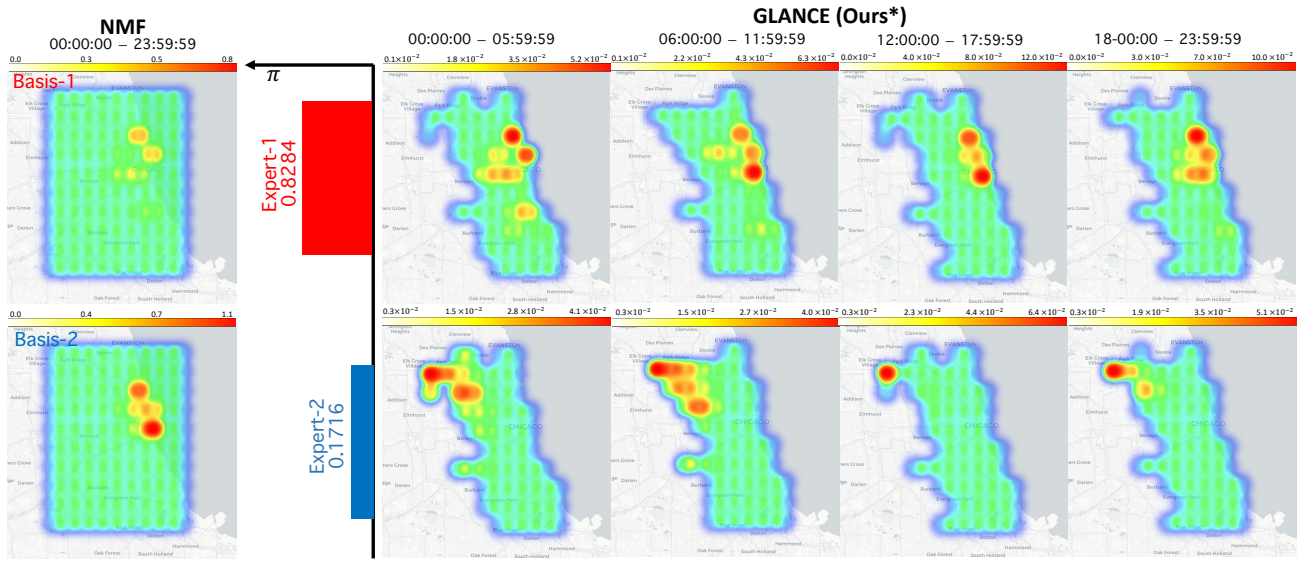

Figure 4: Comparison of the learned expert pattern of our choice model and the basis of non-negative matrix factorization (NMF) on Chicago City crime dataset. To align with the setting of LGCP-NMF, we partition the Chicago area into 10×10 area blocks. **Left**: NMF basis, **Right**: expert patterns learned by our model.

Table 3: Selection of the number of latent classes for Chicago crime dataset. Light purple shadow highlights the current selection of modules. Performance metrics are averaged across three different runs.

# Expert	Metrics		
	Neg. LL ↓	Time Cost (h) ↓	# Params
$H = 1$	$5.64_{\pm 0.03}$	$0.58_{\pm 0.02}$	1.369K
$H = 2$	$5.31_{\pm 0.01}$	$0.65_{\pm 0.01}$	2.732K
$H = 3$	$5.46_{\pm 0.02}$	$0.68_{\pm 0.01}$	4.063K
$H = 4$	$5.40_{\pm 0.02}$	$0.71_{\pm 0.01}$	5.396K

only spatio-temporal features, while crime type enters only through the utilities. This means that theft and battery share the same consideration sets (similar opportunity structures in space–time), but each type has its own utility over those options. The fact that experts mix theft and battery reflects shared high-opportunity regions where both offenses are attractive, with type-specific intensities handled by the utilities. An alternative extension incorporates crime-type embeddings into the gating input, resulting in type-specialized while spatially interpretable experts; we do not explore this variant here.

As shown in Fig. 2, these latent classes reveal distinct crime decision patterns:

- **Theft.** *Class 1 (33%)*: Smaller subgroup operat-

ing 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).

A3: Efficiency, Scalability, and Ablation (Q3). GLANCE balances interpretability with efficiency. As resolution varies (2–12 time slots/day) and training size scales up to 28,164 samples, training times grow smoothly and remain practical (Fig. 3). Compared to deep neural architectures (UniST, NSTPP, et al), GLANCE is faster, while still more expressive than lightweight models such as MNL. Even at the largest scale and largest number of time slots (largest choice set), convergence of our model occurs within one hour—demonstrating scalability to city-scale data.

The ablation study (Tab. 2) confirms that each design choice matters. Using paired embeddings, multiple experts, and individual features yields the best accuracy–efficiency trade-off. Stripping away the mixture structure (single class) degrades performance sharply (Neg. LL, KL, RMSE all rise), underscoring that real-world event distributions require modeling multiple behavioral archetypes. Results in Tab. 6 and Tab. 7 in Appendix C.4 further demonstrate the necessity of the gating function and disentangle the contribution of MoE from the consider–then–choose (CTC) structure, which explicitly models sparse consideration sets followed by expert-specific preferences. Overall, these ablation studies validate our central claim: *heterogeneity is not a nuisance but a structural feature of human spatio-temporal decisions.*

A4: Explaining Other Spatio-Temporal Models (Q4). GLANCE also functions as an interpretive tool for classical models (e.g., Log-Gaussian Cox Process (LGCP) (Diggle et al., 2013)). By aligning with a fitted LGCP through squared-error minimization of probability estimates, GLANCE extracts latent class patterns that explain the LGCP intensity surface. Unlike LGCP-NMF (Miller et al., 2014), which summarizes intensities into broad basis functions, GLANCE reveals fine-grained behavioral clusters that correspond directly to observed crime hotspots. As shown in Fig. 4, while LGCP-NMF identifies broad citywide bases, GLANCE pinpoints concentrated hotspots in the North, West, and Far North Sides—patterns that closely match empirical frequencies. This demonstrates a deeper insight: *classical intensity-based models capture where events are likely to occur, but GLANCE explains why different subgroups generate them.*

Takeaway. Across analyses A1–A4, GLANCE emerges not only as a strong predictor but also as a discovery tool: it disentangles heterogeneity, scales gracefully, and reframes classical spatio-temporal models through the lens of human preference.

6.3 Validation of Experimental Results

Our results are consistent with well-documented socioeconomic patterns in crime research. High-crime regions identified by GLANCE overlap with lower-income neighborhoods, reinforcing inverse income–crime relationship observed by Yildiz et al. (2013). In New York, hotspots in Midtown and Lower Manhattan align with Catlett et al. (2019), while our Chicago findings echo the patterns reported by Linderman and Adams (2014).

Beyond confirming these established correlations, GLANCE contributes a novel perspective: by uncovering distinct latent classes, it highlights the *heterogeneous decision patterns* of offenders—an aspect often overlooked in prior spatio-temporal analyses. This mixture-based view moves beyond aggregate hotspot detection toward a more behavioral understanding of criminal activity.

Finally, our use of demographic attributes follows standard practice in spatio-temporal modeling (Lau, 2021; Niu et al., 2019; Kang and Kang, 2017). All features were processed in line with benchmarks, with no discriminatory intent or implication.

7 CONCLUSION

We introduced GLANCE, a preference-driven framework for spatio-temporal events that interprets aggregate counts as the macro-level outcome of many micro-level *consider–then–choose* decisions. By combining sparse attention, flexible utilities, and a mixture-of-experts, the model captures cognitive limits and heterogeneous preferences in an interpretable way.

Acknowledgements

This work was supported in part by the Key Program of the National Natural Science Foundation of China (NSFC) under Grant No. 72495131; the Shenzhen Stability Science Program 2023; the Shenzhen Science and Technology Program No. JCYJ20250604141038013; and the Longgang District Key Laboratory of Intelligent Digital Economy Security.

References

Akchen, Y.-C. and Mitrofanov, D. (2025). Consider or choose? the role and power of consideration sets. *Management Science*.

An, B., Vahedian, A., Zhou, X., Street, W. N., and Li, Y. (2022). 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.

Aouad, A., Feldman, J., Segev, D., and Zhang, D. (2019). The click-based mnl model: A novel framework for modeling click data in assortment optimization. *Available at SSRN 3340620*.

Araghinejad, S. (2013). Time series modeling. *Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering*, pages 85–137.

Arkoudi, I., Krueger, R., Azevedo, C. L., and Pereira, F. C. (2023). Combining discrete choice models and neural networks through embeddings: Formulation, interpretability and performance. *Transportation research part B: methodological*, 175:102783.

Barron, A. R. (2002). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945.

Bentz, Y. and Merunka, D. (2000). Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting*, 19(3):177–200.

Blanchet, J., Gallego, G., and Goyal, V. (2016). A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905.

Catlett, C., Cesario, E., Talia, D., and Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing*, 53:62–74.

Chen, R. T., Amos, B., and Nickel, M. (2020). Neural spatio-temporal point processes. *arXiv preprint arXiv:2011.04583*.

Correia, G. M., Niculae, V., and Martins, A. F. (2019). Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184.

De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer New York.

Deng, P., Zhao, Y., Liu, J., Jia, X., and Wang, M. (2023). Spatio-temporal neural structural causal models for bike flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4242–4249.

Diggle, P. J. (2006). Spatio-temporal point processes: methods and applications. *Monographs on Statistics and Applied Probability*, 107:1.

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. (2013). Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Farias, V., Jagabathula, S., and Shah, D. (2009). A data-driven approach to modeling choice. *Advances in Neural Information Processing Systems*, 22.

Gallego, G. and Li, A. (2024). A random consideration set model for demand estimation, assortment optimization, and pricing. *Operations Research*.

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, 67(8):1688–1699.

He, P., Zheng, F., Belavina, E., and Girotra, K. (2021). Customer preference and station network in the london bike-share system. *Management Science*, 67(3):1392–1412.

Hu, Y., Simchi-Levi, D., and Yan, Z. (2022). Learning mixed multinomial logits with provable guarantees. *Advances in Neural Information Processing Systems*, 35:9447–9459.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jagabathula, S., Mitrofanov, D., and Vulcano, G. (2022). Personalized retail promotions through a directed acyclic graph–based representation of customer preferences. *Operations Research*, 70(2):641–665.

Jagabathula, S. and Rusmevichientong, P. (2017). A nonparametric joint assortment and price choice model. *Management Science*, 63(9):3128–3145.

Jagabathula, S. and Vulcano, G. (2018). A partial-order-based model to estimate individual preferences using panel data. *Management Science*, 64(4):1609–1628.

Jiang, W., Han, J., Liu, H., Tao, T., Tan, N., and Xiong, H. (2024). Interpretable cascading mixture-of-experts for urban traffic congestion prediction. In

Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 5206–5217.

Kang, H.-W. and Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PloS one*, 12(4):e0176244.

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

Ko, J. and Li, A. A. (2023). Modeling choice via self-attention. *arXiv preprint arXiv:2311.07607*.

Lau, H. (2021). Discovering spatio-temporal pattern of city crime–visual analysis on felony crime in new york. *N/A*.

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

Levin, J. and Milgrom, P. (2004). Introduction to choice theory.

Li, Z., Huang, C., Xia, L., Xu, Y., and Pei, J. (2022). Spatial-temporal hypergraph self-supervised learning for crime prediction. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 2984–2996. IEEE.

Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. In *International conference on machine learning*, pages 1413–1421. PMLR.

Liu, H., Zhang, Y., Wang, X., Wang, B., and Yu, Y. (2023). St-moe: Spatio-temporal mixture of experts for multivariate time series forecasting. In *2023 18th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 562–567. IEEE.

Luce, R. D. (1959). *Individual choice behavior*, volume 4. Wiley New York.

Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.

Maurer, A. (2016). 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.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior. *N/A*.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470.

Miller, A., Bornn, L., Adams, R., and Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. In *International conference on machine learning*, pages 235–243. PMLR.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. Chapman and Hall/CRC.

Niu, X., Elsisy, A., Derzsy, N., and Szymanski, B. K. (2019). Dynamics of crime activities in the network of city community areas. *Applied Network Science*, 4(1):127.

Peters, B., Niculae, V., and Martins, A. F. (2019). Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Rahman, M. H., Rifaat, S. M., Sadeek, S. N., Abrar, M., and Wang, D. (2024). Gated ensemble of spatio-temporal mixture of experts for multi-task learning in ride-hailing system. *Multimodal Transportation*, 3(4):100166.

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

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Wang, H., Li, X., and Talluri, K. (2023). Transformer choice net: A transformer neural network for choice prediction. *arXiv preprint arXiv:2310.08716*.

Wang, R., Zhao, Z., and Ke, C. (2022). Modeling consumer choice and optimizing assortment under the threshold multinomial logit model. *Available at SSRN 4184044*.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. S. (2017). Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C. (2019). Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*.

Yildiz, R., Ocal, O., and Yildirim, E. (2013). The effects of unemployment, income and education on crime: Evidence from individual data. *Journal of Economic & Management Perspectives*, 7(2):32.

Yu, B., Yin, H., and Zhu, Z. (2017a). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Yu, R., Li, Y., Shahabi, C., Demiryurek, U., and Liu, Y. (2017b). Deep learning: A generic approach for extreme condition traffic forecasting. In *Proceedings of the 2017 SIAM international Conference on Data Mining*, pages 777–785. SIAM.

Yuan, Y., Ding, J., Feng, J., Jin, D., and Li, Y. (2024). 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.

Yuan, Y., Ding, J., Shao, C., Jin, D., and Li, Y. (2023). Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3173–3184.

Zhang, Q., Lipani, A., Kirnap, O., and Yilmaz, E. (2020). Self-attentive hawkes process. In *International conference on machine learning*, pages 11183–11193. PMLR.

Zhao, X. and Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 497–506.

Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. (2020). Transformer hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes] We present our proposed model, along with the mathematical setting and assumptions, in Sec. 4. The theoretical analysis and accompanying algorithm are detailed in Section 5 and Appendix A, respectively.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes] The analysis of the scalability and time efficiency is provided in Appendix. C.3.1 and Appendix. C.3.2, paragraph "Efficiency, Scalability, and Ablation Study"
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [No] We commit to releasing the source code upon acceptance of this paper.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes] The statements of the full set of assumptions of all theoretical results are provided in Sec. 5 and Appendix. B.
   (b) Complete proofs of all theoretical results.
   [Yes] We provide the corresponding proofs in Appendix. B.
   (c) Clear explanations of any assumptions.
   [Yes] The explanations of any assumptions is provided in Sec. 5 and Appendix. B.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes] We provide the dataset details and reproducilibity analysis in Sec. 6 of the main text and Appendix. D respectively.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes] The training details are presented in Appendix. C and Appendix. D.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Yes] The computation of the evaluation merics is presented in Sec. 6.1.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   [Yes] Details of the computing infrastructure used is provided in Appendix. D.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   [Not Applicable]

(b) The license information of the assets, if applicable.
[Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable.
[Not Applicable]

(d) Information about consent from data providers/curators.
[Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes] In Appendix. E, we analyze the information of offensive content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
[Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
[Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
[Not Applicable]

From Counts to Preferences: Preference-Driven Models for Spatio-Temporal Event Data — Supplementary Materials

Appendix Overview

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

- Section. A presents the implementation details and pseudocode for our choice-driven spatio-temporal counting process model.
- Section. B establishes the theoretical guarantees for our method, including bounds on both approximation and generalization error, with complete derivations provided.
- Section. C contains additional experimental results and implementation details on the real-world datasets.
- Section. D details the computing infrastructure and hyperparameter selection to ensure reproducibility.
- Section. E validates our findings by comparing them with existing studies, providing strong corroborating evidence for our methodology.
- Section. F discusses the limitations of our work, suggests future research directions, and considers its broader impacts.

A Algorithm Details

A.1 Spatio-Temporal Embedding

To represent alternatives, we construct embeddings for all time–location pairs (t_m, s_m) and relevant covariates. The embedding matrix $X \in \mathbb{R}^{M \times d}$ is built as follows:

- **Spatial features:** The region is divided into disjoint spatial blocks. For each block, we apply linear embeddings of geographic coordinates (latitude/longitude) and add sinusoidal positional encodings to capture relative block positions.
- **Temporal features:** Each time interval t_m is encoded with sinusoidal positional embeddings, following Zuo et al. (2020), to preserve periodicity (e.g., daily or weekly cycles).
- **Contextual features:** Static attributes (e.g., land-use type) are one-hot encoded, while dynamic attributes (e.g., weather, traffic) are linearly embedded.

The embeddings are summed element-wise to form a unified representation for each (t_m, s_m). User- or event-level covariates (e.g., demographics, trip purpose) can be concatenated to these embeddings, so that both the gating and utility functions adapt to context. The resulting matrix X is then projected by class-specific matrices (W_q^h, W_k^h) during model computation.

This embedding design is analogous to positional embeddings in attention models: base spatial, temporal, and contextual encodings are projected into a shared latent space where interactions and preferences are learned.

A.2 Overall Algorithm

The overall training procedure is summarized in Alg. 1.

Algorithm 1 Learning Parameters of the GLANCE Model

Input: Observed events $\{y_{im}\}_{i=1}^{N}$; initial parameters

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

Output: Optimized parameters θ^*
Initialize θ randomly or using heuristics.
repeat
  for each latent class $h \in [H]$ **do**
    Compute interaction scores:

$$E^h = XW_q^h(XW_k^h)^\top$$

    Aggregate into attention logits:

$$\boldsymbol{z}^h = \sigma(E^h)\mathbf{1}$$

    Apply sparse gating:

$$g^h = \alpha^h\text{-entmax}(\boldsymbol{z}^h)$$

  end for
  for each event i and option $m \in [M]$ **do**
    Compute mixture probability:

$$P_{im} = \sum_{h=1}^{H} \pi^h \frac{g_m^h \exp(U_m^h)}{\sum_{m'=1}^{M} g_{m'}^h \exp(U_{m'}^h)}$$

  end for
  Update parameters: Maximize the log-likelihood

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

  using stochastic gradient descent.
until convergence

B Proofs for Theoretical Results

B.1 Preliminaries and Assumptions

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 $g = \text{Entmax}_\alpha(z)$ 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

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

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. Peters et al. (2019); Correia et al. (2019)). Formally, there exists $L_{\mathrm{ent}}(\delta) > 0$ such that

$$\big\|\mathrm{Entmax}_\alpha(z) - \mathrm{Entmax}_\alpha(z')\big\|_2 \;\le\; L_{\mathrm{ent}}(\delta)\,\|z - z'\|_2\,, \qquad \forall z, z' \in \mathbb{R}^M. \tag{2}$$

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

$$\sup_{\{\pi^h, f^h\}} \frac{1}{N}\sum_{i=1}^N \epsilon_i \sum_{h=1}^H \pi^h f^h(x_i) \;\le\; \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \epsilon_i f(x_i), \tag{3}$$

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.2 Proof of Theorem 1 (Universal Approximation)

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

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

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

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

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

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

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

Choosing $H \ge 2(c_1 + c_2 \log(1/\eta))/\epsilon$ yields the stated $\mathcal{O}(1/\epsilon)$ rate. Absorbing constants gives the main-text statement "$H \le 2/\epsilon$" 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.3 Proof of Theorem 2 (Generalization Bound)

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 $P_i(\boldsymbol{\theta}) = (P_{i1}, \dots, P_{iM}) \in \Delta_M$:

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

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 (e.g., Mohri et al. (2018)) and the vector contraction inequality (Maurer, 2016), 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

$$\begin{aligned}
\mathfrak{R}_N(\mathcal{F}) &= \mathbb{E}_\epsilon\left[\sup_{\boldsymbol{\theta}\in\mathcal{W}} \frac{1}{N} \sum_{i=1}^{N}\sum_{m=1}^{M} \epsilon_{im} \sum_{h=1}^{H} \pi^h f_m\big(\boldsymbol{z}_i^h, U^h\big) \right] \\
&\le \mathbb{E}_\epsilon\left[\sup_{\{(\pi^h,\theta^h)\}} \sum_{h=1}^{H} \pi^h \cdot \frac{1}{N} \sum_{i,m} \epsilon_{im}\, f_m\big(\boldsymbol{z}_i^h, U^h\big) \right] \\
&\le \mathbb{E}_\epsilon\left[\sup_{\theta} \; \frac{1}{N} \sum_{i,m} \epsilon_{im}\, f_m\big(\boldsymbol{z}_i(\theta), U(\theta)\big) \right] \;=\; \mathfrak{R}_N(\mathcal{F}_1),
\end{aligned} \tag{4}$$

where $\mathcal{F}_1$ is the single-class function class and we used the convexity bound (3). 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

$$L_f \;\le\; c_0\, e^{C_U}, \tag{5}$$

for an absolute constant $c_0 > 0$.[4] By the chain rule and (2), we obtain

$$\|f(\boldsymbol{z}, U) - f(\boldsymbol{z}', U')\|_2 \;\le\; L_f\Big(L_{\text{ent}}(\delta)\,\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \;+\; \|U - U'\|_2\Big) \;\le\; c_1\,\frac{e^{C_U}}{\delta}\Big(\|\boldsymbol{z} - \boldsymbol{z}'\|_2 + \|U - U'\|_2\Big), \tag{6}$$

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

$$\|\boldsymbol{z}\|_2 \;=\; \|\sigma(E)\mathbf{1}\|_2 \;\le\; \|\sigma(E)\|_F\,\|\mathbf{1}\|_2 \;\le\; \|E\|_F\,\sqrt{M} \;\le\; C_X^2\, C_W\,\sqrt{M}. \tag{7}$$

Similarly, for two parameter settings,

$$\begin{aligned}
\|\boldsymbol{z} - \boldsymbol{z}'\|_2 = \|\sigma(E)\mathbf{1} - \sigma(E')\mathbf{1}\|_2 \;&\le\; \|\sigma(E) - \sigma(E')\|_F\,\|\mathbf{1}\|_2 \;\le\; \|E - E'\|_F\,\sqrt{M} \\
&\le\; \sqrt{M}\,\Big(\|X\|_F^2\big\|W_q - W_q'\big\|_F\,\|W_k\|_F + \|X\|_F^2\|W_q'\|_F\,\big\|W_k - W_k'\big\|_F\Big) \\
&\le\; c_2\, C_X^2\,\sqrt{M}\,\big\|W_q(W_k)^\top - W_q'(W_k')^\top\big\|_F \;\le\; c_2\, C_X^2\,\sqrt{M}\,\cdot 2C_W,
\end{aligned} \tag{8}$$

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 $\mathcal{O}(C_X^2\sqrt{M})$.

Step 5: Putting it together for $\mathcal{F}_1$. Using (6) and (8), 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

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

Absorbing additive constants and C_X into $\tilde{\mathcal{O}}(\cdot)$ and recalling (4), we obtain the advertised form

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

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

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

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

□

[4] 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.

B.4 Remarks on Constants and Independence of Latent Classes

Independence of H. The mixture-to-single-class reduction (4) 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 (7)–(8) (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 $L_{\text{ent}}(\delta) = \mathcal{O}(1/\delta)$; 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* (Araghinejad, 2013): 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* (De Boor and De Boor, 1978): 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* (Diggle et al., 2013; Miller et al., 2014): 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* (Chen et al., 2020): 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* (Yuan et al., 2023): It leverages diffusion models to learn complex spatial-temporal joint distributions. *vi) ST-HSL* (Li et al., 2022): It proposes a Spatial-Temporal Self-Supervised Hypergraph Learning framework for crime prediction. *vii) HintNet* (An et al., 2022): 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* (Deng et al., 2023): A causality-based interpretation model for the bike flow prediction. *ix) UniST* (Yuan et al., 2024): A universal model designed for general urban spatial-temporal prediction across a wide range of scenarios. *x) MNL (Multinomial Logic Choice Model)* (Hu et al., 2022): 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)

Consider a Log-Gaussian Cox Process (LGCP) (Diggle et al., 2013), which is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponentiated Gaussian Process

$$\begin{aligned} &Z(\cdot) \sim GP(0, k(\cdot,\cdot)), \\ &\lambda(\cdot) \sim \exp(Z(\cdot)), \quad x_1, ..., x_N \sim PP(\lambda(\cdot)) \end{aligned}$$

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 (Miller et al., 2014) was proposed to use non-negative matrix decomposition (Lee and Seung, 1999) 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.

$$\Lambda \approx WB$$

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 the LGCP model being well-trained, we fit our model using a new objective function based on

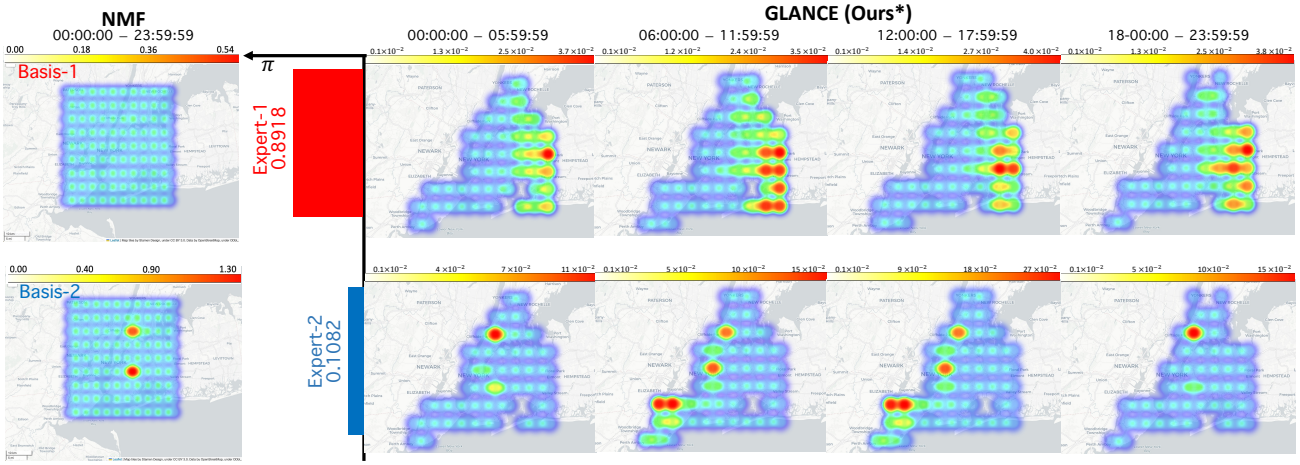

Figure 5: Comparison of the learned expert pattern of our choice model and the basis of non-negative matrix factorization (NMF) on New York City crime dataset. To align with the setting of LGCP-NMF, we partition the NYC area into 10×10 area blocks. **Left**: NMF basis, **Right**: expert patterns learned by our model.

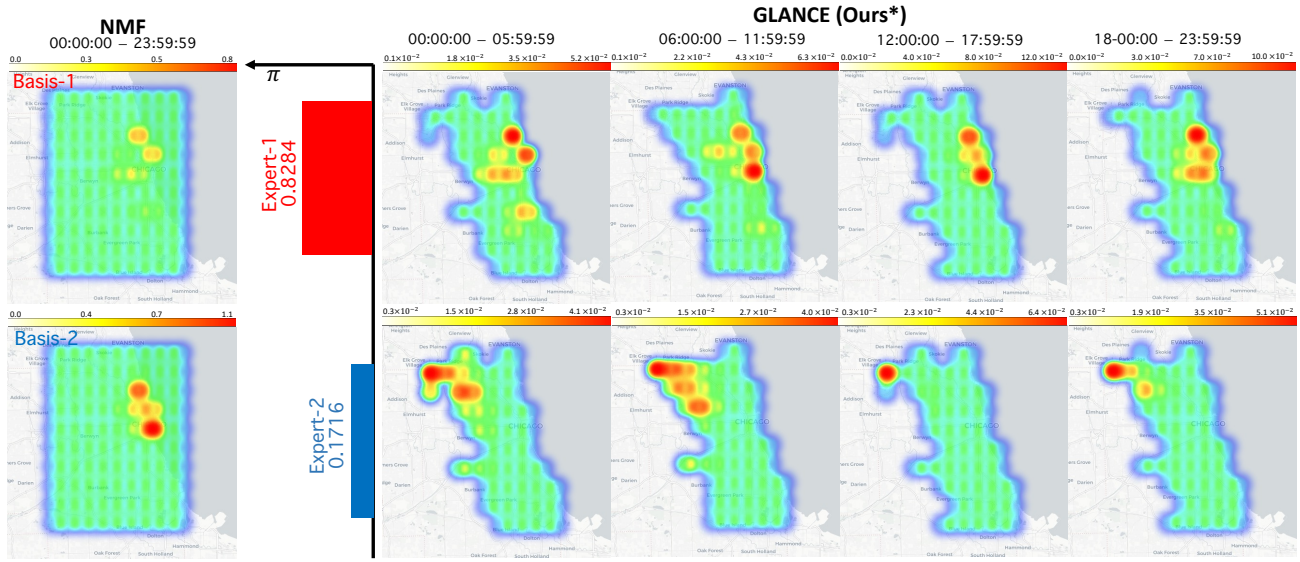

Figure 6: Comparison of the learned expert pattern of our choice model and the basis of non-negative matrix factorization (NMF) on Chicago City crime dataset. To align with the setting of LGCP-NMF, we partition the Chicago area into 10×10 area blocks. **Left**: NMF basis, **Right**: expert patterns learned by our model.

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. 5 for NYC and Fig. 6 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

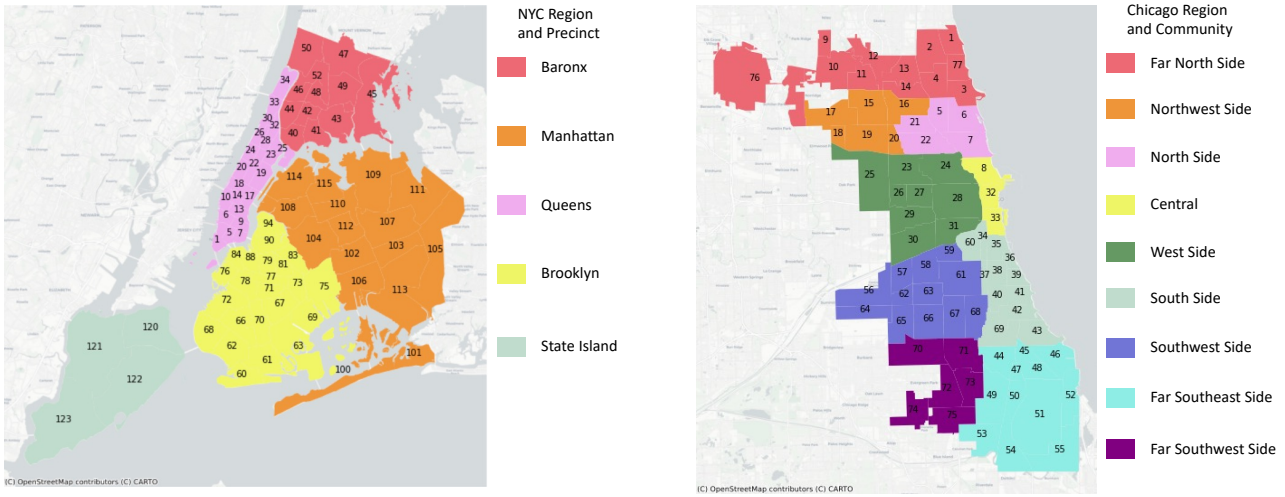

Figure 7: **Left**: NYC region and precincts. **Right**: Chicago region and communities.

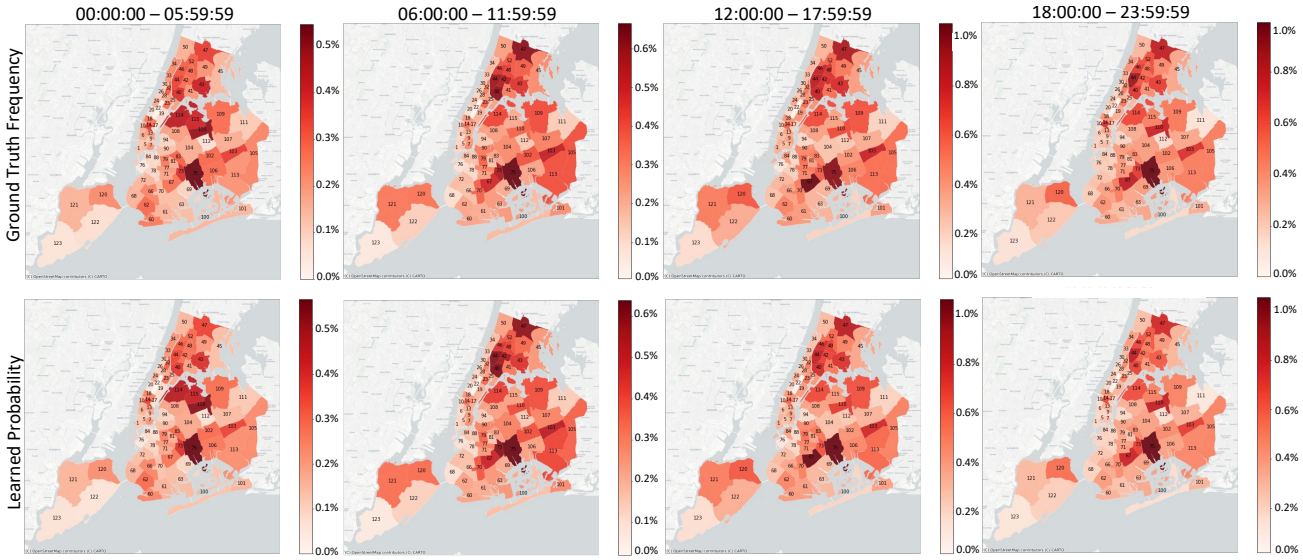

Figure 8: Comparison of the frequency and the modeled probability by precincts and time slots using training dataset from January 1 to January 31, 2024 (16,847 samples), in New York City.

C.3.1 NYC Dataset

In Fig. 7, we provide the purported regions in New York City, covering 77 precincts in total.

Accuracy and Prediction Performance The visualizations for training performance on NYC crime dataset are illustrated in Fig. 8, which indicate that the learned probabilities for each time-location pair closely match

the actual frequencies, with Precinct-75 consistently having the highest crime rates and Precinct-22 the lowest throughout January. Precinct-67 sees a spike from 12:00 - 18:00, while Precinct-40 shows a notable increase from 06:00 - 12:00. Our model effectively captures these time-location patterns. The results in Tab. 1 of main text demonstrate that our model consistently surpasses the majority baseline methods or at least achieves competitive prediction accuracy.

Table 4: Selection of the number of latent classes for NYC crime dataset. Light purple shadow highlights the current selection of modules. Performance metrics are averaged across three different runs, which reported as (Mean ± SD).

# Expert	Metrics		
	Neg. LL ↓	Time Cost (h) ↓	# Params
$H = 1$	$5.42_{\pm 0.08}$	$0.43_{\pm 0.02}$	1.699K
$H = 2$	$5.36_{\pm 0.06}$	$0.46_{\pm 0.02}$	2.724K
$H = 3$	$5.39_{\pm 0.10}$	$0.48_{\pm 0.01}$	3.749K
$H = 4$	$5.37_{\pm 0.07}$	$0.51_{\pm 0.01}$	4.774K

Explain Human Decision Process Empirical results in Tab. 4 suggest that two latent classes for NYC datasets are appropriate. The hyper-parameter selection is detailed in the Appendix. D.2, Tab. 12.

In Fig. 9, we report different crime patterns in NYC based on the gender of crime suspects. For clarity, we assume that higher values of the mixture pattern corresponding to different time-location pairs indicate a greater tendency for suspects to commit crimes in those time-location pairs.

For **Female**, *i)* **Class 1 (61%)**: indicates a tendency for most female suspects to commit crimes in Brooklyn and Queens boroughs, such as Precinct-67, -73, and -75, as well as Precinct-110, -114, and -115, where the values of the mixture pattern are consistently high across all time periods. For specific time slots, some females are less inclined to commit crimes in Precinct-67 between 00:00 and 06:00, but show a higher tendency to commit offenses between 12:00 and 18:00 in the same precinct. Conversely, Precincts-110, -114, and -115 exhibit an opposite trend, with a significantly higher inclination to commit crimes between 00:00 and 06:00 in these precincts compared to the same time slots in other precincts. *ii)* **Class 2 (39%)**: Another pattern of fewer female suspects is inclined to operate in Manhattan and the Bronx, such as Precinct-23, -40, and -44. But they typically avoid selecting Precinct-22 in Manhattan as their crime location across all time slots. Significantly, in this behavioral pattern, there is a notable increase in the inclination of female suspects to commit crimes in Precinct-67 during the 12:00 - 18:00 time slot.

For **Male**, *i)* **Class 1 (73%)**: Most male suspects tend to choose Bronx as their crime location, specifically Precincts-43 and -47, where they exhibit a high inclination across all time slots. The propensity of male suspects to commit crimes in other precincts within the Bronx is generally higher than in other areas. In particular, in the current behavioral pattern, there is a clear increase in the inclination to commit crimes in Precinct-14 during the night, specifically during the time slots of 00:00 - 06:00 and 18:00 - 24:00 compared with other precincts in the same time slots. *ii)* **Class 2 (27%)**: Some men also engage in criminal activities on Queens and Staten Island, particularly in Precinct-103. However, they usually do not tend to choose Manhattan as a location to commit crimes.

Efficiency, Scalability, and Ablation Study To evaluate the scalability of our proposed model, we adjust time slot (divided by hours for a day) resolution within range $\{2, 4, 6, 8, 12\}$ (presuming fixed precincts) and the training sample size within varied by 10 cases (e.g., January 1 to February 6, 2024 in New York City, spanning 19-37 days (up to 20203 samples) with a two-day interval). The results depicted in Fig. 10 affirm the high efficiency and strong adaptability of our model for handling large-scale datasets. The ablation study in Tab. 5 demonstrates that under our current modules combination, for NYC crime dataset, our model strikes a balance between model performance and efficiency.

C.3.2 Chicago Dataset

In Fig. 7, we provide the purported regions in Chicago, covering 77 communities in total.

Table 5: Ablation study using NYC crime dataset with 16,847 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. Our current choice is shaded in light purple.

Embedding		Expert	Utility	Metric				
(w/ prod.)	(w/ feat.)	(w/ multi.)	(w/ feat.)	Neg. LL ↓	KL ↓	RMSE ↓	Time (h) ↓	# Params
✗	✗	✗	✗	$5.50_{\pm 0.08}$	$0.43_{\pm 0.04}$	$0.46_{\pm 0.06}$	$0.39_{\pm 0.00}$	0.850K
✗	✓	✓	✓	$5.47_{\pm 0.12}$	$0.39_{\pm 0.02}$	$0.37_{\pm 0.03}$	$0.47_{\pm 0.01}$	4.011K
✓	✗	✗	✓	$5.42_{\pm 0.08}$	$0.40_{\pm 0.04}$	$0.37_{\pm 0.04}$	$0.43_{\pm 0.00}$	1.699K
✓	✗	✓	✓	$5.36_{\pm 0.06}$	$0.33_{\pm 0.02}$	$0.34_{\pm 0.01}$	$0.46_{\pm 0.02}$	2.724K
✓	✓	✗	✓	$5.47_{\pm 0.08}$	$0.39_{\pm 0.04}$	$0.38_{\pm 0.02}$	$0.44_{\pm 0.01}$	2.309K
✓	✓	✓	✗	$5.46_{\pm 0.09}$	$0.38_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.45_{\pm 0.02}$	4.664K
✓	✓	✓	✓	$5.37_{\pm 0.10}$	$0.37_{\pm 0.03}$	$0.36_{\pm 0.02}$	$0.57_{\pm 0.04}$	5.579K

Accuracy and Prediction Performance The visualizations for training performance on Chicago crime dataset are illustrated in Fig. 11, which indicates that the learned probabilities for each time-location pair closely match the actual frequencies. More analysis can be found in the main text.

Explain Human Decision Process In Tab. 3, we select number of latent classes based on empirical experimental results. In Fig. 2 of main text, we examine different crime patterns in Chicago based on the crime types, that align with perpetrators' anticipatory decision-making patterns. More analysis is provided in the main text.

Efficiency, Scalability, and Ablation Study For Chicago dataset, the results depicted in Fig. 3 (main text) 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. 2 (main text) 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.

C.4 More Ablation Studies

We present more ablation studies on the gating design and the CTC–MoE ensemble. The results support our current model configuration.

C.4.1 Why a Gating Function?

Our design choice is to separate selection from scoring. We use a gating function to "learn to select, then learn utilities": the gate learns a sparse, context-dependent consideration set first, then learns relative preferences within that selected set. This is analogous to variable selection followed by a debiased refit in regression, and it makes the model more interpretable.

As a comparison, if we impose sparsity directly on U_m, like, $L1$ penalization, which shrinks utilities toward zero (introducing bias), while Top-(k) applies a hard threshold and discards all but the largest utilities; both distort preferences and hurt the quality of the learned ranking. To make this concrete, we implemented one-stage "sparse-utility" models ($L1$-penalized and Top-(k) utilities, without gating) and found that they perform substantially worse than our gated model. The results are presented in Tab. 6.

C.4.2 Are gains mainly from ensembling?

We explicitly disentangle the effect of Mixture-of-Experts (MoE) from our consider–then–choose (CTC) structure. In the table below, "CTC = No" removes the whole consideration stage (no gating, dense utilities over all (M) options → a standard context-MNL), while "MoE = No/Yes" toggles the number of experts (1 vs. 2) under the same input features and parameterization.

Table 6: Ablation study of gating vs. direct sparsity on U_m using Chicago crime dataset.

Model	Metrics		
	KL ↓	RMSE ↓	Time Cost (h) ↓
Sparse Utility ($L1$)	$0.52_{\pm 0.05}$	$0.47_{\pm 0.03}$	$0.73_{\pm 0.02}$
Sparse Utility (Top-k)	$0.61_{\pm 0.08}$	$0.53_{\pm 0.06}$	$0.71_{\pm 0.02}$
GLANCE	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.65_{\pm 0.04}$

Table 7: Ablation study of ensembling CTC and MoE using Chicago crime dataset.

Ablation		Metric				
CTC	MoE	Neg. LL ↓	KL ↓	RMSE ↓	Time (h) ↓	# Params
✗	✗	$5.65_{\pm 0.12}$	$0.34_{\pm 0.03}$	$0.35_{\pm 0.05}$	$0.52_{\pm 0.00}$	0.982K
✗	✓	$5.60_{\pm 0.08}$	$0.33_{\pm 0.03}$	$0.32_{\pm 0.03}$	$0.80_{\pm 0.02}$	2.258K
✓	✗	$5.56_{\pm 0.10}$	$0.31_{\pm 0.04}$	$0.29_{\pm 0.02}$	$0.63_{\pm 0.00}$	1.122K
✓	✓	$5.31_{\pm 0.01}$	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.65_{\pm 0.01}$	2.732K

As shown in Tab. 7, comparing "No CTC, No MoE" vs. "Yes CTC, No MoE" shows that adding the consideration stage already yields a clear gain (KL: 0.34 → 0.31, RMSE: 0.35 → 0.29) without ensembling. By contrast, adding MoE without CTC ("No CTC, No MoE" → "No CTC, Yes MoE") gives a smaller improvement (KL: 0.34 → 0.33, RMSE: 0.35 → 0.32). The best performance arises when both are present, but the CTC structure contributes at least as much as ensembling.

In addition to the CTC ablation above, we also compared against one-stage sparse-utility variants (no gating, utilities directly regularized with $L1$ or Top-(k)). These models share the same MoE structure but force U_m to both select and shrink alternatives, and they are substantially worse than GLANCE (e.g., According to results in Tab. 6, KL 0.52/0.61 vs. 0.24; RMSE 0.47/0.53 vs. 0.24). This shows that the gains are not simply due to ensembling or generic sparsity, but to explicit, interpretable modeling of sparse consideration sets followed by expert-specific preferences.

C.5 Long-Term Prediction

The long-term prediction results shown in Tab. 8 and Tab. 9 include comparative accuracy analysis against baseline methods. Despite the inherent challenges of long-term forecasting, our model consistently outperforms or achieves comparable performance compared with baseline methods in terms of prediction accuracy for both NYC and Chicago crime datasets.

C.6 Case Study

Now we provide detailed case studies. Using NYC crime datasets, shown in Tab. 10, for female suspects, both crime patterns uncovered by our model consistently point to the same high-risk time-location pair: (12:00 – 17:59, 67). In contrast, for male suspects, the two identified crime patterns indicate different high-risk locations — (12:00 – 17:59, 47) and (12:00 – 17:59, 103) — during the same time window.

Using Chicago crime datasets, shown in Tab. 11, for theft suspects, Pattern-1 points to the highest-risk time-location pair: (12:00 – 17:59, 30) and Pattern-2 points to the highest-risk time-location pair: (12:00 – 17:59, 41). For battery suspects, after adjustment by preference-driven utility, Pattern-1 point to the highest-risk time-location pair: (18:00 – 23:59, 70). Note that this time-location pair also has highest utility for Pattern-1 of battery crime. In this situation, incorporating the perpetrator's psychological anticipation of criminal outcomes (including factors like crime severity and crime typology), the gated results is different from original results, demonstrating the necessity of adjustment from preference-driven utility. Lastly, Pattern-2 points to the highest-risk time-location pair: (18:00 – 23:59, 24), which shows a different tendency of suspect to yield the battery events.

Table 8: Long-term prediction for future 1-3 days using NYC crime datasets. We use training data from January 1 to January 31, 2024, comprising 16,874 samples for NYC. Performance metrics are averaged across three different runs, which reported as (Mean ± SD).

Model	Future 1 days		Future 2 days		Future 3 days	
	KL ↓	RMSE ↓	KL ↓	RMSE ↓	KL ↓	RMSE ↓
AMAR	$0.65_{\pm 0.06}$	$0.62_{\pm 0.08}$	$0.73_{\pm 0.08}$	$0.77_{\pm 0.10}$	$0.82_{\pm 0.07}$	$0.84_{\pm 0.08}$
CSI	$0.67_{\pm 0.08}$	$0.66_{\pm 0.03}$	$0.83_{\pm 0.07}$	$0.86_{\pm 0.09}$	$0.92_{\pm 0.09}$	$0.95_{\pm 0.10}$
LGCP	$0.67_{\pm 0.11}$	$0.67_{\pm 0.09}$	$0.73_{\pm 0.08}$	$0.76_{\pm 0.10}$	$0.82_{\pm 0.10}$	$0.83_{\pm 0.08}$
NSTPP	$0.51_{\pm 0.06}$	$0.49_{\pm 0.05}$	$0.73_{\pm 0.06}$	$0.74_{\pm 0.08}$	$0.84_{\pm 0.08}$	$0.84_{\pm 0.09}$
DSTPP	$0.47_{\pm 0.04}$	$0.45_{\pm 0.05}$	$0.61_{\pm 0.08}$	$0.64_{\pm 0.07}$	$0.74_{\pm 0.06}$	$0.77_{\pm 0.08}$
ST-HSL	$0.56_{\pm 0.06}$	$0.52_{\pm 0.05}$	$0.76_{\pm 0.08}$	$0.77_{\pm 0.05}$	$0.89_{\pm 0.09}$	$0.83_{\pm 0.10}$
HintNet	$0.38_{\pm 0.03}$	$0.37_{\pm 0.03}$	$0.48_{\pm 0.04}$	$0.48_{\pm 0.06}$	$0.55_{\pm 0.07}$	$0.56_{\pm 0.05}$
STNSCM	$0.37_{\pm 0.02}$	$0.38_{\pm 0.04}$	$0.45_{\pm 0.06}$	$0.46_{\pm 0.05}$	$0.54_{\pm 0.05}$	$0.54_{\pm 0.04}$
UniST	$0.37_{\pm 0.03}$	$0.36_{\pm 0.02}$	$0.45_{\pm 0.04}$	$0.45_{\pm 0.05}$	$0.52_{\pm 0.04}$	$0.52_{\pm 0.03}$
MNL	$0.38_{\pm 0.01}$	$0.38_{\pm 0.01}$	$0.48_{\pm 0.05}$	$0.48_{\pm 0.04}$	$0.51_{\pm 0.04}$	$0.54_{\pm 0.07}$
GLANCE	$0.33_{\pm 0.02}$	$0.34_{\pm 0.01}$	$0.42_{\pm 0.04}$	$0.43_{\pm 0.04}$	$0.49_{\pm 0.07}$	$0.48_{\pm 0.05}$

Table 9: Long-term prediction for future 1-3 days using Chicago crime datasets. We use training data from July 1 to July 31, 2024, comprising 23,545 samples for Chicago City. Performance metrics are averaged across three different runs, which reported as (Mean ± SD).

Model	Future 1 days		Future 2 days		Future 3 days	
	KL ↓	RMSE ↓	KL ↓	RMSE ↓	KL ↓	RMSE ↓
AMAR	$0.70_{\pm 0.10}$	$0.68_{\pm 0.06}$	$0.78_{\pm 0.08}$	$0.86_{\pm 0.05}$	$0.93_{\pm 0.10}$	$0.97_{\pm 0.08}$
CSI	$0.68_{\pm 0.12}$	$0.65_{\pm 0.09}$	$0.73_{\pm 0.07}$	$0.77_{\pm 0.10}$	$0.83_{\pm 0.10}$	$0.85_{\pm 0.12}$
LGCP	$0.69_{\pm 0.09}$	$0.68_{\pm 0.08}$	$0.75_{\pm 0.06}$	$0.73_{\pm 0.07}$	$0.78_{\pm 0.10}$	$0.81_{\pm 0.09}$
NSTPP	$0.42_{\pm 0.07}$	$0.44_{\pm 0.10}$	$0.52_{\pm 0.06}$	$0.51_{\pm 0.07}$	$0.60_{\pm 0.03}$	$0.64_{\pm 0.07}$
DSTPP	$0.47_{\pm 0.04}$	$0.46_{\pm 0.08}$	$0.56_{\pm 0.05}$	$0.58_{\pm 0.08}$	$0.59_{\pm 0.04}$	$0.66_{\pm 0.05}$
ST-HSL	$0.49_{\pm 0.04}$	$0.52_{\pm 0.06}$	$0.54_{\pm 0.05}$	$0.55_{\pm 0.07}$	$0.57_{\pm 0.03}$	$0.60_{\pm 0.05}$
HintNet	$0.26_{\pm 0.04}$	$0.28_{\pm 0.03}$	$0.33_{\pm 0.06}$	$0.37_{\pm 0.07}$	$0.48_{\pm 0.08}$	$0.46_{\pm 0.04}$
STNSCM	$0.27_{\pm 0.01}$	$0.31_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.39_{\pm 0.06}$	$0.45_{\pm 0.06}$	$0.38_{\pm 0.07}$
UniST	$0.27_{\pm 0.05}$	$0.30_{\pm 0.04}$	$0.31_{\pm 0.03}$	$0.35_{\pm 0.02}$	$0.42_{\pm 0.05}$	$0.41_{\pm 0.02}$
MNL	$0.25_{\pm 0.03}$	$0.29_{\pm 0.02}$	$0.36_{\pm 0.03}$	$0.36_{\pm 0.02}$	$0.47_{\pm 0.03}$	$0.50_{\pm 0.04}$
GLANCE	$0.24_{\pm 0.02}$	$0.24_{\pm 0.02}$	$0.32_{\pm 0.02}$	$0.36_{\pm 0.02}$	$0.40_{\pm 0.03}$	$0.45_{\pm 0.02}$

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. 12. The hyper-parameter selection metric is a trade-off between training converged log-likelihood, prediction performance, and time efficiency.

E Validation of Experimental Results

Our experimental results, particularly regarding interpretability, are consistent with prior research in this domain while revealing additional nuances. Previous studies have reported similar spatial patterns, but our empirical

Table 10: Top-3 time-location pairs for NYC dataset based on highest original gating results, learned utility, and refined gating results across different genders (male and female) and crime patterns. We use training data from January 1 to January 31, 2024, comprising 16,874 samples for NYC. The time-location pair is denoted in the format: ("Time Slot", "Precinct Index").

Top-3	**Female: Pattern-1**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 75)	(12:00-17:59, 67)	(12:00-17:59, 67)
No.2	(12:00-17:59, 43)	(18:00-23:59, 110)	(18:00-23:59, 75)
No.3	(12:00-17:59, 47)	(18:00-23:59, 73)	(12:00-17:59, 75)
Top-3	**Female: Pattern-2**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 103)	(06:00-11:59, 44)	(12:00-17:59, 67)
No.2	(00:00-05:59, 23)	(12:00-17:59, 67)	(18:00-23:59, 44)
No.3	(18:00-23:59, 109)	(18:00-23:59, 46)	(06:00-11:59, 44)
Top-3	**Male: Pattern-1**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 47)	(12:00-17:59, 47)	(12:00-17:59, 47)
No.2	(12:00-17:59, 43)	(18:00-23:59, 47)	(18:00-23:59, 47)
No.3	(12:00-17:59, 44)	(18:00-23:59, 14)	(12:00-17:59, 43)
Top-3	**Male: Pattern-2**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 103)	(12:00-17:59, 103)	(12:00-17:59, 103)
No.2	(18:00-23:59, 120)	(12:00-17:59, 109)	(18:00-23:59, 103)
No.3	(18:00-23:59, 109)	(12:00-17:59, 105)	(06:00-11:59, 103)

analysis uncovers finer variations supported by quantitative evidence. Specifically, the high-crime areas identified in both New York City and Chicago largely coincide with low-income regions, consistent with the inverse relationship between income and crime rates demonstrated in Yildiz et al. (2013). The spatial distributions we detect—Midtown and Lower Manhattan (including the Financial District) in New York (Catlett et al., 2019), and corresponding hotspots in Chicago (Linderman and Adams, 2014)—align well with these established findings. Beyond these consistencies, our model further reveals distinct behavioral preferences across different mixture components, a factor that existing studies have largely overlooked.

Incorporating demographic attributes follows prior work (Lau, 2021; Niu et al., 2019; Kang and Kang, 2017) and is not intended for discriminatory analysis. All attributes were processed following standard benchmark protocols to ensure fairness and avoid any bias in interpretation.

F 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

Table 11: Top-3 time-location pairs for Chicago dataset based on highest original gating results, learned utility, and refined gating results across different crime type (theft and battery) and crime patterns. We use training data from July 1 to July 31, 2024, comprising 23,545 samples for Chicago City.

Top-3	**Theft: Pattern-1**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 30)	(12:00-17:59, 30)	(12:00-17:59, 30)
No.2	(06:00-11:59, 30)	(18:00-23:59, 9)	(06:00-11:59, 30)
No.3	(18:00-23:59, 30)	(12:00-17:59, 9)	(18:00-23:59, 27)
Top-3	**Theft: Pattern-2**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(12:00-17:59, 41)	(06:00-11:59, 41)	(12:00-17:59, 41)
No.2	(12:00-17:59, 30)	(12:00-17:59, 41)	(18:00-23:59, 41)
No.3	(00:00-05:59, 41)	(18:00-23:59, 41)	(06:00-11:59, 41)
Top-3	**Battery: Pattern-1**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(18:00-23:59, 47)	(18:00-23:59, 70)	(18:00-23:59, 70)
No.2	(18:00-23:59, 70)	(18:00-23:59, 72)	(12:00-17:59, 70)
No.3	(00:00-05:59, 68)	(12:00-17:59, 70)	(18:00-23:59, 47)
Top-3	**Battery: Pattern-2**		
	Original Gating Results	**Learned Utility**	**Refined Gating Results**
No.1	(18:00-23:59, 24)	(18:00-23:59, 24)	(18:00-23:59, 24)
No.2	(12:00-27:59, 24)	(12:00-17:59, 24)	(18:00-23:59, 27)
No.3	(00:00-05:59, 24)	(12:00-17:59, 28)	(12:00-17:59, 24)

neural architectures advances transparent AI systems that align with human reasoning, benefiting domains like transportation (e.g., ride-sharing demand prediction) 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.

Table 12: Descriptions and values of hyper-parameters used for models trained on the three real-world datasets.

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

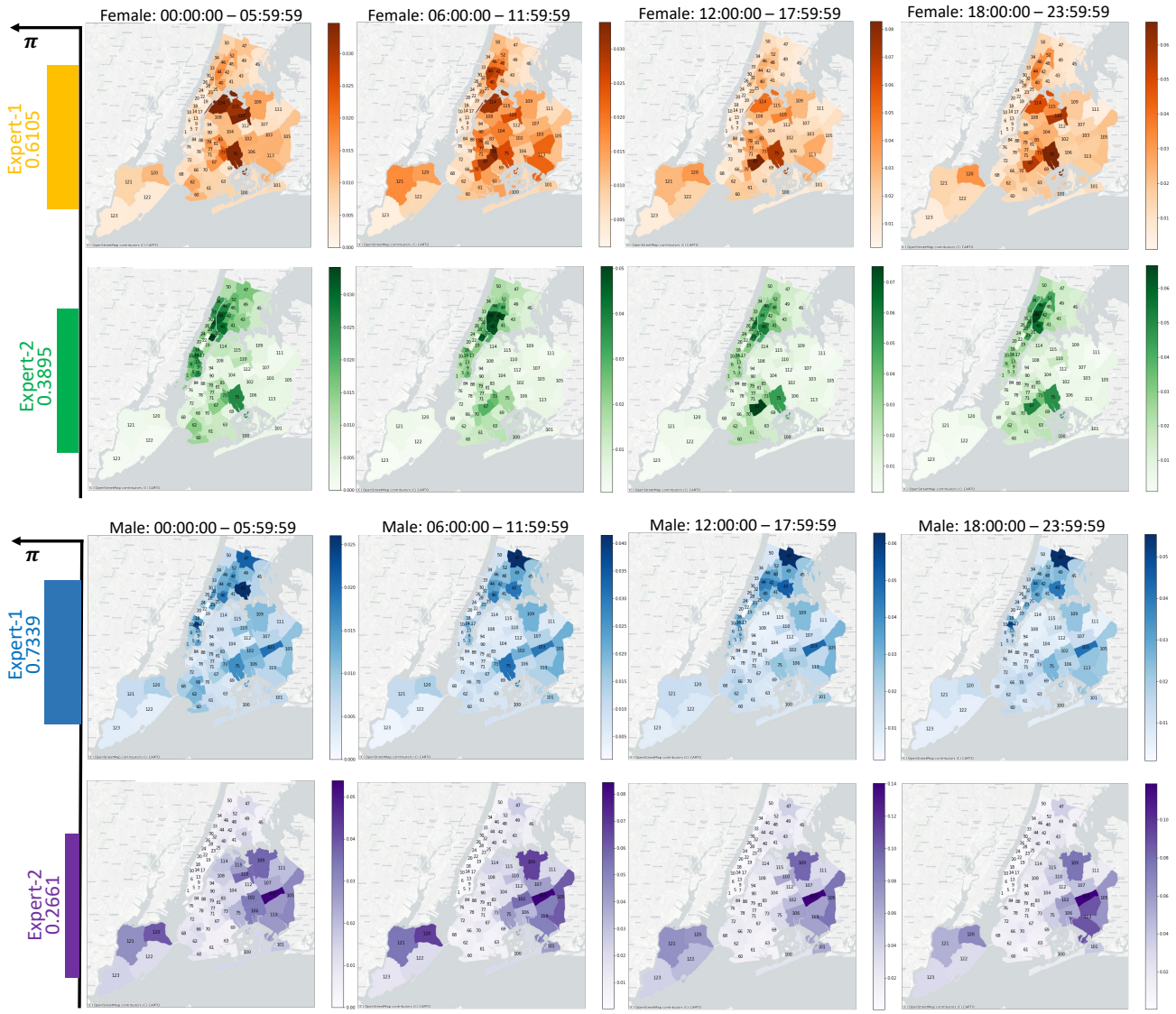

Figure 9: Mixing coefficient π^h (**Left bar plots**) and mixture pattern adjusted by utility score ($g^h \exp(U^h)$) for different latent class-h and different gender (**Right heatmaps**) from January 1 to January 31, 2024 (16,847 samples), in New York City. The selection of the number of experts is based on empirical experiments.

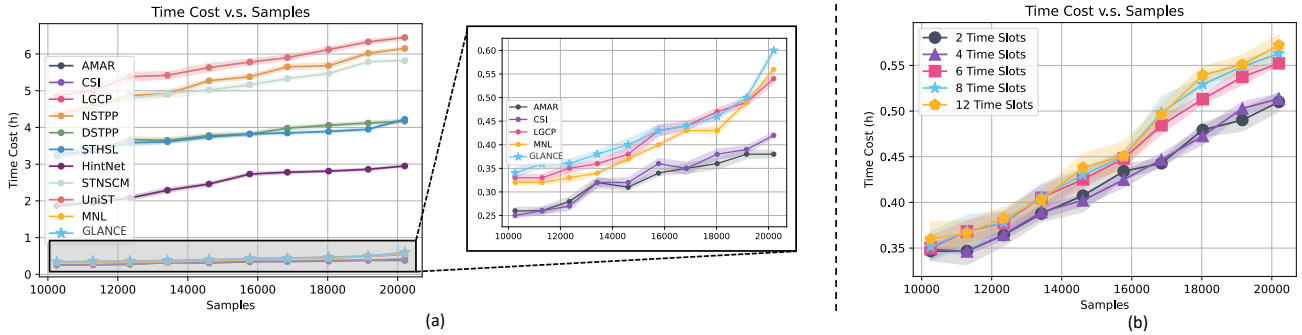

Figure 10: Scalability experiments for NYC crime datasets with varying training samples and time slots. **(a)** Time cost v.s. training samples for all methods with fixed 4 time slots, and **(b)** Time cost v.s. training samples for our proposed method with varying time slots. All the experiments are conducted over three random runs and the standard error is reflected in the shaded areas.

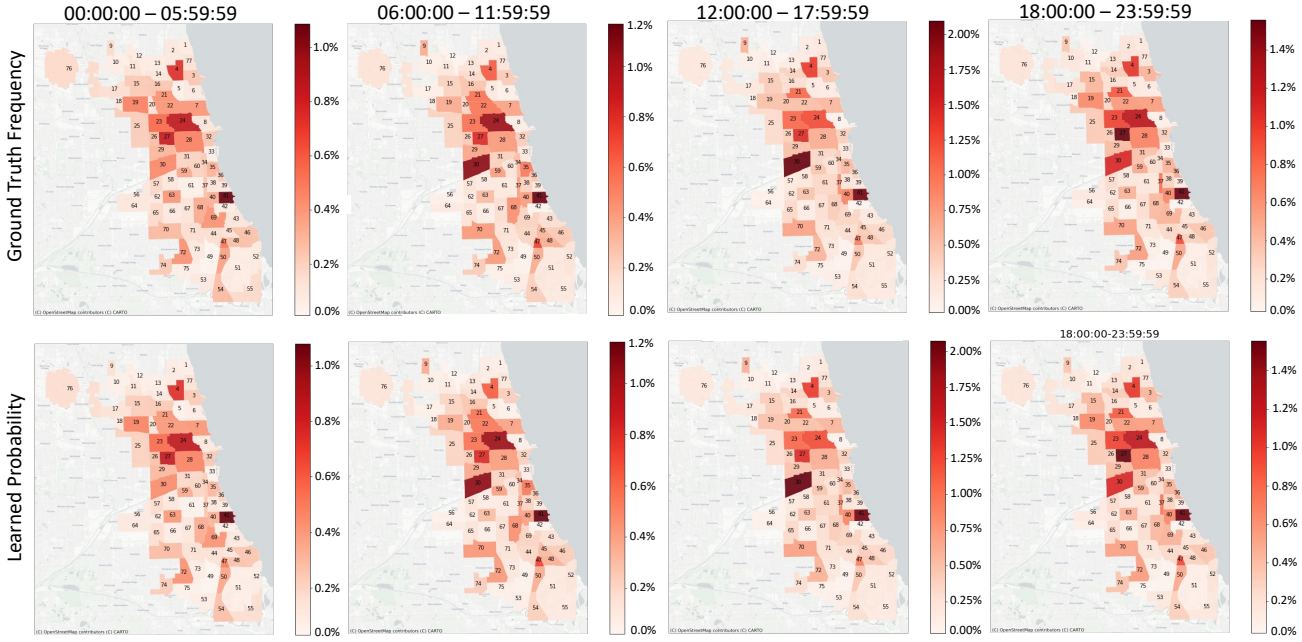

Figure 11: Comparison of the actual crime frequency and the modeled probability by community from July 1 to July 31, 2024 (23,545 samples), in Chicago City.