

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 LONG-RANGE MODELING AND PROCESSING OF MULTIMODAL EVENT SEQUENCES

Anonymous authors

Paper under double-blind review

ABSTRACT

Temporal point processes (TPPs) have emerged as powerful tools for modeling asynchronous event sequences. While recent advances have extended TPPs to handle textual information, existing approaches are limited in their ability to generate rich, multimodal content and reason about event dynamics. A key challenge is that incorporating multimodal data dramatically increases sequence length, hindering the ability of attention-based models to generate coherent, long-form textual descriptions that require long-range understanding. In this paper, we propose a novel framework that extends LLM-based TPPs to the visual modality, positioning text generation as a core capability alongside time and type prediction. Our approach addresses the long-context problem through an adaptive sequence compression mechanism based on temporal similarity, which reduces sequence length while preserving essential patterns. We employ a two-stage paradigm of pre-training on compressed sequences followed by supervised fine-tuning for downstream tasks. Extensive experiments, including on the challenging DanmakuTPP-QA benchmark, demonstrate that our method outperforms state-of-the-art baselines in both predictive accuracy and the quality of its generated textual analyses.

1 INTRODUCTION

Temporal point processes (TPPs) (Daley & Vere-Jones, 2003; 2007) are statistical models designed for analyzing asynchronous event sequences in continuous time. They have found widespread applications across diverse domains including neuroscience (Zhou et al., 2021), finance (Bacry et al., 2015), crime analysis (Mohler, 2013), and epidemiology (Rizoiu et al., 2018). While traditional statistical TPPs can model event timing and dependencies, they often lack the expressiveness needed to capture complex real-world patterns. Recent advances in deep learning have led to the development of deep TPPs (Du et al., 2016; Mei & Eisner, 2017; Zuo et al., 2020; Meng et al., 2024), which demonstrate superior modeling capabilities. Building on the success of large language models (LLMs), Language-TPP (Kong et al., 2025) further extended TPPs to incorporate textual information, showing the potential of integrating natural language processing with temporal modeling.

However, real-world event sequences are increasingly becoming multimodal, extending beyond just temporal and textual information. For instance, in video streaming platforms, user comment streams (known as Danmaku) contain not only timestamps and textual content but also associated visual frames from the video (Jiang et al., 2025). Similarly, traffic monitoring systems capture accident reports that include temporal information, textual descriptions, audio recordings, and visual evidence from surveillance cameras. Current TPP models, including Language-TPP, remain largely confined to limited modalities and lack the capability to effectively model and generate rich multimodal content such as images alongside text and temporal information.

Extending deep TPPs to be fully generative in multimodal settings presents two primary challenges. **First, existing TPP models lack a unified framework for multimodal generation.** While Language-TPP has progressed in incorporating text, it cannot handle visual content or generate text that is conditioned on it. Different modalities require distinct encoding strategies (Lahat et al., 2015), and a truly generative framework must not only encode this heterogeneous data but also produce contextually aware content—such as generating a textual description for a future event that correctly references its associated visual information. To address this, we propose a unified framework

054 that extends Language-TPP’s template-based approach to incorporate visual information, enabling
 055 comprehensive multimodal event modeling and generation.
 056

057 **Second, incorporating multimodal data dramatically increases sequence length, hindering the**
 058 **ability to generate coherent, long-form text that depends on long-range dependencies.** This in-
 059 flation occurs because non-textual modalities are themselves tokenized into sequences; for example,
 060 a single image is commonly decomposed into hundreds of patch-based tokens (Dosovitskiy et al.,
 061 2021). In event sequences where each point is associated with such data, the total length N can be-
 062 come exceptionally large. This makes the $\mathcal{O}(N^2)$ complexity of the Transformer’s (Vaswani, 2017)
 063 self-attention mechanism a critical bottleneck. This prevents the model from accessing the full his-
 064 tory needed to generate a comprehensive analytical report or answer summary questions. To mitigate
 065 this, we propose an adaptive compression mechanism that leverages temporal similarity to reduce
 066 sequence length while preserving the dynamics required for effective long-range text generation.
 067

068 In summary, to address the above challenges, we propose **Multimodal Temporal Point Processes**
 069 (**MM-TPP**), a unified framework that tackles these issues through the following contributions:
 070

- 071 (1) We propose MM-TPP, an extension of Language-TPP to a multimodal framework that predicts
 072 event time and type while generating rich text conditioned on visual, textual, and temporal inputs.
- 073 (2) We introduce an adaptive compression mechanism based on temporal similarity that reduces
 074 sequence length, making the generation of long-form, context-aware text computationally feasible.
- 075 (3) We construct TAXI-PRO, a new multimodal TPP dataset by augmenting the classic NYC Taxi
 076 dataset with visual map patches and textual descriptions, providing a valuable benchmark.
- 077 (4) Comprehensive experiments on two multimodal TPP datasets, including the DanmakuTPP-QA
 078 benchmark, demonstrate the superiority of MM-TPP in predictive accuracy compared to SOTA base-
 079 lines. Furthermore, qualitative analysis reveals that our model generates significantly more insightful
 080 and coherent textual analyses, showcasing its superior reasoning capabilities.

081 2 RELATED WORKS

082 **Deep TPPs.** Deep learning-based TPPs were first introduced by Du et al. (2016) using an RNN-
 083 based framework. Follow-up works improved this approach, including Mei & Eisner (2017); Yang
 084 et al. (2018); Omi et al. (2019); Soen et al. (2021); Chen et al. (2024b). Recurrent models offer
 085 efficient inference, with constant-time prediction and linear training cost, but suffer from limited
 086 parallelism and poor modeling of long-range dependencies. To address these issues, Transformer-
 087 based TPPs emerged, starting with Zuo et al. (2020) and Zhang et al. (2020), and later extended
 088 by works such as Zhu et al. (2021); Zhou et al. (2022); Yang et al. (2022); Meng et al. (2024);
 089 Panos (2024). These models enable parallel training and better capture long-range dependencies
 090 via self-attention, but require higher computational cost—both time and memory scale as $\mathcal{O}(N^2)$.
 091 In contrast to these approaches, MM-TPP not only handles long-range dependencies efficiently but
 092 also extends TPP modeling to generate and reason over rich multimodal content.

093 **Covariate TPPs.** Covariate TPPs enhance traditional models by incorporating contextual informa-
 094 tion to better explain event dynamics and improve predictions. Early works focused on structured
 095 covariates—e.g., Meyer et al. (2012) used demographic data for epidemic modeling, Xiao et al.
 096 (2017) proposed a dual-LSTM for combining covariates and event history, and Adelfio & Chiodi
 097 (2021); Gajardo & Müller (2021) applied similar ideas to earthquake and COVID-19 data. In prac-
 098 tice, many covariates are unstructured, such as text, images, or audio, making their integration into
 099 TPPs challenging. Some studies tackled this by using word frequency features (Zhu & Xie, 2022)
 100 or BERT embeddings (Zhang et al., 2023). More recent efforts leverage LLMs to jointly model
 101 temporal and textual information, as in Liu & Quan (2024); Kong et al. (2025). MM-TPP builds on
 102 this direction by incorporating multiple types of unstructured covariates—specifically, images and
 103 text—into a unified TPP framework, and further enables multimodal generation for future events.

104 3 PRELIMINARIES

105 This section introduces the foundational concepts and background knowledge for our proposed
 106 model, including TPPs and LLM-based TPPs modeling.

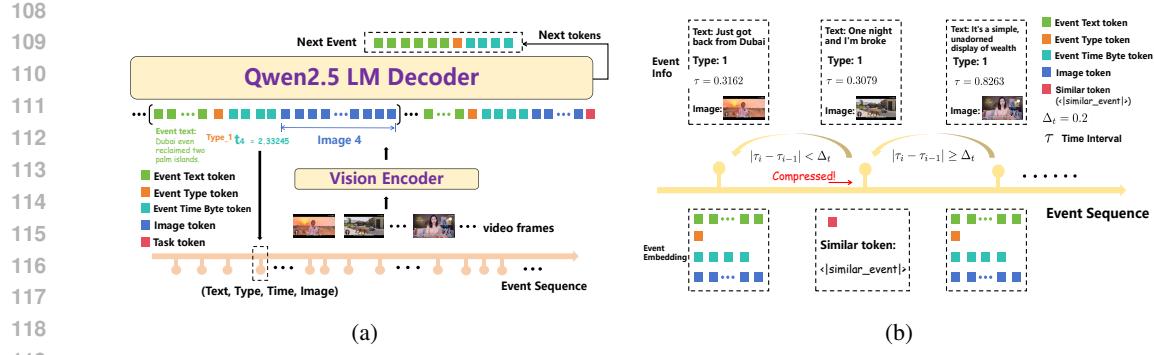


Figure 1: Overview of the MM-TPP framework. (a) Model Architecture: The framework is built upon Qwen2.5-VL, where multimodal events—consisting of time, type, text, and image information—are tokenized into unified token sequences. Visual content is encoded by the vision encoder and fused with other modalities via the language model decoder to perform autoregressive next-event prediction. (b) Adaptive Long Sequence Compression: Events with similar time intervals ($|\tau_i - \tau_{i-1}| < \Delta_t$) are compressed into a special token. This strategy enables efficient processing of long event sequences while preserving temporal structure and key event information.

3.1 TEMPORAL POINT PROCESSES

TPPs are probabilistic models designed for sequences of discrete events occurring in continuous time. A typical realization of a TPP is denoted by $\mathcal{S} = \{(t_i, e_i)\}_{i=1}^N$, where each (t_i, e_i) represents an event of type $e_i \in \mathcal{E} = \{1, \dots, E\}$ that occurs at time $t_i \in (0, T]$, satisfying $0 < t_1 < \dots < t_N \leq T$. Here, N denotes the random number of events observed in the time window $(0, T]$. There are multiple equivalent formulations to characterize TPPs. One common approach is to model the conditional distribution of future events given the past. Let \mathcal{H}_t denote the event history up to and including time t . We define $p(t_{i+1}, e_{i+1} | \mathcal{H}_{t_i})$ as the conditional density function for the next event occurring at time t_{i+1} with type e_{i+1} , given the past history \mathcal{H}_{t_i} . Under this formulation, the probability of observing a sequence \mathcal{S} can be factorized as follows:

$$p(\mathcal{S}) = \prod_{i=1}^N p(t_i, e_i | \mathcal{H}_{t_{i-1}}) \cdot (1 - P(T | \mathcal{H}_{t_N})),$$

where $1 - P(T | \mathcal{H}_{t_N})$ represents the probability that no events occur in $(t_N, T]$, $P(T | \mathcal{H}_{t_N}) = \int_{t_N}^T p(t | \mathcal{H}_{t_N}) dt$ is the cumulative density of the next event time after t_N , and $p(t | \mathcal{H}_{t_i}) = \sum_{e=1}^E p(t, e | \mathcal{H}_{t_i})$. It is standard to decompose the conditional joint density as:

$$p(t, e | \mathcal{H}_{t_i}) = p(t | \mathcal{H}_{t_i}) \cdot p(e | \mathcal{H}_{t_i}, t).$$

Therefore, based on the above formulations, the log-likelihood of a TPP can be derived as follows:

$$\log p(\mathcal{S}) = \sum_{i=1}^N \log p(t_i | \mathcal{H}_{t_{i-1}}) + \log p(e_i | \mathcal{H}_{t_{i-1}}, t_i) + \log (1 - P(T | \mathcal{H}_{t_N})). \quad (1)$$

3.2 LANGUAGE-TPP

Recent advances in LLMs have enabled new approaches to TPP modeling. Early studies adapted pre-trained LLMs to temporal event data using customized tokenization and fine-tuning (Liu & Quan, 2024; Xue et al., 2023; Shi et al., 2024). Following this direction, Kong et al. (2025) proposed Language-TPP, a framework for modeling event sequences of the form $(t_i, e_i, m_i)_{i=1}^N$, where each event includes a timestamp, type, and textual description. The model captures dependencies across the sequence and jointly predicts future event time, type, and description.

To achieve this, Language-TPP adopts a decoder-only causal Transformer and introduces several components. Each event is tokenized by applying a standard language tokenizer to e_i and m_i ,

162 while t_i is encoded using a byte-level strategy with 256 special tokens, allowing each 32-bit timestamp
 163 to be represented by four tokens. The resulting tokens are organized into a structured event
 164 template, where each event is enclosed by $\langle \text{start_of_event} \rangle$ and $\langle \text{end_of_event} \rangle$ tokens,
 165 and each component is prefixed by a modality-specific marker (e.g., $\langle \text{time_prefix} \rangle$
 166 for timestamps). The model then performs autoregressive generation until an end-of-sequence token
 167 is produced, from which the timestamp, type, and description of future events can be decoded.
 168 Language-TPP is trained in two stages. First, event sequences are converted into token sequences
 169 and the backbone LLM is pre-trained using standard next-token prediction. In the second stage,
 170 prompt-response pairs are constructed for each subtask (time prediction, type classification, and
 171 description generation), and fine-tuning is performed on the response portion using next-token loss.

172 While Language-TPP effectively incorporates textual information into TPP modeling, it is limited
 173 to a single modality and faces scalability issues when handling long multimodal sequences. Our
 174 proposed MM-TPP extends this framework by incorporating both visual and textual modalities,
 175 and introduces a temporal similarity-based compression mechanism to address the sequence length
 176 explosion in multimodal settings.

177 4 METHODOLOGY

180 To model multimodal event sequences with rich visual and textual information under long-context
 181 constraints, we propose MM-TPP, an extension of Language-TPP. It jointly models temporal, cat-
 182 egorical, textual, and visual data, and introduces a compression mechanism to enhance long-range
 183 dependency handling.

185 4.1 MODEL FRAMEWORK

187 Fig. 1a provides an overview of the framework of MM-TPP. Our model is built upon Qwen2.5-
 188 VL (Bai et al., 2025), a SOTA pre-trained multimodal LLM. The entire framework follows a
 189 sequence-to-sequence paradigm, taking a multimodal event history sequence as input and autore-
 190 gressively predicting various attributes of future events. Extending the annotations from Kong et al.
 191 (2025), given a multimodal event sequence $(t_i, e_i, m_i, v_i)_{i=1}^N$, where t_i denotes the event timestamp,
 192 e_i represents the event type, m_i is the textual description, and v_i is the associated visual description,
 193 the goal is to model the dependencies within such a sequence and predict the time, type, textual
 194 content of future events.

195 4.2 MULTIMODAL EVENT SEQUENCE TOKENIZATION

197 To handle multimodal event data with language models, we design a tokenization scheme that con-
 198 verts each event tuple (t_i, e_i, m_i, v_i) into a unified token sequence suitable for autoregressive mod-
 199 eling. Temporal information is encoded following the design of Language-TPP, where each 32-bit
 200 time interval is decomposed into four bytes based on its memory layout. These bytes are mapped
 201 to 256 special tokens, $\langle \text{byte_0} \rangle$ to $\langle \text{byte_255} \rangle$, allowing for compact and precise repre-
 202 sentation of time intervals. Categorical information, such as event types, is handled using dedicated
 203 special tokens (e.g., $\langle \text{type_0} \rangle$ to $\langle \text{type_5} \rangle$ for six event categories), enabling unified and
 204 interpretable modeling of discrete labels.

205 For textual information, event descriptions (e.g., user comments) are tokenized using the built-in
 206 tokenizer of Qwen2.5-VL, which converts raw text into standard language tokens that can be directly
 207 processed by the model. Visual information is integrated differently: instead of converting image
 208 pixels into tokens, we insert a placeholder token $\langle \text{image_pad} \rangle$ at the appropriate position in
 209 the sequence. During model execution, the corresponding image is passed through Qwen2.5-VL’s
 210 vision encoder to produce visual embeddings, which are then aligned with the placeholder token.
 211 This design allows for deep fusion of visual, textual, and temporal features while keeping the token
 212 sequence compact and language-model-compatible.

213 To organize these multimodal components, we define a structured event template. Each event is
 214 enclosed by $\langle \text{start_of_event} \rangle$ and $\langle \text{end_of_event} \rangle$ tokens. Within the event, each
 215 modality is marked by a prefix token: $\langle \text{time_start} \rangle$ for time, $\langle \text{type_start} \rangle$ for type,
 $\langle \text{text_start} \rangle$ for text, and $\langle \text{vision_start} \rangle$ for image. This structured template pro-

216 vides a consistent and interpretable format for encoding complex multimodal events. A full example
 217 of the uncompressed event format is shown in Appendix F.
 218

219 **4.3 ADAPTIVE LONG SEQUENCE COMPRESSION**
 220

221 The inclusion of text and image tokens in multimodal event sequences significantly increases input
 222 length, posing challenges for Transformer-based LLMs due to high computational cost and limited
 223 context windows. To mitigate this, we introduce an adaptive compression strategy based on temporal
 224 similarity between events. In many real-world settings—such as video comment streams—events
 225 arrive in bursts or follow periodic rhythms, resulting in similar inter-event intervals. MM-TPP lever-
 226 ages this observation by compressing sequences of temporally similar events, as shown in Fig. 1b.
 227

228 During input construction, we define a temporal similarity threshold Δ and compare the interval
 229 $\tau_i = t_i - t_{i-1}$ of the current event with that of the previous one τ_{i-1} . If the difference between
 230 the two is smaller than the threshold, i.e., $|\tau_i - \tau_{i-1}| < \Delta$, event i is deemed temporally similar to
 231 event $i - 1$. In this case, instead of encoding it using the full multimodal template, it is replaced with
 232 a single special token $\langle \mid \text{similar_event} \mid \rangle$, thereby reducing the number of tokens. Details
 233 on choosing an appropriate value for Δ are discussed in Appendix A.2. Otherwise, if the time
 234 intervals differ significantly, event i is represented in full using the multimodal template defined
 235 in Section 4.2. **For a concrete step-by-step illustration of this sequential compression mechanism
 applied to consecutive events, please refer to Appendix A.1.**
 236

237 This strategy enables compact representation of dense event clusters—potentially requiring hun-
 238 dreds of tokens—using only a few $\langle \mid \text{similar_event} \mid \rangle$ tokens, while preserving key events
 239 with distinctive timing patterns. As a result, MM-TPP can utilize the fixed context window (e.g.,
 240 4096 tokens) more effectively, allowing it to model longer event histories and capture long-range
 241 dependencies with improved efficiency.
 242

243 It is worth noting the distinction between our proposed sequence-level compression (inter-event)
 244 and standard representation-level reduction techniques (intra-event) commonly employed in efficient
 245 MLLMs, such as token pruning or merging (Bolya et al., 2023; Chen et al., 2024a; Dhouib et al.,
 246 2025; Alvar et al., 2025). While these methods effectively reduce computational cost by exploiting
 247 spatial redundancy within individual modalities (e.g., visual patches), they are less applicable to the
 248 structured components of TPPs. Unlike images, key components of TPPs such as timestamps (t_i),
 249 event types (e_i), and text descriptions (m_i) contain dense semantic information. Every time value
 250 and word is precise and necessary for understanding the event sequence, so removing them risks
 251 losing critical logic. Therefore, we prioritize a sequence-level strategy that compresses temporally
 252 similar events. This approach effectively captures the bursty nature of TPPs and allows the model
 253 to accommodate a much longer history within the limited context window. We leave the design of
 254 more advanced, hybrid compression strategies that adapt to more complex and diverse scenarios for
 255 future work.
 256

257 **4.4 TRAINING AND INFERENCE**
 258

259 **Model Training.** To enable the model to understand complex multimodal event structures and the
 260 proposed compression strategy, we adopt a two-stage training framework inspired by Language-TPP.
 261 In the first stage, continued pre-training is conducted on large-scale token sequences constructed
 262 from the defined multimodal event templates, which may include both full and compressed repre-
 263 sentations. This stage helps the model adapt to the new event format and learn the semantics of
 264 special tokens such as $\langle \mid \text{similar_event} \mid \rangle$ and $\langle \mid \text{type_X} \mid \rangle$. The objective follows standard
 265 next-token prediction:
 266

$$\mathcal{L}_{\text{stage1}}(\theta) = -\frac{1}{L} \sum_{i=1}^L \log p_{\theta}(x_i \mid x_{<i}), \quad (2)$$

267 where x_1, \dots, x_L denotes the event token sequence.
 268

269 The second stage involves supervised fine-tuning for downstream tasks, including time prediction,
 270 type classification, and text generation. Training samples are formatted into prompt-response pairs,
 271 where the prompt consists of a compressed historical segment and a task-specific instruction to-
 272 ken (e.g., $\langle \mid \text{time_prediction} \mid \rangle$, $\langle \mid \text{type_prediction} \mid \rangle$, or a natural language query),
 273

270 and the response contains the ground-truth tokens (e.g., byte tokens for time, type token, or text
 271 sequence). The training loss is applied only to the response tokens:
 272

$$273 \quad \mathcal{L}_{\text{stage2}}(\theta) = -\frac{1}{R} \sum_{j=1}^R \log p_{\theta}(r_j \mid \text{Prompt}, r_{<j}), \quad (3)$$

$$274$$

$$275$$

276 where r_1, \dots, r_R denote the response sequence. This two-stage process enables the model to learn
 277 both general sequence modeling and task-specific reasoning over multimodal event data.
 278

279 **Model Inference.** During inference, the model follows an autoregressive decoding strategy sim-
 280 ilar to Language-TPP. Given a historical sequence and a task-specific instruction, it sequentially
 281 generates output tokens. Byte tokens are decoded into floating-point intervals for time prediction,
 282 categorical tokens indicate event types, and language tokens are interpreted as text for generation
 283 or QA tasks. Since Qwen2.5-VL (Bai et al., 2025) does not support image generation, our current
 284 model focuses on time, type, and text outputs. Future extensions may leverage omni-modal models
 285 such as Chameleon (Team, 2025) to enable image generation for predicted events.
 286

287 5 EXPERIMENTS

$$288$$

289 We conduct comprehensive experiments to evaluate the effectiveness of our proposed MM-TPP
 290 across multiple dimensions, including overall performance against SOTA baselines, text generation
 291 quality, the impact of adaptive compression mechanism, and the contribution of multimodal infor-
 292 mation integration.

293 5.1 EXPERIMENTAL SETUP

$$294$$

295 Our MM-TPP model is built upon the Qwen2.5-VL-3B architecture and fine-tuned using LoRA
 296 (Low-Rank Adaptation) (Hu et al., 2021). All experiments are conducted on a single NVIDIA RTX
 297 4090 GPU. For Stage 1 (continued pre-training), we train the model for 5 epochs with a learning
 298 rate of 1×10^{-4} . In Stage 2 (supervised fine-tuning), we also use a learning rate of 1×10^{-4} for 3
 299 epochs. The temporal similarity threshold for our compression mechanism is set to $\Delta = 0.2$, which
 300 we found to offer a good balance between compression and performance. During inference, we set
 301 the decoding temperature to 0.05 and perform each evaluation 5 times with different random seeds
 302 to ensure robustness. Full hyperparameter details can be found in Appendix C.
 303

304 **Baselines.** We compare MM-TPP with several SOTA TPP models, including **NHP** (Mei & Eisner,
 305 2017), **SAHP** (Zhang et al., 2020), **THP** (Zuo et al., 2020), **RMTPP** (Du et al., 2016), **AttNHP** (Mei
 306 et al., 2022), **TPP-LLM** (Liu & Quan, 2024), **S2P2** (Chang et al., 2025), and **Language-TPP** (Kong
 307 et al., 2025). All models except TPP-LLM and Language-TPP are trained and evaluated using the
 308 Easy-TPP framework (Xue et al., 2024), while TPP-LLM and Language-TPP are reproduced using
 309 its official implementation. For a fair comparison, we use a unified maximum epoch setting across
 310 all baselines, select the best-performing checkpoints for testing, and repeat the entire training and
 311 evaluation pipeline 5 times with different random seeds.

312 **Datasets.** Most existing TPP benchmarks are limited to timestamps and event types, with very
 313 few incorporating both textual and visual data. We evaluate MM-TPP on two distinct multimodal
 314 datasets to address this gap:
 315

- 316 • **DanmakuTPP** (Jiang et al., 2025). An established dataset that contains extremely long se-
 317 quences—most with over 1,000 events, and some exceeding 9,600. Each event includes a times-
 318 tamp, textual content (e.g., video comments), and visual information extracted from corresponding
 319 video frames. It also provides the **DanmakuTPP-QA** subset for assessing complex multimodal
 320 reasoning capabilities.
- 321 • **TAXI-PRO**. A new dataset we introduce by augmenting the classic NYC Taxi dataset (Whong,
 322 2014) with rich multimodal content. To create a complementary evaluation scenario, we enrich
 323 each event with 224×224 map image patches centered at each GPS coordinate, as well as natural
 324 language descriptions incorporating landmarks, coordinates, and ride metadata (see Appendix B

324
 325 Table 1: Prediction performance comparison on multimodal TPP datasets for event times and types.
 326 Results reported in terms of RMSE(\downarrow) / ACC(\uparrow) with standard deviations. Best results are in **bold**.
 327
 328
 329

Model	DanmakuTPP		TAXI-PRO	
	RMSE(\downarrow)	ACC(%)(\uparrow)	RMSE(\downarrow)	ACC(%)(\uparrow)
NHP	5.4540 \pm 0.036	30.74 \pm 0.08	0.4494 \pm 0.012	75.93 \pm 0.20
SAHP	5.4245 \pm 0.006	27.38 \pm 0.05	0.3526 \pm 0.006	75.37 \pm 1.30
THP	5.4001 \pm 0.002	24.64 \pm 0.06	0.3736 \pm 0.008	75.31 \pm 1.00
RMTPP	5.4633 \pm 0.076	23.82 \pm 0.50	0.3830 \pm 0.008	76.47 \pm 0.20
AttNHP	5.4384 \pm 0.077	28.92 \pm 0.40	0.4049 \pm 0.007	69.37 \pm 0.80
S2P2	5.5490 \pm 0.016	31.48 \pm 0.13	0.5735 \pm 0.035	75.73 \pm 0.60
TPP-LLM	5.3035 \pm 0.012	24.59 \pm 0.70	0.3336 \pm 0.008	71.09 \pm 0.20
Language-TPP	5.3845 \pm 0.022	22.62 \pm 0.07	0.3376 \pm 0.001	75.27 \pm 0.08
MM-TPP	5.2987 \pm 0.018	27.62 \pm 0.02	0.3310 \pm 0.001	77.56 \pm 0.04

330
 331 for details). In contrast to DanmakuTPP, sequences in TAXI-PRO are significantly shorter (aver-
 332 age length of 36), providing a complementary evaluation scenario where our proposed compres-
 333 sion mechanism is not required.
 334

335
 336 **Metrics.** We evaluate model performance using several task-specific metrics. For timestamp pre-
 337 diction, we report root mean square error (**RMSE**); for event type classification, we use accuracy
 338 (**ACC**); and for text generation, we measure token-level quality using perplexity (**PPL**). For question
 339 answering tasks in DanmakuTPP-QA, we follow the original evaluation protocol—using **ACC** and
 340 **RMSE** for closed-ended questions, and conducting **qualitative analysis** for open-ended responses.
 341

342 5.2 OVERALL PERFORMANCE COMPARISON

343 We first compare MM-TPP against all baseline methods on the complete datasets. Both datasets
 344 are split into training, validation, and test sets with an 8:1:1 ratio. For DanmakuTPP, we constrain
 345 the number of encoded events using a maximum token limit of 4096 tokens during training and
 346 perform evaluation on the first 200 events of each test sequence. For TAXI-PRO, the 4096 token
 347 limit is sufficient to encode complete sequences without truncation. As shown in Table 1, MM-TPP
 348 delivers state-of-the-art performance across both datasets. On DanmakuTPP, it achieves the lowest
 349 RMSE (**5.2987**) for time prediction, while its type accuracy (27.62%) is highly competitive and
 350 substantially improves upon Language-TPP. The advantages are even clearer on TAXI-PRO, where
 351 MM-TPP outperforms all baselines on both metrics, establishing a new benchmark with a leading
 352 RMSE of **0.3310** and an accuracy of **77.56%**.
 353

354 5.3 PERFORMANCE ON MULTIMODAL QUESTION ANSWERING

355 A key strength of our framework lies in its capacity for complex reasoning over multimodal event
 356 streams. To evaluate this, we adopt the **DanmakuTPP-QA** benchmark (Jiang et al., 2025), which
 357 is specifically designed to assess temporal, visual, and textual reasoning. The benchmark comprises
 358 8 closed-ended tasks for quantitative evaluation and 2 open-ended tasks that require generating de-
 359 scriptive analytical reports, evaluated qualitatively. Detailed descriptions of all 10 tasks are provided
 360 in Appendix E.
 361

362 **Closed-ended QA Tasks.** We compare MM-TPP with SOTA LLMs and MLLMs on eight closed-
 363 ended tasks, using baseline results reported in DanmakuTPP-Bench (Jiang et al., 2025). As shown
 364 in Table 2, MM-TPP achieves highly competitive performance across the benchmark and attains
 365 leading results on several tasks. Notably, it consistently outperforms the fine-tuned Qwen2.5-VL-3B
 366 from the original benchmark, despite utilizing the same underlying base model. This performance
 367 gap highlights the strength of our framework, which explicitly captures the structured nature of
 368 multimodal event sequences—unlike conventional fine-tuning approaches.
 369

378 Table 2: Comparative evaluation on the DanmakuTPP-QA closed-ended tasks. Metrics are ACC
 379 (\uparrow) and RMSE (\downarrow). The proposed MM-TPP model is highlighted. Best results are in **bold** and the
 380 second best results are set in underline.

Model / Task Metrics	T-1 ACC \uparrow	T-2 RMSE \downarrow	T-3 RMSE \downarrow	T-4 RMSE \downarrow	T-5 RMSE \downarrow	T-6 RMSE \downarrow	T-7 ACC \uparrow	T-8 ACC \uparrow
LLMs								
Qwen2.5-7B-Instruct	0.33	27.64	134.45	0.65	0.56	0.51	10.67	32.67
Qwen2.5-32B-Instruct	25.33	1.52	122.69	0.36	0.29	0.24	16.67	38.17
Qwen2.5-72B-Instruct	0.67	1.28	123.45	0.30	0.46	0.46	16.00	43.83
Qwen3-8B	6.67	1.80	123.59	0.32	0.41	0.45	<u>19.33</u>	41.50
Qwen3-30B-A3B	0.67	1.33	121.96	<u>0.20</u>	0.33	0.40	23.00	43.67
Qwen3-235B-A22B	8.67	1.39	120.79	0.30	0.31	0.29	10.33	32.50
Llama-3.3-70B-Instruct	1.67	1.11	121.49	0.26	0.27	0.22	17.00	33.33
DeepSeek-V3	25.00	1.30	121.30	0.34	0.26	0.22	13.67	34.5
MLLMs								
Qwen2.5-VL-7B	9.67	11.61	124.99	0.46	0.82	0.66	8.33	22.17
Qwen2.5-VL-32B	8.0	1.26	124.02	0.35	0.51	0.38	12.67	22.17
Qwen2.5-VL-72B	0.33	<u>1.14</u>	<u>121.25</u>	0.28	0.47	0.41	15.98	47.17
Gemma3-27B	0.33	1.33	121.32	0.28	0.27	0.20	15.67	36.17
Finetuned								
Qwen2.5-VL-3B	<u>27.0</u>	1.35	220.43	0.05	<u>0.16</u>	0.08	15.33	43.00
MM-TPP	27.33	1.49	190.45	0.05	0.10	0.07	15.67	<u>44.00</u>

400 **Open-ended QA and Report Generation.** We assess MM-TPP’s capacity for generating high-
 401 quality analytical text—a core capability that sets it apart from purely predictive models. On the
 402 two open-ended tasks from the DanmakuTPP-QA benchmark, MM-TPP consistently produces sig-
 403 nificantly more insightful and coherent text compared to existing MLLMs. To further evaluate its
 404 generalization ability, we prompt MM-TPP to generate descriptive analytical reports on the TAXI-
 405 PRO dataset, which lacks predefined QA tasks and has not been used to fine-tune the model for
 406 generative tasks. Remarkably, the quality of the generated narratives for TAXI-PRO’s temporal data
 407 also surpasses that of leading MLLMs, which are often limited to basic descriptive analysis of the
 408 event sequence. In contrast, MM-TPP uncovers high-level spatio-temporal patterns, demonstrating
 409 a robust capacity to synthesize multimodal event streams into coherent, insightful narratives—even
 410 in a zero-shot setting. Detailed qualitative comparisons of MM-TPP and baseline MLLM outputs
 411 for both datasets are provided in Appendix E.

412 5.4 EFFECTIVENESS OF COMPRESSION STRATEGY

413 To assess the effectiveness of our adaptive compression strategy, we compare the performance of
 414 MM-TPP with and without this mechanism on the DanmakuTPP dataset (note that compression is
 415 not applied to TAXI-PRO due to its shorter sequences). The key advantage of compression lies in its
 416 ability to fit longer event histories within a fixed context window. Under a 4096-token constraint, the
 417 uncompressed model can accommodate an average of 113 events, whereas the compressed variant
 418 extends this to an average of 292 events, with a maximum of 2008. This extended context allows
 419 the model to better capture long-range dependencies, leading to noticeable performance gains. As
 420 a result, the compressed model outperforms its uncompressed counterpart, achieving lower RMSE
 421 (5.2987 vs. 5.5551) and higher accuracy (27.62% vs. 25.87%) on predictive tasks. **To further
 422 verify the effectiveness of compression strategy, we compared MM-TPP against a naive baseline
 423 that randomly drops events. As detailed in Appendix D, the random drop strategy leads to significant
 424 performance degradation, confirming that preserving temporal causality is as critical as extending
 425 context.**

426 In addition, we conduct further experiments to evaluate the effectiveness of our compression method
 427 from a language modeling perspective. Specifically, we measure **PPL** on the encoded event se-
 428 quences from the test set without truncation, a standard and widely-used metric for assessing the
 429 quality of language models (Zhao et al., 2023). As shown in Fig. 2a, perplexity increases with
 430 sequence length—a well-known challenge in sequence modeling. Notably, our compressed model
 431 consistently achieves lower PPL across all lengths, with the gap widening as sequences grow longer.

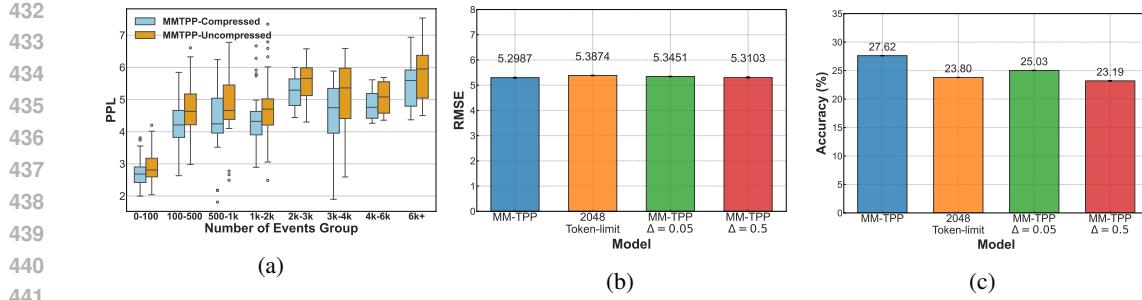


Figure 2: Evaluation of the compression mechanism and key hyperparameters on the DanmakuTPP dataset. (a) PPL comparison on the test set. Our compressed model consistently maintains a lower PPL than the uncompressed version, indicating effective long-range modeling. (b, c) Ablation results for context length and similarity threshold (Δ) on RMSE and ACC, showing optimal performance with a larger context and $\Delta = 0.2$.

Table 3: Ablation study results for MM-TPP. The table evaluates the contribution of visual information (vs. text-only variants) and the effect of model size (vs. a 7B model). Metrics are RMSE(\downarrow) and ACC(\uparrow). Best results are in **bold**.

Model	DanmakuTPP		TAXI-PRO	
	RMSE(\downarrow)	ACC(%)(\uparrow)	RMSE(\downarrow)	ACC(%)(\uparrow)
MM-TPP	5.2987 \pm 0.018	27.62 \pm 0.02	0.3310 \pm 0.001	77.56 \pm 0.04
Language-TPP	5.3845 \pm 0.022	22.62 \pm 0.07	0.3376 \pm 0.001	75.27 \pm 0.08
MM-TPP (text only)	5.4654 \pm 0.008	23.64 \pm 0.06	0.3388 \pm 0.001	76.70 \pm 0.08
MM-TPP (7B)	5.0533 \pm 0.002	26.98 \pm 0.01	0.3337 \pm 0.001	76.16 \pm 0.05

These results highlight a key strength of our approach: by preserving more long-range context within a fixed token budget, the compression mechanism mitigates performance degradation on long and complex event streams.

5.5 ABLATION STUDY

In this section, we present ablation studies to examine the impact of four key design choices in MM-TPP: (i) the incorporation of visual information, (ii) model scale, (iii) context window size (controlled via the maximum token limit), and (iv) the temporal similarity threshold used in the compression mechanism.

Impact of Visual Information. To assess the contribution of visual input, we compare the full MM-TPP model (text + image) with two text-only variants: MM-TPP (text only) and Language-TPP. All models are evaluated under controlled conditions with a fixed number of events in the context window, ensuring that sequence length does not confound the results. As shown in Table 3, MM-TPP consistently outperforms both baselines across datasets, indicating that visual features provide additional predictive value. The integration of image improves both timestamp and type prediction accuracy, highlighting the complementary role of visual context alongside textual information.

Impact of Model Size. To assess the role of model capacity, we replace the base model Qwen2.5-VL-3B with a larger Qwen2.5-VL-7B, and evaluate on DanmakuTPP and TAXI-PRO datasets using identical training procedures and hyperparameters. This experiment investigates whether scaling up the model improves event prediction performance. Results in Table 3 show that scaling to 7B yields mixed results. While the larger model significantly improves time prediction (RMSE) on the complex DanmakuTPP dataset, the original 3B model outperforms it across all other metrics. This suggests a larger model’s capacity is most beneficial for long-range temporal patterns but may hinder performance on simpler tasks, possibly due to overfitting.

486 **Impact of Context Length.** To examine the effect of context window size, we reduce the maximum
 487 token limit from 4096 to 2048 during both training and inference on the DanmakuTPP dataset.
 488 As shown in Figs. 2b and 2c, this reduction results in a clear drop in performance. A shorter context
 489 window limits the model’s access to historical events, hindering its ability to capture long-range de-
 490 pendencies critical for accurate temporal modeling. These results highlight the importance of main-
 491 taining a sufficiently large context window when dealing with long and complex event sequences.
 492

493 **Impact of Temporal Similarity Threshold.** We evaluate how the performance of our compres-
 494 sion mechanism varies with the temporal similarity threshold Δ . Specifically, we compare the de-
 495 fault setting $\Delta = 0.2$ with a stricter threshold ($\Delta = 0.05$) and a looser one ($\Delta = 0.5$), as motivated
 496 in Appendix A.2. As shown in Figs. 2b and 2c, both alternatives lead to performance drops. A small
 497 threshold ($\Delta = 0.05$) results in limited compression, offering little gain in contextual coverage.
 498 In contrast, a large threshold ($\Delta = 0.5$) causes over-compression, merging dissimilar events and
 499 discarding important temporal signals. These results confirm that $\Delta = 0.2$ strikes a good balance
 500 between compression efficiency and temporal fidelity.
 501

502 6 CONCLUSION

503 In this paper, we introduce MM-TPP, a novel framework that advances TPP modeling from a purely
 504 predictive to a generative paradigm. By incorporating visual information alongside temporal, cate-
 505 gorical, and textual data, our model can not only predict future events but also generate rich, descrip-
 506 tive text that analyzes and explains event dynamics. We address the critical challenge of long-context
 507 modeling in multimodal sequences through an adaptive compression strategy that exploits tempo-
 508 ral similarity. Extensive experiments demonstrate that MM-TPP consistently outperforms SOTA
 509 baselines in standard prediction tasks and, more importantly, exhibits superior text generation and
 510 reasoning capabilities on the challenging DanmakuTPP-QA benchmark. While our framework cur-
 511 rently lacks image generation, it opens new avenues for vision-enhanced temporal modeling and
 512 provides a solid foundation for future research in generative multimodal TPPs.
 513

514 ETHICS STATEMENT

515 Our work adheres to the ICLR Code of Ethics. The datasets used are either public benchmarks
 516 (DanmakuTPP) or derived from public data (NYC TLC). For our constructed TAXI-PRO dataset,
 517 we ensured no personally identifiable information was included, focusing only on aggregated spatio-
 518 temporal patterns. We acknowledge the potential for biases inherited from the pre-trained language
 519 model backbone. Our framework is intended for research purposes to analyze event dynamics, and
 520 we foresee no direct negative societal consequences from its use.
 521

524 REPRODUCIBILITY STATEMENT

525 We have made every effort to ensure the reproducibility of our research. The complete source
 526 code for our MM-TPP model, including training and evaluation procedures, will be provided in
 527 the supplementary materials. All crucial hyperparameters for both the pre-training and supervised
 528 fine-tuning stages are documented in Appendix C. Furthermore, details of our experimental setup,
 529 baselines, and evaluation metrics are described in Section 5 to facilitate fair comparison and repli-
 530 cation of our results.
 531

533 REFERENCES

534 Giada Adelfio and Marcello Chiodi. Including covariates in a space-time point process with appli-
 535 cation to seismicity. *Statistical Methods & Applications*, 30(3):947–971, 2021.
 536
 537 Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-
 538 based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision*
 539 and *Pattern Recognition Conference*, pp. 9392–9401, 2025.

540 Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance.
 541 *Market Microstructure and Liquidity*, 1(01):1550005, 2015.
 542

543 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,
 544 Shijie Wang, Jun Tang, et al. Qwen2.5-v1 technical report. *arXiv preprint arXiv:2502.13923*,
 545 2025.

546 Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy
 547 Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

548

549 Yuxin Chang, Alex Boyd, Cao Xiao, Taha Kass-Hout, Parminder Bhatia, Padhraic Smyth, and An-
 550 drew Warrington. Deep continuous-time state-space models for marked event sequences. In
 551 *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

552

553 Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang.
 554 An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-
 555 language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

556

557 Zhiheng Chen, Guanhua Fang, and Wen Yu. On non-asymptotic theory of recurrent neural networks
 558 in temporal point processes. *arXiv preprint arXiv:2406.00630*, 2024b.

559

560 Daryl J Daley and David Vere-Jones. An introduction to the theory of point processes. vol. i. prob-
 561 ability and its applications, 2003.

562

563 Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

564

565 Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and
 566 clustering-based token reduction for faster visual language models. In *Proceedings of the Com-
 567 puter Vision and Pattern Recognition Conference*, pp. 14582–14592, 2025.

568

569 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
 570 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszko-
 571 reit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at
 572 scale, 2021. URL <https://arxiv.org/abs/2010.11929>.

573

574 Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song.
 575 Recurrent marked temporal point processes: embedding event history to vector. In *International
 576 Conference on Knowledge Discovery and Data Mining*, 2016.

577

578 Álvaro Gajardo and Hans-Georg Müller. Point process models for covid-19 cases and deaths. *Jour-
 579 nal of Applied Statistics*, pp. 1–16, 2021.

580

581 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
 582 and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.

583

584 Yue Jiang, Jichu Li, Yang Liu, Dingkang Yang, Feng Zhou, and Quyu Kong. Danmakutppbench: A
 585 multi-modal benchmark for temporal point process modeling and understanding. *arXiv preprint
 586 arXiv:2505.18411*, 2025.

587

588 Quyu Kong, Yixuan Zhang, Yang Liu, Panrong Tong, Enqi Liu, and Feng Zhou. Language-tpp:
 589 Integrating temporal point processes with language models for event analysis. *arXiv preprint
 590 arXiv:2502.07139*, 2025.

591

592 Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods,
 593 challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

594

595 Zefang Liu and Yinzhu Quan. TPP-LLM: Modeling temporal point processes by efficiently fine-
 596 tuning large language models. *arXiv preprint*, 2024.

597

598 Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multi-
 599 variate point process. In *Advances in Neural Information Processing Systems*, 2017.

594 Hongyuan Mei, Chenghao Yang, and Jason Eisner. Transformer embeddings of irregularly spaced
 595 events and their participants. In *International Conference on Learning Representations*, 2022.
 596

597 Zizhuo Meng, Ke Wan, Yadong Huang, Zhidong Li, Yang Wang, and Feng Zhou. Interpretable
 598 Transformer Hawkes processes: Unveiling complex interactions in social networks. In *International
 599 Conference on Knowledge Discovery and Data Mining*, 2024.

600 Sebastian Meyer, Johannes Elias, and Michael Höhle. A space–time conditional intensity model for
 601 invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616, 2012.
 602

603 George Mohler. Modeling and estimation of multi-source clustering in crime and security data. *The
 604 Annals of Applied Statistics*, pp. 1525–1539, 2013.

605 Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. Fully neural network based model for general
 606 temporal point processes. In *Advances in Neural Information Processing Systems*, 2019.
 607

608 Aristeidis Panos. Decomposable Transformer point processes. In *Advances in Neural Information
 609 Processing Systems*, 2024.

610 Marian Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. Sir-hawkes: on
 611 the relationship between epidemic models and Hawkes point processes. In *The Web Conference*,
 612 2018.
 613

614 Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and
 615 Hongyuan Mei. Language models can improve event prediction by few-shot abductive reasoning.
 616 In *Advances in Neural Information Processing Systems*, 2024.

617 Alexander Soen, Alexander Mathews, Daniel Grixti-Cheng, and Lexing Xie. Unipoint: Universally
 618 approximating point processes intensities. In *AAAI conference on artificial intelligence*, 2021.
 619

620 Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL <https://arxiv.org/abs/2405.09818>.
 621

622 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,
 623 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical
 624 report. *arXiv preprint arXiv:2503.19786*, 2025.
 625

626 A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
 627

628 Chris Whong. Foiling NYC’s taxi trip data, 2014.

629 Shuai Xiao, Junch Yan, Xiaokang Yang, Hongyuan Zha, and Stephen Chu. Modeling the inten-
 630 sity function of point process via recurrent neural networks. In *AAAI Conference on Artificial
 631 Intelligence*, 2017.

632 Siqiao Xue, Yan Wang, Zhixuan Chu, Xiaoming Shi, et al. Prompt-augmented temporal point
 633 process for streaming event sequence. In *Advances in Neural Information Processing Systems*,
 634 2023.
 635

636 Siqiao Xue, Xiaoming Shi, Zhixuan Chu, Yan Wang, Hongyan Hao, Fan Zhou, Caigao Jiang, Chen
 637 Pan, James Y. Zhang, Qingsong Wen, Jun Zhou, and Hongyuan Mei. Easytpp: Towards open
 638 benchmarking temporal point processes. In *International Conference on Learning Representa-
 639 tions*, 2024.

640 Chenghao Yang, Hongyuan Mei, and Jason Eisner. Transformer embeddings of irregularly spaced
 641 events and their participants. In *Proceedings of the Tenth International Conference on Learning
 642 Representations*, 2022.

644 Guolei Yang, Ying Cai, and Chandan K Reddy. Recurrent spatio-temporal point process for check-in
 645 time prediction. In *International Conference on Information and Knowledge Management*, 2018.
 646

647 Qiang Zhang, Aldo Lipani, Ömer Kirnap, and Emine Yilmaz. Self-attentive Hawkes process. In
 648 *International Conference on Machine Learning*, 2020.

648 Yixuan Zhang, Quyu Kong, and Feng Zhou. Integration-free training for spatio-temporal multimodal
649 covariate deep kernel point processes. In *Advances in Neural Information Processing Systems*,
650 2023.

651 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min,
652 Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv*
653 preprint *arXiv:2303.18223*, 1(2), 2023.

654 Feng Zhou, Yixuan Zhang, and Jun Zhu. Efficient inference of flexible interaction in spiking-neuron
655 networks. In *International Conference on Learning Representations*, 2021.

656 Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural point process for
657 learning spatiotemporal event dynamics. In *Learning for Dynamics and Control Conference*,
658 2022.

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

661 Shixiang Zhu, Minghe Zhang, Ruyi Ding, and Yao Xie. Deep fourier kernel for self-attentive point
662 processes. In *International Conference on Artificial Intelligence and Statistics*, 2021.

663 Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes
664 process. In *International Conference on Machine Learning*, 2020.

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A DETAILS OF ADAPTIVE SEQUENCE COMPRESSION
703704 A.1 ILLUSTRATION OF COMPRESSION MECHANISM
705706 To clarify how the adaptive compression operates on consecutive events, we provide a step-by-step
707 illustration using a hypothetical sequence of three events: A , B , and C , occurring in that order.
708 Let $\tau_i = t_i - t_{i-1}$ denote the inter-event interval for event i . The compression decision is made
709 sequentially based on the pairwise comparison of adjacent intervals:
710711 • **Step 1 (Processing Event B):** The model compares the interval of event B (τ_B) with the
712 interval of the immediately preceding event A (τ_A). If the absolute difference satisfies
713 $|\tau_B - \tau_A| < \Delta$, event B is considered temporally similar to A and is replaced by a single
714 special token `<|similar_event|>`.
715
716 • **Step 2 (Processing Event C):** The model then proceeds to event C . It compares its in-
717 terval (τ_C) with the interval of the preceding event B (τ_B). Note that this comparison is
718 based on the calculated time intervals regardless of whether B was compressed in the to-
719 ken sequence. If $|\tau_C - \tau_B| < \Delta$, event C is also replaced by the `<|similar_event|>`
720 token.
721722 This sequential process ensures that compression is applied locally based on the dynamic evolution
723 of event density, allowing the model to efficiently encode bursts of events with similar distinct timing
724 patterns.
725726 A.2 TEMPORAL SIMILARITY THRESHOLD SETTING
727728 In our framework, the temporal similarity threshold Δ is a key hyperparameter for adaptive sequence
729 compression. Specifically, for each event, we define the interval $\tau_i = t_i - t_{i-1}$ and compare it with
730 the previous interval τ_{i-1} . If the absolute difference between the two intervals satisfies $|\tau_i - \tau_{i-1}| <$
731 Δ , event i is deemed temporally similar to event $i - 1$ and is eligible for merging. This mechanism
732 directly controls the granularity of sequence modeling and the preservation of fine-grained temporal
733 dynamics, as only sufficiently similar events are compressed into a single representation.
734735 To determine an appropriate value for Δ , we conducted a comprehensive statistical analysis of the
736 absolute differences between adjacent inter-event intervals ($|\tau_i - \tau_{i-1}|$) on the DanmakuTPP training
737 dataset. As illustrated in Fig. 3, the empirical distribution of these differences is highly skewed,
738 with a large proportion of values concentrated near zero. This indicates that event intervals are
739 frequently very similar to their predecessors, validating the potential for a compression strategy
740 based on temporal similarity.
741742 To ground our threshold selection in the data’s structure, we examined the distribution’s quantiles,
743 summarized in Table 4. Based on this analysis, we selected candidate thresholds of 0.05, 0.2, and
744 0.5 seconds. These values correspond to meaningful points in the distribution:
745746
747 • $\Delta = 0.05$ is a conservative threshold, corresponding to approximately the 25th percentile.
748 It compresses only the most similar event pairs, preserving most of the original sequence
749 structure.
750
751 • $\Delta = 0.2$ corresponds to approximately the 50th percentile (the median), offering a balance
752 between compression rate and information preservation.
753
754 • $\Delta = 0.5$ is a more aggressive threshold, near the 75th percentile, which merges a larger
755 portion of the sequence to achieve a higher compression ratio.
756757 The positions of these thresholds relative to the distribution are marked in Fig. 3. This principled,
758 data-driven selection ensures that our adaptive compression mechanism is both statistically grounded
759 and practically effective for multimodal temporal point process modeling.
760

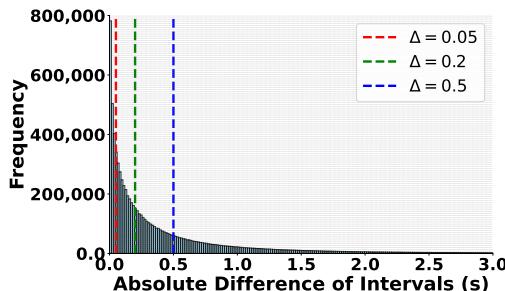


Figure 3: Distribution of the absolute difference between adjacent inter-event intervals on the training set. Vertical lines indicate the tested thresholds Δ .

Table 4: Quantiles of the absolute difference between adjacent inter-event intervals ($|\tau_i - \tau_{i-1}|$).

Percentile	Value (s)
0.050	0.007
0.100	0.017
0.200	0.045
0.250	0.063
0.500	0.214
0.750	0.590
0.900	1.330
0.950	2.115
1.000	1787.000

B TAXI-PRO DATASET CONSTRUCTION

We construct the TAXI-PRO dataset by augmenting the original NYC Taxi dataset with multimodal information (images and text descriptions). This section provides detailed information about the data construction pipeline.

Original Taxi Data. The raw taxi trip data is sourced from the New York City Taxi and Limousine Commission (NYC TLC) public dataset, consisting of files `trip_data_1.csv` through `trip_data_12.csv`. This dataset is identical to the source used by the EasyTPP official taxi dataset and originates from NYC TLC’s open data initiative¹.

We construct the TAXI-PRO dataset with the following setups:

- **Image.** We leverage the coordinates of each event and construct an image crop of the map of the location. High-resolution map tiles are generated from OpenStreetMap data using tools like `contextily`. The base map covers the entire Manhattan area with strict correspondence between pixel coordinates and geographical boundaries (resolution: 6071×6307 pixels). For each event location, we extract a 224×224 pixel patch centered at the event coordinates. Edge cases are handled by padding with gray background to maintain consistent patch dimensions.
- **Text.** We generate the textual mark for each event following a fixed template incorporating landmark information, coordinates, passenger count, and trip distance. For example, “*Picked up at Times Square (40.7580, -73.9855), 2 passengers*”.
- **Event type.** We focus exclusively on Manhattan and subdivide it into three regions: lower Manhattan, midtown Manhattan, and upper Manhattan, resulting in 6 event types for combinations of regions and pickup/dropoff events.

To provide a concrete visualization of our data structure, Fig. 4 showcases the composition of a single event in the TAXI-PRO dataset. It highlights the multimodal nature of each data point, combining the geographical map patch, the templated text description, and the categorical event type. When selecting the final 2000 sequences, we employ a greedy algorithm with variance minimization to ensure balanced event type distribution, avoiding imbalance.

C DETAILED HYPERPARAMETER SETTINGS

Hyperparameters The specific hyperparameters used for each stage of our training procedure are summarized in Table 5. We employed the AdamW optimizer and utilized bfloat16 mixed-precision for all experiments to enhance computational efficiency.

¹FOILing NYC’s Taxi Trip Data: https://chriswhong.com/open-data/foil_nyc_taxi/

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825

```
{
  time_since_last_event: 0.0333,
  event_type: 4
  (pickup at upper_manhattan),
  text: "Picked up at Central Park West (40.800018,
  -73.944778), 1 passenger",
  Image: ----->
}
```



Figure 4: An illustration of a single event in the TAXI-PRO dataset. Each event comprises four key components: (1) a 224×224 map image centered on the event’s coordinates, (2) a generated textual description including landmark and trip details, (3) a categorical event type and (4) a timestamp.

826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847

Table 5: Hyperparameter Setting

Hyperparameter	Value
Base Model	Qwen2.5-VL-3B
LoRA Rank	16
LoRA Alpha	64
LoRA Dropout	0.1
Stage 1 Learning Rate	1e-4
Stage 1 Epochs	5
Stage 2 Learning Rate	1e-4
Stage 2 Epochs	3
Batch Size	1
Max Token Length	4096
Temporal Threshold Δ	0.2s
Inference Temperature	0.05

D COMPARISON WITH RANDOM DROP BASELINE

To rigorously evaluate the effectiveness of our proposed adaptive compression mechanism, we compared it against a naive compression baseline: **Random Drop**. This baseline randomly discards events from the sequence during both training and inference phases.

We conducted experiments on the DanmakuTPP dataset with drop probabilities of $p = 25\%$ and $p = 50\%$. The results, compared with our MM-TPP (Adaptive Compression) and the uncompressed baseline, are presented in Table 6.

As shown in Table 6, the Random Drop strategy results in a clear performance degradation compared to our adaptive approach. Even with a moderate drop rate of 25%, the model’s predictive accuracy and timestamp precision are inferior to MM-TPP. This degradation is likely because event sequences in TPPs contain strict causal dependencies; randomly removing events disrupts the historical context necessary for the model to reason about future dynamics (e.g., self-exciting patterns). In contrast, our adaptive compression strategy selectively merges only temporally similar intervals (which often represent redundant burst signals) while preserving the critical anchor events and the overall semantic continuity of the sequence.

864
 865 Table 6: Performance comparison between MM-TPP’s adaptive compression and naive Random
 866 Drop strategies on the DanmakuTPP dataset. The proposed adaptive method significantly outper-
 867 forms random dropping, highlighting the importance of preserving temporal structure.

868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917	868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917	868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917
Model / Setting	RMSE (↓)	ACC (%) (↑)
MM-TPP	5.2987\pm0.018	27.62\pm0.02
MMTPP-Uncompressed	5.5551 \pm 0.008	25.87 \pm 0.05
MMTPP-Random-Drop ($p = 25\%$)	5.3792 \pm 0.014	24.84 \pm 0.06
MMTPP-Random-Drop ($p = 50\%$)	5.4259 \pm 0.009	24.79 \pm 0.03

E DANMAKUTPP-QA DETAILS

878 Table 7: Detailed descriptions and evaluation metrics for the ten tasks in the DanmakuTPP-QA
 879 benchmark. Closed-ended tasks are evaluated quantitatively (Accuracy or RMSE), while open-
 880 ended tasks are assessed via qualitative analysis.

882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917	882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917	882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917
Task-1	Danmaku burst peak counting	ACC
Task-2	Prediction of the next Danmaku timestamp	RMSE
Task-3	Prediction of the next Danmaku burst peak timestamp	RMSE
Task-4	Assessment of average sentiment polarity	RMSE
Task-5	Sentiment polarity prediction for the next Danmaku	RMSE
Task-6	Sentiment polarity prediction for the next Danmaku burst peak	RMSE
Task-7	Event type inference for the next Danmaku	ACC
Task-8	Prediction of Top-2 triggering event types for the next burst peak	ACC
Task-9	Analysis of global sentiment dynamics and the underlying drivers	Qualitative Analysis
Task-10	Causal attribution analysis for specific Danmaku burst peak formation	Qualitative Analysis

E.1 TEN QUESTION-ANSWERING TASKS

909 This section provides a detailed breakdown of the ten reasoning tasks that constitute the
 910 **DanmakuTPP-QA** benchmark. The benchmark is designed to evaluate a model’s temporal-visual-
 911 textual reasoning capabilities through a series of diverse challenges. As summarized in Table 7,
 912 these tasks are divided into eight closed-ended, quantitatively evaluated tasks and two open-ended
 913 tasks that require qualitative assessment of generated reports.

E.2 MM-TPP’S PERFORMANCE ON OPEN-ENDED TASKS

914 In this section, we conduct a qualitative evaluation of MM-TPP’s ability to generate high-quality
 915 analytical reports for open-ended tasks. We compare its outputs against two SOTA, general-purpose

918 Multimodal Large Language Models (MLLMs): **Qwen2.5-VL-3B** (Bai et al., 2025) and **Gemma3-27B** (Team et al., 2025). Our analysis reveals that MM-TPP consistently produces significantly more
 919 insightful and relevant reports, highlighting the value of its specialized architecture.
 920

921 The qualitative advantage is not only evident on the DanmakuTPP-QA benchmark, where MM-TPP
 922 was fine-tuned, but also generalizes robustly to the TAXI-PRO dataset in a zero-shot setting. To
 923 substantiate these claims, we provide a series of comparative examples in the following sections.
 924 Given the consistent nature of the performance gap, we present two representative examples for
 925 each of the open-ended tasks (Task 9 and Task 10) and for the TAXI-PRO report generation task
 926 to illustrate the consistent superiority of MM-TPP’s analytical capabilities. Here we summarize the
 927 qualitative advantage as follows:
 928

- 929 • **Task 9 (Danmaku sentiment dynamics analysis):** MM-TPP conveys information more
 930 effectively by focusing on essential sentiment dynamics, follows a clear analytical structure
 931 from overall trend to key inflection points and summary, and produces outputs that human
 932 evaluators find more readable and natural.
 933
- 934 • **Task 10 (Danmaku peak burst reason analysis):** Beyond capturing peak bursts, MM-
 935 TPP pinpoints sub-peaks with fine temporal precision, enabling a more rigorous view of
 936 self-exciting dynamics. It also ties specific event types, such as quote/reference and hu-
 937 mor/meme, directly to the triggering mechanisms of bursts, clarifying how content cate-
 938 gories drive cascading responses. Unlike the general-purpose MLLMs, it avoids drifting
 939 into tangential scene descriptions and instead highlights the key drivers, while framing
 940 the explanation explicitly in terms of TPP theory. This balance of precision, focus, and
 941 domain-specific grounding makes the analysis both accessible and academically robust.
 942
- 943 • **TAXI-PRO:** MM-TPP can deliver high-level insights by identifying core TPP patterns in
 944 the driver’s behavior. It moves beyond simple route tracking to model how events mutu-
 945 ally influence one another, revealing strategic responses to spatio-temporal dynamics such
 946 as the formation of self-exciting event clusters in high-demand corridors while MLLMs,
 947 proficient in language generation, tend to produce outputs that are more descriptive than
 948 analytical.
 949

950 **TAXI-PRO Example_1**

951 MM-TPP’s generated text on TAXI-PRO (Example_1)

952 **Prompt:**

953 You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi
 954 driver’s activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence,
 955 please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs. Note that this sequence is NOT VIDEO!
 956 All time intervals are converted to four bytes according to the IEEE 754 standard.
 957 <|TPP_data|>

958 **Answer:**

959 The event sequence reveals a driver’s dynamic movement across multiple iconic New York City (NYC) neighborhoods, primarily centered
 960 around the Upper West Side, Theater District, and surrounding areas. The driver initiates a journey at the Upper West Side, then rapidly
 961 transitions through high-density, culturally rich locations such as the Financial District and Tribeca, indicating a high volume of pedestrian
 962 or dense traffic activity. This pattern aligns with self-exciting TPP dynamics, where initial events trigger subsequent ones through social
 963 or economic stimuli.

964 The driver consistently serves four to six passengers, suggesting a high-visibility or possibly shared-vehicle model, where each shift sees
 965 a mix of pickups and drops-off. The average trip duration is under 10 minutes, with notable short-haul excursions (<1.5 miles), implying
 966 rapid-fire driver turnover or high-frequency service needs.

967 Visually, the map shows tight clustering of points, especially between 40.76–40.78 and 40.69–40.74, aligning with known densest city
 968 blocks like the Theater District and surrounding high-activity zones. The presence of multiple “Dropped off” entries within <1.5 miles
 969 suggests that the driver is likely functioning in a high-density urban environment where such proximity is common.

970 Temporal clustering indicates a high-intensity period, potentially early morning or late-night when public transportation or walking
 971 patterns are denser. The driver’s pattern of serving different numbers of passengers in quick succession implies a shift-based service
 972 model, where each location functions as a micro hub, possibly responding to fluctuating passenger flow.

973 Unexpectedly, despite the high frequency of short trips, there is no significant drop in average speed, suggesting that the driver is
 974 possibly in a city with low-speed zones or high-pedestrian friction. Additionally, the lack of significant negative sentiment correlates
 975 with high-frequency, short-interval events, indicating routine and possibly socially or culturally motivated transportation needs rather
 976 than anomalies.

977 In conclusion, this sequence reflects a high-intensity, short-haul transportation pattern typical in densely populated urban centers, with
 978 temporal and spatial clustering pointing to a structured driver behavior modulated by local geography and timing dynamics.

972
973

Qwen's generated text on TAXI-PRO (Example_1)

974

Prompt:

You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi driver's activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence, please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs.

<|TPP_data|>

977

Answer:

The provided data logs a single taxi driver's activities in New York City over approximately 8.5 hours, capturing a series of pickups and drop-offs with details on time, location, number of passengers, and distance traveled.

978

The sequence begins with a pickup at Times Square, followed by a 1.45-mile trip to Hell's Kitchen. The driver then picks up two more passengers in Hell's Kitchen and drops them off in Tribeca, covering 2.22 miles. Subsequent trips involve short distances within the Theater District and Chelsea, highlighting the high demand for short taxi rides in these areas. Notable longer trips include a 3.67-mile journey from the Theater District to Central Park West and a 5.20-mile trip from Chelsea to Harlem. These longer trips demonstrate the driver's capability to handle diverse travel needs. The majority of the driver's activities, however, focus on short, frequent trips in midtown and lower Manhattan, particularly in areas like Tribeca, Chelsea, and the Theater District. The final recorded activities include a 5.55-mile trip from Tribeca to Midtown East and a 1.90-mile trip from the Theater District to Times Square, concluding the sequence.

979

In summary, the driver operates efficiently in high-demand areas, frequently handling short trips while occasionally taking longer journeys. This pattern reflects the dynamic nature of taxi services in New York City's bustling urban environment.

980

981

982

983

984

985

986

987

988

Gemma's generated text on TAXI-PRO (Example_1)

989

Prompt:

You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi driver's activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence, please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs.

<|TPP_data|>

990

Answer:

This analysis details the activity log of taxi driver over a period of approximately 8.5 minutes. The driver demonstrates a highly active pattern, completing 36 events – encompassing pickups and drop-offs – within this short timeframe. The driver's operations are concentrated within Manhattan, with significant activity in the Midtown, Chelsea, Tribeca, and Theater District areas, with excursions to Harlem and Battery Park. The frequent, short trips suggest a focus on high-demand, quick-turnaround fares, potentially catering to tourists or commuters. The driver consistently picks up and drops off passengers, indicating efficient utilization of time between fares. The initial segment of the log (0-2.5 minutes) shows a rapid sequence of pickups and drop-offs primarily between Times Square, Hell's Kitchen, Tribeca, and Chelsea. This suggests the driver may have been positioned in a high-demand area initially and quickly responded to available requests. The trip distances during this period are relatively short (ranging from 0.06 to 2.22 miles), reinforcing the idea of quick, localized trips. A notable pattern emerges around the 3-5 minute mark, with a shift towards longer trips, including a 3.67-mile journey to Central Park West and a 5.20-mile trip to Harlem. This indicates a willingness to accept longer fares, potentially to maximize earnings or reposition for better opportunities. However, the driver quickly returns to shorter, more frequent trips in the latter part of the log.

991

The driver's passenger load varies between 1 and 2 passengers per trip. The majority of trips carry 2 passengers, suggesting a preference or availability for larger groups. The final few events (6-8.5 minutes) show a return to a pattern of short trips within the Midtown and Tribeca areas, culminating in a drop-off in Times Square. The driver's ability to quickly transition between pickups and drop-offs, even with varying trip lengths, highlights their efficiency and responsiveness to demand. The consistent activity throughout the log suggests a dedicated and active driver, effectively navigating the dynamic landscape of New York City taxi services. Further analysis, including correlating this data with time of day and external events, could reveal more nuanced patterns in the driver's behavior and the demand for taxi services in specific areas.

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

TAXI-PRO Example_2

1009

1010

MM-TPP's generated text on TAXI-PRO (Example_2)

1011

Prompt:

You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi driver's activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence, please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs. Note that this sequence is NOT VIDEO! All time intervals are converted to four bytes according to the IEEE 754 standard.

<|TPP_data|>

1012

Answer:

The event sequence reveals a complex pattern of multi-modal interactions between a single driver and multiple locations in New York City (NYC). The temporal clustering suggests high-frequency activity, likely driven by surge pricing or peak travel times.

1013

The driver initiates a sequence at 40.764385, -73.977257 (Times Square), then repeatedly picks up and dropped off at multiple locations including Hell's Kitchen, Tribeca, Chelsea, and Battery Park, indicating a rapid-fire sequence of events. This pattern aligns with self-exciting TPP dynamics, where initial events trigger subsequent ones through social or economic stimuli.

1014

The majority of trips involve two passengers, suggesting shared rides or multi-destination excursions, while others are single-passenger. This implies dual-activity modes: some are passive observers (e.g., watching from a distance), while others are active participants, possibly due to short-term availability or limited visibility.

1015

The spatial distribution shows a core-to-periphery pattern, with frequent inter-regional movements, such as from central to peripheral city zones. The average 1.66 mile trip suggests relatively short excursions, potentially indicating micro-mobility patterns or commuter behavior.

1016

A notable temporal shift occurs around 14:00–15:00, where multiple high-density drops are observed, coinciding with peak city traffic and surge pricing surges. This temporal clustering supports TPP triggering effects, where past events (e.g., previous drops) excite future ones.

1017

Unexpectedly, no significant negative sentiment correlates with high engagement, suggesting positive reinforcement loops—possibly due

1018

1019

1020

1021

1022

1023

1024

1025

1026
1027
1028
1029
1030

to social interaction or familiarity with the city. The uniformity of event types (all 10%) indicates broad-based excitation, rather than a single driver anomaly. In summary, this sequence reflects a TPP-driven driver behavior, shaped by multi-modal interactions between time, location, and modalities of action. The driver exhibits self-excitation, temporal clustering, and regional dependency, all hallmarks of complex event-triggered processes in urban transportation systems.

1031
1032

Qwen's generated text on TAXI-PRO (Example_2)

1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048

Prompt:

You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi driver's activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence, please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs.

< |TPP_data| >

Answer:

The provided event sequence details the activities of a single taxi driver in New York City over approximately ten hours. The log captures a series of pickups and drop-offs, each recorded with precise timestamps, passenger counts, and geographic coordinates. The driver begins their shift in the Upper West Side, picking up four passengers and dropping them off in the Theater District after 10 minutes. Over the next few hours, the driver operates primarily in the midtown area, making multiple short trips between the Theater District, Herald Square, and Central Park West. These trips range from 0.73 to 1.35 miles and involve one to three passengers. Around the one-hour mark, the driver completes a 1.11-mile trip from the Theater District to Central Park West and then returns to the Theater District. This pattern of short, frequent trips continues until the 2.5-hour mark, when the driver takes a longer 2.39-mile trip from the Theater District to Midtown East. This is followed by a significant gap in activity, suggesting a break or reduced activity. When the driver resumes, they pick up six passengers in Tribeca and drop them off in the Financial District, a 1.79-mile trip. They then complete another long trip from the Financial District to SoHo, covering 2.43 miles. These longer trips indicate a shift to different parts of the city where demand is lower but the trips are more profitable. The driver returns to the midtown area, making several more trips, including a 4.80-mile journey from Herald Square to Central Park West, the longest trip recorded. The final recorded trip is a 0.99-mile journey from Midtown East to Central Park West, bringing the total number of trips to 36 over the approximately ten-hour period.

In summary, the driver primarily operates in the midtown area, making frequent short trips, while occasionally taking longer, potentially more profitable rides to other parts of the city. This balanced approach likely maximizes both efficiency and profitability.

1049
1050

Gemma's generated text on TAXI-PRO (Example_2)

1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070

Prompt:

You are an expert data analyst specializing in spatio-temporal event sequences. Your task is to analyze a chronological log of a single taxi driver's activities in NYC, which includes actions like pickups and drop-offs at various locations. Based on the provided event sequence, please conduct a comprehensive analysis and generate a detailed Analysis in several paragraphs.

< |TPP_data| >

Answer:

This dataset details the activity of a single taxi driver over approximately 10 minutes, revealing a dynamic pattern of pickups and drop-offs across Manhattan. The driver begins in the Upper West Side, completing a relatively quick trip to the Theater District with a full load of four passengers. This initial segment suggests a potential commute or scheduled pickup. Following this, the driver quickly pivots to serving shorter trips, picking up two passengers in the Theater District and dropping them off in Herald Square, then picking up three passengers in the Theater District and dropping them off in Central Park West. This shift indicates a move towards more opportunistic, shorter-distance fares, typical of a driver circulating in a busy area. The frequent changes in passenger count throughout the initial phase (4, 2, 3) suggest the driver isn't catering to a consistent group or pre-booked long-distance rides. The middle portion of the log (between approximately 5 and 8 minutes) shows a significant change in the driver's operational area and passenger volume. A substantial jump in time occurs before the driver picks up six passengers in Tribeca, indicating a longer travel time to reach this southern location. This is followed by a series of pickups and drop-offs primarily concentrated in the Midtown area, including Times Square, Hell's Kitchen, and Herald Square, consistently carrying a full complement of six passengers. This suggests the driver may have accepted a larger group booking or is capitalizing on high-demand areas during peak hours. The distances of these trips vary, ranging from short hops within Midtown to a longer haul to Central Park West, demonstrating adaptability to passenger requests. Finally, the last few events (around 9-10 minutes) show the driver returning to a pattern of shorter trips, picking up six passengers in Midtown East and dropping them off in Central Park West. The consistent high passenger count during this period suggests the driver is actively seeking to maximize occupancy. The overall pattern reveals a driver who is flexible, adapting to varying trip lengths and passenger numbers. The initial phase focuses on shorter, more dispersed trips, while the middle and end phases demonstrate a preference for maximizing passenger capacity, particularly in the Midtown and surrounding areas. The driver's route demonstrates a good understanding of Manhattan's geography and demand hotspots, efficiently navigating between different neighborhoods to secure fares. The data doesn't provide information on wait times or idle periods, which would be valuable for a more complete assessment of the driver's efficiency.

1071

Task 9 Example_1

1072
1073

MM-TPP's generated text on Task 9 (Example_1)

1074
1075
1076
1077
1078
1079

Prompt:

You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to give a sentiment analysis of the time window. All the comments are in Chinese and all the time intervals are converted to four bytes according to the IEEE 754 standard.

< |TPP_data| >

< |question| > Please provide a detailed analysis of the sentiment trend in the bullet comments over time.

Answer:

The video exhibits a predominantly positive sentiment with minor fluctuations. Initially, there's a sharp rise followed by stabilization around a positive polarity. Key inflection points include early peaks at approximately 00:00:30 and 00:01:00, with slight dips around

1080
 1081 00:01:30 and 00:02:00. Significant sentiment shifts occur near 00:03:00, showing a gradual decline but remaining positive. Overall, the
 1082 video maintains a consistently positive emotional trajectory with minimal negative deviations.
 1083

1084 **Qwen's generated text on Task 9 (Example_1)**

1085 **Prompt:**
 1086 You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to
 1087 give a sentiment analysis of the time window. All the comments are in Chinese.
 1088 Question: Please provide a detailed analysis of the sentiment trend in the bullet comments over time.
 1089 <|TPP_data|>
 1090 **Answer:**
 1091 The sentiment trend of the bullet comments for the video generally starts with a neutral to positive tone, as viewers are likely intrigued by
 1092 the title and the initial scenes of the video. As the video progresses, the comment frequency increases, indicating heightened engagement.
 1093 Around the 20-second mark, there is a noticeable spike in comments, suggesting a particularly engaging or exciting moment in the video,
 1094 which could be a dramatic visual or an impactful statement. The sentiment remains largely positive throughout, with occasional spikes
 1095 indicating moments of high excitement or awe. By the mid-point of the video (around 120 seconds), the comments stabilize but continue
 1096 to show enthusiasm, with viewers expressing admiration for the content or sharing their own experiences. Towards the end of the
 1097 video, the comment frequency slightly decreases, but the overall positive sentiment persists, with many viewers leaving final thoughts or
 1098 encouragements. This trend suggests that the video successfully maintains viewer interest and elicits a consistently positive response.
 1099

1100 **Gemma's generated text on Task 9 (Example_1)**

1101 **Prompt:**
 1102 You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to
 1103 give a sentiment analysis of the time window. All the comments are in Chinese.
 1104 Question: Please provide a detailed analysis of the sentiment trend in the bullet comments over time.
 1105 <|TPP_data|>
 1106 **Answer:**
 1107 The bullet comment density indicates consistently high engagement throughout the video, with numerous peaks suggesting moments of
 1108 particular interest or impact. Engagement starts slowly, builds steadily, and then remains very active with frequent spikes from around
 1109 2 seconds onward. There's no clear overall decline in engagement; it remains high until the very end of the data provided. Without the
 1110 comment text, it's impossible to determine what sentiment is being expressed (positive, negative, surprise, etc.). However, the consistent
 1111 high volume suggests the video is generally captivating and elicits a response from viewers, likely related to the exciting content described
 1112 in the title ("Million Volt Voltage Capture!"). The frequent spikes likely correspond to visually or narratively dramatic moments.
 1113

1114 **Task 9 Example_2**

1115 **MM-TPP's generated text on Task 9 (Example_2)**

1116 **Prompt:**
 1117 You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to
 1118 give a sentiment analysis of the time window. All the comments are in Chinese and all the time intervals are converted to four bytes
 1119 according to the IEEE 754 standard.
 1120 <|TPP_data|>
 1121 <|question|> Please provide a detailed analysis of the sentiment trend in the bullet comments over time.
 1122 **Answer:**
 1123 The sentiment polarity trend shows a predominantly positive trajectory with minor fluctuations. Initially, there's a sharp rise followed
 1124 by stabilization around the 0.4 mark. Key inflection points include early peaks and valleys that settle into a consistent positive range.
 1125 Significant sentiment shifts occur at the beginning with a steep increase and towards the end with a slight decline. Overall, the video
 1126 maintains a positive emotional tone with minimal abrupt changes, indicating a stable and uplifting viewer experience.
 1127

1128 **Qwen's generated text on Task 9 (Example_2)**

1129 **Prompt:**
 1130 You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to
 1131 give a sentiment analysis of the time window. All the comments are in Chinese.
 1132 Question: Please provide a detailed analysis of the sentiment trend in the bullet comments over time.
 1133 <|TPP_data|>
 1134 **Answer:**
 1135 The sentiment trend in the bullet comments shows a clear pattern over time. Initially, from 0 to 60 seconds, there is a very high density
 1136 of comments, indicating high excitement and positive reactions as viewers are immediately engaged and enthusiastic. From 60 to 120
 1137 seconds, the comment density moderates, suggesting steady engagement with mixed reactions as viewers remain interested but less
 1138 intensely. In the mid-phase (120 to 300 seconds), the density fluctuates with occasional spikes, reflecting mixed reactions and key scenes
 1139 that spark more comments. From 300 to 480 seconds, the comment density is lower with occasional spikes, indicating more subdued
 1140 reactions and deeper engagement with the content. Finally, from 480 to 633 seconds, there is a higher density of comments with a
 1141 significant spike near the end, showing increased engagement and final thoughts as viewers share their overall impressions and reactions
 1142 to the conclusion. Overall, the video maintains viewer interest throughout, with notable peaks at the beginning and end, and steady
 1143 engagement in between.
 1144

1134
1135

Gemma's generated text on Task 9 (Example_2)

1136

Prompt:

1137

You are an expert model for analyzing bullet comments. Based on the sequence of comments and their timing intervals, your task is to give a sentiment analysis of the time window. All the comments are in Chinese.

1138

Question: Please provide a detailed analysis of the sentiment trend in the bullet comments over time.

1139

<|TPP_data|>

1140

Answer:

1141

The bullet comments show high initial engagement (0-10s), dipping slightly before building to sustained high density from 20s onward. Several peaks in comment frequency (around 20-40s, and 100-150s) suggest impactful moments in the film. The very end (550s+) maintains extremely high comment volume, indicating strong final reactions. Essentially, the audience was consistently engaged, with specific scenes eliciting particularly strong responses throughout the viewing experience. Without the comment content, it's impossible to say if these peaks are positive or negative, only that they represent heightened attention.

1144

1145

Task 10 Example_1

1146

MM-TPP's generated text on Task 10 (Example_1)

1147

Prompt:

1148

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese and all the time intervals are converted to four bytes according to the IEEE 754 standard.

1149

<|TPP_data|>

1150

<|question|> Within the time window: [0, 1.5], please analyze in a concise and fluent paragraph the causes of the burst peaks in bullet comments.

1151

Answer:

1152

The bullet comment burst at timestamp 0 exhibits a sharp initial spike, followed by rapid-fire clusters at 0.08–0.1s, 0.3–0.4s, and 0.6–0.7s, indicating self-exciting TPP dynamics—early comments triggered cascading audience responses. These sub-peaks align with recurring quote/reference and humor/meme event types (e.g., “《百家讲坛》”, “败家讲坛”), suggesting viral textual propagation driven by meme culture and inside jokes within the community.

1153

The average sentiment polarity of 0.42 reflects a mildly positive tone**, yet the diversity in comment types—including critical, emotional, and off-topic remarks—shows heterogeneous audience reactions. Notably, the presence of neutral visual tags (“商务西装”, “演讲”) contrasts with emotionally charged comments, revealing a multi-modal dissonance: formal attire and speech settings likely prompted ironic or humorous reinterpretations, fueling meme generation and engagement.

1154

Temporal alignment reveals that the first sub-peak 0.0–0.1s was dominated by nonsense_text and emotional_reaction, possibly triggered by the sudden visual shift or unexpected content juxtaposition. The second cluster (0.3–0.4s) saw increased social_interaction and personal_experience, indicating viewers began forming collective interpretations or debates. Finally, the third wave (0.6–0.7s) featured more critical_comment and quote/reference, reflecting deeper engagement with thematic comparisons (e.g., “范小勤” vs. “丁真”).

1155

This burst illustrates TPP self-excitation, where early comments acted as triggers, prompting a cascade of related and reactive posts. The multimodal mismatch between formal visuals and humorous/textual responses underscores how audience interpretation, not just on-screen content, drives engagement. The result is a complex interplay of textual virality, temporal clustering, and visual framing, typical of interactive streaming platforms.

1156

1157

Qwen's generated text on Task 10 (Example_1)

1158

Prompt:

1159

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese.

1160

Question: Within the time window: [0, 1.5], please analyze in a concise and fluent paragraph the causes of the burst peaks in bullet comments.

1161

<|TPP_data|>

1162

Answer:

1163

Within the time window [0, 1.5] seconds, the burst peaks in bullet comments are primarily driven by the strong temporal self-excitation in the Temporal Point Process (TPP). These peaks cluster around 0.48–0.58 and 1.06–1.46 seconds, coinciding with the appearance of formal visual cues such as “商务西装”, “演讲”, and “新闻.” These visual elements set a mock-serious or authoritative tone, which catalyzes multi-modal irony and prompts viewers to juxtapose the visuals with satirical or culturally referential commentary. The recurring references to satirical labels like “《百家讲坛》” and “《今日说法》” indicate a feedback loop where one humorous comment inspires similar responses, particularly in the humor/meme and quote/reference categories. Despite a neutral-positive average sentiment (0.384), the bursts of emotional reaction and nonsense text comments suggest sentiment heterogeneity driven by entertainment rather than criticism. The sub-peak intervals, such as 1.25–1.46 seconds, reinforce this pattern, with repeated mentions of “《百家讲坛》” and other meta-commentary further amplifying the engagement. Overall, the burst is best explained by self-exciting humor cascades that resonate culturally and create a continuous stream of ironic and playful interactions.

1164

1165

Gemma's generated text on Task 10 (Example_1)

1166

Prompt:

1167

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese.

1168

Question: Within the time window: [0, 1.5], please analyze in a concise and fluent paragraph the causes of the burst peaks in bullet comments.

1169

<|TPP_data|>

1170

Answer:

1171

The burst of bullet comments is primarily driven by a combination of satirical commentary and ironic juxtaposition, evidenced by the

1188
1189
1190
1191
1192
1193

frequent references to shows like “《百家脑坛》” and “《今日说法》” alongside the formal setting of a debate-like scenario. This creates a strong self-exciting loop where humorous and meme-like comments (around 14.29%) trigger similar responses, amplified by the visual cues of the scene – particularly the presence of figures in “商务西装” engaged in “演讲”. While the sentiment is generally positive, the bursts correlate with moments of playful mockery and cultural references, suggesting viewers are engaging in ironic commentary rather than expressing strong negative reactions. The visual elements, rather than dampening engagement, seem to catalyze it, fostering a dynamic where the formal setting is ironically contrasted with the absurd or culturally resonant commentary, leading to a cascade of humorous responses.

1194
1195

1196 Task 10 Example_2

1197
1198

MM-TPP’s generated text on Task 10 (Example_2)

1199

Prompt:

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese and all the time intervals are converted to four bytes according to the IEEE 754 standard.

<|TPP_data|>

<|question|> Within the time window: [393.5, 396.5], please analyze in detail the causes of the burst peaks in bullet comments.

Answer:

The bullet comment burst peaks at 395 seconds, with dense timestamps revealing sub-peaks around 394.8–395.0 and 395.6–395.8, indicating rapid-fire engagement intervals. These clusters align with humor/meme (e.g., “15!”、“55555”), emotional reactions (“哥们你德玛的啊，你要是在s11遇到我...”), and social interactions, suggesting a self-exciting TPP pattern where early comments trigger cascading responses.

The dominant object tags—*游戏*, *截图*, *电子游戏*—imply a gameplay highlight or screenshot moment that resonated emotionally and humorously with viewers. Sentiment polarity averages at 0.423 (positive), consistent with excitement and meme-like commentary rather than critical or off-topic remarks.

Temporal alignment shows that the initial spike (394.5–394.8) centers on confusion or anticipation (“?”, “我觉得不能的，扣眼珠子”), likely triggered by a visually unexpected event. This is followed by a surge of numeric/humorous texts (“1! 5!”, “5”), indicating real-time voting or scoring rituals common in gaming videos, reinforcing a triggering effect among viewers.

Notably, despite the positive sentiment, some critical or off-topic comments appear (e.g., “不能”), suggesting a mixed but engaged audience. The repeated use of “5” and “15!” reflects a coordinated or meme-driven response pattern, consistent with self-excitation in TPP models—where similar comments cluster closely in time.

Multi-modal correlation reveals that the visual context (gameplay frame tagged 游戏) prompted rapid textual engagement, especially humor and social mimicry, amplifying the burst. No negative sentiment spikes are observed, ruling out controversy-driven engagement. Instead, the burst reflects synchronous emotional and interactive behavior, driven by shared gaming culture and visual triggers.

1215

Qwen’s generated text on Task 10 (Example_2)

1216

Prompt:

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese.

Question: Within the time window: [393.5, 396.5], please analyze in detail the causes of the burst peaks in bullet comments.

<|TPP_data|>

Answer:

Within the time window [393.5, 396.5], the burst peaks in bullet comments are primarily driven by the convergence of visual and textual elements that elicit strong emotional and social interactions. The video frame captures a critical moment in the game, likely a high-difficulty challenge (“地狱级难度翡翠”), which serves as a visual trigger for the audience. This visual cue prompts a surge in comments, characterized by humor and memes (e.g., “扣大雷”, “扣眼珠子”) and social interaction (e.g., numerical repetitions “111111”, “55555”). The mildly positive sentiment (0.386) reflects the excitement and playful banter among viewers, who are anticipating the unpredictable outcomes of the challenging game segment. The rapid-fire clustering of similar phrases and numerals indicates a self-exciting Hawkes process, where initial comments stimulate a cascade of responses, creating a pulse of collective engagement. The temporal alignment of these comments with the critical gameplay moment further reinforces the multi-modal interaction, where visual cues and textual responses work in tandem to drive the burst peaks.

1229

Gemma’s generated text on Task 10 (Example_2)

1230

Prompt:

You are an expert model for analyzing bullet comments. Based on the sequence of comments, their timing interval, and the associated video frames, your task is to analyze in detail the causes of the burst peaks in bullet comments and give a detailed analysis report. All the comments are in Chinese.

Question: Within the time window: [393.5, 396.5], please analyze in detail the causes of the burst peaks in bullet comments.

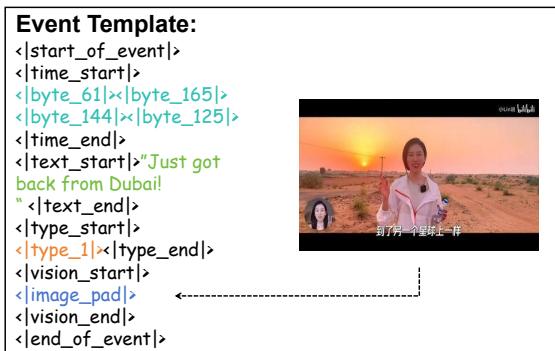
<|TPP_data|>

Answer:

Based on the provided data, the burst of bullet comments between 393.5 and 396.5 seconds is primarily driven by a combination of gameplay events and coordinated social interaction, best explained by option D. The peaks align with moments likely depicting a challenging in-game scenario (“地狱级难度翡翠” - Jade difficulty), triggering emotional reactions and questions from viewers. Simultaneously, a strong element of meme-based coordination is present, evidenced by repetitive numerical expressions (“111111”, “55555”) and phrases like “扣大雷,” indicating viewers are engaging in playful, culturally resonant interaction scripts. This suggests a self-exciting Hawkes process dynamic where initial comments stimulate cascading responses, amplified by the visual cue of the gameplay screenshot acting as a latent trigger for rapid-fire textual engagement. The mildly positive sentiment (0.386) further supports this interpretation of excitement and playful banter rather than negativity or criticism.

1242 **F PROMPT DETAILS**
12431244 Here we present the detailed prompt templates used for preprocessing temporal point process
1245 datasets. Due to the similarity between processing the two datasets, we only present the templates
1246 of TAXI-PRO.
12471248 **Special Tokens** Here we present all special tokens added to the Qwen2.5-VL tokenizer vocabulary
1249 and used in our work to structure the prompts in Table 8. We note that $<|im_start|>$
1250 , $<|im_end|>$, $<|vision_start|>$, $<|vision_end|>$, $<|image_pad|>$ are native special tokens
1251 from the Qwen2.5-VL tokenizer.
12521253 Table 8: Descriptions of Special Tokens
1254

1255 Special Token	1256 Description
$< start_of_event >$	Marks the start of an event.
$< end_of_event >$	Marks the end of an event.
$< time_start >$	Marks the start of a timestamp.
$< time_end >$	Marks the end of a timestamp.
$< type_start >$	Marks the start of an event type.
$< type_end >$	Marks the end of an event type.
$< text_start >$	Marks the start of the text of an event.
$< text_end >$	Marks the end of the text of an event.
$< time_prediction >$	Task token for next event time prediction.
$< type_prediction >$	Task token for next event type prediction.
$< question >$	Task token for Question-Answering.
$< byte_0 > \dots < byte_255 >$	Byte tokens for representing event time intervals as float32 numbers.
$< type_0 > \dots < type_8 >$	Type tokens for representing event types.

1274 **Event Template** Here we present the event template in Fig. 5.
12751288 Figure 5: A complete example of an uncompressed event representation
12891290 **Prompt Template** Here in Table 9 we provide a sample generated event sequence from the TAXI-
1291 PRO dataset used in Stage-1 training process. We also provide two prompt-response pair samples
1292 from the TAXI-PRO dataset used in Stage-2 next event type finetuning and next event time finetuning
1293 in Table 10 and Table 11. The construction logic for all three templates is largely consistent, with
1294 minor variations in the system prompt.
1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

Table 9: Event Sequence sample from TAXI-PRO

Prompt:

<|im_start|>system

You are an intelligent urban mobility analysis model specializing in NYC taxi patterns. Your task is to predict the precise time interval until the next taxi event in Manhattan. You will be given a history of taxi trips, where each event includes its type (e.g.Uptown Pickup, Downtown Drop-off), the time elapsed since the last event, and a map image of the location.

<|im_end|>

Event Sequence:

<|start_of_event|>

<|time_start|><|byte_0|><|byte_0|><|byte_0|><|byte_0|><|time_end|>

<|type_start|><|type_0|><|type_end|>

<|text_start|>Picked up at Tribeca (40.711086, -74.016106), 1 passengers.<|text_end|>

<|vision_start|><|image_pad|><|vision_end|>

<|end_of_event|>

<|start_of_event|>

<|time_start|><|byte_62|><|byte_162|><|byte_34|><|byte_34|><|time_end|>

<|type_start|><|type_3|><|type_end|>

<|text_start|>Dropped off from Tribeca (40.711086, -74.016106) to Times Square (40.757698, -73.982124), 1 passengers, 2.87 miles trip.<|text_end|>

<|vision_start|><|image_pad|><|vision_end|>

<|end_of_event|>

<|start_of_event|>

<|time_start|><|byte_62|><|byte_196|><|byte_68|><|byte_68|><|time_end|>

<|type_start|><|type_4|><|type_end|>

<|text_start|>Picked up at Upper West Side (40.799252, -73.970146), 1 passengers.<|text_end|>

<|vision_start|><|image_pad|><|vision_end|>

<|end_of_event|>

<|start_of_event|>

<|time_start|><|byte_63|><|byte_17|><|byte_16|><|byte_161|><|time_end|>

<|type_start|><|type_1|><|type_end|>

<|text_start|>Dropped off from Upper West Side (40.799252, -73.970146) to Tribeca (40.714455, -74.014008), 1 passengers, 4.37 miles trip.<|text_end|>

<|vision_start|><|image_pad|><|vision_end|>

<|end_of_event|>

1331

1332

1333

Table 10: Prompt-response pair sample for next event time finetuning from TAXI-PRO

Prompt:

<|im_start|>system

You are an intelligent urban mobility analysis model specializing in NYC taxi patterns. Your task is to predict the precise time interval until the next taxi event in Manhattan. You will be given a history of taxi trips, where each event includes its type (e.g. Uptown Pickup, Downtown Drop-off), the time elapsed since the last event, and a map image of the location. Analyze the patterns in event types and timing to forecast the time to the next event. Your response must be exactly four byte tokens representing the time interval.

<|im_end|>

Event Sequence History:

<|start_of_event|>

<|time_start|><|byte_0|><|byte_0|><|byte_0|><|byte_0|><|time_end|>

<|type_start|><|type_0|><|type_end|>

<|text_start|>Picked up at Tribeca (40.711086, -74.016106), 1 passengers.<|text_end|>

<|vision_start|><|image_pad|><|vision_end|>

<|end_of_event|>

<|start_of_event|>

```

1350
1351 < |time_start| >< |byte_62| >< |byte_162| >< |byte_34| >< |byte_34| >< |time_end| >
1352 < |type_start| >< |type_3| >< |type_end| >
1353 < |text_start| >Dropped off from Tribeca (40.711086, -74.016106) to Times Square
1354 (40.757698, -73.982124), 1 passengers, 2.87 miles trip.< |text_end| >
1355 < |vision_start| >< |image_pad| >< |vision_end| >
1356 < |end_of_event| >
1357 < |start_of_event| >
1358 < |time_start| >< |byte_62| >< |byte_196| >< |byte_68| >< |byte_68| >< |time_end| >
1359 < |type_start| >< |type_4| >< |type_end| >
1360 < |text_start| >Picked up at Upper West Side (40.799252, -73.970146), 1
1361 passengers.< |text_end| >
1362 < |vision_start| >< |image_pad| >< |vision_end| >
1363 < |end_of_event| >
1364 < |time_prediction| >

```

Response:

```
< |byte_63| >< |byte_17| >< |byte_16| >< |byte_161| >
```

Table 11: Prompt-response pair sample for next event type finetuning from TAXI-PRO

Prompt:

```
< |im_start| >system
```

You are an intelligent urban mobility analysis model specializing in NYC taxi patterns. Your task is to predict the type of the next taxi event in Manhattan based on a sequence of past events. You will be given a history of taxi trips, where each event includes its type (e.g. Uptown Pickup, Downtown Drop-off), the time elapsed since the last event, and a map image of the location. Analyze the patterns in timing and location to forecast the next action. Your response must be a single token representing one of the six event types.

```
< |im_end| >
```

Event Sequence History:

```
< |start_of_event| >
< |time_start| >< |byte_0| >< |byte_0| >< |byte_0| >< |byte_0| >< |time_end| >
< |type_start| >< |type_0| >< |type_end| >
< |text_start| >Picked up at Tribeca (40.711086, -74.016106), 1 passengers.< |text_end| >
< |vision_start| >< |image_pad| >< |vision_end| >
< |end_of_event| >
< |start_of_event| >
< |time_start| >< |byte_62| >< |byte_162| >< |byte_34| >< |byte_34| >< |time_end| >
< |type_start| >< |type_3| >< |type_end| >
< |text_start| >Dropped off from Tribeca (40.711086, -74.016106) to Times Square
1391 (40.757698, -73.982124), 1 passengers, 2.87 miles trip.< |text_end| >
1392 < |vision_start| >< |image_pad| >< |vision_end| >
1393 < |end_of_event| >
1394 < |start_of_event| >
1395 < |time_start| >< |byte_62| >< |byte_196| >< |byte_68| >< |byte_68| >< |time_end| >
1396 < |type_start| >< |type_4| >< |type_end| >
1397 < |text_start| >Picked up at Upper West Side (40.799252, -73.970146), 1
1398 passengers.< |text_end| >
1399 < |vision_start| >< |image_pad| >< |vision_end| >
1400 < |end_of_event| >
1401 < |type_prediction| >
```

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

Response:

<|type_1|>

G THE USE OF LARGE LANGUAGE MODELS

In the preparation of this manuscript, we utilized LLM as a writing-assistive tool. Its application was strictly limited to improving the language and presentation of the paper, which included correcting grammatical errors, rephrasing sentences for enhanced clarity and conciseness, and refining paragraph structure. The core scientific contributions, including research ideation, methodology, and analysis, were conceived and executed entirely by the human authors. The authors have reviewed and edited all text and take full responsibility for the final content of the paper.