Detecting the Future: All-at-Once Event Sequence Forecasting with Horizon Matching

Ivan Karpukhin[1*], **Andrey Savchenko**[1,2,3]
[1]Sber AI Lab, Moscow, Russia
[2]HSE University, Moscow, Russia
[3]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
iakarpukhin@sberbank.ru, avsavchenko@hse.ru

Abstract

Long-horizon events forecasting is a crucial task across various domains, including retail, finance, healthcare, and social networks. Traditional models for event sequences often extend to forecasting on a horizon using an autoregressive (recursive) multi-step strategy, which has limited effectiveness due to typical convergence to constant or repetitive outputs. To address this limitation, we introduce DEF, a novel approach for simultaneous forecasting of multiple future events on a horizon with high accuracy and diversity. Our method optimally aligns predictions with ground truth events during training by using a novel matching-based loss function. We establish a new state-of-the-art in long-horizon event prediction, achieving up to a 50% relative improvement over existing temporal point processes and event prediction models. Furthermore, we achieve state-of-the-art performance in next-event prediction tasks while demonstrating high computational efficiency during inference.

Code — https://github.com/ivan-chai/hotpp-benchmark
Extended version — https://arxiv.org/abs/2408.13131

Introduction

Data from various domains, such as internet activity, e-commerce transactions, retail operations, and clinical visits, typically consists of timestamps with associated information. When ordered chronologically, these data points form event sequences, which differ fundamentally from other data types. Unlike tabular data (Wang and Sun 2022), events inherently include timestamps and follow a specific temporal order. In contrast to time series data (Lim and Zohren 2021; Kostromina et al. 2025; Savchenko and Kachan 2025), event sequences exhibit irregular time intervals and often contain additional attributes. These unique characteristics necessitate the development of specialized models that can handle complex data streams. One of the primary tasks in the domain of event sequences is predicting future event types and their occurrence times (Xue et al. 2024; McDermott et al. 2024). It may be solved using the apparatus of Marked Temporal Point Processes (MTPP) (Rizoiu et al. 2017) or their extensions for complex data streams that include additional event features (McDermott et al. 2024)

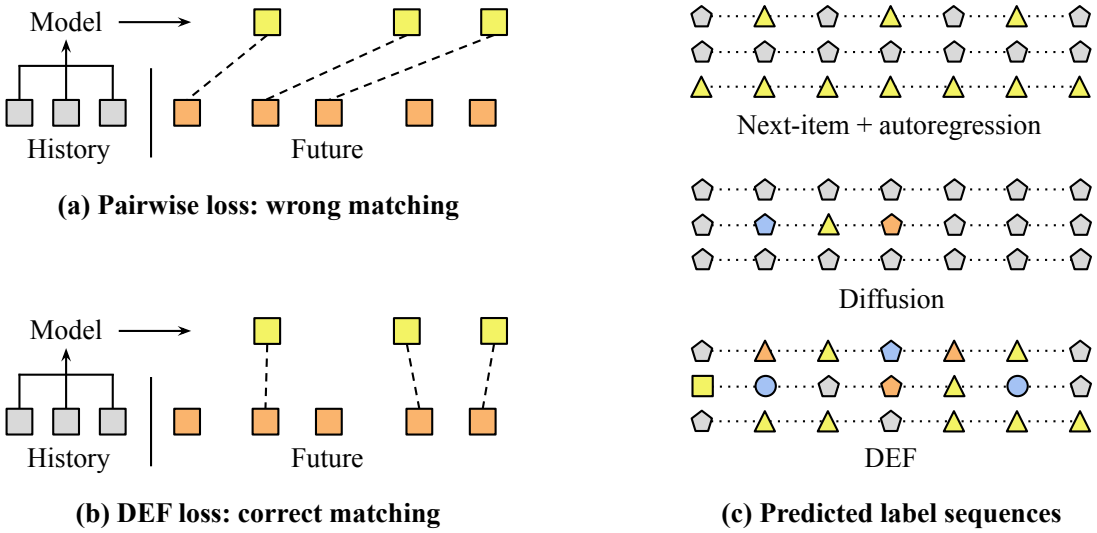
Practical applications often require predicting multiple future events within a specified time horizon, such as forecasting purchases for the next month or making long-term medical prognoses (Xue et al. 2022). This task presents unique challenges that differ from traditional next-event prediction. The conventional approach typically relies on autoregressive models, which predict the next event step by step (Xue et al. 2024; Xiao et al. 2018). While these models are effective for immediate next-event forecasting, their performance tends to deteriorate as the prediction horizon extends (Karpukhin, Shipilov, and Savchenko 2024). The same is true for horizon prediction models, including GAN (Xiao et al. 2018) and diffusion (Zhou et al. 2025), which predict multiple future events at once but use pairwise losses between events on corresponding positions.

In this study, we identify significant limitations of pairwise losses in the context of long-horizon prediction. To address these challenges, we propose DEF (Detection-based Event Forecasting),which detects multiple future events in parallel and employs a novel horizon matching loss, which dynamically aligns predictions with the closest ground-truth events, as illustrated in Fig. 1. This loss function enables the model to capture the full distribution of events within the horizon while remaining robust to outlier events. We demonstrate that our approach establishes a new state-of-the-art in long-horizon prediction, surpassing both autoregressive and horizon prediction approaches in terms of accuracy and prediction diversity. Additionally, our method exhibits high computational efficiency during inference, ranking among the fastest methods.

Related Works

Event Sequences and Marked Temporal Point Processes. MTPP is a stochastic process that consists of a sequence of time-event pairs $(t_1, l_1), (t_2, l_2), \ldots$, where $t_1 < t_2 < \ldots$ denote the times of events, and $l_i \in \{1, \ldots, L\}$ are the corresponding event type labels (Rizoiu et al. 2017). One of the most popular tasks in this domain is predicting the next event in the sequence (Shchur et al. 2021; Zhuzhel et al. 2023). A straightforward approach is to independently predict the time and type of the next event (Shchur, Biloš, and Günnemann 2020; Panos 2024), while more sophisti-

*Corresponding author.



(a) Pairwise loss: wrong matching

(b) DEF loss: correct matching

(c) Predicted label sequences

Figure 1: (a) A typical next-event or pairwise loss compares events at corresponding positions, often resulting in incorrect matching. (b) The proposed matching loss calculates the loss function between the closest events, leading to a more robust and balanced error measure. (c) The proposed DEF method enhances the diversity of predictions. We demonstrate 3 example sequences from the Amazon dataset generated by the autoregressive IFTPP method, Diffusion, and the proposed approach. Each label type is depicted using a distinct shape and color combination. The precise timestamps are omitted for simplicity.

cated methods model the temporal dynamics of each event type separately (Mei and Eisner 2017). Traditional models, such as Poisson and Hawkes processes (Rizoiu et al. 2017), rely on strong assumptions about the underlying generative processes. Recent advancements have shifted towards more flexible and expressive models that leverage neural architectures. These include classical Recurrent Neural Networks (RNNs) (Du et al. 2016; Xiao et al. 2017; Omi, Aihara et al. 2019; Shchur, Biloš, and Günnemann 2020), as well as more advanced architectures like transformers (Zhang et al. 2020; Zuo et al. 2020; Wang and Xiao 2022; Yang, Mei, and Eisner 2022; Panos 2024). Additionally, continuous-time models such as Neural Hawkes Processes (Mei and Eisner 2017), ODE-RNN (Rubanova, Chen, and Duvenaud 2019), and their variants (Jia and Benson 2019; De Brouwer et al. 2019; Kidger et al. 2020; Song et al. 2024; Kuleshov et al. 2024) have been developed to capture the dynamics of event sequences better. Moreover, generative models were adapted for use in MTPPs, including denoising diffusion (Zhou et al. 2025; Zeng, Regol, and Coates 2024) and Generative Adversarial Networks (GANs) (Lin et al. 2022). Some prior works have explored general event sequences that include additional data fields beyond timestamps and labels (McDermott et al. 2024; Padhi et al. 2021).

Long-horizon prediction. In the long-horizon event forecasting task, the goal is to predict future events within a time horizon H, defined as the time interval $(t, t + \mathrm{H})$, where t is the timestamp of the last observed event. A straightforward approach is to use autoregressive inference based on next-event prediction models. Previous research has also explored models that predict multiple future events simultaneously, known as next-K models (Karpukhin, Shipilov, and Savchenko 2024), where K exceeds the typical number of events occurring within the horizon H. These models are trained using pairwise losses that align predicted events with ground truth events at corresponding positions. Notably, next-event models can be viewed as a special case of next-1 models. Some approaches, such as GAN (Xiao et al. 2018) and Diffusion (Zhou et al. 2025; Zeng, Regol, and Coates 2024), incorporate pairwise losses as part of their training objectives. HYPRO (Xue et al. 2022) has also addressed the problem of long-horizon prediction and introduced a technique for selecting the best candidate from a set of generated sequences. HYPRO functions as a meta-algorithm that can enhance the performance of nearly any sequence prediction model. However, its approach requires multiple generation runs for each prediction, significantly reducing training and inference efficiency.

Thus, previous studies have identified several challenges associated with autoregressive models for long-horizon predictions. These models often exhibit reduced prediction diversity and uncertainty over extended horizons, even though the task becomes increasingly complex. As illustrated in Fig. 1.c, the predicted label sequences often have constant or repetitive outputs. This behavior likely stems from the model's reliance on its predictions as input for subsequent predictions, which can amplify errors and lead to repeated events. Even horizon prediction approaches, such as Diffusion (Zhou et al. 2025), exhibit repetitive patterns because each of K outputs predicts the entire distribution of labels, leading to a bias toward the most frequent classes during inference. To address these issues, we propose a novel approach that offers greater diversity in its predictions while maintaining high accuracy for popular event forecasting

tasks. The details are discussed in the following section.

Proposed Approach

The proposed DEF method (Fig. 2) utilizes a backbone model to extract embeddings from historical data. Next, it simultaneously predicts K future event candidates by using multiple prediction heads. Here, the hyperparameter of our method, K, should be larger than the typical sequence length in a specified time horizon H. During training, the model aligns its predictions with ground truth and computes pairwise losses in the novel horizon matching loss function. At inference time, the model simultaneously predicts K events and retains only candidates with high prediction scores. Below, we provide a detailed overview of the event prediction head, the sequence model, and the associated training and inference procedures.

Prediction Head Our method captures the complexity of event sequences by modeling each component of an event using a probabilistic framework that provides a rigorous basis for evaluating the likelihood of ground-truth event sequences. Specifically, we propose predicting the probability of an event occurring, the distribution of event labels, and the distribution of time shifts relative to the last observed event. As depicted in Fig. 2, the probability $\hat{o}$ of an event occurring is modeled using a neural network with a sigmoid activation function. A separate head with softmax activation (SM) models the distribution $\hat{p}(l)$ of event labels. For the time shift, we use a Laplace distribution with a unit scale parameter, similarly to Mixture Density Networks (MDNs) (Bishop 1994) and intensity-free MTPP (Shchur, Biloš, and Günnemann 2020):

$$\mathrm{P}(t) = \frac{1}{2} e^{-|t-\hat{t}|}, \tag{1}$$

where $\hat{t}$ denotes the predicted time shift. This formulation offers a probabilistic interpretation of the MAE loss function. Note that a more rigorous formulation would involve using a truncated distribution to prevent negative time steps. However, preliminary experiments showed slightly worse performance with this approach, likely due to the reliance of most evaluation metrics on MAE.

By combining the predicted probabilities, we can estimate the likelihood of a future event given the output of the model:

$$\log \mathrm{P}(y) = \log \hat{o} + \log \hat{p}(l) - |t - \hat{t}| - \log \mathrm{R}(t). \tag{2}$$

where $y = (t, l)$ represents an event with timestamp t and label l. In Eq 2, we assume that, given the history of events, the timestamp, label, and occurrence of the next event are conditionally independent. The probability of a missed event (no event occurring) is given by:

$$\log \mathrm{P}(\emptyset) = \log(1 - \hat{o}) + C_\emptyset, \tag{3}$$

where $C_\emptyset$ is a constant independent of the model's output, representing the probability associated with a reserved "unknown" time and label values. To compute this loss, we omit $C_\emptyset$ since it does not influence the gradient during training.

Horizon Matching Loss Our approach is designed to predict K future events $\{\hat{y}_i\}_{i=1}^{K}$ within the time horizon H. The set of ground truth events within this horizon is denoted by $\{y_i\}_{i=1}^{T}$, where the number of events T may vary. We propose to align the predicted sequence with the ground truth sequence by finding the matching that minimizes the following loss function, motivated by object detection techniques from computer vision, such as DeTR (Carion et al. 2020):

$$\mathcal{L}(y, \hat{y})_{\text{matching}} = \min_{\sigma \in \mathcal{A}} \left[\sum_{i=1}^{T} \mathcal{L}_{\text{pair}}(y_i, \hat{y}_{\sigma(i)}) + \mathcal{L}_{\text{BCE}}(\sigma, \hat{y}) \right], \tag{4}$$

where $\mathcal{A}$ is the set of all possible alignments between the ground truth and predicted sequences and σ represents a specific alignment. The optimal matching is computed using the Hungarian algorithm (Kuhn 1955), which has a cubic computational complexity with respect to the sequence length. The pairwise loss $\mathcal{L}_{\text{pair}}$ is similar to the negative log-likelihood of the ground truth event y_i given the predicted distribution $\hat{y}_{\sigma(i)}$:

$$\mathcal{L}_{\text{pair}}(y_i, \hat{y}_{\sigma(i)}) = |t_i - \hat{t}_{\sigma(i)}| - \log \hat{p}_{\sigma(i)}(l_i), \tag{5}$$

where $y = (t, l)$ is a ground truth event, $\hat{t}$ is the predicted timestamp, and $\hat{p}(l)$ is the predicted probability of the correct label. The binary cross-entropy $\mathcal{L}_{\text{BCE}}$ trains the model to predict the occurrence probability of events:

$$\mathcal{L}_{\text{BCE}}(\sigma, \hat{y}) = -\sum_{i \in \sigma} \log \hat{o}_i - \sum_{i \notin \sigma} \log(1 - \hat{o}_i), \tag{6}$$

where $\hat{o}_i$ is the predicted probability that the i-th event is matched with some ground truth event.

By minimizing $\mathcal{L}_{\text{matching}}(y, \hat{y})$, we train the model to accurately predict the parameters of the ground truth sequence, adapting to sequences of varying lengths, up to a maximum of K events. Unlike object detection training objectives such as DeTR (Carion et al. 2020), DEF employs the same loss function for both the matching process and model training. Specifically, it integrates the alignment loss $\mathcal{L}_{\text{BCE}}$ into the matching cost.

To enable the method to address both the next-event prediction task and long-horizon forecasting accurately, we incorporate the next event prediction loss into the first output head. The final training objective is defined as follows:

$$\mathcal{L}_{\text{DEF}}(y, \hat{y}) = \mathcal{L}_{\text{matching}}(y, \hat{y}) + \lambda \left[|t_1 - \hat{t}_1| - \log \hat{p}_1(l_1) \right]. \tag{7}$$

Calibration and Inference Inference in our method involves two steps: filtering and sorting. First, predicted events are filtered based on their occurrence probabilities $\hat{o}_i$. The remaining events are sorted according to their predicted timestamps, forming the final output sequence. However, in practice, an additional calibration step is necessary. Without calibration, the model tends to predict a small number of events due to a bias in the predicted occurrence probabilities $\hat{o}_i$ toward the matching frequency of each head, which is typically below the 0.5 threshold. Calibration aims to determine optimal prediction thresholds for all $\hat{o}_i$, aligning the prediction rates with the matching probabilities. This calibration is

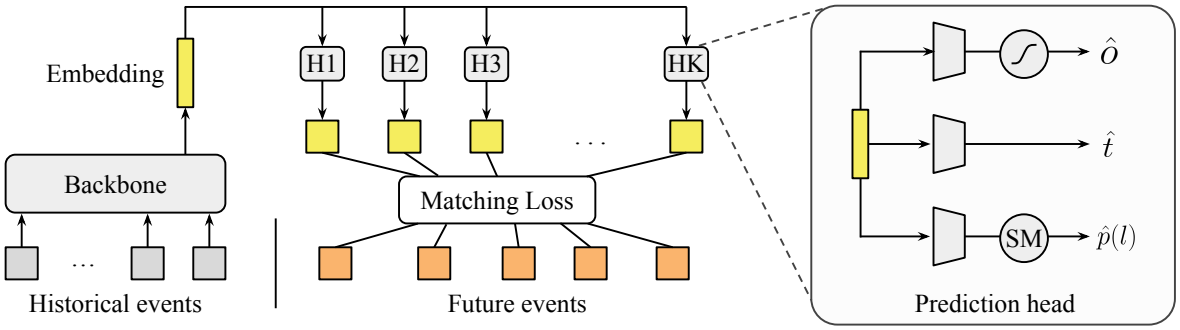


Figure 2: The proposed DEF simultaneously predicts K future events. Each prediction head outputs occurrence probability $\hat{o}$, time $\hat{t}$, and labels distribution $\hat{p}(l)$. During training, a novel matching loss aligns predictions with the ground truth sequence and evaluates its likelihood.

Dataset	Domain	Sequences	Events	Avg. Length	Classes	Time unit	OTD steps	Horizon / Mean length
StackOverflow	Social. net.	2k	138k	64.2	22	Minute	10	10 / 12.0
Amazon	Social. net.	9k	403k	43.6	16	N/A	5	10 / 14.8
Retweet	Social. net.	23k	1.3M	56.4	3	Second	10	180 / 14.7
MIMIC-IV	Medical	120k	2.4M	19.7	34	Day	5	28 / 6.6
Transactions	Financial	50k	43.7M	875	203	Day	5	7 / 9.0

Table 1: Datasets statistics and evaluation parameters

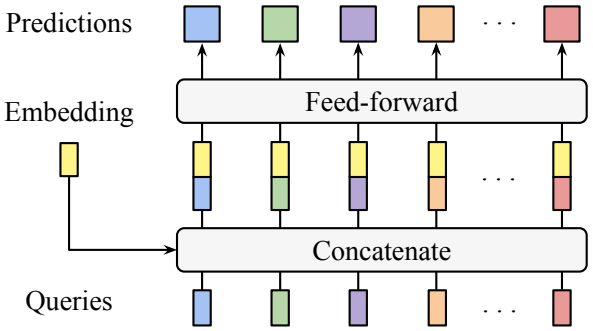


Figure 3: Our conditional prediction head.

performed on the fly during training by tracking matching frequencies and computing the corresponding quantiles using a streaming algorithm. The calibration algorithm is outlined in the extended version of the paper.

Conditional Head Architecture Implementing a separate feed-forward network for each prediction head leads to a large number of parameters and increases the risk of overfitting. Instead, we adopt a conditional approach that uses a single feed-forward network applied to K trainable query vectors, as illustrated in Fig. 3. The model receives two inputs: a query vector and a context vector, which are concatenated and passed through the shared feed-forward network. In this setup, each trainable query vector encodes the specific information required for its corresponding output head. This design enables the generation of all K outputs using a single network, significantly reducing the total number of parameters, accelerating convergence, and improving prediction quality.

Hyperparameter Selection Our method has one key hyperparameter, the maximum number of predictions K, which controls the maximum number of predictions. It typically requires tuning for each dataset, and we recommend setting K to approximately four times the average sequence length in a horizon for which it is required to predict future events. In our experiments, K values ranged from 32 to 64, depending on the characteristics of the dataset.

Moreover, it may be necessary to assign the weights to each component of the proposed loss function. The value of λ from Eq. 7 was set to 4 in all experiments. Additionally, we found it beneficial to adjust the weight of each loss component during alignment to accommodate the number of model outputs, dataset classes, and the average time step. In practice, the optimal weight for $\mathcal{L}_{\text{BCE}}$ is typically around 8 times larger than the weights for the label and timestamp losses. For datasets with larger time steps, such as Retweet, the MAE loss weight should be reduced accordingly. The exact values of hyperparameters used in our experiments are provided in source code.

Experiments

We conducted a series of experiments using the HoTPP benchmark (Karpukhin, Shipilov, and Savchenko 2024) to assess the performance of our approach against several widely used MTPP models: IFTPP (Shchur, Biloš, and

Model	Metrics (OTD ↓ / T-mAP ↑)				
	StackOverflow	Amazon	Retweet	MIMIC-IV	Transactions
IFTPP	13.64 / 8.31% ±0.05 / ±0.50%	6.52 / 22.56% ±0.05 / ±0.52%	172.7 / 31.75% ±4.4 / ±4.44%	11.53 / 21.67% ±0.01 / ±0.21%	6.90 / 5.88% ±0.01 / ±0.13%
IFTPP-T	13.61 / 8.81% ±0.05 / ±0.20%	6.59 / 23.42% ±0.02 / ±0.13%	166.1 / 40.05% ±1.1 / ±1.73%	**11.48** / 23.68% ±0.02 / ±0.29%	6.85 / 5.55% ±0.02 / ±0.19%
RMTPP	13.17 / 12.72% ±0.05 / ±0.16%	6.57 / 20.06% ±0.03 / ±0.33%	166.7 / 44.74% ±3.3 / ±0.94%	13.71 / 21.08% ±0.03 / ±0.29%	6.88 / 6.69% ±0.01 / ±0.12%
NHP	13.24 / 11.96% ±0.02 / ±0.40%	9.02 / 26.29% ±0.35 / ±0.55%	165.8 / 45.07% ±1.6 / ±0.34%	18.60 / 7.32% ±0.19 / ±1.33%	6.98 / 5.61% ±0.01 / ±0.05%
AttNHP	13.30 / 11.13% ±0.02 / ±0.32%	7.30 / 14.62% ±0.06 / ±0.80%	171.6 / 25.85% ±1.0 / ±1.08%	14.68 / 22.46% ±0.08 / ±0.40%	7.50 / 1.48% N/A / N/A
ODE	13.27 / 10.52% ±0.03 / ±0.23%	9.46 / 22.96% ±0.08 / ±0.61%	165.3 / 44.81% ±0.5 / ±0.69%	14.74 / 15.18% ±0.34 / ±0.15%	6.97 / 5.52% ±0.01 / ±0.13%
HYPRO	13.26 / 14.69% N/A / N/A	6.61 / 20.53% N/A / N/A	170.7 / 46.99% N/A / N/A	14.87 / 16.77% N/A / N/A	7.05 / 7.05% N/A / N/A
Diffusion	13.01 / 15.07% ±0.16 / ±0.74%	6.52 / 30.29% ±0.04 / ±0.32%	158.0 / 52.24% ±1.1 / ±0.68%	13.28 / 22.82% ±0.14 / ±0.12%	6.88 / 6.04% ±0.00 / ±0.08%
DEF	**12.14 / 22.72%** ±0.04 / ±0.32%	**5.98 / 37.20%** ±0.04 / ±0.06%	**132.9 / 57.93%** ±0.7 / ±0.33%	12.95 / **30.35%** ±0.32 / ±0.25%	**6.70 / 9.26%** ±0.03 / ±0.09%
Rel. Impr.	+6.7% / +50.8%	+8.3% / +22.8%	+15.9% / +10.9%	−12.8% / +28.2%	+2.2% / +31.3%

Table 2: Evaluation results in the long-horizon prediction task. The best result is shown in bold. Mean and STD values of 5 runs with different random seeds are reported. For HYPRO and AttNHP on the Transactions dataset, we report results using a single seed, as these methods exhibit low computational efficiency.

Günnemann 2020), its transformer-based variant (IFTPP-T), intensity-based RMTPP (Du et al. 2016) and NHP (Mei and Eisner 2017) approaches, ODE-RNN (Rubanova, Chen, and Duvenaud 2019), the transformer-based AttNHP model (Yang, Mei, and Eisner 2022), as well as the long-horizon HYPRO (Xue et al. 2022) and Diffusion (Zhou et al. 2025) methods. Similar to IFTPP and RMTPP, DEF utilizes a Gated Recurrent Unit (GRU) (Cho et al. 2014) as its backbone. We employ the density variant of the IFTPP model and use a Laplace distribution, in place of a log-normal distribution, to better align with the target metrics.

The datasets employed in this study include Retweet (Zhao et al. 2015), Amazon (Jianmo 2018), StackOverflow (Jure 2014), MIMIC-IV (Johnson et al. 2023), and Transactions (AI-Academy for teens 2021), which represent a diverse range of domains and scales. Detailed dataset statistics are provided in Table 1.

MTPP models are typically evaluated based on their accuracy in predicting the next event (Xue et al. 2024). Time and type predictions are often assessed separately, with type prediction quality measured by the error rate and time prediction evaluated using regression metrics such as Mean Absolute Error (MAE). Recent advancements have introduced additional metrics, including Optimal Transport Distance (OTD) (Mei, Qin, and Eisner 2019) and Temporal mAP (T-mAP) (Karpukhin, Shipilov, and Savchenko 2024), which assess long-horizon predictions by comparing predicted sequences to ground truth sequences within a specified horizon. The horizon H should be chosen to align with the requirements of the specific task. In our experiments, we set H to match the horizon used in the T-mAP

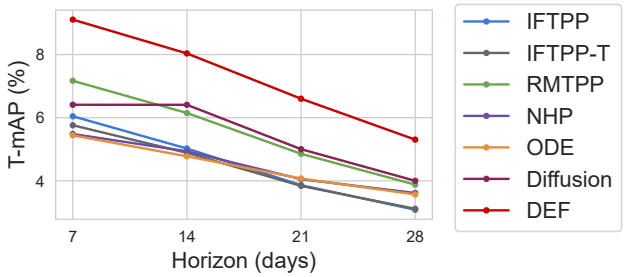


Figure 4: Extended horizons prediction on the Transactions dataset.

metric from the HoTPP benchmark (Karpukhin, Shipilov, and Savchenko 2024), ensuring consistency in evaluation. The selected value is also sufficient to include the necessary number of events for calculating the OTD metric (Mei, Qin, and Eisner 2019). In this work, we employ all of the aforementioned metrics to evaluate the performance of our proposed method. Additionally, we measure prediction diversity using the entropy of the predicted labels distribution.

Note that autoregressive methods typically condition on the ground-truth values of all preceding events when predicting the next event. This assumption becomes unrealistic for long-horizon forecasting. Consequently, we do not include log-likelihood evaluation in our experiments.

Further details on metric computation, ablation studies, additional experiments for general event sequences with multiple data fields, and the training process are provided

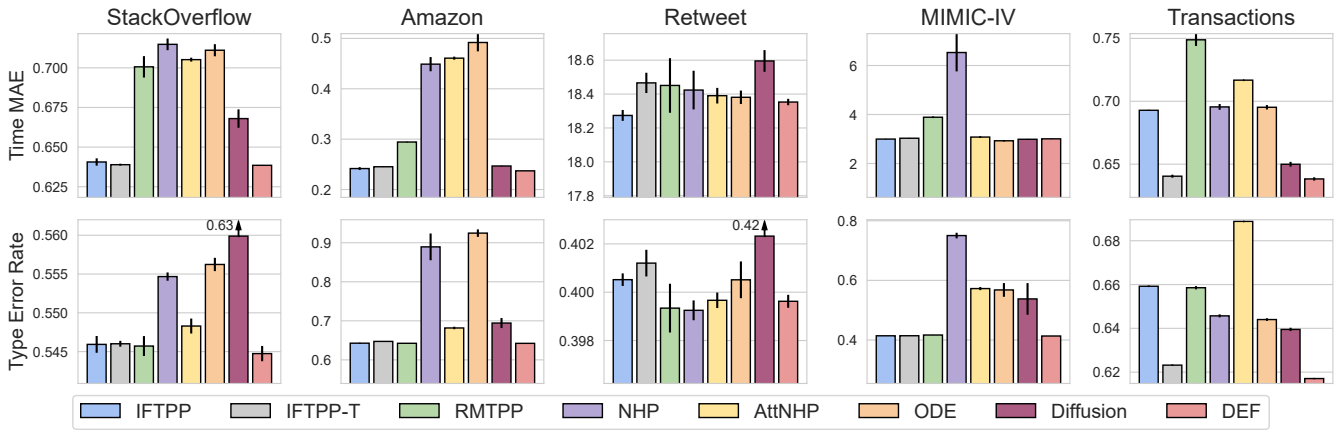


Figure 5: Next event prediction errors: MAE for time and error rate for type.

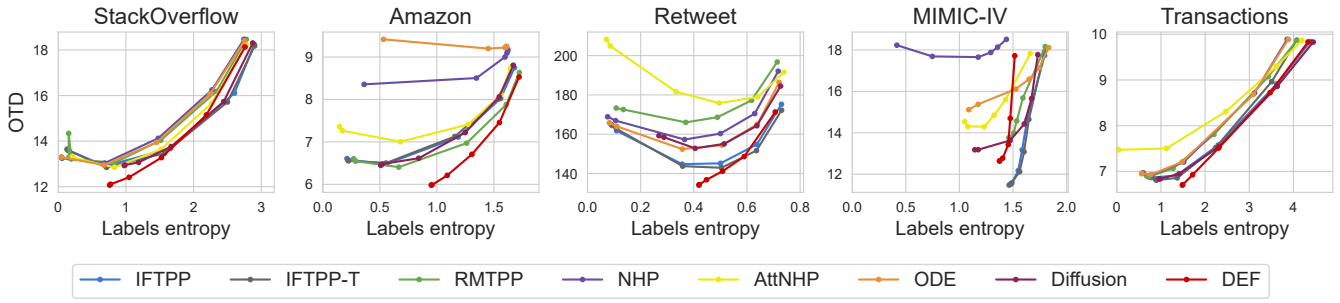


Figure 6: The relation between OTD and predictions' diversity for varying sampling temperature values. The optimal quality (low error and high diversity) corresponds to the bottom right corner.

in the extended version of the paper.

Long-Horizon Events Forecasting We evaluate long-horizon prediction performance using OTD and T-mAP. As shown in Table 2, DEF significantly outperforms existing approaches: it achieves state-of-the-art performance in 9 out of 10 comparisons. The only exception is the OTD metric on the MIMIC-IV dataset, where IFTPP ranks first and DEF second. The high T-mAP scores of our approach can be linked to its training objective, which utilizes matching, similar to T-mAP. However, it also consistently improves the OTD metric, suggesting that its training process enhances overall model performance rather than merely optimizing for a single evaluation criterion.

While the base implementation of our method models a fixed prediction horizon H, this can be seen as a potential limitation. To address this, we explore a hybrid strategy that combines our approach with an autoregressive mechanism. Specifically, predictions over horizon H are appended to the input sequence and used recursively to predict subsequent horizons. A similar autoregressive extension is applied to the Diffusion model for comparison. We evaluate this hybrid approach on the Transactions dataset, as its sequences are sufficiently long to support extended forecasting (see Table 1 for dataset details). Results, shown in Figure 4, demonstrate that the autoregressive variant of our method outperforms both traditional autoregressive baselines and fixed-horizon prediction models in capturing long-range event dynamics.

Next Event Prediction Generative event sequences models and MTPPs are usually evaluated based on the quality of next-event prediction. We measured the next-event type error rate and mean absolute time error (MAE) across various datasets, with the results presented in Fig. 5. The proposed method achieves state-of-the-art results in all comparisons and significantly reduces error on the Transactions dataset, the only dataset where the difference between top methods is significant. Notably, the Transactions dataset also has the largest number of event types. Thus, DEF, although primarily designed for long-horizon prediction, also achieves high-quality performance in the next-event prediction task.

Predictions Diversity As demonstrated qualitatively in Fig. 1, popular autoregressive and horizon prediction methods often produce repetitive outputs. This section provides additional quantitative results to further emphasize the differences between traditional approaches and our method.

A common technique for increasing prediction diversity is adjusting the temperature during sampling. When the temperature approaches zero, the model selects the label with

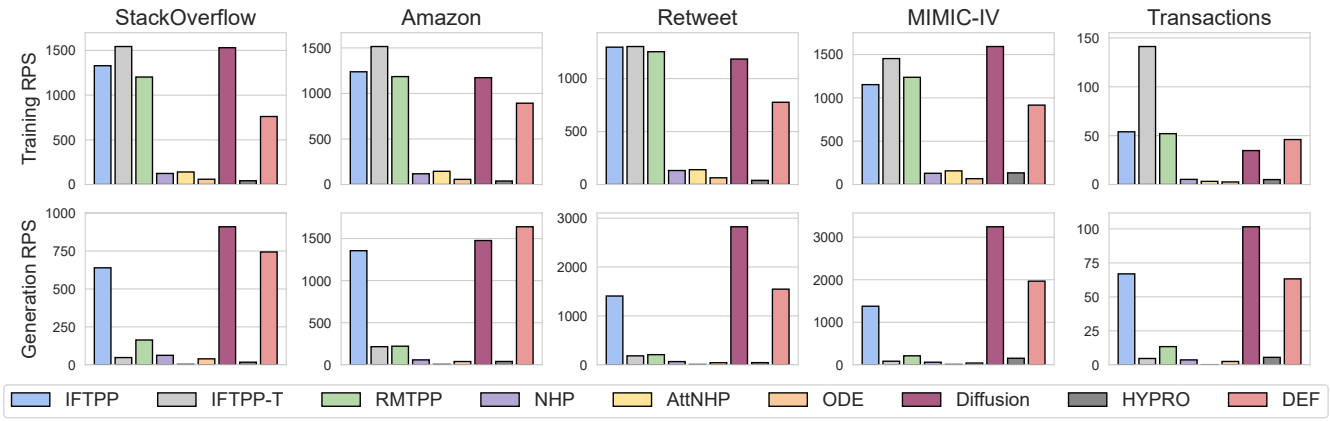


Figure 7: Training and sequence generation computation speed, Requests Per Second (RPS).

the highest probability, resulting in deterministic predictions. As the temperature increases, the model samples from a more uniform label distribution, leading to higher entropy and greater diversity. To analyze the relationship between temperature, prediction diversity, and long-horizon quality, we vary the temperature from 0 to 10, measuring the average entropy of predicted event types within the horizon. Additionally, we assess long-horizon prediction quality using OTD, as T-mAP is not affected by temperature changes.

Fig. 6 indicates that our method achieves the best balance between diversity and prediction accuracy on four out of five datasets. While some methods benefit from sampling-based approaches compared to maximum-probability predictions, they still fail to match DEF's overall performance.

Training and Inference Speed A key practical consideration for any model is its computational efficiency. Fig. 7 compares the training and inference times of various methods in terms of Requests Per Second (RPS), i.e., the number of processed batch elements[1]. The results indicate that our approach exhibits a moderate training time while being one of the fastest methods during inference, alongside Diffusion and IFTPP.

The high inference speed of Diffusion is attributed to the low parameter count in its denoising model, optimized during hyperparameter tuning. In contrast, the proposed method employs a fully connected network for each output prediction, which impacts its RPS. Notably, all models utilizing NHP loss exhibit low training and inference speeds due to the computationally expensive sampling required at each step. Similarly, HYPRO demonstrates low computational efficiency, as it necessitates multiple autoregressive generations per step. Thus, our method ranks among the most computationally efficient methods in the field.

Conclusion and Future Works

In this work, we introduced a novel approach, called DEF, that addresses the challenges of long-horizon event forecasting by leveraging a matching-based training objective. Our experiments demonstrate that the proposed method effectively overcomes the limitations of traditional techniques, including specialized long-horizon models such as HYPRO and Diffusion. Notably, our approach not only achieves substantial improvements in long-horizon prediction accuracy but also sets a new state-of-the-art in next-event modeling, while generating more diverse predictions. Additionally, it offers greater computational efficiency at inference time by predicting multiple future events in parallel. This work advances event sequence modeling and opens new opportunities for a wide range of real-world applications.

We acknowledge several limitations of this work. First, we follow prior works (Shchur, Biloš, and Günnemann 2020; Du et al. 2016) in assuming conditional independence between event attributes. While we focus primarily on the correct alignment of predictions and ground truth, the DEF architecture can be extended to model interdependencies. Second, similar to Diffusion (Zhou et al. 2025), we omit interdependencies between different events within the predicted horizon. Additional techniques, such as beam search or rescoring, could be incorporated to improve the model further. Finally, the training efficiency of our horizon matching loss relies heavily on the Hungarian algorithm, which may benefit from optimization of our custom CUDA kernel.

Future research could investigate the integration of rescoring techniques, such as those in (Xue et al. 2022), or the application of beam search to enhance predictive performance. Additionally, better time modeling might be achieved by integrating intensity-based approaches, like NHP or RMTPP, with DEF, offering another promising direction for research. Furthermore, some techniques from our method could be adapted for object detection in computer vision (Carion et al. 2020). We introduced a probabilistic framework that unifies different loss functions, using the same objective during matching and backpropagation, enhancing optimization robustness. We also employed an occurrence score during matching, which, as shown in our ablation studies, significantly improves the model performance on most datasets.

[1]Experiments were conducted on an Nvidia RTX 4060 GPU

Acknowledgments

The work of A. Savchenko was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

References

AI-Academy for teens. 2021. Age Group Prediction. https://kaggle.com/competitions/clients-age-group. Accessed: 2025-12-05.

Bishop, C. M. 1994. *Mixture density networks*. Aston University.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 213–229. Springer.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.

De Brouwer, E.; Simm, J.; Arany, A.; and Moreau, Y. 2019. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Du, N.; Dai, H.; Trivedi, R.; Upadhyay, U.; Gomez-Rodriguez, M.; and Song, L. 2016. 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*, 1555–1564.

Jia, J.; and Benson, A. R. 2019. Neural jump stochastic differential equations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Jianmo, N. 2018. Amazon Review Data. https://nijianmo.github.io/amazon. Accessed: 2025-12-05.

Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2023. Mimic-IV. https://physionet.org/content/mimiciv/2.2/. Accessed: 2025-12-05.

Jure, L. 2014. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data. Accessed: 2025-12-05.

Karpukhin, I.; Shipilov, F.; and Savchenko, A. 2024. HoTPP Benchmark: Are We Good at the Long Horizon Events Forecasting? *arXiv preprint arXiv:2406.14341*.

Kidger, P.; Morrill, J.; Foster, J.; and Lyons, T. 2020. Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 6696–6707.

Kostromina, A.; Kuvshinova, K.; Yugay, A.; Savchenko, A.; and Simakov, D. 2025. Tsururu: A Python-based Time Series Forecasting Strategies Library. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 11077–11081. Demo Track.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Kuleshov, I.; Boeva, G.; Zhuzhel, V.; Romanenkova, E.; Vorsin, E.; and Zaytsev, A. 2024. COTODE: COntinuous Trajectory neural Ordinary Differential Equations for modelling event sequences. *arXiv preprint arXiv:2408.08055*.

Lim, B.; and Zohren, S. 2021. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194): 20200209.

Lin, H.; Wu, L.; Zhao, G.; Pai, L.; and Li, S. Z. 2022. Exploring Generative Neural Temporal Point Process. *Transactions on Machine Learning Research*.

McDermott, M.; Nestor, B.; Argaw, P.; and Kohane, I. S. 2024. Event Stream GPT: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Mei, H.; and Eisner, J. M. 2017. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Mei, H.; Qin, G.; and Eisner, J. 2019. Imputing missing events in continuous-time event streams. In *Proceedings of the International Conference on Machine Learning (ICML)*, 4475–4485. PMLR.

Omi, T.; Aihara, K.; et al. 2019. Fully neural network based model for general temporal point processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Padhi, I.; Schiff, Y.; Melnyk, I.; Rigotti, M.; Mroueh, Y.; Dognin, P.; Ross, J.; Nair, R.; and Altman, E. 2021. Tabular transformers for modeling multivariate time series. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3565–3569. IEEE.

Panos, A. 2024. Decomposable transformer point processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 88932–88955.

Rizoiu, M.-A.; Lee, Y.; Mishra, S.; and Xie, L. 2017. Hawkes processes for events in social media. In *Frontiers of Multimedia Research*, 191–218.

Rubanova, Y.; Chen, R. T.; and Duvenaud, D. K. 2019. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Savchenko, A.; and Kachan, O. 2025. HN-MVTS: HyperNetwork-based Multivariate Time Series Forecasting. *arXiv preprint arXiv:2511.08340*.

Shchur, O.; Biloš, M.; and Günnemann, S. 2020. Intensity-Free Learning of Temporal Point Processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shchur, O.; Türkmen, A. C.; Januschowski, T.; and Günnemann, S. 2021. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*.

Song, Y.; Donghyun, L.; Meng, R.; and Kim, W. H. 2024. Decoupled Marked Temporal Point Process using Neural Ordinary Differential Equations. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Wang, C.; and Xiao, Z. 2022. A deep learning approach for credit scoring using feature embedded Transformer. *Applied Sciences*, 12(21): 10995.

Wang, Z.; and Sun, J. 2022. TransTab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 2902–2915.

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

Xiao, S.; Yan, J.; Yang, X.; Zha, H.; and Chu, S. 2017. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Xue, S.; Shi, X.; Chu, Z.; Wang, Y.; Zhou, F.; Hao, H.; Jiang, C.; Pan, C.; Xu, Y.; Zhang, J. Y.; et al. 2024. EasyTPP: Towards open benchmarking the temporal point processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xue, S.; Shi, X.; Zhang, J.; and Mei, H. 2022. Hypro: A hybridly normalized probabilistic model for long-horizon prediction of event sequences. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 34641–34650.

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

Zeng, M.; Regol, F.; and Coates, M. 2024. Interacting diffusion processes for event sequence forecasting. In *Proceedings of the International Conference on Machine Learning (ICML)*, 58407–58430.

Zhang, Q.; Lipani, A.; Kirnap, O.; and Yilmaz, E. 2020. Self-attentive Hawkes process. In *Proceedings of the International Conference on Machine Learning (ICML)*, 11183–11193. PMLR.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522.

Zhou, W.-T.; Kang, Z.; Tian, L.; Zhang, J.; and Liu, Y. 2025. Non-autoregressive diffusion-based temporal point processes for continuous-time long-term event prediction. *Expert Systems with Applications*, 267: 126210.

Zhuzhel, V. A.; Grabar, V.; Boeva, G.; Zabolotnyi, A.; Stepikin, A.; Zholobov, V.; Ivanova, M.; Orlov, M.; Kireev, I. A.; Burnaev, E.; et al. 2023. COTIC: Embracing Non-uniformity in Event Sequence Data via Multilayer Continuous Convolution. *arXiv preprint arXiv:2302.06247*.

Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; and Zha, H. 2020. Transformer hawkes process. In *Proceedings of the International Conference on Machine Learning (ICML)*, 11692–11702. PMLR.