多模态事件序列的长程建模与处理

Jichu Li[1*], **Yilun Zhong**[1*], **Zhiting Li**[1], **Feng Zhou**[1, 2†], **Quyu Kong**[3†1]
中国人民大学应用统计中心与统计学院[2]北京未来区块链与隐私计算高级创新中心[3]独立研究者{lijichu52,kongquyu}@gmail.com,
feng.zhou@ruc.edu.cn

arXiv:2602.01125v1 [cs.CL] 1 Feb 2026

摘要

时间点过程（TPP）已成为建模异步事件序列的强大工具。尽管近期研究已将 TPP 扩展至处理文本信息，但现有方法在生成丰富的多模态内容以及推理事件动态方面仍存在局限。一个关键挑战在于，融入多模态数据会大幅增加序列长度，从而妨碍基于注意力的模型生成连贯的长篇文本描述，而这类描述需要长程理解。本文提出一种新颖框架，将基于大语言模型的时间点过程扩展至视觉模态，并将文本生成确立为与时间和类型预测并列的核心能力。我们通过一种基于时间相似性的自适应序列压缩机制解决长上下文问题，在保留关键模式的同时缩短序列长度。我们采用两阶段范式：先在压缩序列上进行预训练，再针对下游任务进行监督微调。大量实验（包括在具有挑战性的 DanmakuTPP-QA 基准上的实验）表明，我们的方法在预测准确率和生成文本分析的质量方面均优于最先进的基线方法。代码已公开发布于 https://github.com/JichuLi/MM-TPP。

1 引言

时间点过程（TPP）（Daley & Vere-Jones, 2003；2007）是用于分析连续时间中异步事件序列的统计模型。它们已广泛应用于多个领域，包括神经科学（Zhou et al., 2021）、金融（Bacry et al., 2015）、犯罪分析（Mohler, 2013）和流行病学（Rizoiu et al., 2018）。传统统计时间点过程虽然能够对事件时间及其依赖关系进行建模，但往往缺乏捕捉复杂现实模式所需的表达能力。近年来，深度学习的发展推动了深度时间点过程的出现（Du et al., 2016；Mei & Eisner, 2017；Zuo et al., 2020；Meng et al., 2024），其建模能力表现更为出色。在大语言模型（LLM）取得成功的基础上，Language-TPP（Kong et al., 2025）进一步扩展了时间点过程，使其能够纳入文本信息，展现了将自然语言处理与时间建模相结合的潜力。

然而，现实世界事件序列正日益呈现多模态特征，不再局限于时间与文本信息。例如，在视频流媒体平台上，用户评论流（称为弹幕）不仅包含时间戳和文本内容，还包含视频中的相关视觉帧（Jiang 等，2025）。同样，交通监控系统会捕获事故报告，其中包括时间信息、文本描述、音频录音以及来自监控摄像头的视觉证据。当前的时序点过程模型，包括 Language-TPP 在内，仍主要局限于有限的模态，缺乏有效建模和生成丰富多模态内容的能力，例如同时生成图像、文本和时间信息。

[*]贡献相同。
[†]通讯作者。

将深度时间点过程扩展为多模态场景下的完全生成式模型，面临两项主要挑战。**首先，现有时序点过程模型缺乏用于多模态生成的统一框架**。尽管 Language-TPP 在融入文本方面取得了进展，但它无法处理视觉内容，也无法生成以视觉内容为条件的文本。不同模态需要采用不同的编码策略（Lahat 等人，2015），而真正的生成式框架不仅必须对这些异构数据进行编码，还必须生成具有上下文意识的内容——例如，为未来事件生成能够正确引用其相关视觉信息的文本描述。为此，我们提出了一个统一框架，将 Language-TPP 基于模板的方法扩展到视觉信息，从而实现全面的多模态事件建模与生成。

其次，融入多模态数据会显著增加序列长度，从而妨碍生成依赖长程依赖的连贯长篇文本。这种膨胀是因为非文本模态本身会被分词为序列；例如，单张图像通常会被分解为数百个基于图像块的词元（Dosovitskiy 等人，2021）。在每个点都与此类数据相关联的事件序列中，总长度 N 可能会变得异常庞大。这使得 Transformer（Vaswani，2017）的自注意力机制的 $\mathcal{O}(N^2)$复杂度成为关键瓶颈，阻碍模型访问生成完整分析报告或回答摘要问题所需的完整历史记录。为缓解这一问题，我们提出了一种利用时间相似性的自适应压缩机制，在保留有效长程文本生成所需动态特征的同时，减少序列长度。

总之，为应对上述挑战，我们提出了 **多模态时间点过程（MM-TPP）**，这是一个通过以下贡献解决这些问题的统一框架：

(1) 我们提出 MM-TPP，它将 Language-TPP 扩展为一个多模态框架，可在视觉、文本和时间输入的条件下预测事件时间和类型，同时生成丰富的文本。

(2) 我们提出一种基于时间相似性的自适应压缩机制，以减少 s seq序列长度，从而使长篇、上下文感知文本的生成在计算上切实可行。 .

(3) 我们构建了 TAXI-PRO，这是一个新的多模态 TPP 数据集，通过为经典的纽约市出租车数据集添加视觉地图图块和文本描述而构成，为研究提供了一个有价值的基准。

(4) 在两个多模态时间点过程数据集（包括 DanmakuTPP-QA 基准）上进行的全面实验表明，与最先进的基线模型相比，MM-TPP 的预测准确率更高。此外，定性分析显示，我们的模型能够生成更具洞察力且连贯性显著更强的文本分析，彰显了其卓越的推理能力。

2 相关工作

深度时间点过程。基于深度学习的时间点过程最早由Du等人（2016）采用基于RNN的框架提出。后续工作改进了该方法，包括Mei和Eisner（2017）；Yang等人（2018）；Omi等人（2019）；Soen等人（2021）；Chen等人（2024b）。循环模型能够实现高效推断，具有常数时间预测和线性训练代价，但并行性有限，且对长程依赖的建模能力较差。为解决这些问题，基于Transformer的时间点过程应运而生，始于Zuo等人（2020）和Zhang等人（2020），随后由Zhu等人（2021）；Zhou等人（2022）；Yang等人（2022）；Meng等人（2024）；Panos（2024）等工作进一步扩展。这些模型支持并行训练，并通过自注意力更好地捕捉长程依赖，但需要更高的计算成本——时间和内存均按 $\mathcal{O}(N^2)$增长。与这些方法不同，MM-TPP不仅能高效处理长程依赖，还将TPP 建模扩展到生成并推理丰富的多模态内容。

协变量时间点过程。协变量时间点过程通过融入上下文信息，增强传统模型，从而更好地解释事件动态并提升预测效果。早期研究主要关注结构化协变量，例如，Meyer等人（2012）将人口统计数据用于流行病建模，Xiao等人（2017）提出了结合协变量和事件历史的双LSTM，而Adelfio和Chiodi（2021）；Gajardo和M¨uller（2021）则将类似方法应用于地震和COVID-19数据。实际上，许多协变量是非结构化的，例如文本、图像或音频，因此难以将其整合到TPP中。一些研究通过使用词频特征（Zhu和Xie，2022）或BERT嵌入（Zhang等人，2023）来应对这一问题。近期研究则利用大语言模型联合建模时间与文本信息，例如Liu和Quan（2024）；Kong等人（2025）。MM-TPP在此基础上

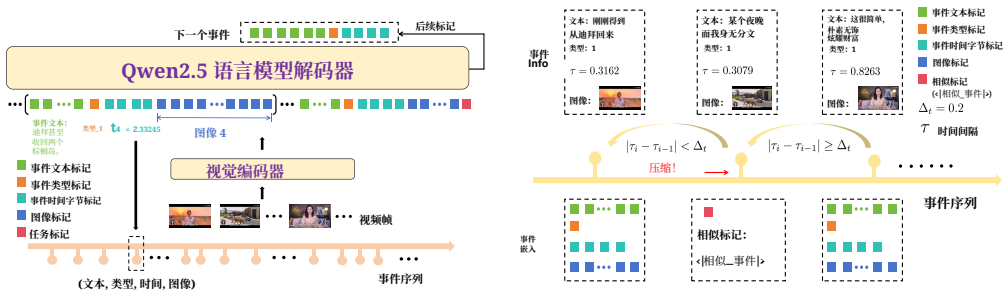


图 1：MM-TPP 框架概览。(a) 模型架构：该框架基于 Qwen2.5-VL 构建，其中由时间、类型、文本和图像信息组成的多模态事件被分词为统一的标记序列。视觉内容由视觉编码器编码，并通过语言模型解码器与其他模态融合，以进行自回归式的下一个事件预测。(b) 自适应长序列压缩：时间间隔相似的事件（$|\tau_i - \tau_{i-1}| < \Delta$）被压缩为一个特殊标记。该策略能够高效处理长事件序列，同时保留时间结构和关键信息。

进一步将多种类型的非结构化协变量——具体而言，图像和文本——纳入统一的TPP框架，并进一步支持对未来事件进行多模态生成。

3 预备知识

本节介绍我们提出的模型所需的基础概念和背景知识，包括 TPP 和基于大语言模型的时间点过程建模。

3.1 时间点过程

TPP 是用于描述在连续时间内发生的离散事件序列的概率模型。一个典型的 TPP 实现记为 $\mathcal{S} = \{(t_i,\ e_i)\}_{i=1}^{N}$，其中每个 $(t_i,\ e_i)$ 表示一个类型为 $e_i \in \mathcal{E} = \{1, \ldots, E\}$、发生时间为 $t_i \in$ 的事件（$0,\ T$]），满足 $0 < t_1 < \cdots < t_N \leq T$。这里，$N$ 表示在时间窗口（$0,\ T$]内观测到的事件随机数量。刻画 TPP 有多种等价的表述方式。一种常见方法是对给定过去信息时未来事件的条件分布进行建模。令 $\mathcal{H}_t$ 表示截至时间 t（含该时刻）的事件历史。我们将 $p(t_{i+1},\ e_{i+1}|\mathcal{H}_{t_i})$ 定义为在给定过去历史 $\mathcal{H}_{t_i}$ 时，下一个事件在时间 t_{i+1} 以类型 e_{i+1} 发生的条件密度函数。根据这种表述，观测到序列 $\mathcal{S}$ 的概率可以分解如下：

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

其中，$1 - P(T \mid \mathcal{H}_{t_N})$ 表示在 $(t_N, T]$ 内没有事件发生的概率，$P(T \mid \mathcal{H}_{t_N}) = \int_{t_N}^{T} p($ $t \mid \mathcal{H}_{t_N})\,\mathrm{d}t$ 是 t_N 之后下一个事件的事件时间累积密度，而 $p(t \mid \mathcal{H}_{t_i}) = \sum_{e=1}^{E} p(t, e \mid \mathcal{H}_{t_i})$。通常将条件联合密度分解为：

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

因此，根据上述公式，可以推导出 TPP 的对数似然如下：

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

3.2 LANGUAGE-TPP

大语言模型的最新进展推动了 TPP 建模的新方法。早期研究通过定制分词和微调，将预训练大语言模型应用于时间事件数据（Liu &

Quan，2024；Xue等人，2023；Shi等人，2024）。沿着这一方向，Kong等人（2025）提出了 Language-TPP，这是一种用于建模以下形式的事件序列的框架（t_i，e_i，m_i）$_{i=1}^{N}$，其中每个事件都包含时间戳、类型和文本描述。该模型捕捉序列中的依赖关系，并联合预测未来的事件时间、类型和描述。

为实现这一目标，Language-TPP 采用仅解码器因果 Transformer，并引入了多个组件。每个事件都通过标准语言分词器对 e_i 和 m_i进行分词，而 t_i 则采用字节级策略进行编码，并使用 256 个特殊标记，使每个 32 位时间戳能够表示为四个词元。生成的词元被组织成结构化事件模板，其中每个事件都由 <|事件开始|> 和 <|事件结束|> 词元包围，每个组件前则添加一个模态专用标记（例如，时间戳前使用 <|时间前缀|>）。随后，模型执行自回归生成，直至生成序列结束词元；通过该词元，可以解码未来事件的时间戳、类型和描述。Language-TPP 的训练分为两个阶段。首先，将事件序列转换为标记序列，并使用标准的下一个词元预测对骨干大语言模型进行预训练。在第二阶段，针对每个子任务（时间预测、类型分类和描述生成）构建提示-响应对，并使用下一标记损失对响应部分进行微调。

尽管 Language-TPP 能够有效地将文本信息融入 TPP 建模，但它仅支持单一模态，在处理较长的多模态序列时还面临可扩展性问题。我们提出的 MM-TPP 通过纳入视觉和文本两种模态扩展了这一框架，并引入基于时间相似性的压缩机制，以应对多模态场景中的序列长度激增。

4 方法

为在长上下文约束下对包含丰富视觉和文本信息的多模态事件序列进行建模，我们提出了 MM-TPP，这是 Language-TPP 的扩展。它联合建模时间、类别、文本和视觉数据，并引入压缩机制，以增强长程依赖处理能力。

4.1 模型框架

图 1a 展示了 MM-TPP 框架的概览。我们的模型基于 Qwen2.5-VL（Bai 等，2025）构建，这是一种最先进的预训练多模态大语言模型。整个框架遵循序列到序列范式，以多模态事件历史序列作为输入，并以自回归方式预测未来事件的各种属性。按照 Kong等人（2025）的标注方式进行扩展，给定一个多模态事件序列 (t_i，e_i，m_i，v_i)$_{i=1}^{N}$，其中 t_i 表示事件时间戳，e_i 表示事件类型，m_i 是文本描述，v_i 是相关的视觉描述，目标是对该序列中的依赖关系建模，并预测未来事件的时间、类型和文本内容。

4.2 多模态事件序列标记化

为了使用语言模型处理多模态事件数据，我们设计了一种标记化方案，将每个事件元组（t_i，e_i，m_i，v_i）转换为适用于自回归建模的统一标记序列。时间信息的编码遵循 Language-TPP 的设计：根据内存布局，将每个 32 位时间间隔分解为四个字节。这些字节映射到 256 个特殊标记，即 <|byte 0|> 至 <|byte 255|>，从而能够紧凑且精确地表示时间间隔。类别信息（例如事件类型）则使用专门的特殊标记处理（例如，用于六种事件类别的 <|type 0|>至 <|type 5|>），以便统一且可解释地建模离散标签。

对于文本信息，事件描述（例如用户评论）会使用 Qwen2.5-VL 的内置分词器进行分词，将原始文本转换为模型可直接处理的标准语言令牌。视觉信息的集成方式则有所不同：我们不会将图像像素转换为词元，而是在序列的适当位置插入占位标记 <|image pad|>。在模型执行过程中，相应的图像会通过 Qwen2.5-VL 的视觉编码器处理，生成视觉嵌入，然后与占位标记对齐。

这种设计能够深度融合视觉、文本和时间特征，同时保持标记序列紧凑并兼容语言模型。

为了组织这些多模态组件，我们定义了一种结构化事件模板。每个事件都由 <|start of event|> 和 <|end of event|> 词元括起来。在事件内部，每种模态都由一个前缀词元标记： <|time start|> 表示时间， <|type start|> 表示类型，<|text start|>表示文本， <|vision start|>表示图像。这种结构化模板为复杂多模态事件的编码提供了一种一致且易于理解的格式。未压缩事件格式的完整示例见 F 节。

4.3 长序列自适应压缩

在多模态事件序列中加入文本和图像词元会显著增加输入长度，由于计算成本高且上下文窗口有限，这给基于 Transformer 的大语言模型带来了挑战。为缓解这一问题，我们提出一种基于事件间时间相似性的自适应压缩策略。在许多现实场景中（例如视频评论流），事件会突发式到达或遵循周期性节律，因此事件间隔相似。MM-TPP 利用这一观察结果，对时间上相似的事件序列进行压缩，如图 1b 所示。

在构造输入时，我们定义一个时间相似度阈值 Δ，并将当前事件的时间间隔 $\tau_i = t_i - t_{i-1}$ 与前一个事件的时间间隔 τ_{i-1}进行比较。如果两者之差小于该阈值，即 $|\tau_i - \tau_{i-1}| < \Delta$，则认为事件 i 与事件 $i-1$在时间上相似。在这种情况下，我们不再使用完整的多模态模板对其进行编码，而是将其替换为单个特殊标记 <|相似事件|>，从而减少词元数量。第 A.2 节讨论了如何选择合适的 Δ 值。反之，如果时间间隔差异显著，则使用第 4.2 节定义的多模态模板完整表示事件 i。有关将此序列压缩机制应用于连续事件的具体分步示例，请参阅第 A.1 节。

该策略仅用少数几个 <|相似事件|> 词元，就能紧凑地表示密集事件簇（这些事件簇可能需要数百个词元），同时保留具有独特时间模式的关键事件。因此，MM-TPP 可以更有效地利用固定的上下文窗口（例如，4096 个词元），从而对更长的事件历史进行建模，并以更高的效率捕捉长程依赖。

值得注意的是，我们提出的序列级压缩（事件间压缩）与高效多模态大语言模型中常用的标准表示级缩减技术（事件内压缩）有所不同，例如词元剪枝或合并（Bolya et al., 2023; Chen et al., 2024a; Dhouib et al., 2025; Alvar et al., 2025）。这些方法通过利用单个模态内部的空间冗余（如视觉图块），有效降低了计算成本，但不太适用于时间点过程的结构化组成部分。与图像不同，时间点过程的关键组成部分，如时间戳（t_i）、事件类型（e_i）和文本描述（m_i），都包含丰富的语义信息。每个时间值和词语都精确且不可或缺，关系到对事件序列的理解，因此删除它们可能会导致关键信息丢失。因此，我们优先采用序列级策略，对时间上相似的事件进行压缩。这种方法能够有效捕捉时间点过程的突发性，使模型在有限的上下文窗口内处理更长的历史序列。我们将更先进、能够适应更复杂多样场景的混合压缩策略设计留待未来工作。

4.4 训练与推断

模型训练。 为了使模型能够理解复杂的多模态事件结构和所提出的压缩策略，我们采用受 Language-TPP 启发的两阶段训练框架。第一阶段，使用根据定义的多模态事件模板构建的大规模标记序列进行持续预训练，其中可能包含完整表示和压缩表示。此阶段有助于模型适应新的事件格式，并学习特殊标记的语义，例如 <|相似事件|> 和 <|类型 X|>。训练目标采用标准的下一个词元预测：

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

其中， $x_1, \ldots, x_L$ 表示事件令牌序列。

第二阶段针对下游任务进行监督微调，包括时间预测、类型分类和文本生成。训练样本被整理为提示-响应对，其中提示由一个压缩历史片段和一个任务专用指令词元组成，例如 <|时间预测|>、 <|类型预测|>或自然语言查询；响应则包含真实标注令牌，例如表示时间的字节令牌、类型令牌或文本序列。训练损失仅作用于响应令牌：

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

其中 $r_1, \ldots, r_R$ 表示响应序列。这一两阶段过程使模型能够学习通用的序列建模能力，以及针对多模态事件数据进行任务专用推理的能力。

模型推理。 推断期间，模型采用与 Language-TPP 类似的自回归解码策略。给定历史序列和任务专用指令，模型会依次生成输出令牌。字节令牌会被解码为浮点时间间隔，用于时间预测；类别令牌表示事件类型；语言令牌则被解释为文本，用于文本生成或问答任务。由于 Qwen2.5-VL（Bai 等，2025）不支持图像生成，当前模型专注于时间、类型和文本输出。未来的扩展可以借助 Chameleon（Team，2025）等全模态模型，为预测事件生成图像。

5 实验

我们进行了全面的实验，从多个维度评估所提出的 MM-TPP 的有效性，包括与最先进基线方法的整体性能对比、文本生成质量、自适应压缩机制的影响，以及多模态信息整合的贡献。

5.1 实验设置

我们的 MM-TPP 模型基于 Qwen2.5-VL-3B 架构构建，并使用低秩适配（LoRA）（Hu et al., 2021）进行微调。所有实验均在单块 NVIDIA RTX 4090 GPU 上进行。对于阶段 1（持续预训练），我们以 1×10^{-4} 的学习率训练模型 5 个 epoch。对于阶段 2（监督微调），我们同样使用 1×10^{-4} 的学习率训练 3 个 epoch。压缩机制的时间相似度阈值设为 $\Delta = 0.2$；我们发现这一设置在压缩率与性能之间实现了良好的平衡。推断时，我们将解码温度设为 0.05，并使用不同的随机种子重复进行每项评估 5 次，以确保结果稳健。完整的超参数详情见 C 节。

基线方法。 我们将 MM-TPP 与多个最先进的 TPP 模型进行比较，包括 **NHP** (Mei & Eisner, 2017)、**SAHP**(Zhang et al., 2020)、**THP** (Zuo et al., 2020)、**RMTPP**(Du et al., 2016)、**AttNHP** (Mei et al., 2022)、**TPP-LLM**(Liu & Quan, 2024)、**S2P2** (Chang et al., 2025)和 **Language-TPP** (Kong et al., 2025)。除 TPP-LLM 和 Language-TPP 外，所有模型均使用 Easy-TPP 框架（Xue et al., 2024）进行训练和评估；TPP-LLM 和 Language-TPP 则使用其官方实现复现。为确保公平比较，我们对所有基线方法采用统一的最大训练轮数设置，选取表现最佳的检查点进行测试，并使用不同的随机种子重复完整的训练与评估流程 5 次。

数据集。 现有的大多数 TPP 基准仅限于时间戳和事件类型，极少有基准同时纳入文本和视觉数据。为弥补这一空白，我们在两个不同的多模态数据集上评估 MM-TPP：

• **DanmakuTPP**（Jiang 等，2025）。这是一个成熟的数据集，包含超长序列——大多数序列包含 1,000 多个事件，部分序列甚至超过 9,600 个。每个事件都包括时间戳、文本内容（例如视频评论）以及从相应视频帧中提取的视觉信息。此外，该数据集还提供 **DanmakuTPP-QA** 子集，用于评估复杂的多模态推理能力。

表 1：多模态 TPP 数据集上的事件时间和类型预测性能对比。结果以均方根误差(↓) / 准确率(↑)及标准差表示。最佳结果以**粗体**显示。

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

• **TAXI-PRO**。这是我们通过在经典的纽约市出租车数据集（Whong，2014）基础上添加丰富的多模态内容而构建的新数据集。为构建互补的评估场景，我们为每个事件添加了以各个 GPS 坐标为中心的 224×224 地图图像块，以及包含地标、坐标和行程元数据的自然语言描述（详见 B 节）。与 DanmakuTPP 相比，TAXI-PRO 中的序列明显更短（平均长度为 36），从而提供了一个互补的评估场景，在该场景中无需使用我们提出的压缩机制。

指标。 我们使用多项任务特定指标评估模型性能。对于时间戳预测，我们报告均方根误差（**RMSE**）；对于事件类型分类，我们使用准确率（**ACC**）；对于文本生成，我们使用困惑度（**PPL**）衡量令牌级质量。对于 DanmakuTPP-QA 中的问答任务，我们遵循原始评估协议——对封闭式问题使用 **ACC** 和**RMSE**，并对开放式回答进行**定性分析** 。

5.2 总体性能比较

我们首先在完整数据集上将 MM-TPP 与所有基线方法进行比较。两个数据集均按照 8:1:1 的比例划分为训练集、验证集和测试集。对于 DanmakuTPP，我们在训练期间使用 4096 个词元的最大词元数限制约束编码事件的数量，并在每个测试序列的前 200 个事件上进行评估。对于 TAXI-PRO，4096 个词元的限制足以编码完整序列，无需截断。如表 1 所示，MM-TPP 在两个数据集上均取得了最先进的性能。在 DanmakuTPP 上，它在时间预测任务中取得了最低的均方根误差（**5.2987**），其类型准确率（27.62%）也极具竞争力，并显著优于 Language-TPP。在 TAXI-PRO 上，这些优势更加明显：MM-TPP 在两项指标上均优于所有基线方法，以领先的均方根误差 **0.3310**和准确率**77.56%**建立了新的基准。

5.3 多模态问答性能

我们的框架的一项关键优势在于能够对多模态事件流进行复杂推理。为评估这一能力，我们采用了**DanmakuTPP-QA** 基准（Jiang 等，2025），该基准专为评估时间、视觉和文本推理而设计。该基准包含 8 项用于定量评估的封闭式任务，以及 2 项需要生成描述性分析报告的开放式任务，并通过定性方式进行评估。E 节详细介绍了全部 10 项任务。

封闭式问答任务。 我们使用 DanmakuTPPBench（Jiang 等，2025）报告的基线结果，将 MM-TPP 与最先进的大语言模型和多模态大语言模型在八项封闭式任务上的表现进行比较。如表 2 所示，MM-TPP 在整个基准测试中均展现出极具竞争力的性能，并在多项任务上取得领先结果。值得注意的是，尽管采用了相同的底层基础模型，MM-TPP 仍始终优于原始基准中的微调后的 Qwen2.5-VL-3B。这一性能

表 2：DanmakuTPP-QA 封闭式任务的对比评估。指标为准确率（↑）和均方根误差（↓）。所提出的 MM-TPP 模型以醒目方式标示。最佳结果以 **粗体**显示，次佳结果以下划线标示。

模型 / 任务 指标	T-1 准确率 ↑	T-2 均方根误差 ↓	T-3 均方根误差 ↓	T-4 均方根误差 ↓	T-5 均方根误差 ↓	T-6 均方根误差 ↓	T-7 准确率 ↑	T-8 准确率 ↑
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	19.33	41.50
Qwen3-30B-A3B	0.67	1.33	121.96	0.20	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
多模态大语言模型								
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	1.14	121.25	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
微调								
Qwen2.5-VL-3B	27.0	1.35	220.43	**0.05**	0.16	0.08	15.33	43.00
MM-TPP	**27.33**	1.49	190.45	**0.05**	**0.10**	**0.07**	15.67	44.00

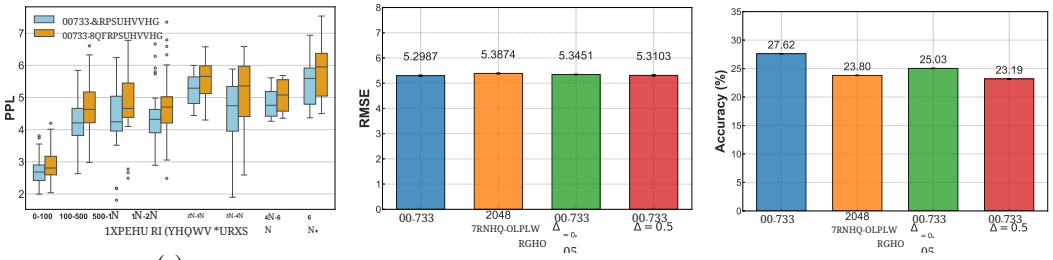
差距凸显了我们框架的优势：该框架能够显式捕捉多模态事件序列的结构化特性，而非常规的微调方法则无法做到这一点。

开放式问答与报告生成。我们评估了 MM-TPP 生成高质量分析性文本的能力——这是其区别于纯预测模型的一项核心能力。在 DanmakuTPP-QA 基准的两项开放式任务中，与现有多模态大语言模型相比，MM-TPP 始终能够生成洞见更深刻、连贯性更强的文本。为进一步评估其泛化能力，我们提示 MM-TPP 在 TAXI-PRO 数据集上生成描述性分析报告；该数据集没有预先定义的问答任务，也未用于针对生成任务微调模型。值得注意的是，对于 TAXI-PRO 的时间数据，MM-TPP 生成的叙述文本质量也超过了领先的多模态大语言模型；后者往往仅限于对事件序列进行基本的描述性分析。相比之下，MM-TPP 能够揭示高层次的时空模式，展现出将多模态事件流综合为连贯且富有洞见的叙述文本的强大能力，即使在零样本设置下亦是如此。E 节提供了 MM-TPP 与基线多模态大语言模型在两个数据集上的详细定性比较。

5.4 压缩策略的有效性

为评估自适应压缩策略的有效性，我们在 DanmakuTPP 数据集上比较了 MM-TPP 使用和不使用该机制时的性能（请注意，由于 TAXI-PRO 的序列较短，因此未对其进行压缩）。压缩的主要优势在于，它能够在固定的上下文窗口内容纳更长的事件历史。在 4096 个词元的限制下，未压缩模型平均可容纳 113 个事件，而压缩模型平均可容纳 292 个事件，最多可达 2008 个。这种上下文扩展使模型能够更好地捕捉长程依赖，从而显著提升性能。因此，在预测任务中，压缩模型优于未压缩模型，均方根误差更低（5.2987 对比 5.5551），准确率更高（27.62% 对比 25.87%）。为进一步验证压缩策略的有效性，我们将 MM-TPP 与一种随机丢弃事件的朴素基线进行了比较。如附录D所述，随机丢弃策略会导致性能显著下降，这证实了保留时间因果性与扩展上下文同等重要。

此外，我们从语言建模的角度进一步开展实验，以评估压缩方法的有效性。具体而言，我们在测试集上未经截断的编码事件序列上测量**困惑度**，这是评估



(a) (b) (c)

图 2：在 DanmakuTPP 数据集上评估压缩机制及关键超参数。(a) 测试集上的困惑度对比。我们的压缩模型始终比未压缩版本保持更低的困惑度，表明其具有有效的长程建模能力。(b、c) 上下文长度和相似度阈值（Δ）对均方根误差和准确率的消融结果，显示较大的上下文和 $\Delta = 0.2$ 能实现最佳性能。

表 3：MM-TPP 的消融研究结果。该表评估了视觉信息的贡献（与纯文本变体相比）以及模型大小的影响（与 70亿模型相比）。指标为均方根误差（↓）和准确率（↑）。最佳结果以**粗体**显示。

模型	DanmakuTPP		TAXI-PRO	
	均方根误差(↓)	准确率(%)(↑)	均方根误差(↓)	准确率(%)(↑)
MM-TPP	5.2987 ±0.018	**27.62** ±0.02	**0.3310** ±0.001	**77.56** ±0.04
Language-TPP	5.3845 ±0.022	22.62 ±0.07	0.3376 ±0.001	75.27 ±0.08
MM-TPP（仅文本）	5.4654 ±0.008	23.64 ±0.06	0.3388 ±0.001	76.70 ±0.08
MM-TPP（70亿）	**5.0533** ±0.002	26。98 ±0.01	0.3337 ±0.001	76.16 ±0.05

语言模型质量（Zhao 等，2023）。如图 2a 所示，困惑度会随着序列长度增加而上升，这是序列建模中众所周知的挑战。值得注意的是，我们的压缩模型在所有序列长度下都始终实现了更低的困惑度，而且随着序列变长，差距也随之扩大。这些结果凸显了我们方法的一项关键优势：通过在固定词元预算内保留更多长程上下文，压缩机制缓解了长而复杂的事件流中的性能下降。

5.5 消融研究

本节通过消融研究，考察 MM-TPP 中四项关键设计选择的影响：(i) 融入视觉信息，(ii) 模型规模，(iii) 上下文窗口大小（通过最大词元数限制控制），以及 (iv) 压缩机制中使用的时间相似度阈值。

视觉信息的影响。为评估视觉输入的贡献，我们将完整的 MM-TPP 模型（文本 + 图像）与两个纯文本变体进行比较：MM-TPP（仅文本）和 Language-TPP。我们在受控条件下评估所有模型，并固定上下文窗口中的事件数量，以确保序列长度不会混淆结果。如表 3 所示，MM-TPP 在各个数据集上始终优于两种基线方法，表明视觉特征提供了额外的预测价值。整合图像提升了时间戳和类型预测的准确率，凸显了视觉上下文与文本信息相辅相成的作用。

模型大小的影响。为评估模型容量的作用，我们将基础模型通义千问2.5-VL-3B替换为更大的通义千问Qwen2.5-VL-7B，并采用相同的训练流程和超参数，在DanmaluTPP和TAXI-PRO数据集上进行评估。本实验旨在考察扩大模型规模能否提升事件预测性能。表 3中的结果显示，扩展至70亿参数后，结果喜忧参半。尽管更大的模型在复杂的DanmakuTPP数据集上显著提升了时间预测（均方根误差）表现，但原始的30亿参数模型在其他所有指标上均优于它。这

表明，较大的模型容量最有利于捕捉长程时间模式，但可能会损害模型在较简单任务上的表现，这可能是过拟合所致。

上下文长度的影响。为考察上下文窗口大小的影响，我们在 DanmakuTPP 数据集上进行训练和推断时，将最大令牌上限从 4096 降至 2048。如图 2b 和图 2c 所示，这一调整导致性能明显下降。较短的上下文窗口限制了模型访问历史事件的能力，妨碍其捕捉对于准确时间建模至关重要的长距离依赖关系。这些结果凸显了在处理长且复杂的事件序列时，维持足够大的上下文窗口的重要性。

时间相似度阈值的影响。我们评估了压缩机制的性能如何随时间相似度阈值 Δ 而变化。具体而言，根据第 A.2 节的设定，我们比较了默认设置 $\Delta = 0.2$、更严格的阈值（$\Delta = 0.05$）和更宽松的阈值（$\Delta = 0.5$）。如图 2b 和图 2c 所示，这两种替代设置都会导致性能下降。较小的阈值（$\Delta = 0.05$）带来的压缩有限，几乎无法提升上下文覆盖范围。相反，较大的阈值（$\Delta = 0.5$）会导致过度压缩，将不相似的事件合并并丢弃重要的时间信号。这些结果证实，$\Delta = 0.2$ 在压缩效率和时间保真度之间取得了良好平衡。

6 结论

本文提出了 MM-TPP，这是一种新颖的框架，将 TPP 建模从纯预测范式推进到了生成式范式。通过在时间、类别和文本数据之外融入视觉信息，我们的模型不仅能够预测未来事件，还能生成丰富且具有描述性的文本，以分析和解释事件动态。针对多模态序列中的长上下文建模这一关键挑战，我们采用了利用时间相似性的自适应压缩策略。大量实验表明，MM-TPP 在标准预测任务中始终优于最先进基线；更重要的是，在具有挑战性的 DanmakuTPP-QA 基准上，它展现出卓越的文本生成和推理能力。尽管我们的框架目前尚不具备图像生成能力，但它为视觉增强时间建模开辟了新方向，并为生成式多模态 TPP 的未来研究奠定了坚实基础。

致谢

本研究得到国家自然科学基金项目（编号：62576346）、教育部人文社会科学重点研究基地项目（22JJD110001）、中央高校基本科研业务费、中国人民大学科研经费（24XNKJ13）以及北京未来区块链与隐私计算高级创新中心的资助。

伦理声明

我们的工作遵循 ICLR 道德准则。所使用的数据集要么是公开基准（DanmakuTPP），要么源自公开数据（纽约市出租车与豪华轿车委员会）。对于我们构建的 TAXI-PRO 数据集，我们确保其中不包含任何个人身份信息，仅关注聚合的时空模式。我们承认预训练语言模型骨干网络可能带来的偏见。我们的框架旨在用于分析事件动态的研究，我们预计其使用不会造成直接的负面社会影响。

可复现性声明

我们已竭尽全力确保研究的可复现性。我们的 MM-TPP 模型的完整源代码，包括训练和评估流程，将在补充材料中提供。预训练和监督微调阶段所需的所有关键超参数均记录在 C 节中。此外，实验设置、基线方法和评估指标的详细信息载于第 5 节，以便公平比较并复现我们的结果。

参考文献

Giada Adelfio 和 Marcello Chiodi。将协变量纳入时空点过程及其在地震活动性中的应用。*Statistical Methods & Applications*，30(3):947–971，2021。

Saeed Ranjbar Alvar、Gursimran Singh、Mohammad Akbari 和 Yong Zhang。Divprune：面向大型多模态模型的多样性视觉词元剪枝。载于计算机视觉与模式识别会议论文集，第 9392–9401 页，2025 年。

Emmanuel Bacry、Iacopo Mastromatteo 和 Jean-François Muzy。《金融中的霍克斯过程》。*Market Microstructure and Liquidity*，1(01):1550005，2015 年。

Shuai Bai、KeqinChen、XuejingLiu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025 年。

Daniel Bolya、Cheng-Yang Fu、Xiaoliang Dai、Peizhao Zhang、Christoph Feichtenhofer 和 Judy Hoffman。标记合并：让你的视觉Transformer运行更快。载于国际学习表征会议，2023 年。

Yuxin Chang、Alex Boyd、Cao Xiao、Taha Kass-Hout、Parminder Bhatia、Padhraic Smyth 和 Andrew Warrington。用于带标记事件序列的深度连续时间状态空间模型。载于神经信息处理系统第*39*届会议（*NeurIPS*）论文集，2025。

Liang Chen、Haozhe Zhao、Tianyu Liu、Shuai Bai、Junyang Lin、Chang Zhou 和 Baobao Chang。第二层之后，一幅图像的价值相当于 1/2 个词元：用于大型视觉语言模型的即插即用推断加速。载于欧洲计算机视觉会议，第 19–35 页。Springer，2024a。

Zhiheng Chen、Guanhua Fang 和 Wen Yu。时间点过程中的循环神经网络非渐近理论。*arXiv* 预印本 *arXiv:2406.00630*，2024b。

Daryl J Daley 和 David Vere-Jones。点过程理论导论。第 I 卷：概率及其应用，2003。

Daryl J Daley 和 David Vere-Jones。 点过程理论导论：第二卷：一般理论与结构。Springer Science & Business Media，2007。

Mohamed Dhouib、Davide Buscaldi、Sonia Vanier 和 Aymen Shabou。Pact：基于剪枝和聚类的词元缩减，可加速视觉语言模型。载于计算机视觉与模式识别会议论文集，第 14582–14592 页，2025 年。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer，2021 年。网址：https://arxiv.org/abs/2010.11929。

Nan Du、Hanjun Dai、Rakshit Trivedi、Utkarsh Upadhyay、Manuel Gomez-Rodriguez 和 Le Song。循环标记时序点过程：将事件历史嵌入向量。载于知识发现与数据挖掘国际会议，2016。

Álvaro Gajardo 和 Hans-Georg M¨uller。COVID-19 病例与死亡人数的点过程模型。应用统计学期刊，第 1–16 页，2021。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。低秩适配：大语言模型的低秩适配，2021。网址：https: //arxiv.org/abs/2106.09685。

Yue Jiang、Jichu Li、Yang Liu、Dingkang Yang、Feng Zhou 和 Quyu Kong。Danmakutppbench：用于时序点过程建模与理解的多模态基准。*arXiv* 预印本 *arXiv:2505.18411*，2025。

Quyu Kong、Yixuan Zhang、Yang Liu、Panrong Tong、Enqi Liu 和 Feng Zhou。Language-tpp：将时间点过程与语言模型相结合，用于事件分析。*arXiv* 预印本 *arXiv:2502.07139*，2025。

Dana Lahat、T¨ulay Adali 和 Christian Jutten。多模态数据融合：方法、挑战与前景综述。*IEEE*会议录，103(9):1449–1477，2015。

Zefang Liu 和 Yinzhu Quan。TPP-LLM：通过高效微调大语言模型来建模时间点过程。*arXiv* 预印本，2024。

Hongyuan Mei 和 Jason Eisner。神经霍克斯过程：一种由神经网络自调节的多变量点过程。载于神经信息处理系统进展，2017。

Hongyuan Mei、Chenghao Yang 和 Jason Eisner。用于不规则间隔事件及其参与者的 Transformer 嵌入。载于国际学习表征会议，2022年。

Zizhuo Meng、Ke Wan、Yadong Huang、Zhidong Li、Yang Wang 和 Feng Zhou。可解释Transformer霍克斯过程：揭示社交网络中的复杂交互。载于知识发现与数据挖掘国际会议，2024。

SebastianMeyer、Johannes Elias 和 Michael H¨ohle。侵袭性脑膜炎球菌病发生的时空条件强度模型。生物计量学，68(2)：607–616，2012。

GeorgeMohler。犯罪与安全数据中多源聚类的建模与估计。应用统计年鉴，第1525–1539页，2013年。

Takahiro Omi、Naonori Ueda和Kazuyuki Aihara。面向一般时间点过程的全神经网络模型。载于神经信息处理系统进展，2019年。

Aristeidis Panos。可分解Transformer点过程。载于神经信息处理系统进展，2024。

Marian Andrei Rizoiu、Swapnil Mishra、Quyu Kong、Mark Carman和Lexing Xie。Sir-Hawkes：论流行病模型与Hawkes点过程之间的关系。载于万维网会议，2018年。

Xiaoming Shi、Siqiao Xue、Kangrui Wang、Fan Zhou、James Zhang、Jun Zhou、Chenhao Tan和Hongyuan Mei。语言模型可通过少样本溯因推理改进事件预测。载于神经信息处理系统进展，2024。

Alexander Soen、Alexander Mathews、Daniel Grixti-Cheng 和 Lexing Xie。《Unipoint：通用逼近点过程强度》。载于*AAAI* 人工智能会议，2021。

Chameleon 团队。Chameleon：混合模态早期融合基础模型，2025。网址：https: // arxiv.org/abs/2405.09818。

Gemma 团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ram´e、MorganeRivi`ere 等。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*，2025。

A Vaswani。《注意力就是一切》。神经信息处理系统进展，2017。

Chris Whong。揭露纽约市出租车行程数据，2014。

Shuai Xiao、Junch Yan、Xiaokang Yang、Hongyuan Zha 和 Stephen Chu。通过循环神经网络对点过程的强度函数进行建模。载于 *AAAI* 人工智能会议，2017。

Siqiao Xue、Yan Wang、Zhixuan Chu、Xiaoming Shi 等。面向流式事件序列的提示增强时间点过程。载于 神经信息处理系统进展，2023。

Siqiao Xue、Xiaoming Shi、Zhixuan Chu、Yan Wang、Hongyan Hao、Fan Zhou、Caigao Jiang、Chen Pan、James Y. Zhang、Qingsong Wen、Jun Zhou 和 Hongyuan Mei。《EasyTPP：迈向时间点过程的开放基准》。载于国际学习表征会议，2024。

Chenghao Yang、Hongyuan Mei 和 Jason Eisner。对不规则间隔事件及其参与者进行 Transformer 嵌入。载于第十届国际学习表征会议论文集，2022。Guolei Yang、Ying Cai 和 Chandan K Reddy。用于签到时间预测的循环时空点过程。载于国际信息与知识管理会议，2018年。¨Qiang Zhang、Aldo Lipani、Omer Kirnap 和 Emine Yilmaz。自注意力霍克斯过程。载于国际机器学习会议，2020年。Yixuan Zhang、Quyu Kong 和 Feng Zhou。时空多模态协变量深度核点过程的无积分训练。载于神经信息处理系统进展，2023。Wayne Xin Zhao、Kun Zhou、Junyi Li、Tianyi Tang、Xiaolei Wang、Yupeng Hou、Yingqian Min、Beichen Zhang、Junjie Zhang、Zican Dong 等。大语言模型综述。*arXiv* 预印本 *arXiv:2303.18223*，1(2)，2023。Feng Zhou、Yixuan Zhang 和 Jun Zhu。脉冲神经元网络中灵活交互的高效推断。载于国际学习表征会议，2021。Zihao Zhou、Xingyi Yang、Ryan Rossi、Handong Zhao 和 Rose Yu。用于学习时空事件动态的神经点过程。载于动力学与控制学习会议，2022。Shixiang Zhu 和 Yao Xie。用于犯罪关联检测的时空文本点过程。应用统计年鉴，16(2)：1151–1170，2022年。Shixiang Zhu、Minghe Zhang、Ruyi Ding 和 Yao Xie。用于自注意力点过程的深度傅里叶核。载于国际人工智能与统计会议，2021。Simiao Zuo、Haoming Jiang、Zichong Li、Tuo Zhao 和 Hongyuan Zha。Transformer 霍克斯过程。载于国际机器学习会议，2020年。

A 自适应序列压缩的详细说明

A.1 压缩机制示意图

为阐明自适应压缩如何作用于连续事件，我们以一个由三个事件组成的假设序列为例，逐步进行说明： A、 B和 C依次发生。令 $\tau_i = t_i - t_{i-1}$ 表示事件 i的事件间隔。压缩决策根据相邻间隔的两两比较依次作出：

- **步骤 1（处理事件 B）：** 模型将事件 B （τ_B）的间隔与紧邻其前的事件 A（τ_A）的间隔进行比较。如果绝对差满足$|\tau_B - \tau_A| < \Delta$，则认为事件 B 与 A 在时间上相似，并将二者替换为一个特殊词元 < |相似事件| >。
- **步骤 2（处理事件 C）：** 随后，模型处理事件 C。它将其间隔（τ_C）与前一个事件 B（τ_B）的间隔进行比较。请注意，无论 B 是否已在词元序列中被压缩，此处的比较均基于计算得出的时间间隔。如果 $|\tau_C - \tau_B| < \Delta$，事件 C 也会被替换为 < |相似事件| >词元。

这一序贯过程确保根据事件密度的动态变化在局部应用压缩，使模型能够高效编码具有相似且鲜明时间模式的事件突发。

A.2 时间相似度阈值设置

在我们的框架中，时间相似度阈值 Δ 是自适应序列压缩的一项关键超参数。具体而言，对于每个事件，我们定义间隔 $\tau_i = t_i - t_{i-1}$ 并将其与前一个间隔 τ_{i-1}进行比较。如果两个间隔的绝对差满足 $|\tau_i - \tau_{i-1}|$ <Δ，则认为事件 i 与事件 $i-1$ 在时间上相似，因而可以合并。这一机制直接控制序列建模的粒度以及细粒度时间动态的保留程度，因为只有足够相似的事件才会被压缩为单一表示。
为了确定 Δ的合适取值，我们在 DanmakuTPP 训练数据集上，对相邻事件间隔（$|\tau_i - \tau_{i-1}|$）绝对差进行了全面的统计分析。如图 3 所示，这些差值的经验分布高度偏斜，大量数值集中在零附近。这表明事件间隔与前序间隔往往非常相似，从而验证了采用基于时间相似性的压缩策略的可行性。

为使阈值选择符合数据结构，我们考察了分布的分位数，结果汇总于表 4。基于这项分析，我们选取了 0.05、0.2 和 0.5 秒作为候选阈值。这些数值对应分布中具有意义的节点：

- $\Delta = 0.05$ 是一个较保守的阈值，对应约第 25 百分位数。它只压缩最相似的事件对，从而保留大部分原始序列结构。
- $\Delta = 0.2$ 对应约第 50 百分位数（中位数），在压缩率与信息保留之间取得平衡。
- $\Delta = 0.5$ 是一个较激进的阈值，接近第 75 百分位数，它会合并序列中更大一部分，以实现更高的压缩比。

这些阈值相对于分布的位置标记于图 3 中。这种有原则且由数据驱动的选择确保我们的自适应压缩机制在多模态时序点过程建模中既具有统计依据，又具备实际有效性。

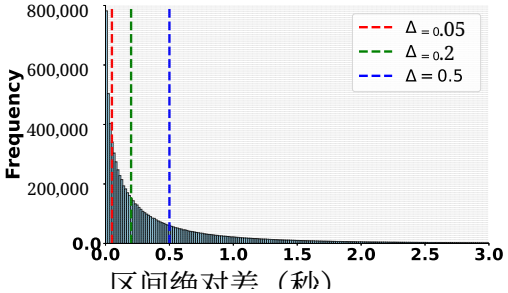


图 3：训练集上相邻事件间隔绝对差值的分布。竖线表示所测试的阈值 Δ。

表 4：相邻事件间隔绝对差值的分位数（$|\tau_i - \tau_{i-1}|$）。

百分位数	值（秒）
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 数据集构建

我们通过在原始纽约市出租车数据集的基础上增加多模态信息（图像和文本描述），构建了 TAXI-PRO 数据集。本节将详细介绍数据构建流程。

原始出租车数据。原始出租车行程数据来自纽约市出租车与豪华轿车委员会（NYC TLC）的公开数据集，由 trip data 1.csv 至 trip data 12.csv 等文件组成。该数据集与 EasyTPP 官方出租车数据集所用的数据源完全相同，均源自纽约市出租车与豪华轿车委员会开放数据计划 [1]。

我们通过以下设置构建 TAXI-PRO 数据集：

- **图像**。我们利用每个事件的坐标，截取其所在位置的地图图像。我们使用 contextily 等工具，根据 OpenStreetMap 数据生成高分辨率地图图块。底图覆盖整个曼哈顿区域，并确保像素坐标与地理边界严格对应（分辨率： 6071×6307像素）。对于每个事件位置，我们截取以事件坐标为中心的 224×224 像素图块。对于边缘情况，我们使用灰色背景填充，以保持图块尺寸一致。

- **文本**。我们按照固定模板为每个事件生成文本标记，模板中包含地标信息、坐标、乘客数量和行程距离。例如，“在时代广场接客（*40.7580, -73.9855*），*2* 名乘客”。

- **事件类型**。我们仅关注曼哈顿，并将其划分为三个区域：下曼哈顿、中城曼哈顿和上曼哈顿，由区域与接客/下客事件组合而成的事件类型共有 6 种。

为了直观展示数据结构，图 4 展示了 TAXI-PRO 数据集中单个事件的组成。图中突出了每个数据点的多模态特性，它结合了地理地图切片、模板化文本描述和类别型事件类型。在选取最终的 2000 个序列时，我们采用方差最小化贪心算法，确保事件类型分布均衡，避免出现偏差。

C 详细超参数设置

超参数 各训练阶段所用的具体超参数汇总于表 5。为提高计算效率，我们在所有实验中均采用 AdamW 优化器和 bfloat16 混合精度。

[1] 揭露纽约市出租车行程数据：https://chriswhong.com/open-data/foil_纽约市_出租车/

{
时间_自_上次_事件:0.0333,

事件_类型:4
(在上_曼哈顿接客),

文本: "在中央公园西侧接客 (40.800018,
-73.944778), 1名乘客",

图像: ---------->
}



图 4：TAXI-PRO 数据集中单个事件的示意图。每个事件包含四个关键组成部分：(1) 以事件坐标为中心的 224×224 地图图像，(2) 包含地标和行程详情的生成文本描述，(3) 类别型事件类型，以及 (4) 时间戳。

表 5：超参数设置

超参数	值
基础模型	Qwen2.5-VL-3B
LoRA 秩	16
LoRA Alpha	64
LoRA 丢弃率	0.1
阶段 1 学习率	1e-4
阶段 1 训练轮数	5
阶段 2 学习率	1e-4
阶段 2 训练轮数	3
批次大小	1
最大词元长度	4096
时间阈值 Δ	0.2s
推理温度	0.05

D 与随机丢弃基线的比较

为严格评估所提出的自适应压缩机制的有效性，我们将其与一种朴素的压缩基线方法进行了比较：**随机丢弃**。该基线方法在训练与推理阶段都会从序列中随机丢弃事件。

我们在 DanmakuTPP 数据集上进行了实验，丢弃概率分别为 $p = 25\%$ 和 $p = 50\%$。结果与我们的 MM-TPP（自适应压缩）以及未压缩基线进行了比较，并呈现在表 6 中。

如表 6 所示，与我们的自适应方法相比，随机丢弃策略会导致明显的性能下降。即使丢弃率适中，仅为 25%，模型的预测准确率和时间戳精度也不及 MM-TPP。这种性能下降很可能是因为时间点过程中的事件序列具有严格的因果依赖；随机移除事件会破坏模型推理未来动态所必需的历史上下文（例如，自激模式）。相比之下，我们的自适应压缩策略只选择性地合并时间相似区间（这些区间通常代表冗余的突发信号），同时保留关键的锚点事件以及序列整体的语义连续性。

表 6：MM-TPP 的自适应压缩与朴素随机丢弃策略在 DanmakuTPP 数据集上的性能比较。所提出的自适应方法显著优于随机丢弃，凸显了保留时间结构的重要性。

模型 / 设置	**均方根误差 (↓)**	**准确率 (%) (↑)**
MM-TPP	**5.2987**±0.018	**27.62**±0.02
MMTPP-Uncompressed	5.5551±0.008	25.87±0.05
MMTPP-Random-Drop（$p = 25\%$）	5.3792±0.014	24.84±0.06
MMTPP-Random-Drop（$p = 50\%$）	5.4259±0.009	24.79±0.03

E DanmakuTPP-QA 详情

表 7：DanmakuTPP-QA 基准中十项任务的详细描述与评估指标。封闭式任务采用定量方式评估（准确率或均方根误差），而开放式任务则通过定性分析进行评估。

	任务描述	**评估指标**	**任务类型**
任务-1	弹幕突发峰值计数	ACC	封闭式
任务-2	预测下一个 弹幕时间戳	RMSE	封闭式
任务-3	下一个弹幕的预测 突发峰值时间戳	RMSE	封闭式
任务-4	平均情感极性评估 情感极性	RMSE	封闭式
任务5	情感极性预测 针对下一条弹幕	RMSE	封闭式
任务6	情感极性预测 下一个弹幕爆发峰值	RMSE	封闭式
任务7	事件类型推断 下一个弹幕	ACC	封闭式
任务-8	预测前2触发事件 类型，用于预测下一个爆发峰值	ACC	封闭式
任务-9	全局情感的分析 动态及其潜在驱动因素	定性分析	开放式
任务-10	因果归因分析 特定的弹幕突发峰值形成	定性分析	开放式

E.1 十项问答任务

本节详细介绍构成**DanmakuTPP-QA**基准的十项推理任务。该基准旨在通过一系列多样化的挑战，评估模型的时序、视觉和文本推理能力。如表 7 所总结，这些任务分为八项封闭式定量评估任务，以及两项需要对生成的报告进行定性评估的开放式任务。

E.2 MM-TPP 在开放式任务上的表现

在本节中，我们对 MM-TPP 为开放式任务生成高质量分析报告的能力进行定性评估。我们将其输出与两个最先进的通用

多模态大语言模型（MLLM）进行比较：**Qwen2.5-VL-3B**（Bai 等，2025）和 **Gemma3-27B**（Team 等，2025）。我们的分析表明，MM-TPP 始终能够生成更有见地且更切题的报告，凸显了其专用架构的价值。

这种定性优势不仅在 MM-TPP 进行微调的 DanmakuTPP-QA 基准上显而易见，而且在零样本设置下推广至 TAXI-PRO 数据集时同样表现稳健。为佐证这些结论，我们将在以下部分提供一系列对比示例。鉴于性能差距始终如一，我们针对开放式任务（任务 9 和任务 10）以及 TAXI-PRO 报告生成任务分别提供两个代表性示例，以展示 MM-TPP 分析能力的一贯优势。我们将定性优势总结如下：

- **任务 9（弹幕情感动态分析）**：MM-TPP 通过聚焦于核心情感动态，更有效地传达信息；其分析结构清晰，遵循从总体趋势到关键拐点再到总结的脉络，并生成了人工评估者认为更易读、更自然的输出。

- **任务 10（弹幕峰值突发原因分析）**：MM-TPP 不仅能够捕捉峰值突发，还能以高时间精度定位次峰值，从而更严谨地呈现自激动态。它还将引用/参考、幽默/梗等具体事件类型，直接联系到突发的触发机制，阐明内容类别如何驱动级联响应。不同于通用多模态大语言模型，它避免偏离主题、陷入无关的场景描述，而是突出关键驱动因素，并明确以 TPP 理论为框架进行解释。这种精确性、聚焦性与领域特定基础之间的平衡，使分析既易于理解，又具备学术严谨性。

- **TAXI-PRO**：MM-TPP 可以通过识别驾驶员行为中的核心 TPP 模式，提供高层次的洞察。它不再局限于简单的路线追踪，而是对事件之间如何相互影响进行建模，从而揭示驾驶员针对时空动态所采取的策略性应对，例如在高需求走廊中形成自激事件簇；相比之下，擅长语言生成的多模态大语言模型往往生成的输出更具描述性，而非分析性。

TAXI-PRO 示例 1 _

MM-TPP 在 TAXI-PRO（示例 1）上生成的文本

提示:
你是一名专门从事时空事件序列分析的数据分析专家。你的任务是分析纽约市一名出租车司机活动的时间顺序日志，其中包括在不同地点接客和送客等行为。请根据所提供的事件序列，进行全面分析，并分多个段落生成一份详细的分析报告。请注意，此序列不是视频！所有时间间隔均根据 IEEE 754 标准转换为四个字节。$< |TPP\ data| >$

-

回答:
Th该事件序列展现了一名司机在纽约市（NYC）多个标志性街区间的动态移动，活动主要集中在上西区、剧院区及周边地区。司机从上西区出发，随后迅速穿行于金融区和翠贝卡等人流密集、文化气息浓厚的地点，这表明行人或交通活动频繁。这种模式与自激型 TPP 动态相符，即初始事件会通过社会或经济刺激引发后续事件。

司机始终服务四至六名乘客，这表明其可能从事曝光度较高的服务，或采用共享车辆模式；每个班次中接客和送客的乘客各不相同。平均行程时长不到 10 分钟，其中不乏短途行程（<1.5 英里），这意味着司机周转迅速，或存在高频服务需求。

从地图上看，各点聚集紧密，尤其集中在 40.76–40.78 和 40.69–40.74 之间，与剧院区及周边活动频繁区域等已知最密集的城市街区相吻合。多个“已送达”条目都位于 <1.5 英里范围内，这表明司机很可能在高密度城市环境中运营，而这种近距离情况十分常见。

时间聚类表明存在一个高强度时段，可能是清晨或深夜，此时公共交通或步行活动模式更为密集。司机接连快速服务不同数量乘客的模式，意味着采用了轮班式服务模式，每个地点都充当一个微型枢纽，可能是为了应对不断变化的乘客流量。

出乎意料的是，尽管短途行程频率很高，平均速度却没有显著下降，这表明司机可能身处限速区域较多或行人阻力较大的城市。此外，未出现显著负面情绪，与高频、短间隔事件相关联，说明这些出行需求可能是日常性的，也可能出于社会或文化因素，而非异常情况。
总之，这一序列反映了人口稠密的城市中心典型的高强度短途运输模式；时间和空间上的聚集表明，司机的行为具有规律性，并受到当地地理环境和时间动态的影响。

通义千问在 TAXI-PRO 上生成的文本（示例 1）

提示：
你是一位专攻时空事件序列的资深数据分析师。你的任务是分析一名出租车司机在纽约市的活动时间日志，其中包括在不同地点接客和送客等行为。请根据提供的事件序列进行全面分析，并分段撰写详细的分析报告。$< |TPP\ data| >$

回答：
所提供的数据记录了一位出租车司机在纽约市约 8.5 小时内的活动，涵盖了一系列接客 and
和送客事件，并包含时间、地点、乘客人数及行驶距离等详细信息。
这段行程始于在时代广场接客，随后行驶 1.45 英里前往地狱厨房。接着，司机在地狱厨房又接上两名乘客，并将他们送至翠贝卡，行驶距离为 2.22 英里。之后的行程都较短，途经剧院区和切尔西，凸显出这些地区对短途出租车行程的旺盛需求。

值得注意的长途行程包括从剧院区到中央公园西侧的 3.67 英里行程，以及从切尔西到哈莱姆区的 5.20 英里行程。这些较长的行程展现了司机满足多样化出行需求的能力。不过，司机的大部分活动都集中在曼哈顿中城和曼哈顿下城的短途、高频行程，尤其是在翠贝卡、切尔西和剧院区等地。记录的最后几段行程包括从翠贝卡到曼哈顿中城东区的 5.55 英里行程，以及从剧院区到时代广场的 1.90 英里行程，为这段行程画上句号。
总之，该司机在高需求区域高效运营，频繁承接短途行程，偶尔也会承接较长行程。这一模式反映了纽约市繁忙城市环境中出租车服务的动态特性。

Gemma在TAXI-PRO上生成的文本（示例1）

提示：
您是一名专精于时空事件序列的数据分析专家。您的任务是分析纽约市一名出租车司机活动的时间顺序日志，其中包括在不同地点的接客和送客等行为。请根据所提供的事件序列进行全面分析，并生成一份详尽的、由若干段落组成的分析。
$< |TPP\ data| >$

回答：
本分析详细说明了该出租车司机在约8.5分钟内的活动日志。该司机表现出高度活跃的模式，在这一短时间内完成了36个事件——包括接客和送客。该司机的运营集中在曼哈顿，在中城、切尔西、翠贝卡和剧院区有大量活动，并前往哈莱姆区和炮台公园。频繁的短途行程表明其侧重于高需求、快速周转的车费，可能服务于游客或通勤者。该司机持续接送乘客，表明其高效利用了各笔车费之间的时间。日志的初始阶段（0–2.5分钟）显示，接客和送客的快速序列主要发生在时代广场、地狱厨房、翠贝卡和切尔西之间。这表明该司机最初可能位于高需求区域，并迅速响应了可用请求。这一时期的行程距离相对较短（从0.06英里到2.22英里不等），进一步印证了快速、局部行程的特点。在约3–5分钟时出现了一个显著模式，转向更长的行程，包括前往中央公园西侧的3.67英里行程，以及前往哈莱姆区的5.20英里行程。这表明其愿意接受更长的车费，可能是为了最大化收入或重新定位以获得更好的机会。然而，在日志后半段，该司机很快又回到了更短、更频繁的行程。

这位司机每趟行程搭载的乘客人数为1至2人。大多数行程搭载2名乘客，这表明较大的团体更受青睐，或司机有能力搭载更多乘客。最后几次事件（6至8.5分钟）显示，司机回到一种模式：在中城和翠贝卡地区进行短途行程，最终在时代广场送客。即使行程时长各不相同，司机也能在接客与送客之间迅速切换，体现出其高效应对需求的能力。整个日志中活动持续不断，表明这是一位敬业且活跃的司机，能够有效穿行于纽约市动态多变的出租车服务环境中。进一步分析这组数据，并结合一天中的时段和外部事件进行关联，或可揭示司机行为以及特定区域出租车服务需求中更细致的规律。

TAXI-PRO 示例2

MM-TPP 在 TAXI-PRO（示例2）上生成的文本

提示：
你是一位专长于时空事件序列的数据分析专家。你的任务是分析一名纽约市出租车司机活动的按时间顺序排列的日志，其中包括在不同地点接客和送客等行为。请根据提供的事件序列进行全面分析，并分段撰写详细分析。请注意，这不是视频！所有时间间隔均已按照 IEEE 754 标准转换为四字节。$< |TPP\ data| >$

答案： Th

该事件序列揭示了纽约市（NYC）一名司机与多个地点之间复杂的多模态交互模式。时间聚类表明活动频率较高，这可能是由动态定价或出行高峰时段所驱动。

司机在坐标 40.764385, -73.977257（时代广场）开启一系列行程，随后反复在多个地点接送乘客 ons
这些地点包括地狱厨房、翠贝卡、切尔西和炮台公园，呈现出一连串密集发生的事件。这种模式与 elf-
令人振奋的 TPP 动力学相符，其中初始事件通过社会或经济刺激引发后续事件。
大多数行程涉及两名乘客，这表明可能是拼车或多目的地出行；其他行程则只有一名乘客 长者。
这意味着存在双重活动模式：有些人是被动观察者（例如远距离观看），而另一些人则是积极的参与者， pos-
可能是由于短期可用性或能见度有限。
空间分布呈现出从核心到外围的模式，并伴有频繁的区域间移动，例如从城市中心区域前往外围区域。平均行程为 1.66 英里，说明出行距离相对较短，可能反映了微出行模式或通勤行为。

14:00–15:00 左右出现了显著的时间变化，期间观测到多次高密度下降，与城市交通高峰和动态定价飙升同时发生。这种时间聚类支持 TPP 触发效应，即过去事件（例如先前的下降）会激发未来事件。

出乎意料的是，没有显著的负 ative情感与高参与度相关，表明积极 强化循环——可能是由于

to 社交互动或对该城市的熟悉。事件类型的均匀性（均为10%）表明广泛激发，而非
sin个驾驶员异常。
总而言之，该序列反映了 TPP 驱动的驾驶员行为，由时间、地点和行动方式之间的多模态交互所塑造。该驾驶员表现出自激、时间聚类和区域依赖，这些都是城市交通系统中复杂事件触发过程的典型特征。

通义千问在 TAXI-PRO 上生成的文本（示例2）

提示： Y

你是一位专精于时空事件序列的资深数据分析师。你的任务是分析一名纽约市出租车司机的活动时间日志，其中包括在不同地点接客和送客等行为。请根据所提供的事件序列进行全面分析，并分段撰写详细分析。$< |TPP\ data| >$

答案：
所提供的事件序列详细记录了一名出租车司机在纽约市约十小时内的活动。该日志记录了一系列接客和送客，每次均附有精确的时间戳、乘客人数和地理坐标。司机从上西区开始班次，接上四名乘客，并在10分钟后将他们送到剧院区。在接下来的几个小时里，司机主要在曼哈顿中城运营，在剧院区、先驱广场和中央公园西侧之间完成多次短途行程。这些行程距离在0.73到1.35英里之间，乘客为一至三人。大约在一小时时，司机完成了从剧院区到中央公园西侧的1.11英里行程，随后返回剧院区。这种短途、频繁的行程模式一直持续到2.5小时时，司机从剧院区前往曼哈顿中城东区，完成了一次较长的2.39英里行程。此后活动出现明显空档，表明司机休息或活动减少。司机恢复运营后，在翠贝卡接上六名乘客，并将他们送到金融区，行程为1.79英里。随后，他们又完成了从金融区到苏豪区的另一次长途行程，全程2.43英里。这些较长的行程表明，司机转向了城市的其他区域，那里需求较低，但行程利润更高。

司机返回曼哈顿中城，继续完成了几趟行程，其中包括从先驱广场到中央公园西侧的一趟4.80英里行程，这是记录中距离最长的一趟。最后记录的一趟行程是从曼哈顿中城东区到中央公园西侧的0.99英里行程。在这段约十小时的时间里，司机总共完成了36趟行程。

总的来说，司机主要在曼哈顿中城一带运营，频繁接载乘客完成短途行程，偶尔也会前往城市其他区域，承接较长且可能利润更高的行程。这种平衡的运营方式很可能兼顾了效率和盈利能力。

Gemma 对 TAXI-PRO（示例2）生成的文本

提示： Y

你是一位专门分析时空事件序列的数据分析专家。你的任务是分析一名出租车司机在纽约市的活动按时间顺序排列的日志，其中包括在不同地点接客和送客等行为。请根据提供的事件序列进行全面分析，并分若干段生成详细的分析。$< |TPP_data| >$

回答：
这份数据集记录了一名出租车司机约 10 分钟内的活动，展现了其在曼哈顿各处接客和送客的动态模式。司机从上西区出发，搭载满 4 名乘客，较快地完成了一趟前往剧院区的行程。这段初始行程可能是通勤，也可能是预约接客。随后，司机很快转而承接较短的行程：先在剧院区接上 2 名乘客，并在先驱广场送下他们；接着又在剧院区接上 3 名乘客，并在中央公园西侧送下他们。这种变化表明，司机开始承接更具临时性的短途车费，这通常是司机在繁忙区域巡游时的模式。在初始阶段，乘客人数频繁变化（4、2、3），说明司机并非一直服务于固定群体，也不是承接预订的长途行程。

日志中段（约第 5 至第 8 分钟）显示，司机的营运区域和乘客数量发生了显著变化。司机接下来在翠贝卡接上 6 名乘客，此前出现了明显的时间间隔，说明他花了更长时间才抵达这一南部地点。之后，司机接连接客和送客，活动主要集中在曼哈顿中城，包括时代广场、地狱厨房和先驱广场，且始终满载 6 名乘客。这表明司机可能接下了一笔规模较大的团体预订，或是在高峰时段利用需求旺盛的区域获利。这些行程的距离各不相同，从曼哈顿中城内的短途行程，到前往中央公园西侧的较长行程，体现了司机对乘客需求的适应能力。

最后几次事件（约 9 至 10 分钟）显示，司机又回到了较短行程的模式，在曼哈顿中城东区接载了六名乘客，并在中央公园西侧让他们下车。这段时间内乘客人数始终较多，表明司机积极寻求提高载客率。整体模式体现出司机灵活应变，能够根据情况调整行程长度和乘客人数。初期主要是较短且分布较分散的行程，而中后期则表现出尽可能提高载客量的倾向，尤其是在中城及周边地区。司机的路线显示出其熟悉曼哈顿的地理环境和需求热点，能够在不同街区之间高效穿行，以获取车费。数据没有提供等待时间或空闲时段的信息，而这些信息有助于更全面地评估司机的效率。

任务 9 示例 1

MM-TPP 在任务 9（示例 1）上生成的文本

提示： Y

你是一位分析弹幕的专家模型。根据评论序列及其时间间隔，你的任务是对该时间窗口进行情感分析。所有评论均为中文，所有时间间隔均按照 IEEE 754 标准转换为四字节。
$< |TPP\ data| >$
$< |question| >$ 请详细分析弹幕中情感趋势随时间的变化。

回答：
The 视频整体呈现出积极情感，仅有轻微波动 . 起初情感急剧上升，随后趋于稳定
aro以及正向极性。关键拐点包括约在00:00:30和00:01:00出现的早期峰值，并在附近略有下降

00:01:30和00:02:00。在00:03:00附近出现显著的情感变化，呈现逐渐下降但仍保持正向。总体而言，
video 的情感轨迹始终保持积极，负面偏差极少。

通义千问在任务 9 中生成的文本（示例 1）

提示：
你是一名弹幕分析专家。请根据评论序列及其时间间隔，对该时间窗口进行情感分析。所有评论均为中文。问题：请详细分析弹幕随时间变化的情感趋势。< |TPP data| >

回答：
The视频弹幕的情感趋势总体上从中性逐渐转向积极，因为观众可能会被视频标题和开场画面吸引。随着视频推进，评论频率不断增加，表明参与度有所提升。在20秒处，评论数量明显激增，说明视频中出现了特别吸引人或令人兴奋的时刻，可能是戏剧性的视觉画面或令人印象深刻的陈述。整体而言，情感始终保持积极，偶尔出现的峰值则表明观众在某些时刻感到格外兴奋或惊叹。视频中段（约120秒处），评论趋于稳定，但观众仍热情不减，纷纷表达对内容的赞赏或分享自己的经历。临近视频结尾时，评论频率略有下降，但整体积极情绪依然持续，许多观众留下了最后的感想或鼓励。这一趋势表明，视频成功地保持了观众的兴趣，并持续引发积极回应。

Gemma 在任务 9 上生成的文本（示例 1）

提示： Y
你是一位分析弹幕的专家模型。根据评论序列及其时间间隔，你的任务是分析该时间窗口内的情感。所有评论均为中文。问题：请详细分析弹幕情感随时间的变化趋势。< |TPP data| >

回答：
弹幕密度表明，整个视频始终保持着较高的参与度，其中多个峰值说明有些时刻尤其引人关注或产生了较大影响。从约 2 秒开始，参与度由低逐渐上升，随后一直十分活跃，并频繁出现峰值。整体参与度并未明显下降；在所提供数据的最后阶段仍然很高。由于没有评论文本，无法判断评论表达了何种情感（积极、消极、惊讶等）。不过，持续较高的评论量表明视频总体上很吸引人，并引发了观众的反应，这可能与标题所描述的精彩内容（“百万伏特电压捕获！”）有关。频繁出现的峰值可能对应着视觉效果或叙事上较为戏剧性的时刻。

任务 9 示例2

MM-TPP 在任务 9（示例2）上生成的文本

提示：
你是一个用于分析弹幕的专家模型。根据评论序列及其时间间隔，你的任务是对时间窗口进行情感分析。所有评论均为中文，所有时间间隔均按照 IEEE 754 标准转换为四个字节。< |TPP data| >
< |question| > 请详细分析弹幕随时间变化的情感趋势。

回答：
情感极性趋势总体呈现积极走向，仅有轻微波动。起初情感值迅速上升，随后稳定在 0.4 左右。关键拐点包括早期的峰值和谷值，之后则趋于稳定的正向区间。显著的情感变化发生在开头，情感值大幅上升；临近结尾时则略有下降。总体而言，视频保持着积极的情感基调，突发变化极少，说明观众体验稳定且令人振奋。

通义千问为任务 9（示例2）生成的文本

提示： 是
你是一名分析弹幕的专家模型。请根据评论的序列及其时间间隔，对该时间窗口进行情感分析。所有评论均为中文。问题：请详细分析弹幕中情感趋势随时间的变化。< |TPP data| >

回答： Th
esis弹幕中的情感趋势随时间呈现出清晰的模式。最初，在0至60秒期间，评论密度非常高，表明观众兴奋度很高、反应积极，因为他们一开始就投入其中并表现出热情。从60至120秒，评论密度有所下降，说明观众参与度稳定，反应不一；观众仍然感兴趣，但热情有所减弱。在中段（120至300秒），评论密度有所波动，偶尔出现峰值，反映出观众反应不一，关键场景会引发更多评论。从300至480秒，评论密度较低，偶尔出现峰值，表明观众反应较为平静，但对内容的投入更深。最后，从480至633秒，评论密度有所上升，并在接近结尾时出现显著峰值，显示观众参与度提高，纷纷分享总体印象以及对结局的看法和反应。总体而言，视频始终维持着观众的兴趣，开头和结尾尤为突出，中间阶段的参与度则较为稳定。

Gemma 在任务 9（示例2）中生成的文本

提示:
你是一名擅长分析弹幕的模型。根据评论序列及其时间间隔，你的任务是对该时间窗口进行情感分析。所有评论均为中文。
问题：请详细分析弹幕随时间变化的情感趋势。$< |TPP\ data| >$

回答:
弹幕显示，观众在最初的 0–10 秒参与度很高，随后略有下降，但从 20 秒起逐渐恢复并持续保持高密度。评论频率出现多个峰值（约在 20–40 秒和 100–150 秒），表明影片中有一些令人印象深刻的时刻。影片结尾处（550 秒+）的评论量仍然极高，说明观众反应强烈。总体而言，观众始终积极参与，其中某些场景在整个观看过程中尤其引发强烈反应。由于缺少评论内容，无法判断这些峰值代表积极还是消极情绪，只能确定它们反映了观众注意力的增强。

任务 10 示例 1

MM-TPP 在任务 10 中生成的文本（示例 1）

提示:
你是一位分析弹幕的专家模型。根据评论序列、评论出现的时间间隔以及相关视频帧，你的任务是详细分析弹幕峰值爆发的原因，并提供一份详细的分析报告。所有评论均为中文，所有时间间隔均按照 IEEE 754 标准转换为四字节。
$< |TPP\ data| >$
$< |question| >$ 在时间窗口： [0, 1。5], 请用简洁流畅的一段话分析弹幕峰值爆发的原因。

回答:
时间戳 0 处的弹幕爆发呈现出明显的初始尖峰，随后在 0.08–0.1 秒、0.3–0.4 秒和 0.6–0.7 秒出现快速密集的评论簇，这表明存在自激型 TPP 动态——早期评论引发了观众的连锁回应。这些次峰值与反复出现的引用/参考和幽默/模因类事件类型（例如“《百家讲唐》”“败家讲坛”）相吻合，说明梗文化和社群内部笑话推动了文本的病毒式传播。
平均情感极性为 0.42，反映出整体语气略偏积极**，但评论类型多种多样，包括批评性评论、情绪化评论和离题言论，体现了异质性受众反应。值得注意的是，中性视觉标签（“商务西装”“演讲”）与情绪饱满的评论形成对比，揭示出一种多模态失调：正式着装和演讲场景可能引发了讽刺性或幽默性重释，进而推动表情包生成和参与度提升。
时间对齐显示，第一个次峰值（0.0–0.1s）以无意义文本和情绪反应为主，可能是由视觉转变突然出现或意外的内容并置所触发。第二个聚集阶段（0.3–0.4s）中，社交互动和个人经历有所增加，表明观众开始形成集体解读或展开讨论。最后，第三波（0.6–0.7s）中，批评性评论和引用/参考有所增多，反映出观众对主题比较（例如，“范小勤”与“丁真”）进行了更深入的参与。Thi s burst ill说明 TPP self-excitati论，h在此 ly 评论 acte d 作为t i触发器，提示 i形成连锁 de of related and 反应i式帖子 . 正式视觉内容与幽默/文本回应之间的多模态不匹配凸显了受众解读，而不仅仅是屏幕上的内容，才是推动参与度的关键。其结果是文本传播性、时间聚类和视觉框架之间复杂的相互作用，这在互动流媒体平台中很常见互动流媒体平台。

通义千问在任务 10 上生成的文本（示例 1）

提示词:
你是一名分析弹幕的专家模型。请根据评论序列、其时间间隔以及相关视频帧，详细分析弹幕峰值爆发的成因，并给出详细的分析报告。所有评论均为中文。
问题：在时间窗口： [0, 1.5], 请用简洁流畅的段落分析弹幕峰值爆发的成因。$< |TPP\ data| >$

回答:
在时间窗口 [0, 1.5] 秒内，弹幕峰值爆发主要由时序点过程（TPP）中强烈的时间自激所驱动。这些峰值集中在0.48–0.58秒和1.06–1.46秒附近，与“商务西装”、“演讲”和“新闻”等正式视觉线索的出现相吻合。这些视觉元素营造出一种戏仿严肃或权威的基调，从而催化多模态反讽，并促使观众将画面与讽刺性或具有文化指涉的评论并置。对“《百家脑坛》”和“《今日说法》”等讽刺性标签的反复提及表明存在一种反馈回路：一条幽默评论会激发类似回应，尤其是在幽默/模因和引用/参考类别中。尽管平均情感呈中性偏正（0.384），情绪反应和无意义文本评论的突发表明，情感异质性由娱乐而非批评所驱动。次峰值区间（如1.25–1.46秒）强化了这一模式，对“《百家讲坛》”及其他元评论的反复提及进一步放大了参与度。总体而言，此次爆发最好解释为具有自激特性的幽默级联，它们在文化上产生共鸣，并形成持续不断的反讽与戏谑互动流。

Gemma 在任务 10 中生成的文本（示例 1）

提示: Y
你是一名分析弹幕的专家模型。
请根据评论序列、评论的时间间隔以及相关视频帧，详细分析弹幕峰值爆发的原因，并给出详细的分析报告。所有评论均为中文。

问题：在时间窗口： [0, 1 .5], 请用简洁流畅的一段话分析弹幕中爆发峰值的原因 子弹评论。
$< |TPP_data| >$

答案:
弹幕的突发主要由讽刺性评论和反讽式并置共同驱动，这一点体现在

人们频繁提及《百家脑坛》和《今日说法》等节目，同时场景又带有辩论般的正式氛围。这形成了一个强烈的自我强化循环：幽默且类似梗的评论（约占 14.29%）会引发类似回应，而场景中的视觉线索——尤其是身着商务西装的人物正在“演讲”——进一步放大了这种效应。整体情感倾向较为积极，但弹幕突发与戏谑式嘲弄和文化典故相关，这表明观众是在进行反讽式评论，而非表达强烈的负面反应。视觉元素非但没有降低参与度，反而似乎起到了催化作用，营造出一种动态氛围：正式场景与荒诞或富有文化共鸣的评论形成反讽式对照，进而引发一连串幽默回应。

任务 10 示例2

MM-TPP 在任务 10（示例2）上生成的文本

提示：
你是一个用于分析弹幕的专家模型。请根据评论序列、其时间间隔以及相关视频帧，详细分析弹幕峰值爆发的原因，并给出详细的分析报告。所有评论均为中文，所有时间间隔均按照 IEEE 754 标准转换为四个字节。$< |TPP\ data| >$

$< |question| >$ 在时间窗口内： [393.5, 396.5], 请详细分析弹幕峰值爆发的原因

答案：
弹幕爆发峰值出现在395秒，密集的时间戳显示394.8–395.0秒和395.6–395.8秒附近存在次峰值，表明参与度在这些时间段内快速集中。这些集群与幽默/模因（例如“1!5!”、“55555”）、情绪反应（“哥们 你德玛的啊,你要是在s11遇到我...”）以及社会互动相吻合，表明这是一种自激式TPP 模式，其中早期评论会触发级联响应。

占主导地位的对象标签——*游戏*、*截图*、*电子游戏*——暗示这是一个在情感和幽默层面引发观众共鸣的游戏精彩片段或截图时刻。情感极性的平均值为0.423（正向），这与兴奋和模因式评论相符，而非批评性或离题言论。

时间对齐显示，初始峰值（394.5–394.8）集中体现了困惑或期待（“?”、“我觉得不能的,扣眼珠 子”），可能是由视觉上出乎意料的事件触发的。随后，数字类/幽默文本（“1!5!”、“5”）大量涌现，表明观众正在进行实时投票或评分，这类仪式在游戏视频中十分常见，也进一步强化了观众之间的触发效应。

值得注意的是，尽管整体情感积极，仍出现了一些批评性或离题评论（例如“不能”），这表明观众虽然多元，但参与度较高。反复使用“5”和“1!5!”反映出一种协调性或由网络梗驱动的回应模式，这与时序点过程模型中的自激现象一致——相似评论会在时间上紧密聚集。

多模态相关性分析表明，视觉上下文（标记为游戏的游戏画面）促发了快速的文本参与，尤其是幽默和社会模仿，从而放大了爆发。未观察到负面情绪峰值，因此可以排除争议驱动的参与。相反，这次爆发反映了同步的情感与互动行为，其驱动因素是共同的游戏文化和视觉触发因素。

通义千问在任务 10（示例2）中生成的文本

提示：
你是一款用于分析弹幕的专家模型。请根据评论序列、评论之间的时间间隔以及相关视频帧，详细分析弹幕峰值爆发的原因，并给出一份详细的分析报告。所有评论均为中文。

问题：在时间窗口： [393.5, 396.5], 内，请详细分析弹幕峰值爆发的原因。$< |TPP\ data| >$

回答：
在时间窗口 [393.5、396.5], 内，弹幕的峰值爆发主要由视觉与文本元素的汇聚所驱动，这些元素引发了强烈的情感与社会互动。视频画面捕捉到游戏中的一个关键时刻，很可能是一项高难度挑战（“地狱级难度翡翠”），成为观众的视觉触发点。这一视觉线索促使评论激增，其特点是幽默与网络迷因（例如“扣大雷”“扣眼珠子”）以及社交互动（例如数字重复“1111111”“55555”）。略微积极的情感（0.386）反映出观众的兴奋与轻松戏谑的互动，他们正期待这一具有挑战性的游戏片段带来不可预测的结果。相似短语与数字的密集快速聚集表明这是一个自激霍克斯过程，最初的评论会刺激响应级联，从而形成集体参与的脉冲。这些评论与关键游戏时刻在时间上的对齐，进一步强化了多模态互动，视觉线索与文本回应相互配合，共同推动峰值爆发。

Gemma 在任务 10（示例2）中生成的文本

提示： Y

您是一名弹幕分析专家。根据弹幕序列、弹幕的时间间隔以及相关视频帧，您的任务是详细分析弹幕出现峰值爆发的原因，并给出详细的分析报告。所有弹幕均为中文。问题：在时间窗口： [393.5, 396.5], 内，请详细分析弹幕峰值爆发的原因。$< |TPP\ data| >$

答案：
根据所提供的数据，393.5 至 396.5 秒之间弹幕骤增，主要由游戏事件与协调性社会互动共同推动，用选项 D 最能解释这一现象。峰值与可能呈现高难度游戏场景的时刻相吻合（“地狱级难度翡翠”），引发了观众的情绪反应和提问。同时，明显存在基于网络迷因的协调，重复出现的数字表达（“1111111”“55555”）以及“扣大雷”等短语表明，观众正通过具有文化共鸣的互动模式进行轻松戏谑的互动。这表明存在一种自激式霍克斯过程动态：初始评论引发级联响应，而游戏截图这一视觉线索则充当潜在触发因素，促使文本互动迅速增加。略偏积极的情感（0.386）进一步支持了这一解读：观众表现出兴奋和轻松戏谑的互动，而非消极情绪或批评。

F 提示词详情

这里展示用于预处理时序点过程数据集的详细提示模板。由于这两个数据集的处理方式相似，我们仅展示 TAXI-PRO 的模板。

特殊标记 这里列出了所有添加到 Qwen2.5-VL 分词器词汇表中、并在我们的工作中用于构建表 8 中提示词的特殊标记。我们注意到，$<|im_start|>$、$<|im_end|>$、$<|vision_start|>$、$<|vision_end|>$、$<|image_pad|>$ 是 Qwen2.5-VL 分词器原生的特殊标记。

表 8：特殊标记说明

特殊标记	**说明**
$<\|start_of_event\|>$	标记事件的开始。
$<\|end_of_event\|>$	标记事件的结束。
$<\|time_start\|>$	标记时间戳的开始。
$<\|time_end\|>$	标记时间戳的结束。
$<\|type_start\|>$	标记事件类型的开始。
$<\|type_end\|>$	标记事件类型的结束。
$<\|text_start\|>$	标记事件文本的开始。
$<\|text_end\|>$	标记事件文本的结束。
$<\|time_prediction\|>$	用于预测下一个事件时间的任务标记。
$<\|type_prediction\|>$	用于下一事件类型预测的任务标记。
$<\|question\|>$	用于问答的任务标记。
$<\|byte_0\|>\cdots<\|byte_255\|>$	用于将事件时间间隔表示为单精度浮点数的字节令牌
	数字 ·
$<\|type_0\|>\cdots<\|type_8\|>$	用于表示事件类型的类型令牌。

事件模板 此处我们在图 5 中展示事件模板。

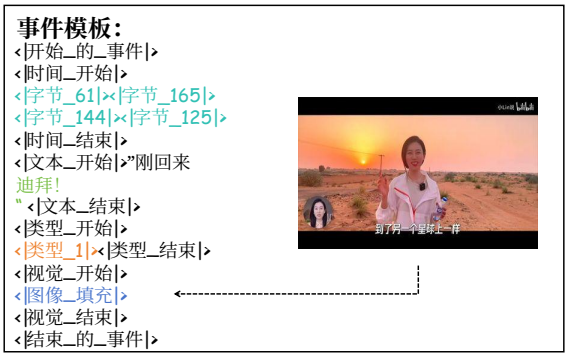


图5：未压缩事件表示的完整示例

提示模板 这里的表9提供了一个来自 TAXI-PRO 数据集的事件序列生成示例，该数据集用于第一阶段训练过程。我们还在表10和表11中分别提供了两个来自 TAXI-PRO 数据集的提示-响应对示例，用于第二阶段的下一事件类型微调和下一事件时间微调。所有三个模板的构建逻辑基本一致，仅系统提示略有不同。

表9：来自 TAXI-PRO 的事件序列示例

提示：
<|im_start|>系统 你是一个智能城市交通分析模型，专门分析纽约市的出租车模式。你的任务是预测曼哈顿下一次出租车事件发生前的精确时间间隔。你将获得出租车行程的历史记录，其中每个事件都包含其类型（例如，上城区上车，市中心下车），自上一个事件以来经过的时间，以及该地点的地图图像。<|im_end|>
事件序列：
<|start_of_event|>
<|time-start|><|byte-0|><|byte-0|><|byte-0|><|byte-0|><|time-end|>
<|type_start|><|type_0|><|type_end|><|text start|>在翠贝卡上车（40.711086,-74.016106），1 名乘客。<|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|>从翠贝卡（40.711086, -74.016106）前往时代广场（40.757698, -73.982124）下车，1 名乘客，行程 2.87 英里。<|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|>在上西区上车（40.799252, -73.970146），1 名乘客。<|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|>从上西区（40.799252, -73.970146）前往翠贝卡（40.714455, -74.014008）下车，1 名乘客，行程 4.37 英里。
<|text_end|><|vision start|><|image pad|><|vision end|><|end of event|>

表10：TAXI-PRO 中用于下一事件时间微调的提示-响应对示例

提示：
<|im_start|>系统
你是一个智能城市出行分析模型，专门分析纽约市的出租车模式。你的任务是预测曼哈顿下一次出租车事件发生前的精确时间间隔。你将获得一段出租车行程历史，其中每个事件都包含其类型（例如，上城区上车、下城下客）、距离上一个事件经过的时间，以及地点的地图图像。请分析事件类型和时间上的模式，预测下一个事件发生的时间。你的回复必须恰好由四个字节令牌组成，用来表示该时间间隔。
<|im_end|>
事件序列历史：
<|start_of_event|>
<|time_start|><|byte_0|><|byte_0|><|byte_0|><|byte_0|><|time_end|>
<|type_start|><|type_0|><|type_end|><|text start|>在翠贝卡接客（40.711086, -74.016106），1 名乘客。<|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| >从翠贝卡（40.711086, -74.016106）下车，前往时代广场（40.757698, -73.982124），1 位乘客，行程 2.87 英里。< |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| >接上 up at 上 West 西区（40.799252, -73.970146), 1 乘客。< |text end| >
< |vision_start| >< |image_pad| >< |vision_end| >
< |end_of_event| >
< |time_prediction| >

响应：
< |byte_63| >< |byte_17| >< |byte_16| >< |byte_161| >

表11：来自 TAXI-PRO 的下一事件类型微调提示-响应对示例

提示：
< |im-start| >系统
你是一个智能城市出行分析模型，专门分析纽约市的出租车模式。你的任务是根据过去一系列事件，预测曼哈顿的下一个出租车事件类型。你将获得出租车行程的历史记录，其中每个事件都包含其类型（例如上城区上车、下城下客）、距上一个事件经过的时间，以及该地点的地图图像。请分析时间和地点模式，预测下一个动作。你的回答必须是一个词元，代表六种事件类型中的一种。
< |im_end| >
事件序列历史：
< |start_of_event| >
< |time_start| >< |byte_0| >< |byte_0| >< |byte_0| >< |byte_0| >< |time_end| >
< |type_start| >< |type_0| >< |type_end| >
< |text-start| >在翠贝卡上车（40.711086, -74.016106），1 位乘客。< |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| >从翠贝卡（40.711086, -74.016106）下车，前往时代广场（40.757698, -73.982124），1 位乘客，行程 2.87 英里。< |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| >接客 up at 上城区 West 侧 (40.799252, -73.970146), 1 乘客.< |text end| >
< |vision_start| >< |image_pad| >< |vision_end| >
< |end_of_event| >
< |type_prediction| >

回应：
$< |type_1| >$

G 大语言模型的使用

在准备本稿件的过程中，我们使用大语言模型作为写作辅助工具。其应用严格限于改进论文的语言和表达，包括纠正语法错误、重新措辞以提升清晰度和简洁性，以及优化段落结构。核心科学贡献，包括研究构思、方法和分析，均由作者独立构思并完成。作者已审阅并编辑所有文本，并对论文的最终内容承担全部责任。