From Marks to Narratives: Language-Augmented Spatio-Temporal Point Processes

Zheng Dong [1] **Xiaoyue Liu** [1]

Abstract

Spatio-temporal point processes (STPPs) provide a principled framework for forecasting discrete events in continuous time and space, yet most existing models represent additional event information as fixed categorical marks. This abstraction is increasingly restrictive for modern event streams, where events are often accompanied by rich free-form textual descriptions and reducing them into categorical labels can lose valuable information about the underlying event dynamics. Large language models emerge as powerful tools for handling textual context. However, they still lack principled mechanisms for modeling complex spatio-temporal dynamics. We introduce language-augmented spatio-temporal point process (LA-STPP), a framework that leverages rich texts in the past events for future event prediction. More importantly, our framework enables free-form text generation as part of the event forecasting. LA-STPP couples an STPP-based forecaster with a fine-tuned language model via a shared history representation that encodes past event times, locations, and textual content. By conditioning language generation on spatio-temporal dynamics, LA-STPP predicts not only the timing and location of future events but also their semantic content. Experiments show that LA-STPP largely improves text prediction quality over text-only baselines while preserving superior spatio-temporal forecasting capability, suggesting a path toward language-capable world models that forecast when, where, and what will happen next.

1. Introduction

Point processes are widely used to model asynchronous event data in real-world applications. These data typically consist of sequences of events, each recording when and where the event occurred, along with additional information that goes beyond the spatio-temporal coordinates, commonly referred to as *marks*, *e.g.*, specific type of offense in crime forecasting (Dong et al., 2024), earthquake magnitudes in seismology (Ogata, 1988), or patient vitals in medical research (Liao et al., 2026). To effectively capture the complex dependencies inherent in marked spatio-temporal events, recent deep learning-based point process models (Mei & Eisner, 2017; Zuo et al., 2020; Chen et al., 2021; Wang et al., 2020; Dong et al., 2023a; Zhu et al., 2022) have been developed. These neural frameworks have demonstrated success in modeling intricate event dynamics and serve as principled tools for event forecasting.

Nonetheless, existing neural point process models face significant challenges when applied to modern applications. The rise of complex systems increasingly involve event marks that take the form of texts, encoding rich information about the underlying event dynamics, *e.g.*, police reports details criminal activity (Zhu & Xie, 2022). This unstructured textual data demands more expressive architectures for capturing the event generation mechanisms. Existing neural point processes cannot accommodate such data because they handle only categorical or continuous scalar marks. Reducing rich textual content to categorical labels discards information that may be critical for accurate forecasting. A more fundamental limitation is that current neural point processes predict only event times, locations, or categorical marks and have no capability to generate free-form text. This hinders their potential for wider applications in modern forecasting tasks, where text generation is indispensable for the development of language-based world models.

Large language models (LLMs) provide powerful text understanding and generation capabilities, yet they lack principled mechanisms for modeling complex spatio-temporal event dynamics. Recent efforts to bridge two paradigms (Liu & Quan, 2024; Kong et al., 2025) improve event representations through language embeddings. However, they still rely on the LLM's innate capability to capture spatio-temporal structure and treat event marks as discrete types without handling rich event-level text.

We introduce LA-STPP, a language-augmented spatio-

[1]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, USA. Correspondence to: Zheng Dong <zdong76@gatech.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

temporal point process to address this gap. LA-STPP enables the model to take full advantage of past textual information in event marks for spatio-temporal forecasting, while simultaneously generating free-form descriptions for future events. The framework achieves this by coupling principled STPP components with a decoder-only LLM through a shared causal history encoding. Our contributions are:

1. We introduce the first spatio-temporal point process that leverage free-form text in event forecasting with free-form event description generation. This extends the mark space from categorical labels to unrestricted natural language.
2. The framework generates time-aware natural language event descriptions that encode complex spatio-temporal dependencies, rather than producing text independently of the event dynamics.
3. Experiments on spatio-temporal event datasets with textual marks demonstrate that our spatio-temporal-conditioned text prediction outperforms text-only LLMs, and that text-driven history encoding achieves superior forecasting performance compared to categorical-mark baselines.

2. Background

Spatio-temporal point processes (STPPs) (Reinhart, 2018) model a sequence of discrete events $\mathcal{H} = \{(t_i, s_i)\}_{i=1}^n$, where $t_i \in [0, T]$ is the time and $s_i \in \mathcal{S} \subset \mathbb{R}^{d_S}$ is the location of ith event. The event number n is also random. Given the history $\mathcal{H}_t = \{(t_i, s_i) \in \mathcal{H} | t_i < t\}$, an STPP is fully characterized by the conditional intensity function

$$\lambda(t, s \mid \mathcal{H}_t) = \lim_{\Delta t \downarrow 0, \Delta s \downarrow 0} \frac{\mathbb{E}\left[\mathbb{N}([t, t+\Delta t] \times B(s, \Delta s)) \mid \mathcal{H}_t\right]}{|B(s, \Delta s)| \Delta t},$$

where $B(s, \Delta s)$ is a ball centered at $s \in \mathbb{R}^{d_S}$ with radius Δs, and the counting measure $\mathbb{N}$ is defined as the number of events occurring in $[t, t+\Delta t] \times B(s, \Delta s) \subset \mathbb{R}^{d_S+1}$. Naturally $\lambda(t, s|\mathcal{H}_t) \geq 0$ for any arbitrary t and s. By conditioning on time aspect, we can re-write the intensity function as $\lambda(t, s|\mathcal{H}_t) = \lambda(t|\mathcal{H}_t) \cdot f(s|t, \mathcal{H}_t)$, where $\lambda(t|\mathcal{H}_t)$ is the ground intensity of the temporal process, and f is the conditional event spatial density. The log-likelihood of observing $\mathcal{H}$ on $[0, T] \times \mathcal{S}$ is given by (Daley & Vere-Jones, 2008)

$$\ell(\mathcal{H}) = \sum_{i=1}^{n} \log \lambda(t_i, s_i | \mathcal{H}_{t_i}) - \int_0^T \lambda(t|\mathcal{H}_t) dt \quad (1)$$

3. Method: language-augmented STPP

We consider event sequences with text information for each event, denoted as $\mathcal{H} = \{(t_i, s_i, \text{text}_i)\}_{i=1}^n$ where text_i is a variable-length free-form textual description from vocabulary $\mathcal{V}$. Our goal is two-fold: (1) model the event spatio-temporal dynamics via the intensity λ, and (2) predict the next event's time, location, and text given history.

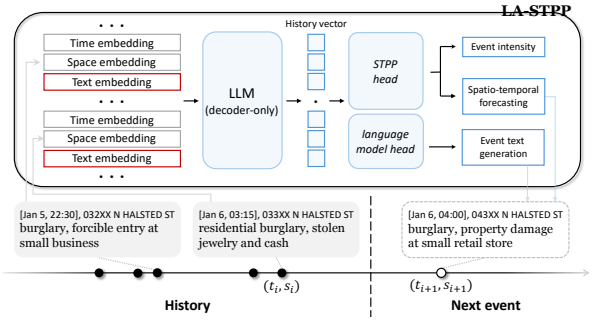

Figure 1. LA-STPP architecture.

Given history, we model the joint event distribution via decomposed marked point process intensity (Reinhart, 2018):

$$\lambda(t, s, \text{text} \mid \mathcal{H}_t) = \lambda(t \mid \mathcal{H}_t) \cdot f(s \mid t, \mathcal{H}_t) \cdot p(\text{text} \mid t, s, \mathcal{H}_t).$$

where $\lambda(t \mid \mathcal{H}_t)$ and f are the conditional temporal intensity and spatial density, respectively, and $p(\text{text} \mid \cdot)$ is the conditional text likelihood. This factorization separates the modeling challenge into three components that operates on a shared history vector representation. Figure 1 shows an overview of the LA-STPP architecture.

3.1. History Vector

The history encoder jointly represent temporal dynamics, spatial patterns, and textual semantics from the event history. We adopt a decoder-only LLM and inject spatio-temporal event information directly at the embedding level. This preserves the continuous nature of times and locations while leveraging the LLM's capacity for sequential reasoning.

Event embedding. Each event $(t_i, s_i, \text{text}_i)$ is mapped to: $\mathbf{E}_i = [\text{TE}(t_i),\ \text{SE}(\mathbf{s}_i),\ \text{LLM\_Emb}(\text{text}_i)]$. The temporal embedding $\text{TE}(t_i) \in \mathbb{R}^D$ encodes the continuous timestamp using sinusoidal positional encoding:

$$\text{TE}(t)_{2j} = \sin(t/10000^{2j/D}),$$
$$\text{TE}(t)_{2j+1} = \cos(t/10000^{2j/D}),$$

where $j = 0, 1, \ldots, D/2 - 1$. The spatial embedding $\text{SE}(s_i) = \mathbf{W}_s s_i + \mathbf{b}_s \in \mathbb{R}^D$ is a learned linear projection, where $\mathbf{W}_s \in \mathbb{R}^{D \times 2}$ and $\mathbf{b}_s \in \mathbb{R}^D$ are trainable parameters. The text embedding $\text{LLM\_Emb}(\text{text}_i) \in \mathbb{R}^{L_i \times D}$ is obtained by tokenizing text_i and passing the token IDs through the LLM's frozen embedding layer. Each event contributes L_i text tokens (truncated to a maximum of $L_{\max}$ tokens).

History vector extraction. We concatenate embeddings of all events in the sequence and get $\mathbf{X} = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_N] \in \mathbb{R}^{(\sum_{i=1}^n L_i + 2n) \times D}$. We then input it to the LLM, producing

hidden states at every position. The history vector $\mathbf{h}_i \in \mathbb{R}^H$ for observed history up to i-th event is defined as the hidden state at the last token position (the final text token) of event i. This vector encodes the observed history and serves as the shared representation for modeling intensity and predicting next event $i{+}1$. Note that we apply causal attention masking to prevent future events from influencing previous events.

3.2. Spatio-Temporal Intensity

Given history vector $\mathbf{h}_i$, the conditional intensity for the next event at time $t > t_i$ is modeled as:

$$\lambda(t \mid \mathcal{H}_t) = \text{softplus}\big(\alpha \cdot (t - t_i) + \mathbf{w}^\top \mathbf{h}_i + b\big)\,,$$

where $\alpha \in \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^H$, and $b \in \mathbb{R}$ are learnable parameters. The softplus activation ensures a non-negative intensity. To model conditional spatial event density, we model it as a K-component Gaussian mixture:

$$f(s \mid t, \mathcal{H}_t) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(s;\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k).$$

To reflect the dependency on the history, all parameters in the mixture model are predicted based the history vector $\mathbf{h}_i$. For each component k, the weight π_k and the mean vector $\boldsymbol{\mu}_k = [\mu_k^{(1)}, \mu_k^{(2)}]^\top \in \mathbb{R}^2$ are directly learned. We model the covariance matrix $\boldsymbol{\Sigma}_k$ via Cholesky decomposition as

$$\boldsymbol{\Sigma}_k = \mathbf{L}_k \mathbf{L}_k^\top, \quad \mathbf{L}_k = \begin{bmatrix} e^{a_k} & 0 \\ b_k & e^{c_k} \end{bmatrix},$$

where (a_k, b_k, c_k) are learnable parameters. We use a single linear network $\mathbf{W}_{\text{gmm}} \in \mathbb{R}^{H \times 6K}$ that outputs $K \times 6$ values as all the mixture parameters.

3.3. Conditional Text Generation

The design of the shared representation in LA-STPP eliminates the need for a separate model or LLM forward pass for the text generation capability. During the forward pass of history vector computation, the LLM produces hidden states at every position in the input sequence, including all text token positions. Thus, we can apply the language model head to project the embeddings to vocabulary at all text token position, predicting the next text token based on:

$$p(x_{i,j+1} \mid x_{i,\le j}, t_i, s_i, \mathcal{H}_{t_i}) = \text{softmax}(\mathbf{W}_{\text{lm}} \cdot \mathbf{h}_{i,j})$$

where $x_{i,j}$ is the j-th text token in i-th event, $\mathbf{h}_{i,j}$ is the hidden state at the corresponding token position, and $\mathbf{W}_{\text{lm}}$ is the language model head. Therefore, the text generation considers not only the text tokens but also the spatio-temporal knowledge of previous events.

3.4. Model Training

The STPP model is trained through Maximum likelihood estimation (MLE) (Reinhart, 2018). The final log-likelihood of observing $\mathcal{H}$ can be expressed as

$$\ell(\mathcal{H}) = \sum_{i=1}^{n} \log \lambda\left(t_i | \mathcal{H}_{t_i}\right) + \sum_{i=1}^{n} \log f(s|t, \mathcal{H}_t) - \int_0^T \lambda(t|\mathcal{H}_t)dt + \sum_{(i,j)\in\mathcal{T}} \log p(x_{i,j+1} \mid x_{i,\le j}, t_i, s_i, \mathcal{H}_{t_i}),$$

where $\mathcal{T}$ is the set of all valid text prediction positions. A higher log-likelihood suggests a better model goodness-of-fit of the event dynamics. We also add auxiliary losses $\mathcal{L}_{\text{time}}$ and $\mathcal{L}_{\text{loc}}$ for time and location prediction (see Appendix B for details). The final training objective combines both as: $\mathcal{L} = -\ell(\mathcal{H}) + \beta_{\text{time}}\mathcal{L}_{\text{time}} + \beta_{\text{loc}}\mathcal{L}_{\text{loc}}$. The weights β_{time} and β_{loc} balance the contribution of the two-fold goal.

Fine-tuning strategy. The LLM backbone is fine-tuned with LoRA (Hu et al., 2022), keeping the majority of parameters frozen while adapting the model to encode domain-specific event dynamics. Other parameters for embedding and intensity are all trained from scratch. The LM head is shared with the pretrained LLM output projection.

4. Experiments

Datasets. We evaluate on three event sequence datasets: (1) *Amazon Reviews*: user review sequences with timestamps and review text; (2) *GitHub Events*: repository activity streams with timestamps and event descriptions; (3) *Chicago Crime*: crime incident reports with timestamps, geographic coordinates, and textual descriptions.

Implementation. We use TinyLlama-1.1B (Zhang et al., 2024) as the default backbone with LoRA (rank 16, α=16). Training uses Adam with learning rate 5×10^{-4}, batch size 4, gradient checkpointing, and 20 MC samples for the compensator. Further details are in Appendix B and C.

Baselines. For event modeling and forecasting, we compare against state-of-the-art neural point process models, including NHP (Mei & Eisner, 2017), THP (Zuo et al., 2020), SAHP (Zhang et al., 2020), NSTPP (Chen et al., 2021), DSTPP (Wang et al., 2020), and TPP-LLM (Liu & Quan, 2024). We also compare against three text-only baselines on model text generation quality: Random, Claude Sonnet 4 (Anthropic, 2024), and LoRA Fine-tuned TinyLlama on event text without point process structure.

Metrics. We report log-likelihood on test set to reflect model goodness-of-fit of the data. We use time and location MAE to measure the model prediction power. For generated text quality assessment, we use a retrieval-based protocol since exact-match on free-form text is unreliable: given history and true timestamp/location, we rank the correct next event text against K=50 distractors sampled from the test split, reporting Mean Reciprocal Rank (MRR), Hits@1, and Hits@5. Higher values suggests closer text prediction to the ground truth. Details are provided in Appendix B.

Table 1. Spatio-temporal event modeling and forecasting performance. Time MAE in hours. Location MAE in km. Best results in **bold**. The numbers in the parentheses are standard deviation from three independent runs.

Method	*Amazon reviews*		*GitHub events*		*Chicago Crime*		
	Testing ℓ (↑)	Time MAE (↓)	Testing ℓ (↑)	Time MAE (↓)	Testing ℓ (↑)	Time MAE (↓)	Loc. MAE (↓)
NHP	-4.597 (0.102)	9.48	0.373 (0.069)	1.41	-5.560 (0.027)	8.87	/
THP	-9.942 (0.084)	8.93	-0.469 (0.055)	1.24	-4.341 (0.032)	7.62	/
SAHP	-6.638 (0.061)	8.25	-0.024 (0.021)	0.83	-4.942 (0.018)	7.49	/
TPP-LLM	-4.127 (0.072)	7.10	0.185 (0.030)	1.17	-4.578 (0.012)	7.02	/
NSTPP	/	/	/	/	-3.050 (0.065)	7.54	0.89
DSTPP	/	/	/	/	**-2.824** (0.037)	7.35	0.73
LA-STPP	**-3.931** (0.039)	**6.42**	**0.535** (0.031)	**0.62**	-3.172 (0.010)	**6.81**	**0.68**

Table 2. Text generation performance. Best results in **bold**.

Dataset	Method	MRR (↑)	Hits@1 (↑)	Hits@5 (↑)
Amazon Reviews	Random	0.090	0.020	0.100
	Claude Sonnet 4	0.278	0.172	0.336
	TinyLlama-Fine-tuned	0.491	0.359	0.447
	LA-STPP	**0.700**	**0.628**	**0.782**
GitHub Events	Random	0.090	0.020	0.100
	Claude Sonnet 4	**0.806**	**0.778**	0.838
	TinyLlama-Fine-tuned	0.723	0.655	0.804
	LA-STPP	0.770	0.700	**0.842**
Chicago Crime	Random	0.090	0.020	0.100
	Claude Sonnet 4	0.665	0.528	0.852
	TinyLlama-Fine-tuned	0.479	0.312	0.583
	LA-STPP	**0.683**	**0.536**	**0.896**

4.1. Baseline Comparisons

Spatio-temporal forecasting. Table 1 presents the quantitative results on spatio-temporal data modeling. On temporal-only textual dataset, LA-STPP achieves the best log-likelihood and time MAE among all methods, outperforming both classical neural point processes and TPP-LLM. This confirms that our text-aware history encoding actively leverages the rich semantic content in event texts, providing informative features for modeling temporal dynamics that categorical marks cannot capture. On the spatio-temporal Chicago Crime dataset, LA-STPP produces competitive goodness-of-fit to the data dynamics against dedicated neural STPP models (NSTPP, DSTPP). While they have no capability to digest textual information, the superior forecasting power of LA-STPP indicates the benefits of incorporating semantic knowledge into event forecasting.

Next-event text generation. Table 2 reports the text generation results. LA-STPP achieves the strongest overall performance and have superior performance on Amazon and Chicago Crime across all metrics. These results demonstrate the benefits of our STPP-principled LLM architecture against text-only approaches in providing more accurate text prediction for events that exhibit intricate spatio-temporal dynamics. The frozen pre-trained LLM (Claude Sonnet 4) performs well on GitHub, where repository-specific terminology provides strong lexical cues even without temporal conditioning. The comparison between the fine-tuned LLM and LA-STPP is also valuable, as it showcases that the lack of spatio-temporal knowledge of the event history can result in consistently lower text prediction accuracy.

Table 3. Ablation study results on Chicago crime data.

Variant	Testing ℓ	MRR
(a) Text fine-tuning ablation		
LA-STPP	-3.172	0.683
LA-STPP (no text fine-tuning)	-3.455	0.309
(b) LLM backbone comparison		
TinyLlama-1.1B	-3.172	0.683
Qwen2.5-1.5B	-3.178	0.677
Phi-3.5-mini	-3.168	0.696

4.2. Ablation Study

We further study two architectural choices in LA-STPP to showcase the benefits of a fine-tuned LLM adapted to event text and the performance robustness over the choice of LLM backbone. We first remove the LoRA LLM fine-tuning, making LLM backbone frozen as a fixed feature extractor, while the spatio-temporal embeddings and prediction heads remain trainable. As shown in Table 3(a), the impact on text prediction is substantial with a significant drop in MRR. This confirms that domain-adapted representations are essential for the model to discriminate between temporally plausible event descriptions. The degradation in testing log-likelihood also suggests the benefits of fine-tuning on learning shared history vector to better capture the underlying event dynamics for intensity modeling. Table 3(b) varies the LLM backbone while keeping other components identical. The three backbones achieve comparable forecasting log-likelihood, suggesting that the STPP component provides a robust mechanism for modeling spatio-temporal dynamics. The comparable MRR scores, with Phi-3.5-mini achieving the best retrieval performance, indicates that stronger backbones can provide modest gains in semantic discrimination.

5. Conclusion

We presented LA-STPP, a language-augmented spatio-temporal point process framework that bridges principled event dynamics modeling with free-form text generation. LA-STPP leverages rich textual information in past events to improve spatio-temporal forecasting, while simultaneously generating semantically meaningful descriptions of future events conditioned on the learned dynamics. Experiments confirm the advantages of LA-STPP in both event forecasting and text generation in spatio-temporal context.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Anthropic. The Claude model family. https://www.anthropic.com, 2024.

Chen, R. T. Q., Amos, B., and Nickel, M. Neural spatio-temporal point processes. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Cheng, X., Dong, Z., and Xie, Y. Deep spatiotemporal point processes: Advances and new directions. *Annual Review of Statistics and Its Application*, 13, 2025.

Daley, D. J. and Vere-Jones, D. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer, 2008.

Dong, Z., Cheng, X., and Xie, Y. Spatio-temporal point processes with deep non-stationary kernels. In *The Eleventh International Conference on Learning Representations*, 2023a.

Dong, Z., Zhu, S., Xie, Y., Mateu, J., and Rodríguez-Cortés, F. J. Non-stationary spatio-temporal point process modeling for high-resolution covid-19 data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72 (2):368–386, 2023b.

Dong, Z., Mateu, J., and Xie, Y. Spatio-temporal-network point processes for modeling crime events with landmarks. *arXiv preprint arXiv:2409.10882*, 2024.

Dong, Z., Fan, Z., and Zhu, S. Conditional generative modeling for high-dimensional marked temporal point processes. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 248–259, 2025.

Dong, Z., Repasky, M., Cheng, X., and Xie, Y. Deep graph kernel point processes over networks. *Journal of Computational and Graphical Statistics*, pp. 1–34, 2026.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1555–1564, 2016.

Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Kong, Q., Zhang, Y., Liu, Y., Tong, P., Liu, E., and Zhou, F. Language-tpp: Integrating temporal point processes with language models for event analysis. *arXiv preprint arXiv:2502.07139*, 2025.

Li, J., Liu, X., Dahan, M., and Montreuil, B. Stochastic service network design with different operational patterns for hyperconnected relay transportation. *Proceedings of 9th International Physical Internet Conference (IPIC)*, 2024.

Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. Learning temporal point processes via reinforcement learning. *Advances in neural information processing systems*, 31, 2018a.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the 6th International Conference on Learning Representations*, 2018b.

Liao, C.-Y., Dong, Z., Garcia, G.-G. P., Paynabar, K., Xie, Y., and Jalali, M. S. Tides need stemmed: A locally operating spatiotemporal mutually exciting point process with dynamic network for improving opioid overdose death prediction. *Manufacturing & Service Operations Management*, 28(2):577–593, 2026.

Lin, H., Wu, L., Zhao, G., Liu, P., and Li, S. Z. Exploring generative neural temporal point process. *arXiv preprint arXiv:2208.01874*, 2022.

Liu, X. and Montreuil, B. Two-echelon delivery vehicle sharing and repositioning in hyperconnected urban logistic networks. *arXiv preprint arXiv:2606.17608*, 2026.

Liu, X., Li, J., and Montreuil, B. Logistics hub capacity deployment in hyperconnected transportation network under uncertainty. In *IISE Annual Conference Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2023.

Liu, X., Li, J., Dahan, M., and Montreuil, B. Multi-period stochastic logistic hub capacity planning for relay transportation. In *Institute of Industrial and Systems Engineers (IISE)*, 2024a.

Liu, X., Xu, Y., and Montreuil, B. Dynamic containerized modular capacity planning and resource allocation in hyperconnected supply chain ecosystems. In *Proceedings of 10th International Physical Internet Conference (IPIC)*, 2024b.

Liu, X., Li, J., Dahan, M., and Montreuil, B. Dynamic hub capacity planning in hyperconnected relay transportation networks under uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, 194:103940, 2025a.

Liu, X., Muthukrishnan, P., and Montreuil, B. Network design and capacity management in hyperconnected urban logistic networks. *Proceedings of 11th International Physical Internet Conference (IPIC)*, 2025b.

Liu, X., Klibi, W., and Montreuil, B. Modular and mobile capacity planning for hyperconnected supply chain networks. *arXiv preprint arXiv:2601.11107*, 2026.

Liu, Z. and Quan, Y. TPP-LLM: Modeling temporal point processes by efficiently fine-tuning large language models. *arXiv preprint arXiv:2410.02062*, 2024.

Mei, H. and Eisner, J. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

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

Sharma, A., Ghosh, A., and Fiterau, M. Generative sequential stochastic model for marked point processes. In *ICML Time Series Workshop*, 2019.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. Neural temporal point processes: A review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 4585–4593. International Joint Conferences on Artificial Intelligence Organization, 2021.

Wang, T., Chen, K., Lin, W., See, J., Zhang, Z., Xu, Q., and Jia, X. Spatio-temporal point process for multiple object tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1777–1788, 2020.

Xiao, S., Xu, H., Yan, J., Farajtabar, M., Yang, X., Song, L., and Zha, H. Learning conditional generative models for temporal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yang, C., Mei, H., and Eisner, J. Transformer embeddings of irregularly spaced events and their participants. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018.

Zhang, P., Zeng, G., Wang, T., and Lu, W. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Zhang, Q., Lipani, A., Kirnap, O., and Yilmaz, E. Self-attentive Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11183–11193, 2020.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Zhu, S. and Xie, Y. Spatiotemporal-textual point processes for crime linkage detection. *The Annals of Applied Statistics*, 16(2):1151–1170, 2022.

Zhu, S., Li, S., Peng, Z., and Xie, Y. Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 34(11): 5391–5402, 2021.

Zhu, S., Wang, H., Dong, Z., Cheng, X., and Xie, Y. Neural spectral marked point processes. In *International Conference on Learning Representations*, 2022.

Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11692–11702, 2020.

A. Related Work

Neural spatio-temporal point processes. Early attempts (Du et al., 2016; Mei & Eisner, 2017) in neural point processes (Shchur et al., 2021) use RNNs for intensity modeling. Subsequent work adopted transformer architectures (Zhang et al., 2020; Zuo et al., 2020; Yang et al., 2022) to model the event intensity and capture long-range event dependencies, overcoming the limitations of RNN-based architectures. In the spatio-temporal setting, NSTPP (Chen et al., 2021) uses neural ODEs to model continuous-space dynamics, while DSTPP (Wang et al., 2020) employs diffusion-based density estimation. All these methods represent event marks as fixed categorical labels. Another line of research (Zhu et al., 2021; 2022; Dong et al., 2023b;a; 2026; Cheng et al., 2025) adopts kernel-based method in point processes for modeling the intricate event dependencies via a so-called influence kernel, leveraging the statistically principled self-exciting architecture in Hawkes processes (Hawkes, 1971). Early attempt on building text-aware influence kernel (Zhu & Xie, 2022) also demonstrates the benefits of incorporating text information for capturing the underlying event dynamics. However, the reliance on the influence kernel restricts the model capability to handle free-form text, with a more fundamental limitation in generating high-dimensional textual event marks. The application of generative models to point processes has received increasing attention. RNN (Li et al., 2018a; Sharma et al., 2019; Xiao et al., 2018) and diffusion models (Lin et al., 2022; Dong et al., 2025) have been adopted for enhancing both the generation capability and scalability of point process models. However, while these approaches have demonstrated success in event sequence generation, none of them have accounted for the event-level narrative text generation in natural language.

Language models for forecasting. TPP-LLM (Liu & Quan, 2024) embeds event type descriptions via an LLM backbone but still classifies into K discrete categories. Language-TPP (Kong et al., 2025) encodes time intervals as byte tokens for seamless LLM integration, yet predicts categorical marks. LA-STPP uniquely occupies the intersection of principled spatio-temporal forecasting and natural language generation. It differs fundamentally from existing models in its capability to handle rich semantic knowledge encapsulated in the complex event data from modern applications. LA-STPP allows the principled modeling of event dynamics while generates unrestricted natural language descriptions conditioned on learned spatio-temporal dynamics, a key step towards the development of language-based world models.

Spatio-temporal event forecasting. Beyond point process modeling, spatio-temporal event forecasting is a broader research area encompassing graph neural network-based approaches for traffic and urban computing (Li et al., 2018b; Yu et al., 2018), attention-based models for multi-variate time series (Zhou et al., 2021), and physics-informed methods for environmental prediction (Raissi et al., 2019). It also serves as a critical input to operational planning under uncertainty, such as stochastic capacity planning (Liu et al., 2023; 2024a; 2025a), resource allocation (Liu et al., 2024b; 2026; Liu & Montreuil, 2026), and network design (Li et al., 2024; Liu et al., 2025b), where richer predictive signals over space and time can directly improve downstream decision quality. Our work contributes to this ecosystem by producing not only quantitative forecasts but also interpretable textual descriptions of predicted events.

B. Experimental Setup

Dataset details. We construct event sequences from three publicly available sources, each exhibiting distinct characteristics in terms of temporal regularity, spatial structure, and textual diversity.

Amazon Reviews consists of product review sequences grouped by individual reviewers. Each event corresponds to a review posted by a user, with the timestamp recording the posting date and the text containing the full review body. We select users with at least 10 reviews and construct sequences ordered chronologically. This dataset is temporal-only (no spatial coordinates) and exhibits irregular inter-event times ranging from hours to months, with review text that varies substantially in length and topic across product categories.

GitHub Events captures repository activity streams from public GitHub repositories. Each event records an action (e.g., push, pull request, issue comment) with a timestamp and an automatically generated or user-written description. Sequences are grouped by repository, producing temporal-only event streams with relatively regular activity patterns. The text in this dataset tends to be concise and repository-specific, making it lexically distinctive across sequences.

Chicago Crime contains crime incident reports from the City of Chicago open data portal. Each event records the timestamp, geographic coordinates (latitude and longitude), and a textual description combining the primary crime type with a detailed narrative of the incident. Sequences are constructed by grouping incidents within geographic regions. This is the only spatio-temporal dataset in our evaluation, and it features relatively homogeneous text across incidents of similar type.

All datasets are split into train, validation, and test sets with an 80/10/10 ratio. Sequences shorter than 5 events are discarded. Event texts are truncated to a maximum of 32 tokens to maintain tractable sequence lengths during training.

Spatio-temporal prediction loss. The time prediction loss is defined as $\mathcal{L}_{\text{time}} = \frac{1}{n-1}\sum_{i=2}^{n}(\log \Delta t_i - \hat{y}_i)^2$, where $\Delta t_i = t_i - t_{i-1}$ and $\hat{y}_i = \mathbf{w}_{\text{time}}^\top \mathbf{h}_{i-1} + b_{\text{time}}$ is the next-event time interval prediction. Similarly, the location prediction loss measures the error between the predicted and true spatial coordinates: $\mathcal{L}_{\text{loc}} = \frac{1}{n-1}\sum_{i=2}^{n}\|s_i - \hat{s}_i\|^2$, where $\hat{s}_i = \mathbf{W}_{\text{loc}}\mathbf{h}_{i-1} + \mathbf{b}_{\text{loc}} \in \mathbb{R}^2$. Both losses are added into the final training objective as weighted components to train the model forecasting capability.

Training configuration. We fine-tune the LLM backbone using LoRA with rank 16, scaling factor $\alpha = 16$, and dropout 0.05, applied to the query, key, value, and output projections of each attention layer. The auxiliary loss weights are set to $\beta_{\text{time}} = 1.0$ and $\beta_{\text{loc}} = 1.0$. The spatial GMM uses $K = 5$ mixture components. We train all models for up to 20 epochs using the Adam optimizer with a learning rate of 5×10^{-4} and apply early stopping based on validation loss with a patience of 5 epochs. Gradient norms are clipped at 1.0 to stabilize training. All experiments are conducted on a single NVIDIA A100 GPU with gradient checkpointing enabled to accommodate the memory requirements of the LLM backbone.

Ablation implementation. In the "no text fine-tuning" ablation, we freeze the entire LLM backbone and disable LoRA, so that the pretrained language model serves purely as a fixed feature extractor. Only the temporal and spatial embedding layers, the intensity head, the spatial GMM head, and the auxiliary prediction heads receive gradient updates. Gradients still propagate through the frozen backbone to reach the input embedding layers, ensuring that the temporal and spatial embeddings can adapt to the data. For the backbone comparison experiment, we evaluate TinyLlama-1.1B, Qwen2.5-1.5B, and Phi-3.5-mini under identical LoRA configurations, head architectures, and training schedules to isolate the effect of the pretrained backbone on downstream performance.

Text generation evaluation protocol. For each test event, we construct a candidate pool of $K{=}50$ texts: the ground-truth next event text plus 49 distractors randomly sampled without replacement from all unique texts in the test split. Each model scores all candidates by computing the conditional log-likelihood of each text given the event history and true timestamp/location, then ranks them in descending order. We report three metrics: Mean Reciprocal Rank (MRR $= \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{\text{rank}_q}$), Hits@1 (fraction of queries where the correct text is ranked first), and Hits@5 (fraction where the correct text appears in the top 5). The Random baseline assigns uniform scores to all candidates, yielding Hits@$k = k/K$ and MRR $= H_K/K$ (where $H_K = \sum_{j=1}^{K} 1/j$ is the K-th harmonic number), producing identical performance across datasets. The difficulty of the retrieval task varies by dataset due to differences in text pool diversity: datasets with highly distinctive per-sequence patterns (e.g., GitHub repository-specific terminology) are easier to rank correctly than datasets with more homogeneous text (e.g., Chicago crime descriptions that share similar phrasing across incidents).

C. Additional Results

C.1. Hyperparameter Sensitivity Analysis

We investigate the sensitivity of LA-STPP to key hyperparameters on the Chicago Crime dataset, varying one parameter at a time while holding all others at their default values. Results are reported in Table 4.

LoRA rank. The LoRA rank r controls the expressiveness of the low-rank adaptation applied to the LLM backbone. Performance improves substantially as rank increases from 4 to 16, reflecting the model's growing capacity to encode domain-specific event patterns into its hidden representations. However, increasing the rank beyond 16 yields slightly worse results, suggesting that excessive adaptation capacity leads to overfitting on the training sequences. The default rank of 16 provides a favorable balance between adaptation expressiveness and generalization.

GMM components. The number of Gaussian mixture components K governs the flexibility of the spatial density model. A single component produces notably higher location error, as it cannot capture the multi-modal spatial distribution of crime incidents across different neighborhoods. Increasing K from 1 to 5 steadily improves spatial prediction, but further increases to 7 yield diminishing returns with comparable performance. The log-likelihood and time MAE remain largely unaffected by K, confirming that the spatial head operates independently of the temporal intensity component.

Loss weights. The auxiliary loss weights β_{time} and β_{loc} exhibit complementary trade-offs. For β_{time}, removing auxiliary time supervision entirely preserves log-likelihood but degrades time prediction accuracy, while over-weighting it begins to interfere with the intensity-based log-likelihood optimization. For β_{loc}, the location MAE improves monotonically with

Table 4. Hyperparameter sensitivity analysis on Chicago Crime. Each row varies one parameter from its default (marked with *). We report testing log-likelihood, time MAE, location MAE, and text retrieval MRR.

Parameter	Value	Testing ℓ	Time MAE	Loc. MAE
LoRA rank (r)	4	-4.819	8.71	0.84
	8	-3.540	7.29	0.75
	16*	-3.172	6.81	0.68
	32	-3.284	6.92	0.70
GMM components (K)	1	-3.183	6.86	0.98
	3	-3.177	6.83	0.78
	5*	-3.172	6.81	0.68
	7	-3.175	6.80	0.69
β_{time}	0.0	-3.166	7.98	0.63
	0.5	-3.170	7.23	0.67
	1.0*	-3.172	6.81	0.68
	2.0	-3.389	6.49	0.82
β_{loc}	0.0	-3.168	6.76	0.85
	0.5	-3.171	6.79	0.74
	1.0*	-3.172	6.81	0.68
	2.0	-3.179	6.82	0.66

increasing weight, but aggressive weighting marginally degrades the log-likelihood. The default setting of $\beta_{\text{time}} = \beta_{\text{loc}} = 1.0$ achieves a balanced trade-off across all metrics without substantially compromising any single objective.