RELY: BEYOND DETERMINISTIC GRAPHS FOR UNCERTAINTY-AWARE CRIME FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting crime across urban regions requires knowing which regions influence one another. This structure is not observed and must be inferred from sparse event records, leaving much of it uncertain. Existing models build a single graph, either by hand or by learning, and then treat it as ground truth, so errors in the graph pass directly into the forecast. Uncertainty about the graph is a second source of forecast error, separate from the randomness of the counts themselves, and a model that represents only the latter cannot correct for the former. We introduce RELY, which keeps the graph uncertain rather than fixing it: it assigns each relation a probability that is updated for every forecast window, and estimates how much each region can rely on each kind of relation. This reliability then controls how much each relation contributes to the two components of a zero-inflated count forecast: the expected number of events, and the probability that none are observed. On Chicago and New York, RELY reduces MAE by 6.9% and 4.5% over the strongest baselines, with the largest gains in the sparsest regions; its reliability estimates track forecast error, and the relations that matter for the expected number of events differ from those that matter for whether any are observed.

1 INTRODUCTION

Reliable crime forecasts support risk assessment and the timely allocation of public-safety resources (Kounadi et al., 2020; Sun et al., 2020). Crime in one region of a city is informative about crime in others: neighbouring areas, non-adjacent areas with similar urban functions, and groups of regions exposed to common events all tend to move together. Graph models have therefore become the standard way to represent these spatial, semantic, and higher-order dependencies (Sun et al., 2020; Wang et al., 2022; Xia et al., 2022; Li et al., 2022; Zhou et al., 2024). These models, however, assume that every relation in the graph they are given is real and relevant to the window being forecast, and this assumption is rarely examined.

In practice the graph is uncertain, and existing models discard that uncertainty. Adjacency, urban function, and event histories all provide evidence about how regions are related, but none of them shows directly which regions influence one another. Regional crime counts are sparse, zero-inflated, and overdispersed (Li et al., 2022; Wang et al., 2025): most region–day counts are zero and the rest vary widely. Each relation is therefore supported by only a few events, and two regions may record incidents in the same window by chance without one affecting the other. Collapsing this uncertainty into a single edge weight, whether set by hand or learned, hides whether the relation should take part in message passing at all. Modelling the distribution of the counts, as zero-inflated models do, does not help here: it captures how random the counts are, not whether the graph used to predict them is right.

Existing approaches either fix the graph in advance from adjacency or historical similarity (Sun et al., 2020; Xia et al., 2022; Li et al., 2022) or learn association strengths from the data (Wang et al., 2022; Zhou et al., 2024); even ST-HHOL (Du et al., 2026), which learns hierarchical hypergraph relations and updates them online, uses the learned strengths as given, so no later stage knows how much to trust them. Probabilistic structure learning does treat the graph as a random variable (Shang et al., 2021; Manenti et al., 2024), but it has been developed for a single homogeneous relation and stops once the structure is inferred, without carrying its uncertainty into the prediction. Bringing these ideas to crime forecasting therefore raises three challenges.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The first challenge is that no labels exist for the relations themselves, so the model must infer which relations hold in each window from indirect evidence alone. Prior urban knowledge can propose plausible relations and narrow the candidates, but treating it as ground truth carries every error in it into the forecast. Learning the relations from the current window alone is equally problematic: with so few events, a correlation that appears in one window cannot be told apart from a lasting one. A suitable model must combine prior knowledge with the current state of each region, retain uncertainty where evidence is weak, and revise the relations without adding ones that the data do not support.

The second challenge is to turn the uncertainty in the inferred relations into a quantity that later stages can use. A posterior over relations says how likely each one is, but a region needs to know something coarser: whether a whole kind of relation, adjacency or a shared POI profile for instance, can be relied upon in the current window. This reliability must be derived without ground truth and must actually track forecast error, otherwise it is only another learned weight.

The third challenge is that the forecast has two parts that need different relations. A zero-inflated negative binomial (ZINB) model, the standard choice for such counts, separates the expected count, μ, from the probability that the region is in an excess-zero state, π (Yau et al., 2003; Feng, 2021; Wang et al., 2025). The relations that explain how many crimes occur need not be those that indicate whether a region stays at zero, and a single fusion rule that ignores reliability forces both parameters to use the same evidence.

We propose RELY, named for the structural reliability that governs how it uses relational evidence. For the first challenge, RELY neither trusts prior knowledge nor learns the graph from scratch: each candidate relation starts from a prior probability derived from urban evidence, and the model learns only a correction to it from the current window, so structure is revised but never invented, and the result is kept as a distribution rather than a fixed graph. Relations of different kinds—spatial, semantic, and higher-order—are kept as separate channels. For the second challenge, reliability is not a learned weight but a function of the posterior itself, computed from how much support a channel has and how decisive its probabilities are, so nothing is trained to produce it and it cannot be optimised to fit the loss. For the third challenge, two routers with identical inputs, one serving μ and one serving π, weight the channels with reliability as a soft bias, so that the division of labour between the two parameters emerges from their objectives rather than being imposed.

Our contributions are threefold:

1. **Structural uncertainty as a distinct source of error.** We identify uncertainty about which relations hold in a given window as a source of forecast error separate from the randomness of the counts, and show that modelling the output distribution alone does not remove it.
2. **RELY.** We propose a forecasting model that (i) infers a posterior over relations of several kinds by learning only corrections to evidence-derived priors, so that the graph is corrected rather than replaced; (ii) derives from that posterior a reliability score for each region and kind of relation, without training it as a weight; and (iii) uses reliability to route relational evidence separately into the intensity and excess-zero components of a zero-inflated forecast.
3. **Findings.** Gains over strong baselines concentrate in the sparsest regions and categories, where structure is most uncertain; the reliability estimates decrease monotonically with forecast error; and the intensity and excess-zero components are shown to use different relations.

2 Related Work

Graph-based crime forecasting has moved from predefined pairwise relations, as in DeepCrime and CrimeForecaster (Huang et al., 2018; Sun et al., 2020), to hypergraphs (Xia et al., 2022; Li et al., 2022), adaptive or heterogeneous relations (Wang et al., 2022; Zhou et al., 2024), and hierarchical hypergraphs updated online (Du et al., 2026). Whether the graph is predefined or learned, the resulting weights are used as given for propagation, and no later stage is told how much to trust them.

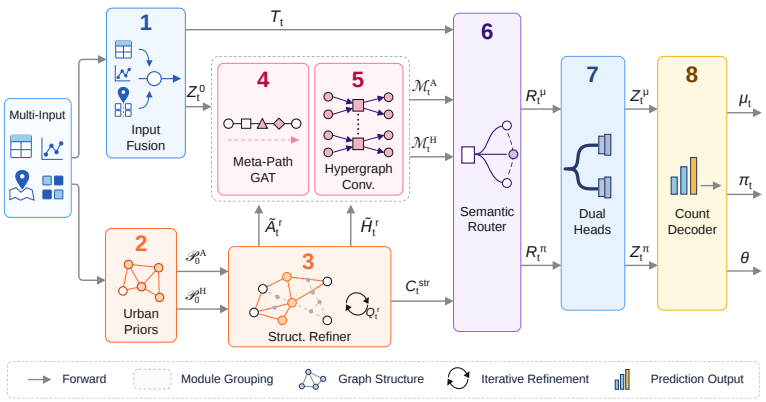


Figure 1: Overview of RELY. Urban knowledge defines typed relation priors, which are refined into window-conditioned posteriors and propagated through six relation channels; structural reliability derived from the posteriors biases two routers that select evidence for the ZINB intensity (μ) and excess-zero (π) parameters.

Two lines of work model uncertainty, but not this kind. STMGNN-ZINB uses a zero-inflated negative binomial output (Wang et al., 2025), which captures how random the counts are, not whether the graph used to predict them is right. Probabilistic structure learning treats the graph as a random variable (Shang et al., 2021), and Manenti et al. (2024) show that prediction loss alone does not recover relational uncertainty; but this line addresses a single homogeneous relation and stops once the structure is inferred. RELY combines the two: it infers a posterior over heterogeneous crime relations, derives a reliability score from that posterior, and uses it to route relational evidence separately to the intensity and excess-zero components of a zero-inflated count model.

3 Methodology

Figure 1 shows the architecture of RELY. An input-fusion module (1) turns each region's recent observations and static context into a window-specific state. Urban evidence is converted into prior probabilities over candidate relations (2), which a structure refiner turns into posteriors conditioned on the regional states (3). The inferred relations are propagated through six channels: four graph channels with meta-path attention (4) and two hypergraph channels (5). A reliability score derived from the posterior biases a semantic router (6) that assembles separate relational mixtures for the intensity and excess-zero branches, each with its own head (7), and a count decoder produces the ZINB forecast (8). All modules are trained jointly; the following subsections describe them in this order, and full derivations of the components summarised here are given in Appendix B.

Problem formulation. A city is partitioned into N regions. At forecast origin t, the model observes a length-L history of crime counts $Y_t^{\text{his}} \in \mathbb{N}_0^{N\times L\times C}$ over C categories, auxiliary urban-activity counts $O_t^{\text{his}} \in \mathbb{N}_0^{N\times L\times O}$ over O types (in our data, 311 service requests, i.e. non-emergency municipal reports), and city-level weather $W_t^{\text{his}} \in \mathbb{R}^{L\times W}$, together with static socioeconomic attributes $X^{\text{sta}} \in \mathbb{R}^{N\times D_s}$, POI counts $X^{\text{poi}} \in \mathbb{N}_0^{N\times P}$, and spatial adjacency $A^{\text{sp}} \in \{0,1\}^{N\times N}$. The task is to forecast the crime counts $Y_t^{\text{fut}} \in \mathbb{N}_0^{N\times H\times C}$ over the next H steps. RELY outputs, for every region, horizon, and category, the parameters of a ZINB distribution, from which point forecasts and predictive uncertainty are derived. The relations that couple regions are not observed: the inputs above provide evidence for them, and RELY treats that evidence as priors over latent structure rather than as the structure itself.

3.1 Input Fusion

Each region has a dynamic view, its recent crime, activity, and weather sequences, and a static view, its socioeconomic attributes and POI composition. A dilated temporal convolutional network shared across regions encodes the dynamic view into $T_t \in \mathbb{R}^{N\times d_t}$, a shared MLP encodes the static view

into $S_t \in \mathbb{R}^{N\times d_t}$, and the two are fused into the initial regional state

$$Z_t^0 = \mathrm{LN}(f_{\mathrm{init}}([T_t, S_t])) \in \mathbb{R}^{N\times d_g}, \tag{1}$$

which conditions structure inference and initialises propagation. Input preprocessing and encoder details are given in Appendix A.2.

3.2 Urban Priors and Structure Refinement

In this study, we assume observed urban relations provide useful inductive bias but are incomplete and may vary in relevance across prediction windows. RELY therefore distinguishes an evidence-derived prior structure from a learned posterior structure.

To address this, we construct four heterogeneous relation types, $\mathcal{R}_A = \{RR, RP, RC, RO\}$, representing region-region spatial adjacency, region-POI relations, region-crime-type relations, and region-urban-activity-type relations, respectively. For RR, the spatial adjacency matrix is symmetrised and self-loops are removed. Observed spatial adjacencies are assigned the prior probability $p_{\max}$, while unsupported relations are initially assigned zero probability. For the remaining relations, let $s_{ij}^r \in [0,1]$ denote the log-normalised intensity of type j in region i: POI counts for RP, and crime and activity counts summed over the historical window for RC and RO (Appendix B.1). The prior probability is

$$P_{0,ij}^r = \begin{cases} p_{\min} + (p_{\max} - p_{\min})\, s_{ij}^r, & s_{ij}^r > 0, \\ 0, & s_{ij}^r = 0. \end{cases} \tag{2}$$

This affine map places every supported relation in $[p_{\min}, p_{\max}]$ with $0 < p_{\min} < p_{\max} < 1$: observational support alone never yields a prior below $p_{\min}$, and no prior saturates at one, so the prior log-odds used by the residual refinement below remain finite and revisable. Probabilities of exactly one are reserved for the fixed anchor self-memberships of the hypergraphs, and the same map is applied to cosine similarity when constructing the hypergraph priors. Alongside P_0^r, each relation type keeps a binary evidence mask M_0^r marking entries with direct observational support; for $r \in \{RP, RC, RO\}$, $M_{0,ij}^r = \mathbb{I}[s_{ij}^r > 0]$.

To capture higher-order regional dependencies, we additionally construct two anchor-based hypergraphs, $\mathcal{R}_H = \{HF, HE\}$. The functional hypergraph (HF) is derived from static socioeconomic attributes and POI composition, while the event hypergraph (HE) uses recent crime and activity trajectories. Each anchor region and its k_0 most similar other regions form a hyperedge, with cosine similarity mapped to prior membership probabilities in $[p_{\min}, p_{\max}]$ and anchor self-membership fixed at one (Appendix B.1). Dynamic priors use only observations from the current historical window, while static priors use spatial relations and regional attributes available at prediction time.

Because the observed priors may omit meaningful dependencies, we enlarge the search space with representation-dependent candidates. Specifically, we construct a semantic k_s-nearest-neighbour graph over the normalised regional states,

$$\bar{Z}_t^0 = \mathrm{normalize}(\mathrm{sg}[Z_t^0]), \qquad M_{\mathrm{sem}} = M_{\mathrm{kNN}}(\bar{Z}_t^0) \vee M_{\mathrm{kNN}}(\bar{Z}_t^0)^\top, \tag{3}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient and $\vee$ element-wise union. The stop-gradient is deliberate: the candidate set follows the current regional states but is not itself optimised, so the model cannot lower its loss by enlarging the search space, and the number of admissible relations stays bounded by k_s per region. For the region–region relation and the two hypergraphs, the candidate mask combines the observed support with semantic neighbours, $M_{\mathrm{can}}^r = M_0^r \vee M_{\mathrm{sem}}$; for the region–type relations, all region–type pairs are candidates.

Over this candidate space we define a completed prior $\bar{P}_0^r$: relations with direct observational support keep their prior probability, newly admitted candidates receive a weak prior p_{new}, and relations outside the candidate set receive zero:

$$\bar{P}_{0,ij}^r = \begin{cases} P_{0,ij}^r, & M_{0,ij}^r = 1, \\ p_{\mathrm{new}}, & M_{0,ij}^r = 0,\ M_{\mathrm{can},ij}^r = 1, \\ 0, & M_{\mathrm{can},ij}^r = 0. \end{cases} \tag{4}$$

Thus M_{can}^r determines the search space of the structure refiner and $\bar{P}_0^r$ its initial belief over that space. Because p_{new} is small, a newly admitted relation is absent by default and must be supported

by the current window to obtain appreciable posterior probability; together with the bounded candidate set, this is what prevents structure learning from expanding freely under sparse observations.

Given the candidate space, we refine each prior into a posterior relation probability. For every candidate pair, a relation-specific scorer g_r (an MLP with scalar output) predicts a correction $\Delta^r_{t,ij}$ to the prior log-odds from a pair representation $\phi^r_{t,ij}$ built from the current regional states, yielding

$$Q^r_{t,ij} = M^r_{\text{can},ij}\,\sigma\big(\text{logit}(\bar{P}^r_{0,ij}) + \Delta^r_{t,ij}\big)\,, \qquad \Delta^r_{t,ij} = g_r\big(\phi^r_{t,ij}\big)\,. \tag{5}$$

This residual parameterisation allows the model to revise, rather than discard, prior urban knowledge: with $\Delta = 0$ the posterior equals the prior, and relations outside the candidate set can never be introduced. Pair representations combine regional states with neighbouring regions, prior-conditioned type representations, or hyperedge representations, using relation-specific scorers (Appendix B.2). For the hypergraphs, each anchor retains deterministic membership in its own hyperedge ($Q^r_{t,jj} = 1$), and only the remaining candidate memberships are refined.

During training, we use Binary-Concrete samples of the posterior structures for propagation, with an annealed temperature τ; at inference, we use the posterior probabilities directly. We denote the resulting graph and hypergraph structures by $\widetilde{A}^r_t$ and $\widetilde{H}^r_t$, respectively. Candidate masks, anchor self-memberships, and the symmetry and absence of self-loops in RR are preserved (Appendix B.3).

3.3 Relation-specific Propagation

The six learned structures are propagated independently so that their relational semantics are preserved before fusion: the four graph relations by meta-path attention (module 4 in Figure 1) and the two hypergraphs by hypergraph convolution (module 5). For the region–region channel, probability-gated attention combines semantic compatibility with inferred structural support; relations with zero posterior probability do not participate. Each region–type relation is propagated through a two-hop region-type-region path with structure-weighted attention, capturing dependencies mediated by POI functions, crime composition, and urban activity. Each hypergraph uses degree-normalised region-to-hyperedge and hyperedge-to-region aggregation with non-linear transforms. This produces six relation-specific regional messages, $\mathcal{M}_t = \{M^{RR}_t, M^{RPR}_t, M^{RCR}_t, M^{ROR}_t, M^{HF}_t, M^{HE}_t\}$. The full operators are given in Appendix B.4.

3.4 Structural Reliability and Parameter-specific Routing

Edge-wise posteriors state how likely each relation is, but routing needs a coarser quantity: how far a region can rely on a whole channel. We therefore summarise each channel's posterior into a reliability score. Let q denote the row of the posterior Q^r_t for region i, i.e. its candidate relations in channel r. We combine its support strength $a(q)$ and its posterior decisiveness $d(q)$ into the structural reliability $c(q)$,

$$a(q) = 1 - \exp\Big(-\sum_j q_j\Big), \qquad d(q) = \frac{\sum_j q_j\,(1 - h(q_j)/\log 2)}{\max\Big(\sum_j q_j, \epsilon\Big)}, \qquad c(q) = a(q)\,d(q), \tag{6}$$

where $h(q_j) = -q_j \log q_j - (1-q_j)\log(1-q_j)$ is the Bernoulli entropy. Applying this operation to the six relation channels produces $C^{\text{str}}_t \in [0,1]^{N\times 6}$. A channel therefore receives high reliability only when it contains sufficient relational support and its posterior relations are comparatively decisive. The two factors are deliberately separate: a channel with few but well-determined relations, such as spatial adjacency, can be highly reliable, whereas a channel whose candidate set is large but whose posteriors remain close to the prior, such as a region–type relation, has ample support yet low decisiveness and therefore low reliability.

The relations that govern crime intensity may differ from those that govern excess zeros, so the semantic router (module 6) has two independent branches, one for μ and one for π. The six messages are stacked as $M_t \in \mathbb{R}^{N\times 6\times d_g}$, and for branch $b \in \{\mu, \pi\}$, region i, and relation channel k, the routing score and weights are

$$\rho^b_{t,i,k} = f^b_{\text{score}}([T_{t,i}, M_{t,i,k}]) + \text{softplus}(\gamma^b)\log(C^{\text{str}}_{t,i,k} + \epsilon), \qquad w^b_{t,i,k} = \text{softmax}_k(\rho^b_{t,i,k}). \tag{7}$$

Routing is thus determined jointly by the current temporal state, relation-specific message content, and the reliability of the inferred relational structure; reliability enters as a soft bias whose strength

$\mathrm{softplus}(\gamma^b)$ is learned, so low-reliability channels are down-weighted rather than removed. The messages are fused and projected by the dual heads (module 7) into two branch-specific representations,

$$R^b_{t,i} = f^b_{\mathrm{route}}\Big(\sum_{k=1}^{6} w^b_{t,i,k} M_{t,i,k}\Big), \qquad Z^b_t = \mathrm{LN}\big(f^b_{\mathrm{head}}(R^b_t)\big), \qquad b \in \{\mu, \pi\}. \tag{8}$$

Z^μ_t is used to model crime-count intensity, whereas Z^π_t captures factors associated with excess-zero generation and the urban observation process.

3.5 Probabilistic Decoding and Joint Optimisation

Crime counts are discrete, overdispersed, and frequently zero, so the count decoder (module 8) parameterises a zero-inflated negative binomial distribution rather than predicting a single count. Let e_c denote the learnable embedding of crime category c and e_h that of horizon h, and let $\psi^b_{i,h,c} = [Z^b_{t,i}, e_h, e_c]$ for $b \in \{\mu, \pi\}$. The ZINB parameters are

$$\mu_{i,h,c} = \mathrm{softplus}\big(f^\mu_{\mathrm{dec}}(\psi^\mu_{i,h,c})\big) + \epsilon, \qquad \pi_{i,h,c} = \sigma\big(f^\pi_{\mathrm{dec}}(\psi^\pi_{i,h,c})\big), \qquad \theta_c = \mathrm{softplus}\big(f_\theta(e_c)\big) + \epsilon, \tag{9}$$

where μ is the negative-binomial mean, which governs the expected number of events, θ_c a category-specific dispersion, and π the probability of a structural zero, which governs whether any events are observed. The resulting likelihood $p_{\mathrm{ZINB}}(y \mid \mu, \theta, \pi)$ is given in Appendix B.5. During training, the differentiable conditional mean $\widehat{Y}^{\mathrm{mean}}_t = (1-\pi_t) \odot \mu_t$ is used in the MAE term of the prediction loss. At inference, point forecasts are obtained from the median of the predicted ZINB distribution. To encourage Z^π_t to preserve information about activity and reporting conditions, an auxiliary decoder predicts future activity counts, $\widehat{O}^{\mathrm{fut}}_t = \mathrm{softplus}(f_{\mathrm{proxy}}(Z^\pi_t))$, trained with a Poisson negative log-likelihood $\mathcal{L}_{\mathrm{aux}}$. This output is used only for supervision and does not contribute to the crime forecast.

All modules are trained jointly with a single objective that combines the forecast loss, the auxiliary loss, and three structural regularisers. The prediction loss combines the ZINB negative log-likelihood with an MAE term computed from the conditional mean,

$$\mathcal{L}_{\mathrm{pred}} = -\,\mathrm{Mean} \log p_{\mathrm{ZINB}}(Y^{\mathrm{fut}}_t) + \lambda_{\mathrm{MAE}}\,\mathrm{Mean}\,\big|(1-\pi_t) \odot \mu_t - Y^{\mathrm{fut}}_t\big|. \tag{10}$$

Three structural regularisers constrain the learned posteriors (definitions in Appendix B.5). The prior-preserving term $\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(Q_t \| \bar{P}_0)$ discourages the posterior from deviating excessively from the evidence-derived prior. The sparsity term $\mathcal{L}_{\mathrm{sparse}}$ penalises high posterior probability on newly introduced candidates without observational support. The diversity term $\mathcal{L}_{\mathrm{div}}$ projects every channel into a common region–region space and penalises pairwise cosine similarity above a margin m, so that the six channels do not collapse to the same structure. The first two terms are what keeps refinement close to the evidence, as required in Section 1. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \eta_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}} + \lambda_{\mathrm{KL}}\beta_e\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{sparse}}\mathcal{L}_{\mathrm{sparse}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}, \tag{11}$$

where β_e increases linearly from 0 to 1 during a warm-up period to introduce the prior constraint gradually. The loss weights are given in Appendix A.2.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. We construct two datasets from public records of Chicago (N=79 regions) and New York City (N=148), partitioning each city into 3 km grid cells and aggregating crime incidents into daily counts of four categories (Theft, Robbery, Violent Crime, Property Crime). Inputs comprise the crime history, six types of 311 service requests, city-level weather, ten POI categories, four socioeconomic attributes, and Queen adjacency. Models observe L=7 days and forecast the next day. Forecast windows are randomly split 70% /15% /15% into training, validation, and test sets. We report MAE, RMSE, and MAPE on the count scale, averaged with equal weight over the four categories; MAPE uses a lower-bounded denominator $\max(|y|, 2)$. Results are means and standard deviations over five random seeds. Dataset construction and category definitions are detailed in Appendix A.1.

Table 1: Performance Comparison on Chicago and New York Datasets

Method	Chicago			New York		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
EX-Crime	1.28±0.0003	2.03±0.0023	0.41±0.0014	1.50±0.0860	2.95±0.0352	0.32±0.0476
HGAurban	1.17±0.0088	1.47±0.0104	0.49±0.0045	1.41±0.1773	1.78±0.2086	0.60±0.0876
HYSTL	1.09±0.0306	1.65±0.0326	0.39±0.0112	1.09±0.0251	2.09±0.0507	0.33±0.0396
ST-DN	0.88±0.0017	1.46±0.0039	0.32±0.0014	0.63±0.0011	1.28±0.0023	0.19±0.0008
ST-HHOL	1.16±0.0235	1.56±0.0260	0.44±0.0097	1.36±0.0252	1.74±0.0362	0.54±0.0095
MSTHG	0.86±0.0039	1.39±0.0145	0.31±0.0010	0.63±0.0029	1.27±0.0160	0.19±0.0004
ST-FiT	0.88±0.0017	1.39±0.0088	0.32±0.0023	0.65±0.0137	1.29±0.0148	0.19±0.0060
PatchSTG	0.97±0.0042	1.39±0.0051	0.37±0.0026	1.01±0.0915	1.50±0.0159	0.36±0.0549
Ours	**0.80**±0.0021	**1.35**±0.0331	**0.27**±0.0030	**0.60**±0.0023	**1.23**±0.0082	**0.17**±0.0016

Baselines and implementation. We compare against eight recent models: four crime- or urban-event forecasters with graph or hypergraph relations (EX-Crime (Fu et al., 2025), HGAurban (Zhang et al., 2025), HYSTL (Karimova et al., 2025), ST-HHOL (Du et al., 2026)) and four general-purpose spatiotemporal forecasters with learned or adaptive structure (ST-DN (Cao et al., 2025), MSTHG (Zhang et al., 2026), ST-FiT (Lei et al., 2025), PatchSTG (Fang et al., 2025)). All models receive the same inputs and adjacency, and hyperparameters are tuned on the validation set. RELY is trained with AdamW and early stopping on the validation loss. For evaluation, MAE, RMSE, and MAPE are computed using the median of the predicted ZINB distribution as the point forecast. Hyperparameters, hardware, and a sensitivity analysis are given in Appendix A.2 and A.5.

4.2 Overall Comparison

RELY achieves the lowest error on all six metric–city combinations (Table 1), reducing MAE, RMSE, and MAPE relative to the strongest baseline by 6.9%, 2.7%, and 11.7% on Chicago and by 4.5%, 3.2%, and 6.4% on New York. More informative than the average gain is where it concentrates, and three views agree that it is largest where the evidence for any particular relation is weakest. Across metrics, the gain is largest on MAPE, which is dominated by low-count region–type pairs, and smallest on RMSE, which is dominated by occasional large counts. Across categories (Appendix A.3), the gain is largest on the sparsest category, Chicago Robbery (MAE 0.40 to 0.32), and the only metrics on which RELY is not best are RMSE values on high-volume categories. Across region–category pairs stratified by training occurrence frequency (Appendix A.6), the margin over the next-best model is largest in the sparsest bin (MAE 0.117 vs. 0.137). A model that merely fitted peaks better would show the opposite pattern. The concentration of gains in sparse regimes is what Section 1 predicts: sparse histories make the structure, not only the counts, uncertain, and a model that represents this uncertainty gains most exactly there.

Two further observations follow from Table 1. General-purpose forecasters with adaptive structure (MSTHG, ST-DN, ST-FiT) are markedly stronger than the crime-specific graph models, so how relations are learned matters more than which domain-specific relation types are supplied. None of RELY's gain comes from model size: it has 1.16M parameters and the lowest FLOPs and peak memory of the compared models (Appendix A.7).

4.3 Ablation Study

The ablations on Chicago (Figure 2) show that neither modelling the output distribution nor modelling structural uncertainty absorbs the error that the other addresses. Replacing the ZINB decoder with a negative-binomial decoder, which removes the explicit excess-zero component while keeping structural uncertainty, raises MAE from 0.803 to 0.832. Removing structural uncertainty while keeping the ZINB decoder costs at least as much: fixing the completed prior instead of refining it gives 0.828, using the posterior deterministically instead of sampling gives 0.831, and dropping the reliability bias from the routers gives 0.837. Each ablation removes only part of the gain, and the full model is better than every variant on all three metrics, so the two sources of uncertainty are complementary rather than redundant, which is the first contribution of Section 1. Among the relational inputs, removing the four heterogeneous graph channels is the single most harmful change (MAE

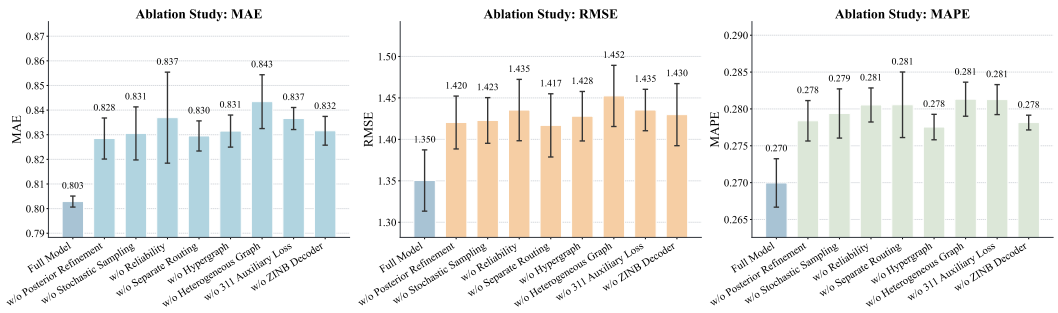


Figure 2: Ablation study on Chicago (MAE, RMSE, and MAPE; mean and standard deviation over seeds). Removing the reliability bias from the routers or removing the heterogeneous graph channels causes the largest degradation.

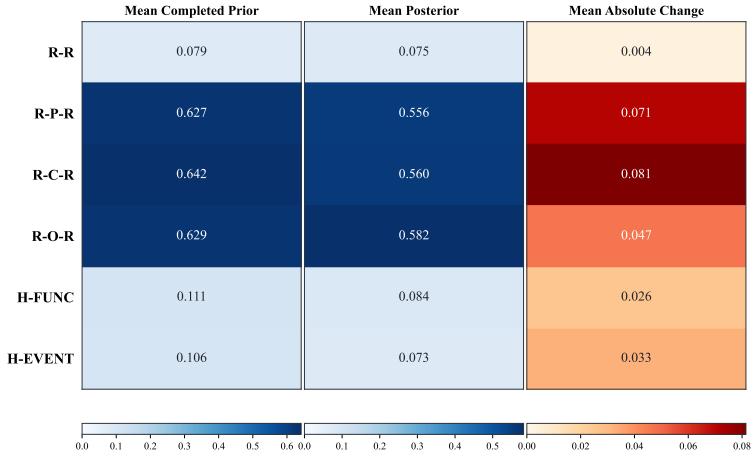


Figure 3: Prior-to-posterior refinement on Chicago. For each of the six relation channels: mean completed prior, mean posterior, and their absolute difference, averaged over all test windows.

0.843), whereas removing the two hypergraph channels costs less (0.831), so the pairwise channels carry the primary relational evidence and the hypergraphs add complementary higher-order information. Sharing one router between μ and π (0.830) and dropping the auxiliary 311 objective (0.837) also degrade all metrics. Reliability is therefore not only diagnostic but load-bearing, since removing it is the largest degradation among the uncertainty-related components. Ablations of the regularisers and of the individual hypergraphs have smaller, mainly stabilising effects (Appendix A.4).

4.4 Structure Refinement and Channel Reliability

Refinement removes rather than invents relations, and it does so selectively. Figure 3 compares, per channel, the mean completed prior with the mean posterior over all test windows on Chicago. Every channel contracts, as the sparsity constraint intends, but the contraction concentrates on the co-occurrence-based channels: the POI-, crime-, and 311-induced relations drop by 0.05–0.08, whereas spatial adjacency is almost untouched (0.079 to 0.075). The model thus treats geographic proximity as a stable prior and re-evaluates semantic relations in every window, which is the behaviour the prior-anchored design of Section 3.2 was meant to produce. The resulting reliabilities differ by almost an order of magnitude, from 0.70 for adjacency to 0.34–0.35 for the two hypergraphs and 0.13–0.18 for the three type-induced paths. This ordering follows from the definition in Section 3.4: adjacency has few candidates but near-decisive posteriors, whereas the type-induced relations admit every region–type pair as a candidate and their posteriors stay close to the prior, so their support is ample but their decisiveness is low. A single fixed structure, or a fusion rule that treats all channels alike, cannot express differences of this size, and the reliability-aware router down-weights the low-reliability channels without discarding them.

4.5 Reliability and Forecast Error

The reliability that RELY assigns to its structures tracks forecast error, which is the second empirical claim of Section 1. For each test window–region pair we compute an effective reliability R_{eff}, the channel reliabilities weighted by the two routers, and sort pairs into five equal-size quantile groups. MAE decreases monotonically from 1.34 in the lowest quantile to 0.31 in the highest (Figure 4), and RMSE (1.95 to 0.68) and MAPE (0.38 to 0.14) follow the same trend (Appendix A.8). Reliability therefore identifies windows and regions whose forecasts are likely to be poor, without any ground-truth structure and before the outcome is observed. One caveat applies: part of the relationship may reflect count level, since low-reliability regions tend to be busier, and disentangling the two is left for future work.

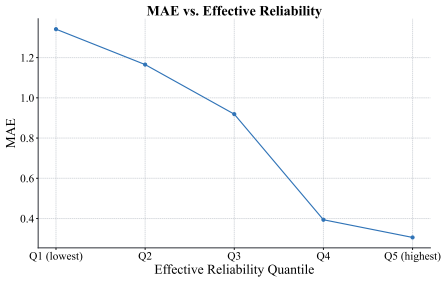


Figure 4: MAE vs. effective reliability quantile on Chicago (Q1 lowest, Q5 highest).

4.6 Parameter-specific Routing

Intensity and excess-zero prediction draw on different relational evidence, which is the third empirical claim of Section 1. Figure 5 shows the mean routing weight of each channel under the two routers on Chicago. The routers agree on the extremes, weighting spatial adjacency highly and the crime-type path lowest, but diverge systematically in between. The intensity router places more weight on the two hypergraphs (functional 0.24 vs. 0.19, event-trajectory 0.17 vs. 0.10), that is, on evidence about which groups of regions share activity patterns. The excess-zero router places more weight on the 311- and POI-induced paths (0.21 vs. 0.16 and 0.16 vs. 0.12), that is, on evidence about activity and reporting conditions. This division is not imposed: the two routers receive identical inputs and differ only in their objective, with the auxiliary 311 objective merely encouraging the π branch to retain activity-related context. The direction of the divergence is interpretable, since what determines how many events to expect differs from what determines whether a region is in an observed state at all. Together with the shared-router and auxiliary-loss ablations of Section 4.3, this confirms that a single fusion rule would force a compromise between the two mechanisms.

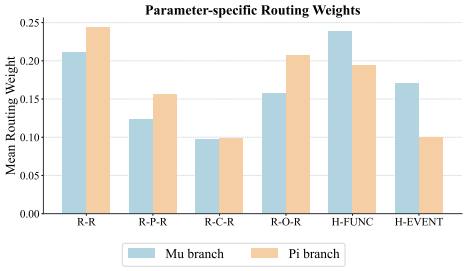


Figure 5: Mean routing weight per relation channel for the intensity (μ) and excess-zero (π) routers on Chicago.

5 Conclusion

We argued that uncertainty about which relations hold between regions is a source of forecast error distinct from the randomness of the counts, and that modelling the output distribution alone does not remove it. RELY addresses this by keeping the graph as a posterior anchored to urban priors, deriving a reliability score from that posterior, and routing relational evidence by reliability separately into the intensity and excess-zero components of a zero-inflated forecast. On Chicago and New York it improves on strong baselines, with the largest gains where structure is most uncertain; its reliability estimates decrease monotonically with forecast error; and the two components learn to use different relations. Two limitations remain: the relationship between reliability and count level has not been disentangled, and abrupt surges in crime counts are still poorly captured. Both are directions for future work.

AI use statement

During the preparation of this manuscript, large language models (LLMs) were used solely to assist with language editing and improve the clarity and readability of the text. All research ideas, method-

ological designs, experiments, analyses, and interpretations of the results were independently developed and carried out by the authors. The authors reviewed and revised all AI-assisted text and take full responsibility for the accuracy, integrity, and originality of the manuscript.

Ethics statement

This work develops a spatiotemporal crime prediction framework to model aggregated crime patterns across regions and crime types. The framework is intended for academic research, analysis of regional crime trends, and providing references for high-level public safety planning. It is not designed for identifying potential offenders, assessing individual criminal risk, or directly guiding operational law enforcement decisions.

We acknowledge that recorded crime data may contain biases associated with reporting practices, law enforcement activity, and socioeconomic conditions. Although this study focuses on aggregated regional patterns, this setting does not eliminate the possibility of reproducing existing biases or stigmatizing particular communities. Model predictions should therefore be interpreted with consideration of data limitations and predictive uncertainty, rather than as definitive judgments about the safety of a region or the characteristics of its residents.

Any practical application should involve context-specific evaluation, appropriate privacy safeguards, and meaningful human oversight, with particular attention to performance disparities across regions and potential inequities in resource allocation. Predictions should not serve as the sole basis for decisions affecting individuals or communities.

Reproducibility statement

To facilitate reproducibility, we make the source code for our proposed method available in an anonymous repository at https://anonymous.4open.science/r/RELY/. The released implementation provides a reference for understanding the model and reproducing the experiments reported in this paper.

References

Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. Spatiotemporal-aware trend-seasonality decomposition network for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11463–11471, 2025.

Keqing Du, Yufan Kang, Xinyu Yang, and Wei Shao. St-hhol: Spatio-temporal hierarchical hypergraph online learning for crime prediction. In *International Conference on Learning Representations*, volume 2026, pp. 143229–143250, 2026.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 307–317. ACM, 2025. doi: 10.1145/3690624.3709177.

C. X. Feng. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications*, 8(8), 2021. doi: 10.1186/s40488-021-00121-4. URL https://doi.org/10.1186/s40488-021-00121-4.

Hongzhu Fu, Yutao Wei, Gege Chen, Xing He, Qiang Gao, and Fan Zhou. Augmented graph information bottleneck with type-aware periodicity heterogeneity for explainable crime prediction. *Information Processing & Management*, 62(6):104227, 2025.

Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 1423–1432, 2018.

Fidan Karimova, Tong Chen, Yu Yang, and Shazia Sadiq. Learning a universal crime predictor with knowledge-guided hypernetworks. *arXiv preprint arXiv:2511.02336*, 2025.

Ourania Kounadi, Alina Ristea, A. Araujo, et al. A systematic review on spatial crime forecasting. *Crime Science*, 9(7), 2020. doi: 10.1186/s40163-020-00116-7. URL https://doi.org/10.1186/s40163-020-00116-7.

Zhenyu Lei, Yushun Dong, Jundong Li, and Chen Chen. ST-FiT: Inductive spatial-temporal forecasting with limited training data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12031–12039, 2025. doi: 10.1609/aaai.v39i11.33310.

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

Alessandro Manenti, Daniele Zambon, and Cesare Alippi. Learning latent graph structures and their uncertainty. *arXiv preprint arXiv:2405.19933*, 2024.

Chao Shang, Jie Chen, and Jinbo Bi. Discrete graph structure learning for forecasting multiple time series. *arXiv preprint arXiv:2101.06861*, 2021.

Jiao Sun, Mingxuan Yue, Zongyu Lin, Xiaochen Yang, Luciano Nocera, Gabriel Kahn, and Cyrus Shahabi. Crimeforecaster: Crime prediction by exploiting the geographical neighborhoods' spatiotemporal dependencies. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 52–67. Springer, 2020.

Chenyu Wang, Zongyu Lin, Xiaochen Yang, Jiao Sun, Mingxuan Yue, and Cyrus Shahabi. Hagen: Homophily-aware graph convolutional recurrent network for crime forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4193–4200, 2022.

Zepu Wang, Xiaobo Ma, Huajie Yang, Weimin Lyu, Yang Liu, Peng Sun, and Sharath Chandra Guntuku. Uncertainty-aware crime prediction with spatial temporal multivariate graph neural networks. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Liefeng Bo, Xiyue Zhang, and Tianyi Chen. Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning. *arXiv preprint arXiv:2201.02435*, 2022.

K. K. W. Yau, K. Wang, and A. H. Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452, 2003. doi: 10.1002/bimj.200390024. URL https://doi.org/10.1002/bimj.200390024.

Ping Zhang, Jiayu Leng, Liang Yang, Anchen Li, Xiaochun Cao, and Riting Xia. Multi-source information driven spatio-temporal hypergraph learning for traffic forecasting. In *Proceedings of the ACM Web Conference 2026 (WWW)*, pp. 580–591. ACM, 2026. doi: 10.1145/3774904.3792119.

Qianru Zhang, Xinyi Gao, Haixin Wang, Dong Huang, Siu-Ming Yiu, and Hongzhi Yin. HGAurban: Heterogeneous graph autoencoding for urban spatial-temporal learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 4139–4148. ACM, 2025. doi: 10.1145/3746252.3761383.

Binbin Zhou, Hang Zhou, Weikun Wang, Liming Chen, Jianhua Ma, and Zengwei Zheng. Hdm-gnn: A heterogeneous dynamic multi-view graph neural network for crime prediction. *ACM Transactions on Sensor Networks*, 2024.

A Appendix

A.1 Dataset Construction

Each city is partitioned into 3 km square grid cells in the EPSG:5070 projection, retaining cells whose overlap with the city boundary is at least 25%, which yields 79 regions for Chicago (607 km^2) and 148 for New York (1,219 km^2). Crime records (Chicago Data Portal; NYC Open Data) are

first mapped to six intermediate categories (assault, burglary, theft, motor-vehicle theft, robbery, drugs/narcotics) to reconcile the two cities' label systems, then aggregated into four targets: Theft, Robbery, Violent Crime (assault and drugs/narcotics), and Property Crime (burglary and motor-vehicle theft), giving 177,136 incidents in Chicago and 311,942 in New York. 311 service requests are mapped to six types (graffiti/vandalism, abandoned vehicles, lighting outages, traffic-signal issues, sanitation, rodents/pests; 258,392 and 352,113 requests). Both are aggregated into daily region-level counts. POIs come from an OpenStreetMap snapshot filtered by a tag whitelist and grouped into ten categories (nightlife/alcohol, transport hubs, parking/fuel, retail/grocery, cash/finance, education, healthcare, public safety, lodging, leisure/entertainment; 16,051 and 24,136 POIs). Static attributes (mean years of education, per-capita income, population density, unemployment rate) are derived from ACS five-year estimates by area-weighting block groups into grid cells, with missing values imputed by the city median. Weather comprises six daily city-level variables (mean, maximum, and minimum temperature, precipitation, snowfall, snow depth). Spatial adjacency uses Queen contiguity.

A.2 Implementation Details

Input encoding. For the dynamic view, log-transformed crime and activity counts are concatenated with standardised city-level weather,

$$X_t^{\mathrm{tem}} = \left[\log(1+Y_t^{\mathrm{his}}), \log(1+O_t^{\mathrm{his}}), \widetilde{W}_t^{\mathrm{his}}\right] \in \mathbb{R}^{N\times L\times(C+O+W)}, \tag{12}$$

and the TCN's hidden state at the end of the window gives $T_t = f_{\mathrm{TCN}}(X_t^{\mathrm{tem}})$. For the static view, standardised attributes $\widetilde{X}^{\mathrm{sta}}$ and log-transformed, column-normalised POI counts $\widetilde{X}^{\mathrm{poi}} = \mathrm{NormCol}(\log(1+X^{\mathrm{poi}}))$ are projected by a shared MLP, $S_t = \mathrm{LN}(f_{\mathrm{static}}([\widetilde{X}^{\mathrm{sta}}, \widetilde{X}^{\mathrm{poi}}]))$, where $\mathrm{NormCol}$ divides each column by its maximum (all-zero columns are left unchanged) and LN denotes layer normalisation.

Training. RELY is implemented in PyTorch 2.5.1 and trained on a single NVIDIA RTX 4090. We use AdamW with learning rate 10^{-3}, weight decay 10^{-3}, gradient-norm clipping at 2.0, one window per batch, at most 20 epochs, and early stopping with patience 4 on the total validation loss; the best validation checkpoint is restored for testing. Representation dimensions are d_t=32, d_g=64, d_h=128, d_r=128, d_μ=128, d_π=64, with 16-dimensional category and horizon embeddings; the TCN has three blocks with kernel size 2, exponentially increasing dilation $d_\ell = 2^\ell$, two dilated causal convolutions per block each followed by GELU and dropout (0.05), a residual connection with a 1×1 convolution when channel widths differ, and layer normalisation. Structure learning uses k_0=k_s=8, $p_{\min}$=0.5, $p_{\max}$=0.95, p_{new}=0.05, and a structure temperature annealed linearly from 1.0 to 0.2. When computing reliability, posteriors are clipped to $[\epsilon, 1-\epsilon]$ inside the entropy and the score $c(q)$ is clipped to $[\epsilon, 1]$; the router MLP f^b_{route} consists of a linear layer, GELU, dropout, and layer normalisation. Loss weights are λ_{MAE}=0.5, η_{aux}=0.05, λ_{KL}=0.001, $\lambda_{\mathrm{sparse}}$=$\lambda_{\mathrm{div}}$=0.01, diversity margin m=0.9, and a KL warm-up over the first 30% of epochs.

A.3 Per-category Results

Table 2: Performance Comparison on New York Dataset by Crime Type

Method	Theft			Robbery			Violent Crime			Property Crime		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
EX-Crime	3.26±0.0119	6.80±0.0098	0.49±0.0046	0.51±0.1481	0.82±0.0300	0.20±0.0759	1.79±0.0255	3.51±0.1027	0.39±0.0288	0.45±0.1606	0.67±0.0106	0.18±0.0812
HGAurban	2.11±0.3382	2.95±0.3848	0.70±0.1723	0.95±0.0727	1.06±0.0685	0.48±0.0360	1.67±0.3115	2.09±0.4090	0.74±0.1419	0.93±0.1487	1.01±0.1423	0.47±0.0747
HYSTL	2.23±0.0253	4.55±0.0942	0.57±0.0167	0.37±0.0158	0.70±0.0075	0.17±0.0761	1.41±0.0128	2.51±0.0277	0.44±0.0168	0.34±0.0465	0.61±0.0734	0.15±0.0486
ST-DN	1.19±0.0121	2.21±0.0190	0.28±0.0046	0.26±0.0037	0.65±0.0082	**0.11±0.0013**	0.86±0.0046	1.68±0.0093	0.25±0.0037	**0.22±0.0022**	**0.58±0.0030**	0.11±0.0010
ST-HHOL	2.00±0.0531	2.87±0.0574	0.66±0.0229	0.94±0.0143	1.03±0.0138	0.45±0.0072	1.56±0.0299	2.07±0.0706	0.59±0.0080	0.92±0.0050	1.00±0.0046	0.45±0.0025
MSTHG	1.18±0.0005	2.20±0.0045	0.28±0.0010	0.26±0.0010	0.66±0.0037	**0.11±0.0007**	0.85±0.0008	**1.62±0.0043**	0.25±0.0015	0.23±0.0116	0.61±0.0620	0.11±0.0034
ST-FiT	1.23±0.0458	2.23±0.0640	0.29±0.0194	0.27±0.0027	0.66±0.0073	0.12±0.0011	0.87±0.0102	1.66±0.0144	0.26±0.0077	0.23±0.0016	0.62±0.0062	0.11±0.0007
PatchSTG	1.36±0.0457	2.35±0.0456	0.33±0.0436	0.90±0.1488	1.00±0.0827	0.43±0.0796	0.99±0.0653	1.80±0.0734	0.28±0.0322	0.78±0.2763	0.86±0.1848	0.38±0.1394
Ours	**1.13±0.0158**	**2.09±0.0249**	**0.26±0.0027**	**0.24±0.0031**	**0.62±0.0071**	**0.11±0.0009**	**0.81±0.0197**	**1.62±0.0664**	**0.23±0.0066**	**0.22±0.0005**	0.59±0.0027	**0.10±0.0008**

Table 3: Performance Comparison on Chicago Dataset by Crime Type

Method	Theft			Robbery			Violent Crime			Property Crime		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
EX-Crime	1.66±0.0003	2.88±0.0008	0.46±0.0003	0.41±0.0002	0.94±0.0001	0.17±0.0001	2.02±0.0004	2.79±0.0100	0.58±0.0058	1.06±0.0006	1.48±0.0007	0.41±0.0004
HGAurban	1.26±0.0402	1.69±0.0242	0.50±0.0269	0.86±0.0206	0.99±0.0317	0.43±0.0108	1.37±0.0155	1.77±0.0064	0.54±0.0099	1.20±0.0533	1.42±0.0343	0.51±0.0326
HYSTL	1.38±0.0275	2.28±0.0216	0.46±0.0184	0.47±0.0459	0.82±0.0297	0.24±0.0123	1.57±0.0317	2.16±0.0552	0.48±0.0124	0.94±0.0171	1.35±0.0239	0.37±0.0015
ST-DN	1.07±0.0087	1.82±0.0110	0.38±0.0076	0.40±0.0029	0.89±0.0076	0.17±0.0009	1.21±0.0047	1.79±0.0118	0.39±0.0011	0.86±0.0015	1.35±0.0023	0.33±0.0027
ST-HHOL	1.25±0.0451	1.91±0.0348	0.44±0.0212	0.87±0.0084	1.04±0.0065	0.41±0.0044	1.42±0.0333	1.90±0.0455	0.46±0.0117	1.10±0.0221	1.40±0.0331	0.46±0.0096
MSTHG	1.07±0.0035	1.63±0.0076	0.37±0.0014	0.41±0.0172	0.94±0.0637	0.17±0.0061	1.17±0.0019	**1.75±0.0023**	0.37±0.0008	0.80±0.0005	**1.25±0.0039**	0.31±0.0008
ST-FiT	1.10±0.0049	1.67±0.0177	0.37±0.0019	0.40±0.0008	0.85±0.0067	0.18±0.0005	1.19±0.0033	1.77±0.0096	0.39±0.0034	0.83±0.0029	1.27±0.0023	0.33±0.0042
PatchSTG	1.15±0.0083	1.61±0.0071	0.41±0.0103	0.50±0.0165	0.91±0.0483	0.22±0.0088	1.28±0.0131	1.75±0.0091	0.44±0.0087	0.96±0.0162	1.28±0.0264	0.39±0.0088
Ours	**0.98±0.0208**	**1.60±0.1419**	**0.32±0.0075**	**0.32±0.0058**	**0.75±0.0238**	**0.14±0.0007**	**1.13±0.0158**	1.78±0.0114	**0.33±0.0078**	**0.78±0.0152**	1.28±0.0089	**0.29±0.0077**

A.4 Additional Ablations

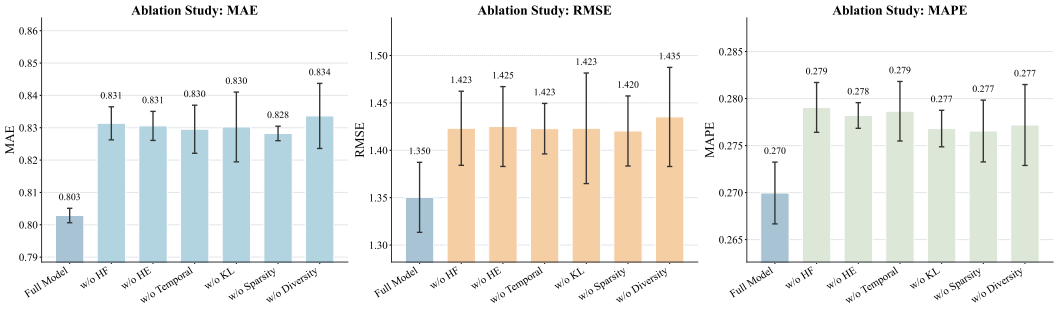


Figure 6: Ablation Appendix

A.5 Hyperparameter Sensitivity

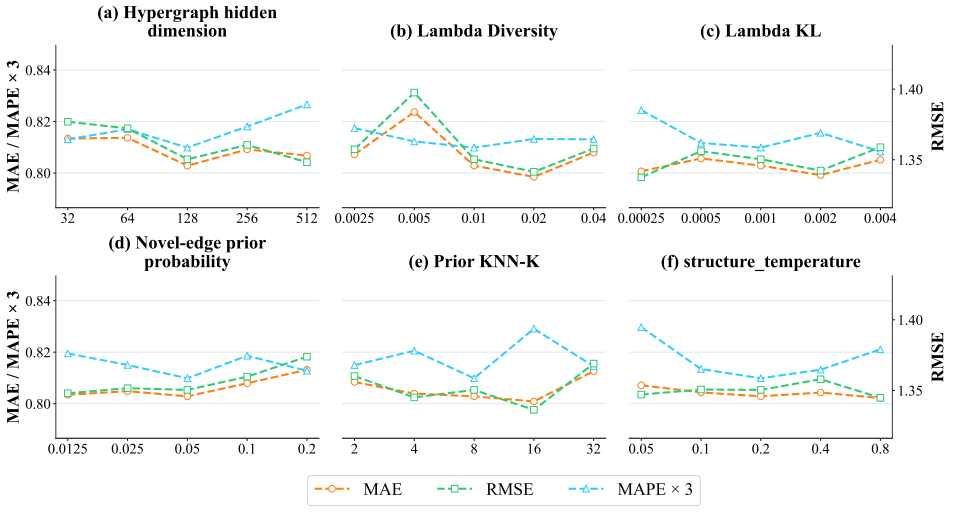


Figure 7: Sensitivity of RELY on Chicago to six hyperparameters (MAE, RMSE, and MAPE×3). Default values: hypergraph hidden dimension 128, $\lambda_{\text{div}} = 0.01$, $\lambda_{\text{KL}} = 0.001$, $p_{\text{new}} = 0.05$, $k_0 = 8$, structure temperature 0.2.

A.6 Low-frequency Evaluation

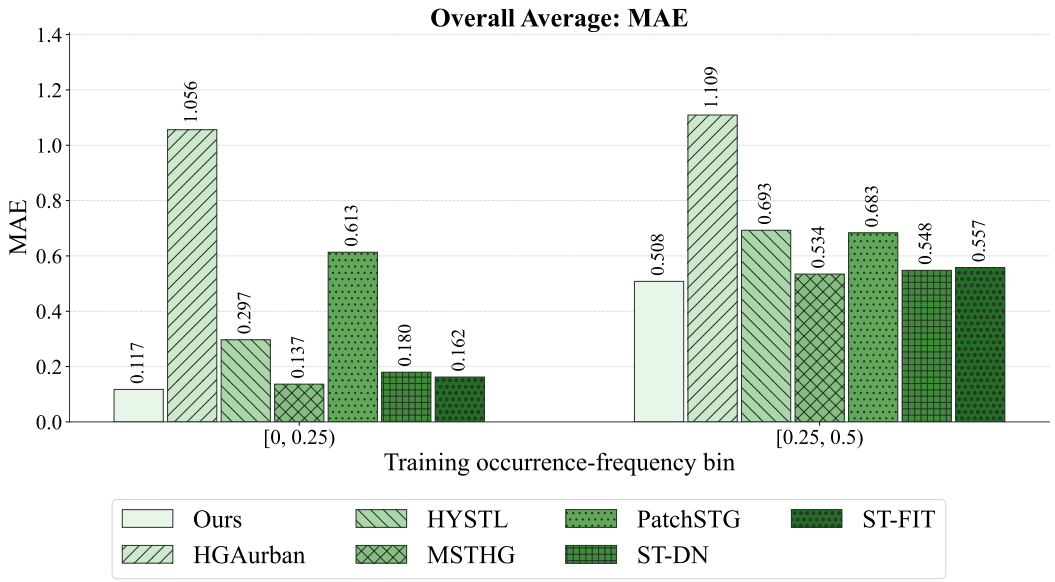


Figure 8: Low Frequency Overall Average MAE

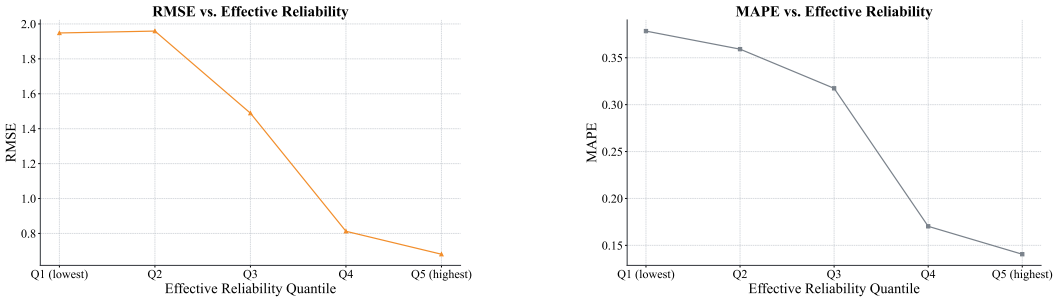
A.7 Efficiency

Table 4: Model Complexity and Efficiency Comparison

Model	Params (M)	FLOPs (G)	Train time (min)	Model-only latency (ms/window)	Warm pipeline (ms/window)	Peak CUDA memory (GiB)
PatchSTG	8.96	27.23	2.27 ± 0.30	19.95 ± 6.32	22.10 ± 0.68	0.71 ± 0.078
ST-DN	18.33	177.59	1.79 ± 0.12	72.81 ± 17.74	65.49 ± 13.29	6.10 ± 0.081
ST-FiT	0.75	3.33	7.33 ± 0.88	68.07 ± 5.66	72.44 ± 3.86	0.50 ± 0.053
ST-HHOL	54.49	77.82	0.69 ± 0.05	7.06 ± 0.33	7.14 ± 0.34	1.31 ± 0.136
Ours	1.16	2.61	1.90 ± 0.32	10.11 ± 1.62	10.10 ± 1.41	0.12 ± 0.000

Table 4 compares model size and cost on Chicago. RELY has 1.16M parameters and 2.6 GFLOPs per window, the lowest FLOPs among the compared models and 47× fewer parameters than ST-HHOL, with a model-only inference latency of 10.1 ms per window and a peak memory of 0.12 GiB. Its training time (1.9 min) is comparable to ST-DN and shorter than PatchSTG and ST-FiT.

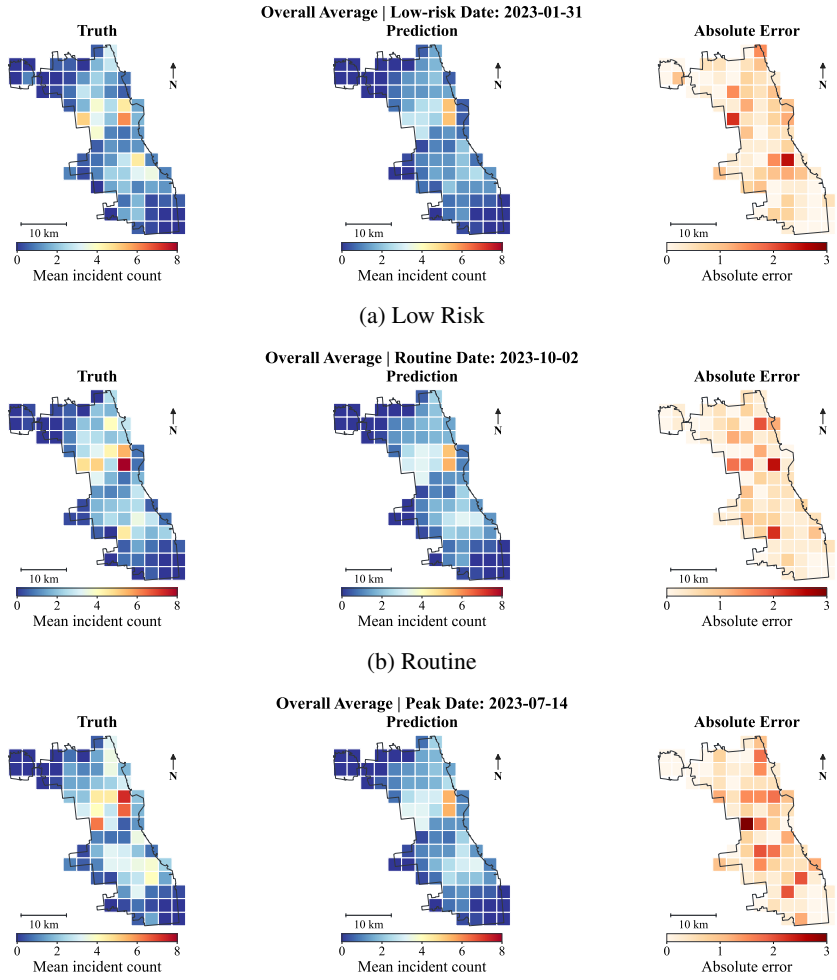
A.8 Reliability and Forecast Error



(a) RMSE (b) MAPE

Figure 9: RMSE and MAPE as a function of effective reliability quantile on Chicago, complementing Figure 4. Per-quantile values (Q1 to Q5): MAE 1.34, 1.17, 0.92, 0.39, 0.31; RMSE 1.95, 1.96, 1.49, 0.81, 0.68; MAPE 0.38, 0.36, 0.32, 0.17, 0.14.

A.9 Prediction Visualization



(a) Low Risk

(b) Routine

(c) High Risk

Figure 10: Prediction Visualization

B METHOD DETAILS

B.1 PRIOR CONSTRUCTION

Region–type priors. The regional profiles used for the region–type relations are

$$V^{RP} = X^{\text{poi}}, \qquad V_t^{RC} = \sum_{\ell=1}^{L} Y_{t,\ell}^{\text{his}}, \qquad V_t^{RO} = \sum_{\ell=1}^{L} O_{t,\ell}^{\text{his}}, \tag{13}$$

where X^{poi} contains untransformed POI counts and $Y_{t,\ell}^{\text{his}}$, $O_{t,\ell}^{\text{his}}$ denote untransformed crime and activity counts at step ℓ of the historical window. The normalised relation intensity is

$$s_{ij}^r = \frac{\log(1 + V_{ij}^r)}{\max_{i'} \log(1 + V_{i'j}^r)}, \tag{14}$$

where columns with a zero denominator are set to zero, and the evidence mask is $M_{0,ij}^r = \mathbb{I}[s_{ij}^r > 0]$.

Anchor-based hypergraph priors. To capture higher-order regional dependencies, we additionally construct two anchor-based hypergraphs, $\mathcal{R}_H = \{HF, HE\}$. The functional hypergraph (HF) is derived from relatively stable regional characteristics,

$$F^{HF} = [\widetilde{X}^{\text{sta}}, \widetilde{X}^{\text{poi}}]. \tag{15}$$

The event hypergraph (HE) is constructed from recent crime and activity trajectories,

$$F_{t,i}^{HE} = \left[\text{vec}\big(\log(1 + Y_{t,i}^{\text{his}})\big), \text{vec}\big(\log(1 + O_{t,i}^{\text{his}})\big)\right]. \tag{16}$$

For each anchor region j, cosine similarity is used to select the k_0 most similar other regions, which together with the anchor itself form an anchor-centred hyperedge; the evidence mask $M_{0,ij}^r$ is one exactly for these members. For $r \in \mathcal{R}_H$, the incidence matrix has shape $N \times N$, with column j corresponding to the hyperedge anchored at region j. Let F_i^r denote the regional features used to construct hypergraph r. Its prior is

$$P_{0,ij}^r = \begin{cases} 1, & i = j, \\ p_{\min} + (p_{\max} - p_{\min}) \dfrac{1 + \cos(F_i^r, F_j^r)}{2}, & i \neq j,\ M_{0,ij}^r = 1, \\ 0, & M_{0,ij}^r = 0. \end{cases} \tag{17}$$

Dynamic priors use only observations from the current historical window, while static priors use spatial relations and regional attributes available at prediction time.

B.2 POSTERIOR REFINEMENT

This section gives the pair representations and scorers summarised in Section 3.2. For RR, we use a symmetric pair representation,

$$\phi_{t,ij}^{RR} = [Z_{t,i}^0 + Z_{t,j}^0,\ |Z_{t,i}^0 - Z_{t,j}^0|,\ Z_{t,i}^0 \odot Z_{t,j}^0]. \tag{18}$$

For a region–type relation $r \in \{RP, RC, RO\}$,

$$\phi_{t,ij}^r = [Z_{t,i}^0,\ E_{t,j}^r,\ |Z_{t,i}^0 - E_{t,j}^r|,\ Z_{t,i}^0 \odot E_{t,j}^r], \tag{19}$$

where the type representation combines a learnable embedding with prior-weighted regional context,

$$E_{t,j}^r = e_{j,\text{ref}}^r + W_{\text{ctx}}^r \left(\frac{\sum_i \bar{P}_{0,ij}^r Z_{t,i}^0}{\sum_i \bar{P}_{0,ij}^r + \epsilon} \right). \tag{20}$$

The scorer g_r is applied to $\phi_{t,ij}^r$ as in Eq. 5.

For each hypergraph relation $r \in \mathcal{R}_H$, we infer posterior membership probabilities between regions and anchor-centred hyperedges. Each column j of the completed incidence prior $\bar{P}_0^r \in [0,1]^{N \times N}$

corresponds to the hyperedge anchored at region j. We first construct a prior-conditioned hyperedge representation by aggregating the initial regional states:

$$B^r_{t,j} = W^r_{\mathrm{ctx}} \left(\frac{\sum_{i=1}^{N} \bar{P}^r_{0,ij} Z^0_{t,i}}{\max\left(\sum_{i=1}^{N} \bar{P}^r_{0,ij}, \epsilon\right)} \right), \qquad r \in \mathcal{R}_H, \tag{21}$$

where W^r_{ctx} is a relation-specific learnable affine projection. Although each hyperedge is indexed by its anchor region, its representation summarises the regions weighted by the completed incidence prior, rather than using only the anchor representation.

A relation-specific scorer then predicts a correction to the prior log-odds:

$$\Delta^r_{t,ij} = g_r\big([Z^0_{t,i}, B^r_{t,j}, |Z^0_{t,i} - B^r_{t,j}|, Z^0_{t,i} \odot B^r_{t,j}]\big), \tag{22}$$

where g_r is an MLP with a scalar output. The posterior incidence probability is defined as

$$Q^r_{t,ij} = \begin{cases} 1, & i = j, \\ \sigma\big(\mathrm{logit}\big(\mathrm{clip}(\bar{P}^r_{0,ij}, \epsilon, 1-\epsilon)\big) + \Delta^r_{t,ij}\big), & i \neq j,\ M^r_{\mathrm{can},ij} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Thus, each anchor retains deterministic membership in its own hyperedge, while the remaining candidate memberships are refined using the current regional representations. The two hypergraph relations use separate context projections and membership scorers.

B.3 Structure Sampling

During training, we apply a Binary-Concrete relaxation to the posterior probabilities:

$$\widetilde{Q}^r_{t,ij} = \sigma\left(\frac{\mathrm{logit}(Q^r_{t,ij}) + \log u^r_{ij} - \log(1 - u^r_{ij})}{\tau} \right), \qquad u^r_{ij} \sim \mathcal{U}(0,1), \tag{24}$$

with an annealed temperature τ. The sampled structures are used for propagation during training, $\widetilde{A}^r_t = \widetilde{Q}^r_t$ for $r \in \mathcal{R}_A$ and $\widetilde{H}^r_t = \widetilde{Q}^r_t$ for $r \in \mathcal{R}_H$; at inference time the posterior probabilities are used directly, $\widetilde{A}^r_t = Q^r_t$ and $\widetilde{H}^r_t = Q^r_t$. Candidate masks and anchor self-memberships are preserved under sampling, and the RR posterior remains symmetric without self-loops.

B.4 Propagation Operators

Region–region. For the region–region relation, we employ probability-gated attention. Query-key compatibility is computed as

$$e^{RR}_{ij} = \frac{(W^{RR}_Q Z^0_{t,i})^\top (W^{RR}_K Z^0_{t,j})}{\sqrt{d_g}}, \tag{25}$$

and the learned structural probability is incorporated as an additive log-prior:

$$\alpha^{RR}_{ij} = \frac{\mathbb{I}[\widetilde{A}^{RR}_{t,ij} > 0]\, \widetilde{A}^{RR}_{t,ij} \exp(e^{RR}_{ij})}{\sum_\ell \mathbb{I}[\widetilde{A}^{RR}_{t,i\ell} > 0]\, \widetilde{A}^{RR}_{t,i\ell} \exp(e^{RR}_{i\ell})}. \tag{26}$$

For regions without supported neighbours, all attention weights are set to zero. The resulting regional message is

$$M^{RR}_{t,i} = \mathrm{LN}\left(Z^0_{t,i} + W^{RR}_{\mathrm{out}} \sum_j \alpha^{RR}_{ij} W^{RR}_V Z^0_{t,j} \right). \tag{27}$$

Thus, message propagation is jointly determined by latent semantic compatibility and inferred structural support.

Region–type–region. For $r \in \{RP, RC, RO\}$, propagation is performed through the corresponding type nodes.

Let $e^r_{j,\text{prop}}$ denote a learnable type embedding used for propagation, with parameters separate from those used for structure refinement. Both propagation directions use structure-weighted attention:

$$e^{R\to T,r}_{ij} = \frac{(W^{r,1}_Q Z^0_{t,i})^\top (W^{r,1}_K e^r_{j,\text{prop}})}{\sqrt{d_g}}, \tag{28}$$

$$\alpha^{R\to T,r}_{ij} = \frac{\widetilde{A}^r_{t,ij} \exp(e^{R\to T,r}_{ij})}{\sum_\ell \widetilde{A}^r_{t,\ell j} \exp(e^{R\to T,r}_{\ell j})}, \tag{29}$$

$$U^r_{t,j} = e^r_{j,\text{prop}} + \sum_i \alpha^{R\to T,r}_{ij} W^{r,1}_V Z^0_{t,i}, \tag{30}$$

$$e^{T\to R,r}_{ij} = \frac{(W^{r,2}_Q Z^0_{t,i})^\top (W^{r,2}_K U^r_{t,j})}{\sqrt{d_g}}, \tag{31}$$

$$\alpha^{T\to R,r}_{ij} = \frac{\widetilde{A}^r_{t,ij} \exp(e^{T\to R,r}_{ij})}{\sum_\ell \widetilde{A}^r_{t,i\ell} \exp(e^{T\to R,r}_{i\ell})}. \tag{32}$$

Attention weights are set to zero when the corresponding denominator is zero.

Regions are first aggregated into type representations, which are then propagated back to regions:

$$M^{\text{path}(r)}_{t,i} = \text{LN}\left(Z^0_{t,i} + W^r_{\text{out}} \sum_j \alpha^{T\to R,r}_{ij} W^{r,2}_V U^r_{t,j} \right), \tag{33}$$

$$\text{path}(RP) = RPR, \qquad \text{path}(RC) = RCR, \qquad \text{path}(RO) = ROR. \tag{34}$$

This induces region-level dependencies mediated by POI functions, crime composition, and 311 activity.

Hypergraph. For each hypergraph relation $r \in \mathcal{R}_H$, let $\widetilde{H}^r_t$ denote the learned soft incidence matrix. The hyperedge and region degree matrices are

$$D^r_e = \text{diag}\Big((\widetilde{H}^r_t)^\top \mathbf{1}\Big), \qquad D^r_v = \text{diag}\Big(\widetilde{H}^r_t \mathbf{1}\Big). \tag{35}$$

Region representations are first aggregated into hyperedges and then transformed:

$$K^r_t = \text{LN}\Big(\text{GELU}\Big(f^r_{v2e}\Big((D^r_e)^{-1}(\widetilde{H}^r_t)^\top Z^0_t\Big)\Big)\Big). \tag{36}$$

The hyperedge states are subsequently aggregated back into regions, followed by an affine transformation, activation, and normalisation:

$$U^r_t = \text{LN}\Big(\text{GELU}\Big(f^r_{e2v}\Big((D^r_v)^{-1}\widetilde{H}^r_t K^r_t\Big)\Big)\Big). \tag{37}$$

Finally, the regional messages are projected into the common relational space:

$$M^r_t = f^r_{\text{hg-out}}(U^r_t). \tag{38}$$

Here, f^r_{v2e}, f^r_{e2v}, and $f^r_{\text{hg-out}}$ are learnable affine projections. Dropout is applied to the hyperedge states before region aggregation and to the regional messages before the final projection; it is omitted from the equations for clarity.

B.5 Likelihood and Regularisers

The negative-binomial likelihood is

$$p_{\text{NB}}(y \mid \mu, \theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)} \left(\frac{\theta}{\theta+\mu}\right)^\theta \left(\frac{\mu}{\theta+\mu}\right)^y, \tag{39}$$

and the resulting ZINB distribution is

$$p_{\mathrm{ZINB}}(y \mid \mu, \theta, \pi) = \begin{cases} \pi + (1-\pi)p_{\mathrm{NB}}(0 \mid \mu, \theta), & y = 0, \\ (1-\pi)p_{\mathrm{NB}}(y \mid \mu, \theta), & y > 0. \end{cases} \tag{40}$$

For inference, the point forecast is the discrete median of the predicted ZINB distribution:

$$\widehat{Y}^{\mathrm{med}}_{t,i,h,c} = \min\left\{k \in \mathbb{N}_0 : F_{\mathrm{ZINB}}(k \mid \mu_{i,h,c}, \theta_c, \pi_{i,h,c}) \geq \tfrac{1}{2}\right\}, \tag{41}$$

where F_{ZINB} denotes the cumulative distribution function of the predicted ZINB distribution. This discrete quantile is used for inference and metric computation; the MAE term in the training objective is computed from the differentiable conditional mean.

The auxiliary activity decoder is trained with Poisson negative log-likelihood,

$$\mathcal{L}_{\mathrm{aux}} = \mathrm{Mean}\left[\widehat{O}^{\mathrm{fut}}_t - O^{\mathrm{fut}}_t \odot \log(\widehat{O}^{\mathrm{fut}}_t + \epsilon)\right]. \tag{42}$$

We regularise the learned structures with three complementary terms. The prior-preserving term $\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(Q_t \| \bar{P}_0)$ discourages the learned posterior structure from deviating excessively from the evidence-derived prior. The sparsity term $\mathcal{L}_{\mathrm{sparse}} = \mathrm{Mean}_{r,\,(i,j)\in\Omega^r_{\mathrm{new}}} Q^r_{t,ij}$ penalises newly introduced relations that receive high posterior probability without direct observational support, where

$$\Omega^r_{\mathrm{new}} = \{(i,j) : M^r_{\mathrm{can},ij} = 1,\ M^r_{0,ij} = 0\}. \tag{43}$$

To compare heterogeneous graph and hypergraph relations in a common region–region space, we define

$$G^r = \begin{cases} \dfrac{Q^{RR}_t + (Q^{RR}_t)^\top}{2}, & r = RR, \\ Q^r_t (Q^r_t)^\top, & r \in \{RP, RC, RO, HF, HE\}, \end{cases} \qquad \bar{G}^r = \mathrm{OffDiag}(G^r), \tag{44}$$

where $\mathrm{OffDiag}(\cdot)$ sets the diagonal entries to zero. The diversity term

$$\mathcal{L}_{\mathrm{div}} = \frac{2}{|\mathcal{R}|(|\mathcal{R}|-1)} \sum_{r<r'} \max\Big(0, \cos\Big(\mathrm{vec}(\bar{G}^r), \mathrm{vec}(\bar{G}^{r'})\Big) - m\Big), \tag{45}$$

where $\mathcal{R} = \mathcal{R}_A \cup \mathcal{R}_H = \{RR, RP, RC, RO, HF, HE\}$, discourages different relation channels from collapsing to overly similar structures.