# Novelty Assessment Report

**Paper**: Long-range Modeling and Processing of Multimodal Event Sequences
**PDF URL**: https://openreview.net/pdf?id=Krxt7wCnig
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-27

## 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.



## Core Task Landscape

This paper addresses: **Multimodal Temporal Point Process Modeling with Long-Range Dependencies**
A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Temporal Point Process Architectures and Mechanisms**
- **Multimodal Fusion and Integration Strategies**
- **Long-Horizon and Sequential Task Modeling**
- **Spatiotemporal and Convolutional Temporal Modeling**
- **Domain-Specific Multimodal Temporal Applications**
- **Temporal Modeling Enhancements and Specialized Mechanisms**

### Complete Taxonomy Tree

- Multimodal Temporal Point Process Modeling with Long-Range Dependencies Survey Taxonomy
- Temporal Point Process Architectures and Mechanisms
  - Attention-Based Temporal Point Process Models (3 papers)
  - [1] Transformers for mixed-type event sequences (F Draxler, 2025) View paper
  - [11] Spatio-temporal point processes with attention for traffic congestion event modeling (Shixiang Zhu, 2021) View paper
  - [42] Shock-Biased Attention: Enhancing Transformer Hawkes Processes with Amplitude-Driven Temporal Kernels (S Hwang, 2025) View paper
  - Recurrent and Hybrid Temporal Point Process Models (2 papers)
  - [32] Learning time series associated event sequences with recurrent point process networks (Shuai Xiao, 2019) View paper
  - [33] Multi-modal Generative Modeling of Event Sequences and Time Series for Solar PV Systems (Jiayu Huang, 2025) View paper
  - Case-Based and Memory-Augmented Temporal Models (2 papers)
  - [20] A Case-based reasoning and explaining model for temporal point process (Liu, 2024) View paper
  - [49] Predicting text features of social temporal point process (Di, 2025) View paper
- Multimodal Fusion and Integration Strategies
  - Transformer-Based Multimodal Fusion (3 papers)
  - [6] Multimodal fusion method with spatiotemporal sequences and relationship learning for valence-arousal estimation (Yu Jun, 2024) View paper
  - [13] HGTFM: Hierarchical Gating-Driven Transformer Fusion Model for Robust Multimodal Sentiment Analysis (Chengcheng Yang, 2025) View paper
  - [18] Transformer Models for Multimodal Sequence Learning (Ragavenderan, 2025) View paper
  - Hierarchical and Gated Multimodal Fusion (2 papers)
  - [26] MHSCNET: A Multimodal Hierarchical Shot-Aware Convolutional Network for Video Summarization (Wujiang Xu, 2022) View paper
  - [27] Bilevel Relational Graph Representation Learning-based Multimodal Emotion Recognition in Conversation (Huan Zhao, 2024) View paper
  - Adaptive and Dynamic Multimodal Alignment (2 papers)
  - [16] BALM-TSF: Balanced Multimodal Alignment for LLM-Based Time Series Forecasting (Zhou Shi-qiao, 2025) View paper
  - [34] SpeechCARE: dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts (Hossein Azadmaleki, 2025) View paper

- Mixture-of-Experts and Modular Multimodal Architectures (3 papers)
- [10] M3Rec: Selective State Space Models with Mixture-of-Modality Experts for Multi-Modal Sequential Recommendation (Xu Guo, 2025) View paper
- [14] UMI-Rec: A Unified Multi-modal Intent Fusion Framework with State-Space Models and Large Language Models for Recommendation (Zare, 2025) View paper
- [21] MELON: Multimodal Mixture-of-Experts with Spectral-Temporal Fusion for Long-Term Mobility Estimation in Critical Care (Zhang Jia-qing, 2025) View paper

• Long-Horizon and Sequential Task Modeling
- Vision-Language-Action Models for Long-Horizon Manipulation ★ (4 papers)
- [0] Long-range Modeling and Processing of Multimodal Event Sequences (Anon et al., 2026) View paper
- [9] From Watch to Imagine: Steering Long-horizon Manipulation via Human Demonstration and Future Envisionment (Ye, 2025) View paper
- [29] Preference-Based Long-Horizon Robotic Stacking with Multimodal Large Language Models (Yu, 2025) View paper
- [36] Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation (Ding, 2025) View paper
- Memory-Driven and Chain-of-Thought Long-Horizon Planning (2 papers)
- [2] Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory (Long Lin, 2025) View paper
- [17] Memory-driven multimodal chain of thought for embodied long-horizon task planning (X Liang, 2025) View paper
- Autoregressive and Phase-Aware Long-Horizon Generation (2 papers)
- [12] Longvie: Multimodal-guided controllable ultra-long video generation (Gao Jian-xiong, 2025) View paper
- [38] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation (Yang Liu-di, 2025) View paper
- Trajectory and Motion Prediction for Extended Horizons (2 papers)
- [5] Label-free long-horizon 3d uav trajectory prediction via motion-aligned rgb and event cues (Yuan, 2025) View paper
- [41] Exploring complex dependencies for multi-modal semantic trajectory prediction (Jie Liu, 2022) View paper

• Spatiotemporal and Convolutional Temporal Modeling
- Dual-Branch Spatiotemporal Architectures (2 papers)
- [4] Domain-collaborative multimodal transformer for fault diagnosis of rotating machines under noisy environments (Seon-Gyu Kim, 2025) View paper
- [8] SkyNet: A deep learning architecture for intra-hour multimodal solar forecasting with ground-based sky images (Guoping Ruan, 2025) View paper
- Temporal Convolutional and Recurrent Spatiotemporal Models (2 papers)
- [7] Multimodal Fall Detection Using Spatial-Temporal Attention and Bi-LSTM-Based Feature Fusion (Jungpil Shin, 2025) View paper
- [15] Automatic depression recognition with an ensemble of multimodal spatio-temporal routing features (Yaowei Wang, 2025) View paper
- Graph-Based Spatiotemporal Reasoning (2 papers)
- [28] Capturing Spectral and Long-term Contextual Information for Speech Emotion Recognition Using Deep Learning Techniques (Samiul, 2023) View paper
- [47] Scene-aware Graph-enhanced Multimodal Collaboration for Video-grounded Dialogue (Shanshan Du, 2025) View paper

• Domain-Specific Multimodal Temporal Applications
- Healthcare and Clinical Temporal Modeling (3 papers)
- [3] Multimodal integration of physiological signals clinical data and medical imaging for ICU outcome prediction (Wang Qingquan, 2025) View paper
- [45] AI-Driven Multimodal Prognosis in Stroke: Integrating Neuroimaging and Longitudinal Clinical Data for Enhanced Detection and Long-Term Outcome Prediction (Ch. Avaneesh, 2025) View paper
- [50] Edge-AI Enabled IoT Framework for Real-Time Stroke Risk Prediction Using Multimodal Biosignals (Sujatha Krishna, 2025) View paper
- Video Understanding and Captioning (3 papers)
- [22] Toward long form audio-visual video understanding (Wenxuan Hou, 2024) View paper
- [25] Joint multi-scale information and long-range dependence for video captioning (Zhongyi Zhai, 2023) View paper
- [40] LongInsightBench: A Comprehensive Benchmark for Evaluating Omni-Modal Models on Human-Centric Long-Video Understanding (Han, 2025) View paper
- Action Recognition and Human Activity Analysis (2 papers)
- [19] The Journey of Action Recognition (Xi Ding, 2025) View paper
- [24] Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions (David Curto, 2021) View paper
- Time Series Forecasting with Multimodal Context (2 papers)
- [30] A multi-modal approach for mixed-frequency time series forecasting. (Leopoldo Lusquino Filho, 2024) View paper
- [39] Hybrid informerâ□□BiLSTM Model for Long-Term Forecasting of Urban Heat Islands: A Multimodal Approach Integrating Remote Sensing and Climate Data (Bilikis Arinola Alege-Ibrahim, 2025) View paper
- Specialized Temporal Applications (3 papers)
- [37] MT-DPCQA: A Multimodal Time-aware Learning Approach for No-Reference Dynamic Point Cloud Quality Assessment (Swarna Chakraborty, 2025) View paper
- [46] Attention-Based Multiscale Temporal Fusion Network for Uncertain-Mode Fault Diagnosis in Multimode Processes (Li Guang-qiang, 2025) View paper
- [48] SPAN: Continuous Modeling of Suspicion Progression for Temporal Intention Localization (Hu Xinyi, 2025) View paper

• Temporal Modeling Enhancements and Specialized Mechanisms
- Sequence Compression and Efficiency Mechanisms (1 papers)
- [23] TFF-temporal fusion framework for advancing video retrieval through long-range dependencies and multi-modal intent (Pratibha Singh, 2025) View paper
- Multi-Stage and Dual-Stage Attention Architectures (2 papers)
- [43] Dual-Stage Attention-Augmented ACT Model: Optimization and Validation for Bimanual Fine Manipulation Tasks (Shuhang Liang, 2025) View paper
- [44] Robot Confirmation Generation and Action Planning Using Long-context Q-Former Integrated with Multimodal LLM (Chiori Hori, 2025) View paper
- Temporal Multimodal Learning Foundations (2 papers)

- [31] Temporal multimodal learning in audiovisual speech recognition (Di Hu, 2016) View paper
- [35] Text-enhanced multi-granularity temporal graph learning for event prediction (Xiaoxue Han, 2022) View paper

### Narrative

Core task: Multimodal temporal point process modeling with long-range dependencies. This field addresses the challenge of capturing event sequences that unfold over time and involve multiple modalities—such as vision, language, and sensor data—while maintaining sensitivity to dependencies that span extended temporal horizons. The taxonomy reflects a diverse landscape organized into six main branches. Temporal Point Process Architectures and Mechanisms focuses on foundational modeling techniques, including transformer-based approaches for mixed event types (Transformers Mixed Events[1]) and neural architectures for event sequences (Event Sequences Networks[32]). Multimodal Fusion and Integration Strategies examines how to combine heterogeneous data streams, with applications ranging from ICU prediction (Multimodal ICU Prediction[3]) to fault diagnosis (Multimodal Fault Diagnosis[4]). Long-Horizon and Sequential Task Modeling emphasizes extended temporal reasoning, particularly in vision-language-action settings for robotic manipulation and planning. Spatiotemporal and Convolutional Temporal Modeling addresses spatial dynamics alongside temporal patterns, as seen in trajectory prediction (UAV Trajectory Prediction[5], Trajectory Prediction Dependencies[41]) and urban analytics. Domain-Specific Multimodal Temporal Applications showcases specialized use cases in healthcare, energy forecasting, and affective computing, while Temporal Modeling Enhancements and Specialized Mechanisms explores refinements such as attention mechanisms and memory structures (Multimodal Agent Memory[2]).

A particularly active line of work centers on long-horizon sequential tasks, where models must integrate multimodal observations over extended episodes to guide decision-making or manipulation. Within this branch, vision-language-action models for robotic manipulation represent a dense cluster, with papers like Long-VLA[36] and Robotic Stacking Preferences[29] exploring how to ground language instructions in visual perception and action sequences. The original paper, Multimodal Event Sequences[0], sits naturally within this cluster, emphasizing the modeling of event sequences that span long temporal ranges and multiple modalities. Compared to Long-VLA[36], which focuses on robotic manipulation tasks, Multimodal Event Sequences[0] appears to take a broader view of temporal point processes, potentially addressing a wider variety of event-driven scenarios beyond embodied agents. Meanwhile, works like Watch to Imagine[9] highlight the role of predictive modeling in long-horizon settings, contrasting with the more direct event-sequence framing of Multimodal Event Sequences[0]. Open questions remain around scalability, the trade-offs between specialized domain models and general-purpose architectures, and the effective integration of symbolic event representations with continuous multimodal streams.

## Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. From Watch to Imagine: Steering Long-horizon Manipulation via Human Demonstration and Future Envisionment

**Authors**: Ye, Ke, Zhou Jia-ming, Qiu Yuan-feng, Liu Jiayi, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Generalizing to long-horizon manipulation tasks in a zero-shot setting remains a central challenge in robotics. Current multimodal foundation based approaches, despite their capabilities, typically fail to decompose high-level commands into executable action sequences from static visual input alone. To address this challenge, we introduce Super-Mimic, a hierarchical framework that enables zero-shot robotic imitation by directly inferring procedural intent from unscripted human demonstration vide...

#### Relationship Analysis

Both papers belong to the Vision-Language-Action Models for Long-Horizon Manipulation category, addressing extended robotic manipulation tasks through multimodal integration. While the original paper focuses on modeling multimodal temporal point processes (TPPs) for event sequences with long-range dependencies using LLMs and adaptive compression mechanisms, the candidate paper (Super-Mimic) addresses long-horizon robotic manipulation by translating human demonstration videos into executable action sequences through hierarchical planning and future dynamics prediction. The key difference is that the original paper models asynchronous event sequences with temporal dynamics, whereas the candidate paper focuses on zero-shot robotic imitation learning from human videos.

---

### 2. Preference-Based Long-Horizon Robotic Stacking with Multimodal Large Language Models

**Authors**: Yu, Wanming, RÃ¶fer, Adrian, Valada, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Pretrained large language models (LLMs) can work as high-level robotic planners by reasoning over abstract task descriptions and natural language instructions, etc. However, they have shown a lack of knowledge and effectiveness in planning long-horizon robotic manipulation tasks where the physical properties of the objects are essential. An example is the stacking of containers with hidden objects inside, which involves reasoning over hidden physics properties such as weight and stability. To th...

#### Relationship Analysis

Both papers belong to the Vision-Language-Action Models for Long-Horizon Manipulation category, focusing on integrating multimodal information for extended robotic tasks. The original paper addresses multimodal temporal point process modeling for event sequences (e.g., video comments with timestamps and images), emphasizing long-range dependencies and text generation through adaptive compression mechanisms. The candidate paper focuses on robotic stacking tasks using multimodal LLMs to reason over hidden physical properties (weight, stability) obtained through force/audio sensing, with a preference-based planning approach rather than temporal event sequence modeling.

---

### 3. Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation

**Authors**: Ding, Pengxiang, Tong Xin-yang, Zhu Yu-yang, Lu Hongchao, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Vision-Language-Action (VLA) models have become a cornerstone in robotic policy learning, leveraging large-scale multimodal data for robust and scalable control. However, existing VLA frameworks primarily address short-horizon tasks, and their effectiveness on long-horizon, multi-step robotic manipulation remains limited due to challenges in skill chaining and subtask dependencies. In this work, we introduce Long-VLA, the first end-to-end VLA model specifically designed for long-horizon robotic ...

#### Relationship Analysis

Both papers belong to the Vision-Language-Action Models for Long-Horizon Manipulation category, focusing on integrating multimodal inputs for extended robotic task execution. The original paper addresses long-range dependencies in multimodal temporal point process modeling through adaptive sequence compression and LLM-based event prediction, while the candidate paper tackles long-horizon robotic manipulation through phase-aware input masking and end-to-end VLA training. The key difference is that the original paper

focuses on modeling and analyzing asynchronous event sequences with temporal point processes, whereas the candidate paper focuses on robotic policy learning and action generation for sequential manipulation tasks.

## Contributions Analysis

**Overall novelty summary.** The paper proposes a multimodal temporal point process framework that extends LLM-based TPPs to visual modality and positions text generation as a core capability alongside time and type prediction. It resides in the 'Vision-Language-Action Models for Long-Horizon Manipulation' leaf, which contains four papers total including the original work. This leaf sits within the broader 'Long-Horizon and Sequential Task Modeling' branch, indicating a moderately populated research direction focused on extended temporal reasoning. The taxonomy reveals this is an active but not overcrowded area, with sibling papers addressing robotic manipulation and planning tasks.

The paper's position connects it to several neighboring research directions. Adjacent leaves include 'Memory-Driven and Chain-of-Thought Long-Horizon Planning' (2 papers) and 'Autoregressive and Phase-Aware Long-Horizon Generation' (2 papers), suggesting the broader branch emphasizes extended temporal horizons through diverse mechanisms. The parent branch excludes short-horizon prediction and non-sequential applications, clarifying that this work's focus on long-range dependencies distinguishes it from standard temporal point process architectures. Nearby branches like 'Multimodal Fusion and Integration Strategies' (9 papers across 4 leaves) and 'Temporal Point Process Architectures and Mechanisms' (7 papers across 3 leaves) provide complementary perspectives on fusion techniques and core modeling approaches.

Among 28 candidates examined across three contributions, no clearly refuting prior work was identified. The MM-TPP framework examined 10 candidates with 0 refutable matches, suggesting limited direct overlap in the specific combination of LLM-based TPPs with visual modality and text generation. The adaptive compression mechanism based on temporal similarity also examined 10 candidates with no refutations, indicating this particular approach to addressing long-context challenges may be relatively unexplored. The TAXI-PRO benchmark examined 8 candidates with no refutations, though benchmark novelty depends heavily on domain-specific requirements not fully captured in semantic search.

Based on the limited search scope of 28 top-K semantic matches, the work appears to occupy a relatively sparse intersection of multimodal TPPs, LLM integration, and long-horizon modeling. The taxonomy structure confirms this sits at a junction between temporal point processes, multimodal fusion, and long-horizon reasoning—areas that individually are well-studied but whose combination remains less densely explored. The analysis cannot assess exhaustive novelty but suggests the specific technical approach and application context may offer meaningful differentiation from existing work within the examined candidate set.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

### Contribution 1: MM-TPP: Multimodal Temporal Point Process Framework

**Description**: The authors introduce MM-TPP, a unified framework that extends temporal point processes to handle multimodal data (visual, textual, and temporal information). Unlike prior work limited to text, MM-TPP jointly models and generates content across multiple modalities, positioning text generation as a core capability alongside traditional time and type prediction.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Does Multimodality Lead to Better Time Series Forecasting?

**URL**: View paper

**Brief Assessment**

Multimodality Time Series[70] focuses on time series forecasting with textual and visual inputs, not temporal point processes for event sequence modeling. The candidate addresses forecasting future values in continuous time series, while the original paper models discrete asynchronous event sequences with temporal point processes.

---

#### 2. Mm-forecast: A multimodal approach to temporal event forecasting with large language models

**URL**: View paper

**Brief Assessment**

Mm-forecast[75] focuses on temporal event forecasting using image function identification (highlighting/complementary) with MLLMs, not on extending temporal point processes to jointly model and generate multimodal content. The technical approaches differ fundamentally.

---

#### 3. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting

**URL**: View paper

**Brief Assessment**

Gpt4mts[69] focuses on multimodal time-series forecasting by combining numerical time series with textual summaries for prediction tasks. This differs fundamentally from MM-TPP's focus on modeling asynchronous event sequences with temporal point processes that jointly predict event timing, types, and generate multimodal content.

---

#### 4. DanmakuTPPBench: A Multi-modal Benchmark for Temporal Point Process Modeling and Understanding

**URL**: View paper

**Brief Assessment**

DanmakuTPPBench[64] focuses on benchmark construction and evaluation rather than proposing a unified multimodal TPP framework. It does not present a model architecture that jointly generates content across modalities or position text generation as a core capability alongside time/type prediction.

---

#### 5. Language-TPP: Integrating Temporal Point Processes with Language Models for Event Analysis

**URL**: View paper

**Brief Assessment**

Language-TPP[76] focuses on integrating temporal point processes with language models for textual event descriptions, not multimodal data with visual inputs. The candidate does not address visual modality or multimodal generation capabilities.

---

#### 6. Spatio-temporal Event Prediction via Deep Point Processes

**URL**: View paper

**Brief Assessment**

Spatio-temporal Event Prediction[77] focuses on spatio-temporal event prediction with spatial coordinates, not on multimodal data fusion with visual and textual inputs for content generation as in MM-TPP.

---

### 7. Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery

**URL**: View paper

**Brief Assessment**

Supply Chain Forecasting[72] addresses demand forecasting in supply chains using multimodal transformers for text, time series, and satellite imagery. This is fundamentally different from MM-TPP's focus on temporal point processes for asynchronous event sequences with visual and textual information, where the core task is modeling event timing, types, and generating contextual descriptions rather than forecasting demand patterns.

---

### 8. EventTSF: Event-Aware Non-Stationary Time Series Forecasting

**URL**: View paper

**Brief Assessment**

EventTSF[71] focuses on time series forecasting with textual events, not temporal point processes for event sequence modeling. The candidate addresses non-stationary forecasting with external events, while the original models asynchronous event sequences with multimodal generation capabilities.

---

### 9. Multi-modal news event detection with external knowledge

**URL**: View paper

**Brief Assessment**

News Event Detection[74] focuses on detecting discrete news events from social media posts using text-image pairs, not on modeling continuous-time event sequences with temporal point processes for generative multimodal content.

---

### 10. Spatio-temporal wildfire prediction using multi-modal data

**URL**: View paper

**Brief Assessment**

Wildfire Prediction Multimodal[73] focuses on wildfire risk prediction using spatio-temporal point processes with static/dynamic environmental marks (weather, vegetation), not on general multimodal event sequence modeling with visual and textual content generation as in MM-TPP.

---

## Contribution 2: Adaptive Compression Mechanism Based on Temporal Similarity

**Description**: The authors propose a novel sequence compression strategy that exploits temporal similarity between events. When consecutive events have similar inter-event intervals, they are compressed using special tokens, enabling the model to fit longer event histories within fixed context windows and capture long-range dependencies more effectively.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Spatio-temporal segmentation based adaptive compression of dynamic mesh sequences

**URL**: View paper

**Brief Assessment**

Mesh Compression Adaptive[58] focuses on spatio-temporal segmentation for 3D mesh animation compression, not event sequence compression using special tokens for similar inter-event intervals in temporal point processes.

---

### 2. Longer: Scaling up long sequence modeling in industrial recommenders

**URL**: View paper

**Brief Assessment**

Longer Recommenders[53] focuses on token merge strategies for user behavior sequences in recommender systems, not temporal point processes. The compression is based on grouping adjacent tokens in recommendation sequences, not exploiting temporal similarity between event inter-arrival intervals as in the original paper.

---

### 3. Compressive transformers for long-range sequence modelling

**URL**: View paper

**Brief Assessment**

Compressive Transformers[55] compresses past activations in transformer memory using fixed compression functions (pooling, convolution) applied uniformly to old memories, not based on temporal similarity between events. The original paper's contribution specifically exploits temporal patterns in event sequences (similar inter-event intervals) for adaptive compression.

---

### 4. Periodicity decoupling framework for long-term series forecasting

**URL**: View paper

**Brief Assessment**

Periodicity Decoupling[54] focuses on decomposing time series into periodic components for forecasting, not on compressing event sequences based on temporal similarity for long-range dependency modeling in multimodal event streams.

---

### 5. Nonrecurrent neural structure for long-term dependence

**URL**: View paper

**Brief Assessment**

Nonrecurrent Long-term[60] focuses on feedforward neural networks with memory blocks for sequence modeling in speech/language tasks, not on adaptive compression of event sequences based on temporal similarity patterns.

---

### 6. Multi-Domain Spatial-Temporal Redundancy Mining for Efficient Learned Video Compression

**URL**: View paper

**Brief Assessment**

Video Compression Redundancy[56] focuses on spatial-temporal redundancy in video compression using neural networks for encoding visual frames. The original paper proposes sequence compression for event streams based on inter-event temporal intervals to fit longer histories in language models—a fundamentally different application domain and technical approach.

---

### 7. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models

**URL**: View paper

**Brief Assessment**

DyCoke[52] focuses on video token compression for visual large language models, not event sequence modeling in temporal point processes. The temporal similarity concept applies to video frames, not inter-event intervals in TPP sequences.

---

### 8. Temporal patterns decomposition and Legendre projection for long-term time series forecasting.

**URL**: View paper

**Brief Assessment**

Temporal Patterns Decomposition[57] focuses on decomposing temporal patterns and using Legendre projection for long-term time series forecasting. This is fundamentally different from the original paper's adaptive sequence compression mechanism that exploits temporal similarity between consecutive events in multimodal event sequences to fit longer histories within fixed context windows.

---

### 9. Trajgat: A graph-based long-term dependency modeling approach for trajectory similarity computation

**URL**: View paper

**Brief Assessment**

Trajgat[59] focuses on trajectory similarity computation using graph-based methods for spatial sequences, not on compressing multimodal event sequences using temporal similarity between inter-event intervals for LLM context management.

---

### 10. Longvu: Spatiotemporal adaptive compression for long video-language understanding

**URL**: View paper

**Brief Assessment**

Longvu[51] focuses on video-language understanding using frame-level similarity for temporal reduction in visual sequences, not event sequence compression for temporal point processes. The technical domains and compression targets differ fundamentally.

---

## Contribution 3: TAXI-PRO: Multimodal TPP Benchmark Dataset

**Description**: The authors create TAXI-PRO, a new benchmark dataset that enriches the classic NYC Taxi data with multimodal content including map image patches and natural language descriptions. This dataset provides a complementary evaluation scenario with shorter sequences compared to existing benchmarks.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding

**URL**: View paper

**Brief Assessment**

CoCa-CXR[65] focuses on chest X-ray vision-language understanding with temporal progression analysis in medical imaging, not on temporal point process evaluation with multimodal event sequences.

---

### 2. Retrieval of Temporal Event Sequences from Textual Descriptions

**URL**: View paper

**Brief Assessment**

Temporal Event Retrieval[68] focuses on retrieval tasks with textual descriptions and introduces TESRBench for sequence retrieval evaluation, not multimodal TPP benchmarks with visual and textual data for event prediction tasks.

---

### 3. Lost in Time: A New Temporal Benchmark for VideoLLMs

**URL**: View paper

**Brief Assessment**

Lost in Time[67] focuses on video understanding benchmarks for VideoLLMs with temporal reasoning evaluation, not multimodal temporal point process datasets with event sequences, map images, and natural language descriptions for TPP modeling.

---

### 4. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing

**URL**: View paper

**Brief Assessment**

Biomedical Vision-Language[62] focuses on temporal chest X-ray imaging with medical reports for clinical progression tasks, not multimodal temporal point process evaluation with map images and taxi trajectory descriptions.

---

### 5. A survey on video temporal grounding with multimodal large language model

**URL**: View paper

**Brief Assessment**

Video Temporal Grounding[63] focuses on video-language understanding benchmarks for temporal grounding tasks (moment retrieval, dense captioning), not temporal point process evaluation with multimodal data for event sequence modeling.

---

### 6. DanmakuTPPBench: A Multi-modal Benchmark for Temporal Point Process Modeling and Understanding

**URL**: View paper

**Brief Assessment**

DanmakuTPPBench[64] introduces DanmakuTPP-Events and DanmakuTPP-QA datasets from Bilibili's danmaku system, not NYC Taxi data. The datasets serve different purposes and domains, with no overlap in construction methodology or data sources.

---

### 7. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering

**URL**: View paper

**Brief Assessment**

Mtbench[61] focuses on multimodal time-series benchmarks for financial and weather forecasting with news-driven QA tasks, not temporal point process (TPP) evaluation. The datasets serve fundamentally different modeling paradigms and evaluation objectives.

---

### 8. Localizing moments in video with temporal language

**URL**: View paper

**Brief Assessment**

Localizing Moments Video[66] focuses on temporal localization of moments in video using natural language queries, not on temporal point process modeling. The datasets are fundamentally different in purpose and structure.

## Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 5 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. DanmakuTPPBench: A Multi-modal Benchmark for Temporal Point Process Modeling and Understanding

**Detected in**: Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

### 2. Language-TPP: Integrating Temporal Point Processes with Language Models for Event Analysis

**Detected in**: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References


- [0] Long-range Modeling and Processing of Multimodal Event Sequences View paper
- [1] Transformers for mixed-type event sequences View paper
- [2] Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory View paper
- [3] Multimodal integration of physiological signals clinical data and medical imaging for ICU outcome prediction View paper
- [4] Domain-collaborative multimodal transformer for fault diagnosis of rotating machines under noisy environments View paper
- [5] Label-free long-horizon 3d uav trajectory prediction via motion-aligned rgb and event cues View paper
- [6] Multimodal fusion method with spatiotemporal sequences and relationship learning for valence-arousal estimation View paper
- [7] Multimodal Fall Detection Using Spatial-Temporal Attention and Bi-LSTM-Based Feature Fusion View paper
- [8] SkyNet: A deep learning architecture for intra-hour multimodal solar forecasting with ground-based sky images View paper
- [9] From Watch to Imagine: Steering Long-horizon Manipulation via Human Demonstration and Future Envisionment View paper
- [10] M3Rec: Selective State Space Models with Mixture-of-Modality Experts for Multi-Modal Sequential Recommendation View paper
- [11] Spatio-temporal point processes with attention for traffic congestion event modeling View paper
- [12] Longvie: Multimodal-guided controllable ultra-long video generation View paper
- [13] HGTFM: Hierarchical Gating-Driven Transformer Fusion Model for Robust Multimodal Sentiment Analysis View paper
- [14] UMI-Rec: A Unified Multi-modal Intent Fusion Framework with State-Space Models and Large Language Models for Recommendation View paper
- [15] Automatic depression recognition with an ensemble of multimodal spatio-temporal routing features View paper
- [16] BALM-TSF: Balanced Multimodal Alignment for LLM-Based Time Series Forecasting View paper
- [17] Memory-driven multimodal chain of thought for embodied long-horizon task planning View paper
- [18] Transformer Models for Multimodal Sequence Learning View paper
- [19] The Journey of Action Recognition View paper
- [20] A Case-based reasoning and explaining model for temporal point process View paper
- [21] MELON: Multimodal Mixture-of-Experts with Spectral-Temporal Fusion for Long-Term Mobility Estimation in Critical Care View paper
- [22] Toward long form audio-visual video understanding View paper
- [23] TFF-temporal fusion framework for advancing video retrieval through long-range dependencies and multi-modal intent View paper
- [24] Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions View paper
- [25] Joint multi-scale information and long-range dependence for video captioning View paper
- [26] MHSCNET: A Multimodal Hierarchical Shot-Aware Convolutional Network for Video Summarization View paper
- [27] Bilevel Relational Graph Representation Learning-based Multimodal Emotion Recognition in Conversation View paper
- [28] Capturing Spectral and Long-term Contextual Information for Speech Emotion Recognition Using Deep Learning Techniques View paper
- [29] Preference-Based Long-Horizon Robotic Stacking with Multimodal Large Language Models View paper
- [30] A multi-modal approach for mixed-frequency time series forecasting. View paper
- [31] Temporal multimodal learning in audiovisual speech recognition View paper
- [32] Learning time series associated event sequences with recurrent point process networks View paper
- [33] Multi-modal Generative Modeling of Event Sequences and Time Series for Solar PV Systems View paper
- [34] SpeechCARE: dynamic multimodal modeling for cognitive screening in diverse linguistic and speech task contexts View paper
- [35] Text-enhanced multi-granularity temporal graph learning for event prediction View paper
- [36] Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation View paper
- [37] MT-DPCQA: A Multimodal Time-aware Learning Approach for No-Reference Dynamic Point Cloud Quality Assessment View paper
- [38] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation View paper
- [39] Hybrid informerâ□□BiLSTM Model for Long-Term Forecasting of Urban Heat Islands: A Multimodal Approach Integrating Remote Sensing and Climate Data View paper
- [40] LongInsightBench: A Comprehensive Benchmark for Evaluating Omni-Modal Models on Human-Centric Long-Video Understanding View paper
- [41] Exploring complex dependencies for multi-modal semantic trajectory prediction View paper
- [42] Shock-Biased Attention: Enhancing Transformer Hawkes Processes with Amplitude-Driven Temporal Kernels View paper
- [43] Dual-Stage Attention-Augmented ACT Model: Optimization and Validation for Bimanual Fine Manipulation Tasks View paper
- [44] Robot Confirmation Generation and Action Planning Using Long-context Q-Former Integrated with Multimodal LLM View paper
- [45] AI-Driven Multimodal Prognosis in Stroke: Integrating Neuroimaging and Longitudinal Clinical Data for Enhanced Detection and Long-Term Outcome Prediction View paper

- [46] Attention-Based Multiscale Temporal Fusion Network for Uncertain-Mode Fault Diagnosis in Multimode Processes View paper
- [47] Scene-aware Graph-enhanced Multimodal Collaboration for Video-grounded Dialogue View paper
- [48] SPAN: Continuous Modeling of Suspicion Progression for Temporal Intention Localization View paper
- [49] Predicting text features of social temporal point process View paper
- [50] Edge-AI Enabled IoT Framework for Real-Time Stroke Risk Prediction Using Multimodal Biosignals View paper
- [51] Longvu: Spatiotemporal adaptive compression for long video-language understanding View paper
- [52] DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models View paper
- [53] Longer: Scaling up long sequence modeling in industrial recommenders View paper
- [54] Periodicity decoupling framework for long-term series forecasting View paper
- [55] Compressive transformers for long-range sequence modelling View paper
- [56] Multi-Domain Spatial-Temporal Redundancy Mining for Efficient Learned Video Compression View paper
- [57] Temporal patterns decomposition and Legendre projection for long-term time series forecasting. View paper
- [58] Spatio-temporal segmentation based adaptive compression of dynamic mesh sequences View paper
- [59] Trajgat: A graph-based long-term dependency modeling approach for trajectory similarity computation View paper
- [60] Nonrecurrent neural structure for long-term dependence View paper
- [61] Mtbench: A multimodal time series benchmark for temporal reasoning and question answering View paper
- [62] Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing View paper
- [63] A survey on video temporal grounding with multimodal large language model View paper
- [64] DanmakuTPPBench: A Multi-modal Benchmark for Temporal Point Process Modeling and Understanding View paper
- [65] CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding View paper
- [66] Localizing moments in video with temporal language View paper
- [67] Lost in Time: A New Temporal Benchmark for VideoLLMs View paper
- [68] Retrieval of Temporal Event Sequences from Textual Descriptions View paper
- [69] Gpt4mts: Prompt-based large language model for multimodal time-series forecasting View paper
- [70] Does Multimodality Lead to Better Time Series Forecasting? View paper
- [71] EventTSF: Event-Aware Non-Stationary Time Series Forecasting View paper
- [72] Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery View paper
- [73] Spatio-temporal wildfire prediction using multi-modal data View paper
- [74] Multi-modal news event detection with external knowledge View paper
- [75] Mm-forecast: A multimodal approach to temporal event forecasting with large language models View paper
- [76] Language-TPP: Integrating Temporal Point Processes with Language Models for Event Analysis View paper
- [77] Spatio-temporal Event Prediction via Deep Point Processes View paper