CONTINUOUS-TIME LIMIT ORDER BOOK FORECASTING WITH MAMBA STATE SPACE MODELS

FRANCESCO PAPINI AND MARTIN LOTZ

ABSTRACT. We study continuous-time forecasting of limit order book event streams using a selective state-space backbone (Mamba) coupled with a mixture-of-exponentials temporal point-process head and hierarchical mark decoders. The model produces calibrated probabilities for near-horizon arrivals and structured distributions over event marks (price move, size, side, event type, book level). We describe the data preprocessing, binning, hierarchical smoothing, and training objectives needed to make state-space models competitive on noisy market message streams, and we outline an evaluation protocol.

CONTENTS

1. INTRODUCTION

Modern electronic markets produce high-frequency streams of limit order book (LOB) messages: submissions, cancellations, and executions arriving at millisecond granularity. Trading systems increasingly need calibrated forecasts of both *when* the next event will arrive and *what* its attributes (marks) will be, in order to manage inventory, routing, and queue position. While discrete-time sequence models capture local patterns, continuous-time structure drives queue dynamics, especially around sharp changes in activity.

Date: 2nd December 2025.

This paper introduces a continuous-time forecasting architecture that couples a selective state-space backbone (Mamba, [GD23]) with a mixture-of-exponentials temporal point-process (TPP) head and hierarchical decoders for event marks (price move, size, time bucket, event type, side, and book-level proxy).

The key ingredients are:

- **Continuous-time modeling**: A mixture-of-exponentials intensity yields closed-form survival and arrival probabilities over arbitrary horizons, enabling calibrated risk estimates.
- **Selective state-space backbone**: Mamba processes long streams efficiently and can incorporate both inter-arrival times and absolute time-of-day signals.
- **Hierarchical marks**: Coarse/residual binning with neighbor-aware smoothing produces sharp yet stable distributions over marks, aligning with microstructure semantics (e.g., tails beyond the spread).
- **Practical training recipe**: Quantile-based bin fitting on the training set, censor-aware losses, cosine warmup with AdamW, and optional soft binning for heavy-tailed features.

We focus on specifying the model, data processing, and training objectives. Experimental results and ablations will be added when broader evaluations across instruments and market regimes are complete.

2. Preliminaries

2.1. **Limit order book dynamics.**

A limit order book (LOB) is an electronic record of buy and sell queues indexed by price levels. Let $b_k(t)$ and $a_k(t)$ denote queue sizes at the k-th level at time $t$from the best bid and best ask, respectively, with best prices $p^{\text{bid}}(t)$ and $p^{\text{ask}}(t)$ such that $p^{\text{bid}}(t) < p^{\text{ask}}(t)$. Events update the state of the LOB via:

- **Limit add**: insert volume s at price p on side $\sigma \in \{\text{buy}, \text{sell}\}$, increasing the corresponding queue: $b_k \leftarrow b_k + s$ or $a_k \leftarrow a_k + s$ when p matches level k.
- **Cancel/amend**: remove or reduce standing volume at a given level.
- **Marketable order**: consume volume against the opposite queue. If volume sweeps multiple levels, best prices update. Trades are recorded as executions.

The midprice is $m(t) = \frac{1}{2}\big(p^{\text{bid}}(t) + p^{\text{ask}}(t)\big)$ and the spread is $p^{\text{ask}}(t) - p^{\text{bid}}(t)$. Queue evolution is piecewise-constant with jumps at event times; see [GPW$^+$13, Bou08, CDL13, ACJMT16] for formal treatments. Most electronic venues operate price-time priority: orders are first ranked by price (best bid/ask at the front), and within each price level by arrival time.

A market maker maintains resting bids and asks, earning the spread while bearing inventory and adverse-selection risk. Key controls include (i) *quote placement* (choosing price levels and sizes relative to the spread and queue depth), (ii) *quote lifetime* (cancelling or repricing as flow risk changes), and (iii) *inventory targets* (skewing quotes or size to neutralize exposure). Short-horizon forecasts of arrival intensity and mark distributions are central to these decisions.

2.2. **Temporal point processes.**

We model the arrival times of LOB events using temporal point processes (TPPs). Let $\{t_i\}_{i\geq 1}$ be a strictly increasing sequence of event times. The history up to time t is denoted by $\mathcal{H}_t = \{(t_i, m_i) : t_i < t\}$, comprising the times and marks m_i (e.g., price, size, type) of all events occurring before t. A TPP is fully characterized by its conditional intensity function $\lambda(t \mid \mathcal{H}_t)$, which represents the instantaneous rate of a new event arrival given the history:

$$\lambda(t \mid \mathcal{H}_t) = \lim_{\Delta t \to 0^+} \frac{P(N(t+\Delta t) - N(t) = 1 \mid \mathcal{H}_t)}{\Delta t},$$

where $N(t)$ counts the number of events up to time t. Given the most recent event at t_n, the probability density function of the next inter-arrival time $\Delta t = t_{n+1} - t_n$ is

$$f(\Delta t \mid \mathcal{H}_{t_n}) = \lambda(t_n + \Delta t \mid \mathcal{H}_{t_n}) \exp\left(- \int_{t_n}^{t_n+\Delta t} \lambda(u \mid \mathcal{H}_u)\, du \right).$$

For a mixture-of-exponentials intensity with non-negative weights α_k and rates β_k,

$$\lambda(\Delta t) = \sum_{k=1}^{K} \alpha_k \beta_k \exp(-\beta_k \Delta t), \qquad \Lambda(\Delta t) = \sum_{k=1}^{K} \alpha_k \big(1 - e^{-\beta_k \Delta t}\big)/\beta_k, \tag{2.1}$$

which yields closed-form survival $S(\Delta t) = \exp(-\Lambda(\Delta t))$ and horizon probability $P(\Delta t \le \tau) = 1 - \exp(-\Lambda(\tau))$. This tractability motivates the head used in our architecture.

2.3. **State-space sequence models.** Continuous-time linear state-space models (SSMs) define a latent $x(t) \in \mathbb{R}^d$ driven by input $u(t)$:

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.2}$$

$$y(t) = Cx(t) + Du(t), \tag{2.3}$$

with learned A, B, C, D. Discretization with step Δ yields a recurrence $x_{n+1} = \bar{A}x_n + \bar{B}u_n$ and output $y_n = Cx_n + Du_n$, equivalent to a convolution with a kernel derived from $(\bar{A}, \bar{B}, C, D)$. This gives linear-time, memory-efficient inference for long sequences, in contrast to the quadratic cost of attention.

Mamba [GD23] is a *selective* SSM: the effective transition and input matrices are modulated by the current token, gating information flow based on content while retaining the fast scan/convolution implementation. The result combines long-context capacity with low latency and footprint, making it suitable for high-frequency LOB streams where both history length and computational budget are critical.

3. Mixture-of-exponentials temporal point processes

Mixtures of exponentials offer a tractable family for modeling inter-arrival densities while retaining flexibility beyond a single-scale Poisson process. Let K denote the number of components. Given non-negative weights α_k (not necessarily normalized) and rates $\beta_k > 0$, the intensity and cumulative intensity after the last event are

$$\lambda(\Delta t) = \sum_{k=1}^{K} \alpha_k \beta_k \exp(-\beta_k \Delta t), \tag{3.1}$$

$$\Lambda(\Delta t) = \sum_{k=1}^{K} \alpha_k \big(1 - e^{-\beta_k \Delta t}\big)/\beta_k. \tag{3.2}$$

The survival and horizon probabilities follow immediately:

$$S(\Delta t) = \exp\big(-\Lambda(\Delta t)\big), \qquad P(\Delta t \le \tau) = 1 - \exp\big(-\Lambda(\tau)\big).$$

This yields closed-form log-likelihoods for observed arrivals ($\log \lambda - \Lambda$) and for censored observations ($-\Lambda$), making the family attractive for maximum-likelihood training and calibration at arbitrary horizons.

Connections to prior work. Mixture intensities and hazard mixtures have long been used in survival analysis as semi-parametric approximations [DVJ03]. In neural point process literature, mixture or basis expansions of intensities provide a balance between flexibility and closed-form integration (e.g., DDT$^+$16, ME17, TJNR21). Our choice mirrors this design: the model learns (α_k, β_k) from the Mamba hidden state via linear projections followed by softplus activations (to ensure positivity), preserving analytical integrals, inverse transform sampling, and efficient evaluation of $P(\Delta t \le \tau)$ needed for trading risk controls.

4. Model

Our architecture combines engineered representations for LOB messages with a Mamba backbone and specialized output heads for time and marks.

4.1. **Feature extraction and binning.** We load raw LOB messages (time, type, order id, size, price, direction) and compute derived features x_i for each event i:

- **Price change**: $\Delta p_i = (p_i - p_{i-1})/\delta$, where δ is the tick size. This is discretized into a centered window $[\ell, h]$ (e.g., $[-2, +2]$ ticks) plus two tail bins for larger moves.
- **Size and Time**: Log-transformed size $\log(1 + s_i)$ and inter-arrival time $\log(1 + \Delta t_i/\tau_{\text{scale}})$, where τ_{scale} is the median positive inter-arrival time in the training set.
- **Categorical**: Event type (limit, cancel, deletion, execution), side (ask/bid), and a level proxy derived from $|\Delta p_i|$ (at-spread vs. away).
- **Time-of-day**: Absolute time t_i^{abs} (seconds from midnight) is encoded cyclically as

$$[\sin(2\pi h/24), \cos(2\pi h/24), \sin(2\pi m/60), \cos(2\pi m/60)]$$

For continuous features (size, time), we employ a soft binning strategy to preserve local information. Given bin edges $e_0, \ldots, e_B$, we compute the distance of a value x to bin centers $c_j = (e_j + e_{j+1})/2$. A softmax with temperature T yields weights $w_j \propto \exp(-|x - c_j|/T)$, which are used to compute a weighted sum of learnable bin embeddings $E \in \mathbb{R}^{B\times d}$:

$$\mathbf{e}_{\text{soft}}(x) = \sum_{j=1}^{B} w_j(x)\mathbf{E}_j.$$

This allows the model to interpolate between bins, mitigating boundary effects inherent in hard quantization.

TABLE 1. Example of raw LOB messages and corresponding derived features.

Field	Time	Type	Side	Price	Size
Raw Value	34200.123	1 (Limit)	-1 (Ask)	100.05	100
Derived Feature	$t^{\text{abs}} = 34200.123$	Type Code 0	Side Code 0	$\Delta p = 0$ ticks	$\log(1 + s) \approx 4.62$
					$\Delta t_{\text{prev}} = 0.050\text{s}$

For marks with meaningful locality (price, size, time), we structure the label space hierarchically. A field with N fine bins is partitioned into C coarse groups. For price, these groups typically represent "negative", "zero", and "positive" moves (or finer granularities like "small negative", "large negative"). Each fine bin k maps to a coarse index $c(k)$ and a residual index $r(k)$. This structure informs both the model architecture and the loss function.

4.2. **Backbone.** For an input segment of length L, we embed each field and concatenate the embeddings. A learned projection maps the concatenated vector to the model dimension d. An additional projection of $\log(1+\Delta t_i)$ injects inter-arrival information; cyclical encodings of absolute time-of-day are added when enabled. The resulting sequence is passed through n Mamba layers, followed by LayerNorm and dropout.

4.3. **Temporal head: mixture of exponentials.** Given hidden states $h_i \in \mathbb{R}^d$, the temporal head outputs non-negative weights and rates $(\alpha_{i,k}, \beta_{i,k})_{k=1}^{K}$ via linear projections and softplus activations. Using (2.1), the log-likelihood for an observed Δt_i is

$$\ell_{\text{time}} = \log \lambda(\Delta t_i) - \Lambda(\Delta t_i), \tag{4.1}$$

while censored observations (no arrival before a cap) contribute $-\Lambda(\Delta t_{\max})$. The probability of at least one arrival within a horizon τ is $1 - e^{-\Lambda(\tau)}$, used for calibration and downstream risk flags.

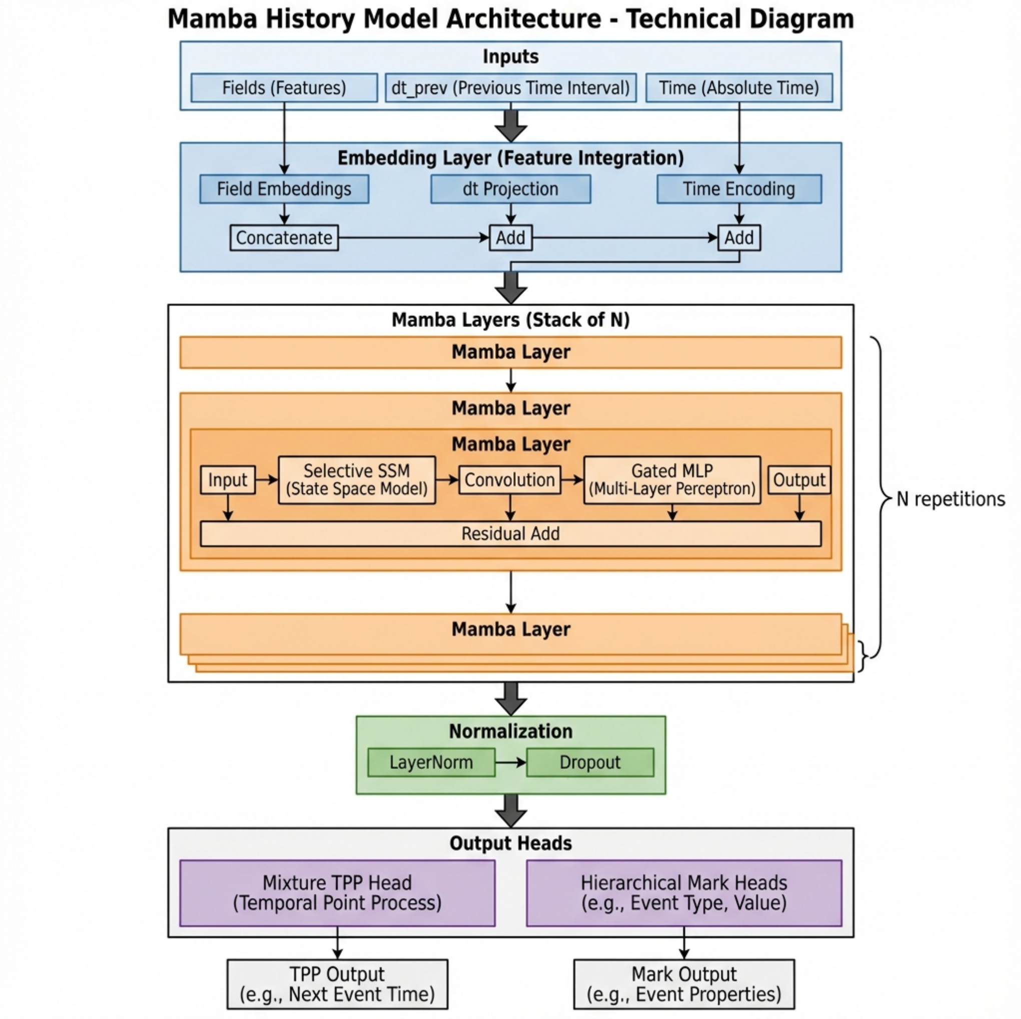
Mamba History Model Architecture - Technical Diagram
Inputs
Fields (Features)
dt_prev (Previous Time Interval)
Time (Absolute Time)
Embedding Layer (Feature Integration)
Field Embeddings
dt Projection
Time Encoding
Concatenate
Add
Add
Mamba Layers (Stack of N)
Mamba Layer
Mamba Layer
Mamba Layer
Input
Selective SSM (State Space Model)
Convolution
Gated MLP (Multi-Layer Perceptron)
Output
Residual Add
Mamba Layer
N repetitions
Normalization
LayerNorm
Dropout
Output Heads
Mixture TPP Head (Temporal Point Process)
Hierarchical Mark Heads (e.g., Event Type, Value)
TPP Output (e.g., Next Event Time)
Mark Output (e.g., Event Properties)

FIGURE 1. Detailed view of the Mamba History Model. The architecture ingests a sequence of derived features, processes them through a stack of Mamba layers, and branches into a temporal head (for arrival times) and hierarchical mark heads (for event attributes).

4.4. **Hierarchical mark heads.** For each mark field, the prediction is factorized into a coarse selection and a residual refinement:

$$p(m) = p(c \mid h_i)\, p(r \mid c, h_i).$$

The coarse probability $p(c \mid h_i)$ is obtained via a linear layer and softmax. To condition the residual prediction on the coarse choice, we compute a context vector $\mathbf{c}_{\text{ctx}}$ by aggregating learnable coarse embeddings $\mathbf{E}^{\text{coarse}}$ weighted by the predicted coarse probabilities (soft conditioning):

$$\mathbf{c}_{\text{ctx}} = \sum_j p(j \mid h_i)\mathbf{E}_j^{\text{coarse}}.$$

The residual logits are then computed from the concatenation $[h_i; \mathbf{c}_{\text{ctx}}]$ via a specialized linear layer that outputs logits for all possible (c, r) pairs, reshaped and masked to valid combinations. This design allows the model to share information across the coarse group while refining the specific bin prediction.

4.5. **Sampling.** Sampling a next event involves drawing Δt from the mixture via inverse transform (with Newton refinement) and sampling marks by first drawing a coarse bin and then a residual, mapping back to fine indices. This allows joint Monte Carlo scenarios for stress testing or simulation.

5. Training setup

5.1. **Data splits and bin fitting.** We fit quantile-based bin edges for size and time on the training portion (default 80%), avoiding look-ahead. Price bins use a fixed centered window around the spread with two tails. Validation uses the remaining held-out segment (default 10%), with the final 10% reserved for future test reporting. Sliding windows of length L and stride s build sequences; the last position in each window is masked for mark supervision.

5.2. **Loss functions.** The total objective is a weighted sum:

$$\mathcal{L} = \lambda_{\text{time}}\,\mathcal{L}_{\text{TPP}} + \sum_{f}\big(\lambda_f^{\text{coarse}}\mathcal{L}_f^{\text{coarse}} + \lambda_f^{\text{resid}}\mathcal{L}_f^{\text{resid}}\big),$$

where f ranges over marks (price, size, time, type, side, level).

- **Temporal loss** $\mathcal{L}_{\text{TPP}}$ is the negative log-likelihood from the mixture-of-exponentials head. Observed events use $\log\lambda - \Lambda$; censored events contribute $\Lambda(\Delta t_{\max})$.
- **Mark losses** use cross-entropy with neighbor-aware label smoothing. We construct a smoothing matrix S based on the adjacency of bins. For a target bin y, the smoothed target distribution $q(k)$ is:

$$q(k) = \begin{cases} 1-\epsilon & \text{if } k = y \\ \epsilon/|N(y)| & \text{if } k \in N(y) \\ 0 & \text{otherwise} \end{cases}$$

  where $N(y)$ is the set of immediate neighbors of y in the ordered bin sequence, and ϵ is a smoothing parameter. This is applied to both coarse and residual targets, penalizing "near misses" less than far-off errors.

Continuous targets that exceed a censoring cap are masked, aligning the supervision with the near-term horizon of interest.

5.3. **Optimization.** We use the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight decay of 0.1. The learning rate follows a cosine decay schedule with a linear warmup phase. Specifically, for the first W steps, the learning rate increases linearly to η_{peak}, then decays according to:

$$\eta_t = \frac{\eta_{\text{peak}}}{2}\left(1 + \cos\left(\frac{\pi(t-W)}{T-W}\right)\right)$$

where T is the total number of steps. Gradient clipping (norm 1.0) stabilizes training. Dropout is applied to embeddings (0.15) and MLP blocks (0.32). Mixed precision (FP16/BF16) is enabled for efficiency. Key hyperparameters include mixture size $K = 3$, horizon $\tau = 0.75$s, and model dimension $d = 512$.

5.4. **Inference and calibration.** At inference, $P(\Delta t \le \tau)$ provides a calibrated near-horizon arrival probability. Mark posteriors combine coarse and residual distributions. Reliability is monitored via Brier score and calibration bins; sampling utilities enable scenario analysis and downstream simulators.

6. Evaluation plan

We report both temporal and mark-centric metrics, emphasizing calibration for trading use-cases.

- **Temporal metrics**: negative log-likelihood of Δt, Brier score for $P(\Delta t \le \tau)$, expected calibration error (ECE), and coverage of calibration bins.
- **Mark metrics**: coarse and residual accuracy/entropy per field; joint mark accuracy; hierarchical perplexity when applicable.

- **Sampling diagnostics**: empirical mean and variance of sampled Δt; histograms of sampled coarse marks compared to validation frequencies.

Experiments will be added after completing broader evaluations across instruments and days. We will focus on (i) ablations over mixture size K, binning schemes, and smoothing strengths; (ii) the impact of time-of-day and soft binning; and (iii) robustness to regime shifts (open/close, news events).

7. Related work

Neural temporal point processes. Recurrent and attention-based neural TPPs learn history-dependent intensities without hand-crafted kernels [DDT+16, ME17, SBMD21, TJNR21]. Many models trade off expressiveness and tractable integrals; basis expansions or mixtures (including exponentials) keep survival functions closed-form. Our approach uses a mixture-of-exponentials head to retain analytical likelihoods while delegating history modeling to a state-space backbone.

State-space sequence models. Selective state-space models such as Mamba [GD23] provide linear-time sequence processing with competitive long-context capacity. Compared to transformers, they reduce quadratic cost and memory, making them attractive for long LOB streams. We combine Mamba with continuous-time conditioning to capture both inter-arrival structure and latent regime shifts (e.g., time-of-day).

Limit order book modeling. Deep models for LOB data have focused on price movement classification, depth forecasting, or point processes over order events. Prior work often uses Hawkes processes or RNN/transformer architectures with discrete time steps. By coupling a continuous-time head with hierarchical mark decoders, we aim to obtain calibrated arrival probabilities and structured mark distributions aligned with microstructure (spread-centered price bins, heavy-tailed sizes).

8. Conclusion and next steps

We presented a continuous-time LOB forecasting architecture that integrates a mixture-of-exponentials TPP head with hierarchical mark decoders on top of a Mamba state-space backbone. The design emphasizes calibrated near-horizon risk estimates and structured mark distributions while remaining computationally efficient for long streams.

Future work will extend the empirical study to multiple assets and regimes, explore robustness under distribution shift, and compare against transformer-based TPP baselines. Additional extensions include richer mark taxonomies (e.g., depth snapshots), adaptive binning under drift, and deployment-oriented latency profiling.

References

[ACJMT16] Frédéric Abergel, Anirban Chakraborti, Amin Jedidi, and Ioane Muni Toke. *Limit Order Books*. Cambridge University Press, 2016.

[Bou08] Jean-Philippe Bouchaud. Markets as a collective phenomenon. *Appendix in Les Houches School of Physics: Complex Systems*, 2008.

[CDL13] Rama Cont and Adrien De Larrard. A stochastic model for order book dynamics. *Operations Research*, 61(6):1243–1256, 2013.

[DDT+16] Nan Du, Hanjun Dai, Rakshit Trivedi, Uttara Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*, 2016.

[DVJ03] Daryl John Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, 2003.

[GD23] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[GPW+13] Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[ME17] Hongyuan Mei and Jason Eisner. Neural Hawkes process: A neurally self-modulating multivariate point process. In *NeurIPS*, 2017.

[SBMD21] Oleksandr Shchur, Jùlia Bilo, Iasonas Markou, and Nathalie Dupuy. Foundations of neural temporal point processes. *arXiv preprint arXiv:2104.03528*, 2021.

[TJNR21] Yingtao Tan, Gustav Jiang, Yu Nardi, and Kannan Ramchandran. Relaxing the non-explosion assumption for neural point processes. In *ICML*, 2021.

(Papini) Warwick Mathematical Institute, University of Warwick, UK
Email address: `martin.lotz@warwick.ac.uk`

(Lotz) Warwick Mathematical Institute, University of Warwick, UK
Email address: `martin.lotz@warwick.ac.uk`