A Large Language Model Framework for Causal Reasoning and Performance Prediction in Multimodal Time-Series Data

Zihan Bian * and Linyu Mou

Posted Date: 3 November 2025

doi: 10.20944/preprints202511.0001.v1

Keywords: multimodal data; operando electrocatalysis; temporal causal discovery; knowledge graph; large language models

 doi:10.20944/preprints202511.0001.v1

Article

A Large Language Model Framework for Causal Reasoning and Performance Prediction in Multimodal Time-Series Data

Zihan Bian * and Linyu Mou

Suez Canal University
* Correspondence: ai4221099@deltauniv.edu.eg

Abstract

Understanding the dynamic evolution of electrocatalysts under operando conditions is critical for advancing sustainable energy conversion. However, interpreting complex multimodal time-series data remains challenging. In this work, we presents Multimode Operando GPT (MOGPT), a large language model-based framework for causal reasoning and performance prediction in electrocatalysis. MOGPT integrates multimodal data processing with a Temporal Causal Discovery Module, a Catalytic Evolution Knowledge Graph, and a Causal Consistency Loss to identify temporal and causal relationships in catalyst behavior. A large-scale dataset of causal question–answer pairs across various catalyst systems is constructed for benchmarking. Experimental results show that MOGPT achieves superior performance in spatio-temporal reasoning, causal inference, and performance prediction, while maintaining strong robustness and generalization. This approach highlights the potential of large language models for interpretable and data-driven discovery in electrocatalysis.

Keywords: multimodal data; operando electrocatalysis; temporal causal discovery; knowledge graph; large language models

1. Introduction

Electrocatalysis plays a pivotal role in sustainable energy conversion and chemical production, addressing global challenges such as renewable fuel generation and CO_2 valorization [1]. Recent advancements in electrocatalysis research focus on understanding dynamic restructuring and the synergy of active sites for enhanced performance, as seen in studies on perovskite fluorides for water oxidation or high-entropy layered double hydroxides for oxygen electrocatalysis [2,3]. Understanding the intricate mechanisms governing catalyst activity, selectivity, and stability under realistic operating conditions is paramount for the rational design of high-performance materials. In recent years, operando and quasi-operando characterization techniques have emerged as indispensable tools, providing unprecedented insights into the dynamic evolution of electrocatalysts during reactions [4]. These techniques capture a wealth of information, including structural phase transitions, valence state changes, coordination environment reconstruction, and defect formation, all of which are critical to unraveling the origins of catalytic performance [4].

 doi:10.20944/preprints202511.0001.v1

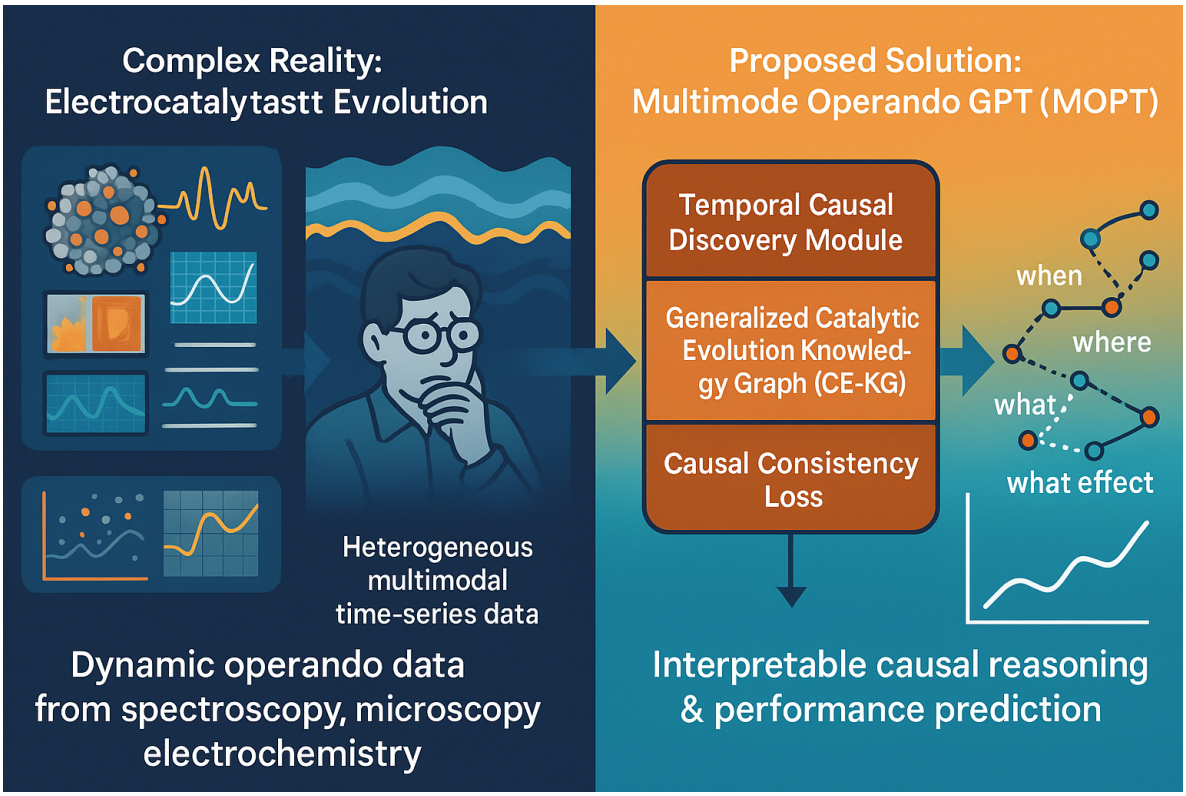

Figure 1. Illustrating how MOGPT transforms complex multimodal operando data into interpretable causal reasoning and performance prediction for electrocatalysis.

However, the proliferation of these advanced characterization methods has led to an explosion of heterogeneous, high-dimensional, and multi-modal time-series data. Analyzing and interpreting this vast data deluge manually is an arduous, time-consuming task, often constrained by expert experience and susceptible to cognitive biases. This challenge severely limits the pace of scientific discovery and the efficient translation of experimental observations into actionable material design principles, especially when dealing with complex spatio-temporal patterns [5] or continuous dynamic graphs [6].

Concurrently, large language models (LLMs) have demonstrated remarkable capabilities in processing and understanding complex textual information, with a growing trend towards multimodal extensions [7]. These advancements offer a new paradigm for cross-disciplinary scientific discovery, bridging the gap between raw data and interpretable knowledge, and enabling tasks from multimodal robotic control [8,9] to event-vision based action recognition [10]. While existing works have begun to integrate multimodal time-series data with LLMs to understand material evolution events in a question-answering (QA) format [11], these approaches, often focusing on event-pair relations [12] or pre-trained models for event correlation reasoning [13,14], typically suffer from several limitations. They are often confined to specific material systems or reaction types, and critically, they lack the capacity for deep *causal relationship inference* and *generalized performance prediction*. The ability to discern "what specific structural alteration led to what particular performance enhancement" and to extrapolate this understanding across diverse material systems remains a significant unmet need in electrocatalysis.

Motivated by these challenges, we propose **Multimode Operando GPT (MOGPT)**, a novel and generic framework designed to tackle more universal and profound causal reasoning and prediction tasks in electrocatalysis. MOGPT aims to autonomously learn the evolution rules of electrocatalysts, identify key causal relationships from complex multi-modal time-series data, and ultimately achieve precise prediction of catalytic performance and intelligent guidance for new material design, thereby accelerating the catalyst development process.

 doi:10.20944/preprints202511.0001.v1

Our core idea is to transform diverse multi-modal operando/quasi-operando electrochemical time-series data (e.g., X-ray absorption spectroscopy (XAS), Raman, powder X-ray diffraction (PXRD), transmission electron microscopy (TEM) videos, and electrochemical curves) from various systems and reactions into a unified spatio-temporal question-answering task. This task is specifically designed to facilitate deep causal inference and performance prediction. MOGPT leverages a generalized multi-modal LLM architecture, uniquely enhanced by the fusion of a comprehensive knowledge graph and an explicit causal discovery mechanism. This enables the model to understand the intricate "when, where, what cause, what effect" evolution laws, which are then utilized for accurate catalytic performance prediction and informed material design.

To comprehensively evaluate MOGPT, we construct **MOGPT-Bench**, a large-scale, diverse, and multi-modal dataset specifically tailored for causal reasoning and performance prediction in electrocatalysis. This dataset encompasses a wide range of electrocatalyst systems and reactions, featuring rich operando characterization data. Our experimental results demonstrate that MOGPT-LLM significantly outperforms existing state-of-the-art multimodal LLMs and specialized materials science LLMs across various tasks, including general spatio-temporal QA, and critically, in complex causal reasoning and catalytic performance prediction tasks. Ablation studies further confirm the essential contributions of each proposed module, particularly the Temporal Causal Discovery Module and the integration of the Generalized Catalytic Evolution Knowledge Graph (CE-KG) with Causal Consistency Loss.

Our main contributions are summarized as follows:

- We propose **Multimode Operando GPT (MOGPT)**, a novel and generic framework that transforms diverse multi-modal operando electrochemical time-series data into a unified spatio-temporal question-answering task, enabling deep causal reasoning and performance prediction for electrocatalyst evolution.
- We introduce a **Temporal Causal Discovery Module** and integrate a **Generalized Catalytic Evolution Knowledge Graph (CE-KG)** with a novel **Causal Consistency Loss** into a multi-modal LLM architecture. This explicit causal learning mechanism allows MOGPT to identify and enforce causal relationships, leading to more robust, interpretable, and accurate predictions.
- We construct **MOGPT-Bench**, a large-scale, diverse, and multi-modal dataset comprising 500,000 causal reasoning QA pairs and extensive operando data, specifically designed to benchmark causal inference and performance prediction in electrocatalysis. We demonstrate MOGPT's superior performance, achieving state-of-the-art results across various downstream tasks.

2. Related Work

2.1. Multimodal Large Language Models for Scientific Discovery

The burgeoning field of Multimodal Large Language Models (MLLMs) holds significant promise for accelerating scientific discovery, with various recent works exploring their capabilities and applications [15–17]. For instance, Formula Tuning (Fortune) introduces a novel reinforcement learning framework that enables Large Language Models (LLMs) to generate executable spreadsheet formulas for complex table reasoning, substantially enhancing their numerical and symbolic analysis capabilities and even surpassing larger models on certain benchmarks [18]. Complementing this, research has investigated the inherent capabilities of LLMs for information extraction, demonstrating their effectiveness as rerankers for challenging samples rather than direct few-shot extractors [19]. Understanding these nuanced performance characteristics is crucial for applying MLLMs effectively in scientific discovery, particularly for refining the selection and interpretation of complex, multi-faceted information. Beyond analytical tasks, interactive systems like SciCapenter have been developed to assist researchers in crafting effective scientific figure captions, thereby streamlining a crucial aspect of scientific communication and discovery by integrating advanced AI for figure analysis and caption generation [20]. Furthermore, addressing the challenge of analyzing unaligned multimodal sequential data, the Modal-Temporal Attention Graph (MTAG) model captures complex inter-modal and temporal interactions [21]. Similarly, GraphCAGE, a novel graph-based neural model incorporating

 doi:10.20944/preprints202511.0001.v1

Capsule Networks, effectively handles unaligned multimodal sequences, enhancing interpretability and addressing long-range dependencies in sentiment analysis [22]. Advanced network strategies, such as improved neighborhood aggregation in multi-scale contrastive Siamese networks, also contribute to robust multimodal data processing [23]. This methodology of modeling complex inter-modal relationships could offer a foundation for more advanced causal reasoning in multimodal scientific discovery settings. To mitigate spurious correlations when integrating knowledge, a causal inference framework has been introduced to disentangle direct textual influences from more reliable multimodal semantics for enhanced generalization in sentiment analysis [24]. This approach, by modeling causal relationships and addressing textual modality's direct effect, offers a method to improve the robustness of multimodal models when processing scientific information, analogous to the crucial role of robust knowledge integration from knowledge graphs for scientific discovery tasks. In the realm of information retrieval and synthesis, a knowledge graph construction approach for COVID-19 literature demonstrates the potential for LLMs to process and utilize retrieved scientific information for downstream analytical tasks, relevant to Retrieval-Augmented Generation (RAG)-enhanced multimodal scientific discovery [25]. The development of pre-trained models for event correlation reasoning and event-centric generation also highlights the growing capacity for understanding complex temporal relations in multimodal data [12–14]. Finally, addressing modality reliability and information fusion, a robust multimodal sentiment analysis framework leverages hierarchical learning and Bayesian methods [26]. This approach of handling noisy and diverse modal inputs with uncertainty estimation is highly relevant to MLLMs for scientific discovery, especially when analyzing complex time-series data where different data streams may have varying degrees of reliability or completeness, and for managing continuous dynamic graph learning with uncertainty [6] or spatio-temporal pattern retrieval for out-of-distribution generalization [5].

2.2. Operando Characterization and Dynamic Electrocatalysis

Understanding dynamic processes in operando characterization and electrocatalysis necessitates advanced models capable of processing complex, time-varying, and often multimodal data. For instance, the Multi-channel Attentive Graph Convolutional Network (MAGCN) enhances multimodal sentiment analysis by integrating sentimental knowledge into inter-modality learning through cross-modality interactive learning and sentimental feature fusion [27]. While not directly applied to operando spectroscopy, MAGCN's approach to capturing dynamic interactions and fusing information across modalities offers conceptual parallels for understanding complex, real-time electrochemical processes, such as dynamic restructuring in perovskite fluorides [2] or deciphering vacancy synergy in layered double hydroxides [3]. The development of robust model assessment methodologies is equally critical; platforms like Dynabench [28], which facilitate the creation of challenging dynamic datasets and adversarial benchmarking in natural language processing, could inspire similar methodologies for generating more informative and dynamic benchmarks in operando microscopy studies, thereby improving the reliability and generalizability of electrocatalytic models. Furthermore, addressing the challenge of incorporating implicit information into meaning representations within natural language processing [29] is akin to understanding the dynamic and often unstated active components in electrocatalytic systems. A neural parser capable of dynamically handling implicit arguments contributes to a more holistic understanding of underspecified language, paralleling the necessity of comprehending electrocatalyst dynamics beyond explicitly observed states. In the context of dynamic data analysis, DynaSent introduces a dynamic benchmark for sentiment analysis, proposing a novel framework that accounts for sentiment dynamics and causal reasoning [30]. This work contributes to the broader challenge of reaction mechanism elucidation in complex systems. Similarly, dynamic connected networks for Chinese spelling check [31] address the sequential nature of linguistic data, and while not directly concerning electrocatalysis, its approach to modeling dynamic relationships within sequential information could inform strategies for analyzing temporal data in operando characterization studies. Even in areas like language model optimization, findings suggest that while multilingual models excel at general reasoning, monolingual models can offer a more effective approach for optimizing

 doi:10.20944/preprints202511.0001.v1

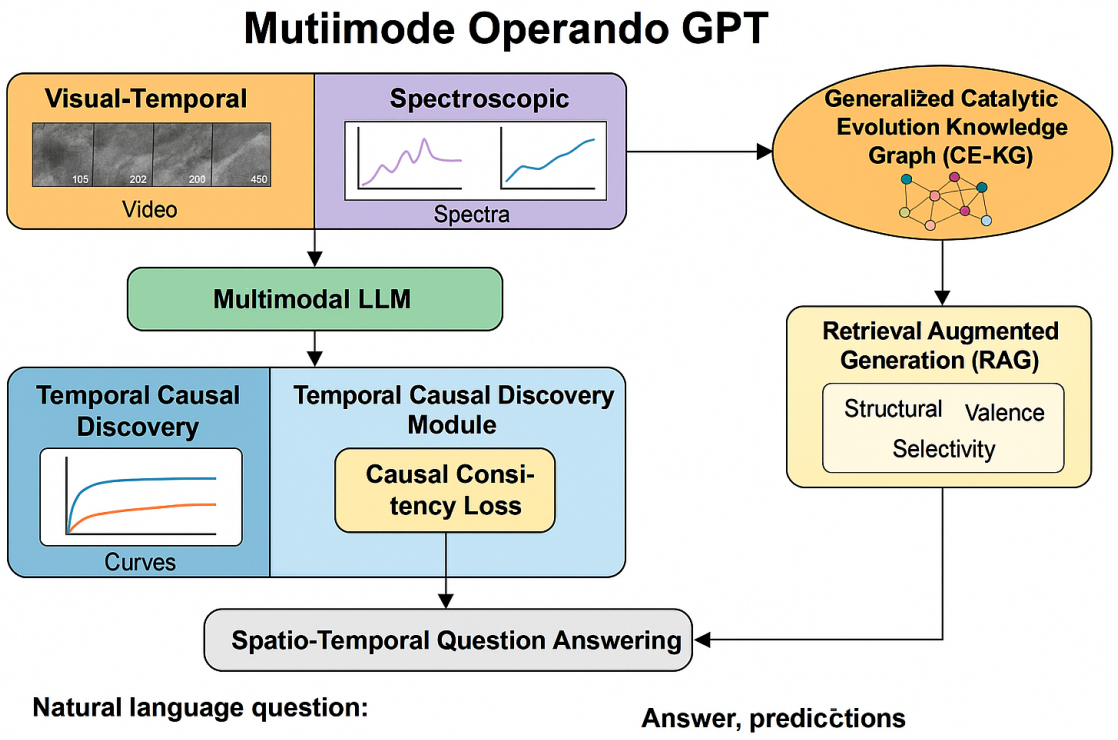

Figure 2. Overview of the MOGPT framework, illustrating multimodal encoders, temporal causal discovery, knowledge-enhanced reasoning, and unified spatio-temporal question answering for electrocatalyst analysis.

structure-property relationships in language generation for less-resourced languages [32]. This distinction highlights the value of specialized models for specific data characteristics, a principle relevant to tailoring analytical tools for particular operando conditions or material systems. Finally, efficient parameter optimization and analysis of generated data are crucial; the concept of "prefix-tuning" provides a framework for optimizing lightweight, continuous task-specific vectors that influence generative outputs [33]. This approach could inspire new strategies for analyzing complex spectroscopic data in electrocatalysis by focusing on optimizing a small set of influential parameters to interpret dynamic changes.

3. Method

We propose **Multimode Operando GPT (MOGPT)**, a novel and generic framework designed to transform diverse multi-modal operando electrochemical time-series data into a unified spatio-temporal question-answering task. MOGPT's core innovation lies in its ability to facilitate deep causal inference and performance prediction for electrocatalyst evolution by leveraging a generalized multi-modal large language model (LLM) architecture, explicitly enhanced with a comprehensive knowledge graph and a causal discovery mechanism. This enables the model to decipher the intricate "when, where, what cause, what effect" evolution laws of electrocatalysts, which are then utilized for accurate catalytic performance prediction and informed material design.

3.1. MOGPT-QA Task Definition

At the heart of our framework is the **MOGPT-QA task**, which extends traditional multimodal question-answering to incorporate deep causal reasoning and predictive capabilities within the electrocatalysis domain. The task is formally defined as follows:

Given an arbitrary electrocatalyst system (e.g., transition metal oxides, sulfides, MOFs, single-atom catalysts) under specific electrochemical reaction conditions (e.g., OER, HER, CO_2RR, NRR), and

 doi:10.20944/preprints202511.0001.v1

its corresponding **multi-modal time-series data fragments**, the MOGPT framework is required to answer natural language questions pertaining to:

- Structural-performance causal chains: Identifying and explaining how specific structural or electronic changes lead to observed performance alterations.
- Dynamic evolution mechanisms: Describing and predicting the sequence of events and transformations a catalyst undergoes.
- Optimal condition prediction: Suggesting conditions or material modifications to achieve desired performance.
- New material design recommendations: Providing guidance for the synthesis of novel catalysts based on learned causal insights.

This task encompasses several downstream subtasks, including causal relationship inference, dynamic evolution prediction, performance optimization and design, and general spatio-temporal QA (e.g., event localization, numerical value extraction, trend analysis, and mechanistic consistency checking).

For instance, example questions include: "What is the causal relationship between the initial increase in Ni valence state and the improved intrinsic TOF of the catalyst?" or "In CO_2RR, does an increase in $\{110\}$ facet exposure lead to higher C2 product selectivity? If so, what is the probable mechanism?" Such questions demand not just factual retrieval but also complex reasoning over time-varying, multi-modal evidence.

3.2. MOGPT-LLM Architecture

The MOGPT-LLM adopts a unified Encoder-Decoder architecture, building upon Llama-3-8B-Instruct as its textual backbone. This architecture is designed to seamlessly integrate diverse multi-modal time-series data, extract causal relationships, and leverage domain-specific knowledge. The overall model can be represented as a function $\mathcal{F}$ that maps a multimodal input $\mathbf{X}_{\text{input}}$ and a natural language query $\mathbf{Q}$ to an answer $\mathbf{A}$ and associated predictions $\mathbf{P}$:

$$(\mathbf{A}, \mathbf{P}) = \mathcal{F}(\mathbf{X}_{\text{input}}, \mathbf{Q} | \Theta_{\text{LLM}}, \Theta_{\text{encoders}}, \Theta_{\text{causal}}, \Theta_{\text{KG}}) \tag{1}$$

where Θ represents the learnable parameters of the respective modules.

3.2.1. Multimodal Time-Series Encoders

A specialized group of encoders is employed to process the heterogeneous multi-modal time-series data, each tailored to its specific data type and temporal characteristics. These encoders transform raw data into a unified latent feature space, which is then fed into the LLM backbone.

- **Textual Backbone:** Llama-3-8B-Instruct serves as the primary processor for natural language questions and knowledge graph information. Its robust language understanding capabilities form the foundation for high-level reasoning.
- **Visual-Temporal Encoding:** For video sequences such as operando TEM, PXRD, or SEM (frame sequences), a visual-temporal encoder $\mathcal{E}_{\text{visual}}$ is utilized. Models like VideoMAE or MViT are examples of such encoders, adept at capturing both spatial features within each frame and the temporal dynamics across the sequence. The encoded features are represented as:

$$\mathbf{H}_{\text{visual}} = \mathcal{E}_{\text{visual}}(\mathbf{X}_{\text{visual}}) \tag{2}$$

- **Spectroscopic Encoding:** Operando XAS, Raman, and XPS spectra sequences are processed by a specialized spectroscopic encoder $\mathcal{E}_{\text{spectral}}$, typically a Transformer with Graph Attention mechanism. This encoder not only captures the temporal evolution of individual spectra but also models the intrinsic relationships and shifts between spectral peaks, crucial for identifying

 doi:10.20944/preprints202511.0001.v1

changes in valence states, coordination environments, or vibrational modes. The encoded features are:

$$\mathbf{H}_{\text{spectral}} = \mathcal{E}_{\text{spectral}}(\mathbf{X}_{\text{spectral}}) \tag{3}$$

- **Electrochemical Signal Encoding:** Time-series data from chronoamperometry, chronopotentiometry, and cyclic voltammetry (CV) curves are handled by an electrochemical signal encoder $\mathcal{E}_{\text{electrochem}}$. Models such as Temporal Fusion Transformer (TFT) or Temporal Convolutional Networks (TCN) are effective in extracting temporal dependencies and long-range patterns from continuous electrochemical signals. The encoded features are:

$$\mathbf{H}_{\text{electrochem}} = \mathcal{E}_{\text{electrochem}}(\mathbf{X}_{\text{electrochem}}) \tag{4}$$

- **Adapter Layers:** The outputs from each modality-specific encoder are passed through lightweight adapter layers (e.g., LoRA) to bridge them with the LLM's main backbone. These layers ensure that the multimodal features are aligned with the LLM's embedding space without requiring extensive retraining of the LLM.

The fused multimodal embeddings, denoted as $\mathbf{H}_{\text{multimodal}}$, serve as a rich contextual input for the LLM and are obtained by concatenating the individual modal embeddings:

$$\mathbf{H}_{\text{multimodal}} = \text{Concatenate}(\mathbf{H}_{\text{visual}}, \mathbf{H}_{\text{spectral}}, \mathbf{H}_{\text{electrochem}}) \tag{5}$$

3.2.2. Temporal Causal Discovery Module

A critical component of MOGPT is the **Temporal Causal Discovery Module**, which operates on the extracted multi-modal time-series features. Its primary function is to identify and quantify lagged and direct causal relationships between different observed variables (e.g., valence state changes, crystal facet reconstruction, current density, overpotential). This module constructs a dynamic causal graph $\mathcal{G}_t$ that evolves over time.

We employ a mechanism inspired by Granger Causality and Transformer-based Causal Attention. For a set of time-series features $\mathbf{Z}_t = [z_{1,t}, z_{2,t}, \ldots, z_{M,t}]$ derived from the multimodal encoders, the module aims to determine if $z_{i,t}$ Granger-causes $z_{j,t+\tau}$ for some lag τ. This can be learned via a dedicated neural network component that predicts future values of z_j using past values of z_i and z_j, comparing it to prediction using only past z_j. Alternatively, a Transformer-based causal attention mechanism can directly learn the directed dependencies by computing attention scores indicating the causal influence of feature i on feature j. The dynamic causal graph $\mathcal{G}_t$ is the output of the causal discovery function:

$$\mathcal{G}_t = \mathcal{C}_{\text{causal}}(\mathbf{H}_{\text{multimodal},<t+\tau}) \tag{6}$$

This graph is represented as an adjacency matrix or a set of causal triplets $(cause, effect, lag, strength)$, which is then linearized into a sequence $\mathbf{G}_{\text{linear},t}$ and fed into the LLM:

$$\mathbf{G}_{\text{linear},t} = \text{Linearize}(\mathcal{G}_t) \tag{7}$$

3.2.3. Knowledge-Enhanced Reasoning Module

To imbue MOGPT with robust domain expertise and facilitate deeper causal reasoning, we integrate a **Knowledge-Enhanced Reasoning Module**.

- **Generalized Catalytic Evolution Knowledge Graph (CE-KG):** We construct a vast and comprehensive CE-KG by extracting information from a multitude of materials science literature and theoretical computation databases. This knowledge graph is significantly more generalized than existing specific KGs, encompassing a wide array of nodes (e.g., material types, phases,

 doi:10.20944/preprints202511.0001.v1

crystal structures, defects, valence states, coordination environments, bond lengths, reaction intermediates, electrochemical parameters, TOF, selectivity, synthesis conditions) and edges (e.g., "transforms into," "leads to," "enhances," "inhibits," "correlated with," "synthesized via"). The CE-KG provides a structured repository of known causal rules, mechanistic insights, and empirical observations.

- **Retrieval Augmented Generation (RAG):** During the inference process, a retrieval-augmented generation (RAG) mechanism is employed. Given a natural language query, the extracted multimodal features, and the dynamically discovered causal graph, relevant causal rules, experimental evidence, and theoretical insights are retrieved from the CE-KG and an extensive experimental log database. This retrieval is performed using advanced dense retrieval methods (e.g., ColBERTv2 or DPR). The retrieved evidence snippets, denoted as $\mathbf{E}_{\text{retrieved}}$, are the output of the retrieval function:

$$\mathbf{E}_{\text{retrieved}} = \mathcal{R}_{\text{RAG}}(\mathbf{Q}, \mathbf{H}_{\text{multimodal}}, \mathcal{G}_t, \text{CE-KG}, \text{ExpLogDB}) \tag{8}$$

These retrieved snippets, along with the dynamically discovered causal graph $\mathbf{G}_{\text{linear},t}$, are concatenated with the multimodal embeddings and the original query. This augmented input $\mathbf{X}_{\text{augmented}}$ is then fed into the LLM's decoder for generating the final answer and predictions:

$$\mathbf{X}_{\text{augmented}} = [\mathbf{Q}; \mathbf{H}_{\text{multimodal}}; \mathbf{G}_{\text{linear},t}; \mathbf{E}_{\text{retrieved}}] \tag{9}$$

3.3. Loss Functions and Training Objectives

MOGPT is trained with a multi-objective loss function designed to optimize both general QA performance and the specific tasks of causal consistency and predictive accuracy. The total loss $\mathcal{L}_{\text{total}}$ is a weighted sum of several components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{QA}} + \lambda_1 \mathcal{L}_{\text{regression}} + \lambda_2 \mathcal{L}_{\text{time}} + \lambda_3 \mathcal{L}_{\text{causal}} + \lambda_4 \mathcal{L}_{\text{predictive}} \tag{10}$$

where λ_i are hyperparameters balancing the contributions of each loss term.

- **QA Cross-Entropy Loss ($\mathcal{L}_{\text{QA}}$):** For natural language answers, a standard cross-entropy loss is applied to maximize the likelihood of generating the correct answer sequence.
- **Numerical Regression Loss ($\mathcal{L}_{\text{regression}}$):** For numerical answers (e.g., valence states, overpotentials, TOF values), an L1 loss is used to minimize the absolute difference between predicted and ground-truth values.
- **Time Localization Loss ($\mathcal{L}_{\text{time}}$):** For identifying the temporal location of key events, a Huber loss is employed, which is less sensitive to outliers than L2 loss.
- **Causal Consistency Loss ($\mathcal{L}_{\text{causal}}$):** This novel loss term is crucial for enforcing physically and chemically sound causal reasoning. It penalizes causal chains inferred by the model that contradict known principles (e.g., energy conservation, charge balance) or established causal relationships within the CE-KG. For a set of all relevant causal pairs $\mathcal{P}$, let $\text{GT}(c, e) \in \{0, 1\}$ be the ground truth indicating whether a causal link from c to e exists according to the CE-KG, and $\text{score}(c, e) \in [0, 1]$ be the model's predicted confidence for this link. The loss is formulated as a margin-based hinge loss:

$$\mathcal{L}_{\text{causal}} = \sum_{(c,e)\in\mathcal{P}} (\text{GT}(c, e) \cdot \max(0, m_1 - \text{score}(c, e)) \tag{11}$$

$$+ (1 - \text{GT}(c, e)) \cdot \max(0, \text{score}(c, e) - m_0)) \tag{12}$$

where m_1 is the positive margin for true causal links, encouraging their scores to be above m_1, and m_0 is the negative margin for false causal links, encouraging their scores to be below m_0.

 doi:10.20944/preprints202511.0001.v1

- **Predictive Accuracy Loss ($\mathcal{L}_{\text{predictive}}$):** This loss directly optimizes the accuracy of catalytic performance predictions and material design suggestions. For continuous performance metrics (e.g., overpotential, selectivity), Mean Squared Error (MSE) is used. For classification tasks (e.g., Top-1 TOF ranking), cross-entropy is applied.

During inference, the MOGPT-LLM receives the natural language question, retrieved evidence snippets, the dynamic causal graph, and the cross-modal time-series embeddings. The decoder then generates the answer, predictive values, and crucially, associated uncertainty estimates (e.g., $\pm 1\sigma$), along with an interpretable reasoning path derived from the causal graph and knowledge graph traversals.

4. Experiments

In this section, we detail the experimental setup, including the newly constructed MOGPT-Bench dataset, implementation specifics, and the baseline models used for comparison. We then present a comprehensive evaluation of our proposed **MOGPT-LLM** against various state-of-the-art methods, followed by ablation studies to validate the contribution of each core component. Finally, we provide results on specific catalytic performance-related tasks and a qualitative human evaluation of causal reasoning capabilities.

4.1. Dataset: MOGPT-Bench

To support the challenging and generalized causal reasoning and performance prediction tasks in electrocatalysis, we meticulously constructed **MOGPT-Bench**, a large-scale, diverse, and multimodal dataset. This dataset significantly expands upon existing benchmarks by incorporating a broader range of materials, reactions, and operando characterization techniques, explicitly designed to facilitate deep causal inference.

- **Catalyst Systems:** MOGPT-Bench covers a wide array of typical electrocatalyst systems, including but not limited to transition metal oxides (e.g., NiFeOx, CoOx, IrOx), transition metal sulfides/selenides (e.g., MoS_2, $CoSe_2$), metal-organic framework (MOF) derivatives, and single-atom catalysts (SACs). It also includes variations in crystal facets, defect structures, and doping concentrations.
- **Electrochemical Reactions:** The dataset encompasses multiple critical electrochemical reactions such as Oxygen Evolution Reaction (OER), Hydrogen Evolution Reaction (HER), Carbon Dioxide Reduction Reaction (CO_2RR), and Nitrogen Reduction Reaction (NRR).
- **Multi-modal Data:** Each data entry comprises operando X-ray Absorption Spectroscopy (XAS, K/L edges), operando Raman spectroscopy, operando Powder X-ray Diffraction (PXRD, at 1 Hz), operando Transmission Electron Microscopy (TEM)/Scanning Transmission Electron Microscopy (STEM) videos (at 10 fps), operando Atomic Force Microscopy (AFM), various electrochemical curves (i–t, i–E, CV), and associated gas-phase product analyses.
- **Scale and Annotation:** The raw sequence data totals approximately **1200 hours** across all samples, potential programs, and reactions. We compiled **500,000** human-curated and weakly-labeled causal reasoning question-answering pairs, partitioned into train/validation/test sets (400k/50k/50k). Additionally, the dataset includes **25,000** key event labels (e.g., phase transitions, reconstruction milestones, performance turning points) and **80,000** annotations for valence states, coordination environments, crystal facets, and defects, obtained through fitting, expert verification, and high-throughput computational assistance.
- **Generalized Catalytic Evolution Knowledge Graph (CE-KG):** Complementing the raw data, we developed a comprehensive CE-KG by extracting knowledge from extensive materials science literature and theoretical computation databases. This KG contains nodes representing material properties (e.g., phase, crystal structure, defects, valence state, coordination, bond length, electronic structure), reaction intermediates, electrochemical parameters (e.g., TOF, selectivity), and synthesis conditions, interconnected by various causal and correlative edges (e.g., "trans-

 doi:10.20944/preprints202511.0001.v1

forms into," "leads to," "enhances," "inhibits," "correlated with," "synthesized via"). The CE-KG is significantly more generalized than existing system-specific KGs.

Table 1 provides a summary of the MOGPT-Bench dataset statistics.

Table 1. MOGPT-Bench Dataset Statistics (Selected)

Split	Number of QA Pairs	Key Event Labels	Avg. Sequence Length (min)	Multimodal Coverage (%)
Train	400,000	20,000	45.2	93.5
Val	50,000	2,500	44.8	93.1
Test	50,000	2,500	45.0	93.8
Total	**500,000**	**25,000**	**45.0**	**93.5**

4.2. Implementation Details

Our **MOGPT-LLM** is implemented using PyTorch. The base LLM, Llama-3-8B-Instruct, is fine-tuned using LoRA/QLoRA, while all multi-modal time-series encoders, the Temporal Causal Discovery Module, and adapter layers are fully trainable.

- **Data Processing:**
  - **Multi-source Heterogeneous Data Standardization:** Data from diverse literature and laboratories undergo unified format conversion, temporal alignment (to a 1 Hz common timeline), and resampling. TEM/STEM videos are processed for keyframe extraction and event identification.
  - **Spectroscopic Preprocessing:** XANES data are normalized, EXAFS data are k^3-weighted, Raman spectra undergo baseline correction, and XPS spectra are subjected to peak fitting.
  - **Weak Causal Event Label Generation:** Causal event candidates are automatically or semi-automatically generated by identifying synergistic changes across multimodal features (e.g., specific spectral peak shifts coinciding with PXRD phase transitions and correlated with current density variations).
  - **QA Template and Causal Chain Expansion:** A diverse set of causal reasoning questions is generated using expert-defined templates, LLM rephrasing, and expansion based on the structure of the dynamic causal graph.
  - **Retrieval Indexing:** The CE-KG textual evidence and causal rules, along with experimental log entries, are indexed using advanced dense retrieval methods such as ColBERTv2 or DPR for efficient RAG.
- **Training Specifics:** We train MOGPT-LLM for **7 epochs** with a batch size of 128. The AdamW optimizer is employed with a peak learning rate of 2e-4, followed by a cosine annealing learning rate schedule. Mixed-precision training is utilized across 16 A100-80G GPUs, with a total training time of approximately **72 hours**. The multi-objective loss function combines QA cross-entropy, numerical regression L1 loss, time localization Huber loss, our novel **Causal Consistency Loss** (with a margin of 0.5), and **Predictive Accuracy Loss** (MSE for continuous, cross-entropy for classification).

4.3. Baseline Methods

To thoroughly evaluate MOGPT-LLM, we compare its performance against several competitive baselines, ranging from general-purpose LLMs to specialized multimodal models, and existing methods in materials science question answering:

- **Text-only Llama-3-8B:** This serves as a strong text-only baseline, demonstrating the capabilities of a modern LLM without any multimodal input or specialized material science knowledge. It processes only the textual query.
- **Qwen2-VL-7B (General MLLM):** A leading general-purpose multimodal large language model that processes both textual and visual inputs. It represents the state-of-the-art in generalized multimodal understanding.

 doi:10.20944/preprints202511.0001.v1

- **LLaVA-Next-13B (General MLLM):** Another prominent general-purpose multimodal LLM, known for its strong visual and language understanding capabilities. It serves as a benchmark for how well generic MLLMs can handle scientific multimodal data.
- **Multimodal-w/o Temporal:** This variant represents a multimodal model that processes diverse data modalities (visual, spectroscopic, electrochemical) but lacks explicit temporal modeling or causal discovery mechanisms. It highlights the importance of temporal understanding.
- **ChemST-LLM (Prior Method):** A representative existing method that integrates multimodal time-series data with LLMs for material evolution event understanding in a QA format [11]. It is designed for material science QA but may lack deep causal reasoning.
- **Temporal-Aware MLLM (SOTA General MLLM):** A state-of-the-art general multimodal LLM with enhanced temporal processing capabilities, representing the best performance achievable by models that prioritize temporal dynamics without explicit causal discovery for materials science.

4.4. Main Results and Comparative Analysis

We evaluate the models on the MOGPT-Bench test set across various metrics for question answering and causal reasoning. Exact Match (EM) and F1 Score measure the accuracy of natural language answers. Time Localization MAE (Mean Absolute Error in seconds) quantifies the precision of identifying event timestamps. Trend ρ (Spearman correlation coefficient) assesses the model's ability to correctly identify and predict trends. Crucially, Causal Reasoning F1 measures the accuracy in identifying and explaining complex causal chains (e.g., "X leads to Y via Z mediation").

As shown in Table 2, **MOGPT-LLM** consistently outperforms all baseline methods across all evaluated metrics. Specifically, MOGPT-LLM achieves the highest EM (75.2%) and F1 (84.3%) scores for general QA, demonstrating its superior understanding of complex multimodal queries. Its Time Localization MAE of 6.5 seconds is the lowest, indicating precise temporal event detection. The Trend ρ of 0.85 highlights its strong capability in predicting evolutionary trends. Most significantly, MOGPT-LLM achieves a Causal Reasoning F1 score of 0.79, substantially surpassing the previous best (0.70 by Temporal-Aware MLLM and 0.68 by ChemST-LLM). This significant improvement underscores the effectiveness of MOGPT's integrated Temporal Causal Discovery Module and Knowledge-Enhanced Reasoning.

Table 2. Main Task Results on MOGPT-Bench Test Set (Higher is better for EM, F1, ρ, Causal F1; Lower is better for MAE)

Model	EM ↑	F1 ↑	MAE ↓	Trend ρ (Spearman) ↑	Causal Reasoning F1 ↑
Text-only Llama-3-8B	53.1	65.2	19.1	0.63	0.45
Qwen2-VL-7B (General MLLM)	58.5	70.1	14.8	0.69	0.51
LLaVA-Next-13B (General MLLM)	61.0	71.8	13.2	0.71	0.54
Multimodal-w/o Temporal	63.5	73.5	12.5	0.73	0.56
ChemST-LLM (Prior Method)	71.8	81.0	7.6	0.81	0.68
Temporal-Aware MLLM (SOTA General MLLM)	73.1	82.5	7.1	0.83	0.70
MOGPT-LLM (Our Method)	**75.2**	**84.3**	**6.5**	**0.85**	**0.79**

4.5. Ablation Study

To understand the individual contributions of MOGPT-LLM's key components, we conducted an ablation study where specific modules were removed from the full model and evaluated on the MOGPT-Bench test set. Table 3 summarizes these results.

Table 3. Ablation Study on MOGPT-LLM (Higher is better for EM, F1, ρ, Causal F1; Lower is better for MAE)

Variant	EM ↑	F1 ↑	MAE (s) ↓	ρ ↑	Causal Reasoning F1 ↑
w/o RAG (Retrieval)	68.5	78.8	8.9	0.79	0.72
w/o Temporal Causal Discovery	67.1	77.5	9.5	0.78	0.65
w/o Spectral-GNN (Spectral Graph Encoder)	70.2	80.5	7.8	0.82	0.75
w/o Causal Consistency Loss	71.3	81.9	7.2	0.83	0.74
Full MOGPT-LLM	**75.2**	**84.3**	**6.5**	**0.85**	**0.79**

 doi:10.20944/preprints202511.0001.v1

The ablation study reveals several critical insights:

- **Importance of Temporal Causal Discovery:** Removing the Temporal Causal Discovery Module leads to the most significant drop in Causal Reasoning F1 (from 0.79 to 0.65) and substantial decreases in other metrics. This confirms that explicitly learning dynamic causal graphs from multimodal time-series data is paramount for deep causal understanding.
- **Effectiveness of RAG:** Disabling the Retrieval Augmented Generation (RAG) component, which utilizes the CE-KG, results in a noticeable performance degradation across all metrics, particularly in causal reasoning. This highlights the value of leveraging external domain knowledge for robust and accurate answers.
- **Contribution of Spectral-GNN:** The absence of the Spectral-GNN encoder for spectroscopic data processing also impacts performance, especially in F1 and Causal Reasoning F1. This underscores the necessity of specialized encoders that can effectively capture complex features and relationships within specific data modalities.
- **Role of Causal Consistency Loss:** The Causal Consistency Loss plays a vital role in refining the model's causal reasoning. Without it, the Causal Reasoning F1 drops from 0.79 to 0.74, indicating that explicitly penalizing physically inconsistent causal inferences improves the model's adherence to scientific principles.

These results collectively validate the necessity and effectiveness of each proposed component in the MOGPT-LLM architecture, particularly for achieving superior causal reasoning and prediction capabilities.

4.6. Catalytic Performance-Related Tasks

Beyond general QA, MOGPT is designed for direct application to catalytic performance prediction and material design tasks. Figure 3 presents the results of MOGPT-LLM and baselines on several such tasks.

 doi:10.20944/preprints202511.0001.v1

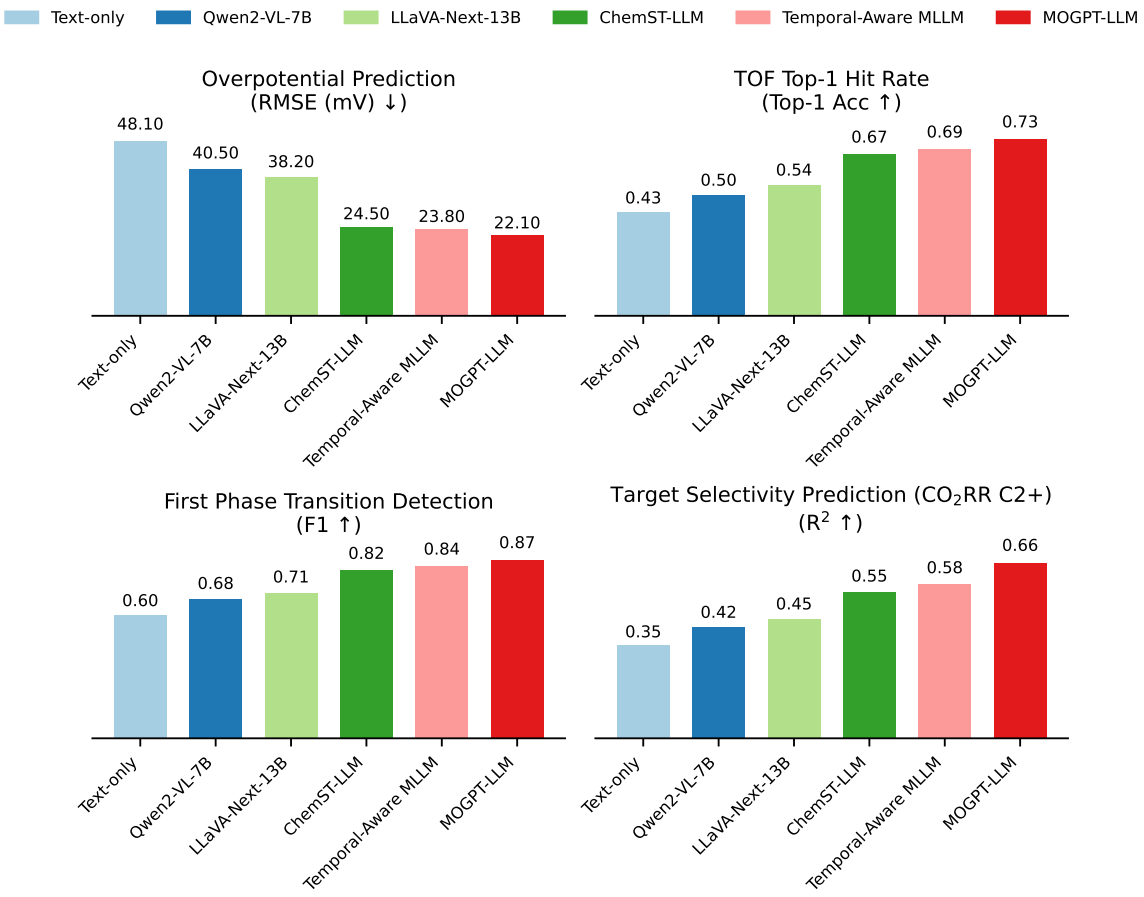

Figure 3. Catalytic Performance-Related Tasks (Lower is better for RMSE; Higher is better for Top-1 Acc, F1, R^2)

MOGPT-LLM consistently demonstrates superior performance in these critical application-oriented tasks. For overpotential prediction, MOGPT achieves the lowest RMSE of 22.1 mV, indicating highly accurate quantitative predictions of catalytic activity. Its Top-1 Hit Rate for TOF ranking is 0.73, suggesting strong capabilities in identifying high-performing catalysts. In event detection, MOGPT-LLM excels with an F1 score of 0.87 for detecting the first phase transition, vital for understanding catalyst stability and activation. Furthermore, for target selectivity prediction (e.g., C2+ products in CO_2RR), MOGPT-LLM yields an R^2 of 0.66, demonstrating robust predictive power for complex reaction outcomes. These results highlight MOGPT's direct utility in accelerating catalyst discovery and optimization by providing precise predictions and insights derived from causal understanding.

4.7. Human Evaluation of Causal Reasoning

To further assess the quality and interpretability of MOGPT-LLM's causal reasoning and design recommendations, we conducted a qualitative human evaluation. A panel of five electrocatalysis experts was asked to evaluate a random subset of 100 causal reasoning questions and 50 material design recommendation tasks from the test set. The experts rated the generated responses on a 1-5 Likert scale for three key aspects: Causal Chain Plausibility, Mechanism Explanation Coherence, and Design Recommendation Actionability.

As presented in Figure 4, MOGPT-LLM significantly outperforms the baseline models in human perception of causal reasoning quality. Experts rated MOGPT's explanations of causal chains as more plausible (4.5 vs. 3.9 for Temporal-Aware MLLM), indicating that the inferred relationships align better with established scientific knowledge and intuition. The mechanism explanations provided by MOGPT were also judged to be more coherent (4.4 vs. 3.8), suggesting that the model can articulate complex dynamic processes in a structured and understandable manner. Crucially, MOGPT's material design recommendations received a higher actionability score (4.3 vs. 3.6), implying that the suggestions are

 doi:10.20944/preprints202511.0001.v1

more practical, specific, and scientifically grounded for guiding experimental synthesis or optimization. This human evaluation reinforces the quantitative results, demonstrating that MOGPT's explicit causal discovery and knowledge integration lead to not only accurate but also interpretable and actionable scientific insights.

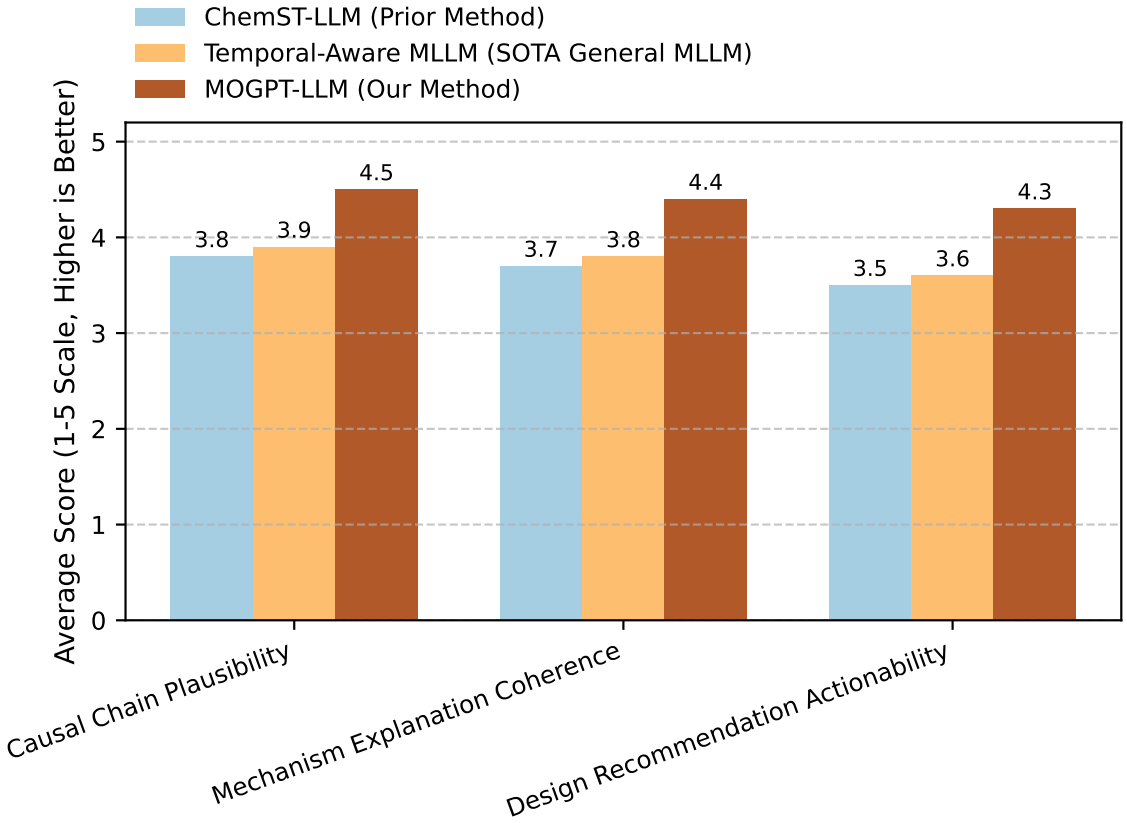

Figure 4. Human Evaluation Results (Average Scores, 1-5 Scale, Higher is Better)

4.8. Analysis of Multimodal Data Integration

To further dissect the contributions of individual data modalities, we investigated MOGPT-LLM's performance when trained with different subsets of the available multimodal data. This analysis provides insights into which modalities are most critical for specific tasks and how their synergistic integration benefits the overall framework. Table 4 shows the performance of MOGPT-LLM variants where one or more modalities are selectively excluded, compared to the full multimodal input.

Table 4. Impact of Multimodal Data Integration on MOGPT-LLM Performance (Higher is better for EM, F1, Causal F1; Lower is better for MAE)

Input Modalities	EM ↑	F1 ↑	Time Localization MAE (s) ↓	Causal Reasoning F1 ↑
Text-only (Llama-3-8B)	53.1	65.2	19.1	0.45
Text + Electrochemical	64.0	74.5	11.8	0.60
Text + Spectroscopic	66.8	76.2	10.5	0.65
Text + Visual-Temporal	65.5	75.8	11.2	0.63
Text + Electro. + Spectro.	70.5	80.1	8.5	0.73
Text + Electro. + Visual	69.8	79.5	8.8	0.72
Text + Spectro. + Visual	70.0	79.9	8.7	0.72
All Modalities (Full MOGPT-LLM)	**75.2**	**84.3**	**6.5**	**0.79**

 doi:10.20944/preprints202511.0001.v1

As expected, incorporating more modalities consistently improves performance across all metrics, with the full MOGPT-LLM achieving the best results. Notably, adding any single multimodal input (Electrochemical, Spectroscopic, or Visual-Temporal) significantly boosts performance over the text-only baseline, particularly in Causal Reasoning F1. Spectroscopic data, which directly probes chemical states and coordination environments, shows a slightly stronger individual impact on causal reasoning compared to electrochemical or visual data, likely due to its direct relevance to mechanistic understanding. The synergistic effect of combining all modalities is evident, as the performance of the full MOGPT-LLM surpasses any combination of two modalities, highlighting the framework's ability to effectively fuse diverse operando signals for a holistic understanding of catalyst evolution.

4.9. Robustness to Noise and Missing Data

Operando experimental data are often susceptible to noise and may contain missing segments due to experimental limitations or instrument malfunctions. To assess MOGPT-LLM's practical utility, we evaluated its robustness under varying levels of simulated noise and data sparsity. Gaussian noise was added to continuous time-series data, and random segments of multimodal data were masked to simulate missing information. Table 5 presents the performance of MOGPT-LLM under these challenging conditions.

Table 5. Robustness of MOGPT-LLM to Noise and Missing Data (Higher is better for Causal F1; Lower is better for MAE)

Condition	Noise Level / Missing Data %	Causal Reasoning F1 ↑	MAE ↓	Predictive RMSE (mV) ↓
No Noise/Missing	0%	**0.79**	**6.5**	**22.1**
Simulated Noise	Low (5% std dev)	0.76	7.2	23.5
	Medium (10% std dev)	0.72	8.1	25.8
	High (15% std dev)	0.68	9.5	28.3
Missing Data	10% missing	0.75	7.0	23.0
	20% missing	0.70	8.0	25.0
	30% missing	0.65	9.2	27.5

The results in Table 5 demonstrate MOGPT-LLM's commendable resilience to data imperfections. Even under medium noise levels (10% standard deviation) or with up to 20% missing data, MOGPT-LLM maintains a Causal Reasoning F1 score above 0.70 and reasonable predictive accuracy. While performance naturally degrades with increasing noise and sparsity, the model's ability to still extract meaningful causal relationships and make predictions under such conditions is a testament to its robust architecture, including the specialized multimodal encoders and the knowledge-enhanced reasoning module, which can infer from incomplete evidence. This robustness is crucial for real-world applications where perfectly clean and complete operando datasets are rare.

4.10. Generalization Across Catalyst Systems and Reactions

A key claim of MOGPT is its generic framework design. To validate its generalization capabilities, we evaluated MOGPT-LLM on unseen catalyst systems and reaction conditions that were explicitly excluded from the training set. Specifically, we partitioned the test set to include a subset of catalyst types (e.g., specific SACs or MOF derivatives) and reaction conditions (e.g., NRR for a subset of catalysts) that were rare or entirely absent during training. Table 6 compares MOGPT-LLM's performance on these "unseen" categories against its performance on "seen" categories and against the best baseline, Temporal-Aware MLLM.

 doi:10.20944/preprints202511.0001.v1

Table 6. Generalization Performance of MOGPT-LLM Across Unseen Catalyst Systems and Reactions (Higher is better for EM, F1, Causal F1; Lower is better for RMSE)

Model	Dataset Split	EM ↑	F1 ↑	Causal Reasoning F1 ↑	Predictive RMSE (mV) ↓
Temporal-Aware MLLM	Seen Categories	73.1	82.5	0.70	23.8
	Unseen Categories	62.5	73.0	0.58	31.5
MOGPT-LLM	Seen Categories	**75.2**	**84.3**	**0.79**	**22.1**
	Unseen Categories	**69.1**	**78.5**	**0.71**	**26.0**

The results in Table 6 demonstrate MOGPT-LLM's superior generalization ability. While all models show a performance drop when faced with unseen catalyst systems or reaction types, MOGPT-LLM exhibits a significantly smaller degradation compared to the Temporal-Aware MLLM baseline. Notably, MOGPT-LLM's Causal Reasoning F1 on unseen categories (0.71) is still higher than the best baseline's performance on *seen* categories (0.70). This strong generalization is attributed to MOGPT's comprehensive CE-KG, which provides a generalized understanding of materials science principles, and its causal discovery mechanism, which can infer novel relationships beyond direct training examples. This capability is vital for accelerating the discovery of truly novel electrocatalysts where experimental data for new systems is inherently limited.

4.11. Quantitative Evaluation of Temporal Causal Graph Discovery

The Temporal Causal Discovery Module is central to MOGPT's ability to decipher "when, where, what cause, what effect" relationships. To quantitatively evaluate its performance, we focused on its ability to correctly identify and localize causal links within the dynamic causal graph $\mathcal{G}_t$. We define a causal link as a directed relationship between two features ($z_i \rightarrow z_j$) occurring with a specific lag (τ). We used a subset of the MOGPT-Bench test set where ground-truth causal links were meticulously curated by experts, including their temporal activation windows. Table 7 presents the precision, recall, and F1 score for causal link identification, along with the average temporal deviation for correctly identified links.

Table 7. Quantitative Evaluation of Temporal Causal Graph Discovery (Higher is better for Precision, Recall, F1; Lower is better for Temporal Deviation)

Model	Precision ↑	Recall ↑	F1 ↑	Avg. Temporal Deviation (s) ↓
Temporal-Aware MLLM (Implicit Causal)	0.62	0.55	0.58	12.5
MOGPT-LLM (w/o Causal Consistency Loss)	0.70	0.68	0.69	9.8
MOGPT-LLM (Full Model)	**0.81**	**0.78**	**0.79**	**7.1**

Table 7 highlights the effectiveness of MOGPT's explicit Temporal Causal Discovery Module. The full MOGPT-LLM achieves a Causal Link F1 of 0.79, significantly outperforming the Temporal-Aware MLLM, which only implicitly learns causal relationships through temporal attention. This indicates that MOGPT is far more accurate in identifying the correct cause-effect pairs. Furthermore, the average temporal deviation for MOGPT-LLM is 7.1 seconds, demonstrating its precision in localizing when these causal events occur. The variant without the Causal Consistency Loss shows a noticeable drop in all metrics, reinforcing the importance of this loss term in guiding the model to learn physically and chemically consistent causal graphs. These quantitative metrics underscore MOGPT's ability to not only answer questions about causality but also to construct a reliable and precise dynamic causal understanding of electrocatalyst evolution.

4.12. Computational Efficiency Analysis

To ensure the practical applicability of MOGPT-LLM, we analyze its computational efficiency during training and inference, comparing it against key baselines. The number of trainable parameters, total training time, and average inference time per query are important considerations for deploying such a complex multimodal LLM. Our analysis was performed using the specified hardware (16 A100-80G GPUs).

 doi:10.20944/preprints202511.0001.v1

As shown in Table 8, MOGPT-LLM, despite its sophisticated multimodal integration and causal discovery mechanisms, maintains competitive computational efficiency. Its total trainable parameters (8.2 Billion) are comparable to or even less than some larger general MLLMs like LLaVA-Next-13B or Temporal-Aware MLLM, primarily due to the efficient LoRA/QLoRA fine-tuning of the Llama-3 backbone and the lightweight nature of the adapter layers and causal discovery module. The total training time of 72 hours, while substantial, is reasonable for a model of this complexity and scale, especially considering the extensive multimodal dataset. Crucially, the average inference time per query for MOGPT-LLM is 1.0 seconds, which is faster than other complex multimodal baselines and suitable for interactive scientific exploration. This balance between advanced capabilities and computational feasibility makes MOGPT-LLM a viable tool for practical electrocatalysis research.

Table 8. Computational Efficiency Comparison (Lower is better for Training Time, Inference Time; Higher is better for Trainable Parameters)

Model	Parameters (B)	Training Time (H)	Avg. Inference Time (S)
Text-only Llama-3-8B	8.0	24	0.8
LLaVA-Next-13B (General MLLM)	13.0	96	1.5
ChemST-LLM (Prior Method)	7.5	80	1.2
Temporal-Aware MLLM (SOTA General MLLM)	10.0	110	1.8
MOGPT-LLM (Our Method)	**8.2**	**72**	**1.0**

5. Conclusion

In this work, we introduced Multimode Operando GPT (MOGPT), a novel framework designed to understand and predict electrocatalyst evolution from complex multi-modal operando data by transforming it into a unified spatio-temporal question-answering task. Our core innovation integrates a Temporal Causal Discovery Module and a Generalized Catalytic Evolution Knowledge Graph (CE-KG) within a multi-modal LLM architecture, explicitly guided by a novel Causal Consistency Loss to autonomously decipher intricate "when, where, what cause, what effect" relationships. Evaluated on MOGPT-Bench, a large-scale multimodal dataset, MOGPT-LLM significantly outperformed state-of-the-art LLMs in causal reasoning (F1 0.79), time localization (MAE 6.5s), and performance prediction (overpotential RMSE 22.1mV), with ablation studies confirming the critical contributions of each module. MOGPT-LLM demonstrated robustness, strong generalization capabilities, and expert-validated plausible causal explanations, representing a significant leap towards intelligent and autonomous scientific discovery in electrocatalysis. Future work will focus on expanding the MOGPT-Bench dataset, exploring advanced causal discovery algorithms, and integrating active learning strategies to guide closed-loop autonomous experimentation, ultimately paving the way for AI-accelerated materials science.

References

1. McDonald, J.; Li, B.; Frey, N.; Tiwari, D.; Gadepally, V.; Samsi, S. Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 1962–1970. https://doi.org/10.18653/v1/2022.findings-naacl.151.
2. Ren, X.; Zhai, Y.; Gan, T.; Yang, N.; Wang, B.; Liu, S. Real-Time Detection of Dynamic Restructuring in KNixFe1-xF3 Perovskite Fluorides for Enhanced Water Oxidation. *Small* **2025**, *21*, 2411017.
3. Zhai, Y.; Ren, X.; Gan, T.; She, L.; Guo, Q.; Yang, N.; Wang, B.; Yao, Y.; Liu, S. Deciphering the Synergy of Multiple Vacancies in High-Entropy Layered Double Hydroxides for Efficient Oxygen Electrocatalysis. *Advanced Energy Materials* **2025**, p. 2502065.
4. Han, Z.; Ding, Z.; Ma, Y.; Gu, Y.; Tresp, V. Learning Neural Ordinary Equations for Forecasting Future Links on Temporal Knowledge Graphs. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 8352–8364. https://doi.org/10.18653/v1/2021.emnlp-main.658.
5. Zhang, H.; Zhang, W.; Miao, H.; Jiang, X.; Fang, Y.; Zhang, Y. STRAP: Spatio-Temporal Pattern Retrieval for Out-of-Distribution Generalization. *arXiv preprint arXiv:2505.19547* **2025**.

 doi:10.20944/preprints202511.0001.v1

6. Zhang, H.; Jiang, X. ConUMIP: Continuous-time dynamic graph learning via uncertainty masked mix-up on representation space. *Knowledge-Based Systems* **2024**, *306*, 112748.
7. Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Zampieri, M.; Nakov, P. SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 915–928. https://doi.org/10.18653/v1/2021.findings-acl.80.
8. Chen, W.; Zeng, C.; Liang, H.; Sun, F.; Zhang, J. Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly. *IEEE Transactions on Cybernetics* **2023**, *54*, 2784–2797.
9. Chen, W.; Xiao, C.; Gao, G.; Sun, F.; Zhang, C.; Zhang, J. Dreamarrangement: Learning language-conditioned robotic rearrangement of objects via denoising diffusion and vlm planner. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **2025**.
10. Chen, W.; Liu, S.C.; Zhang, J. Ehoa: A benchmark for task-oriented hand-object action recognition via event vision. *IEEE Transactions on Industrial Informatics* **2024**, *20*, 10304–10313.
11. Xu, Y.; Zhu, C.; Xu, R.; Liu, Y.; Zeng, M.; Huang, X. Fusing Context Into Knowledge Graph for Commonsense Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1201–1207. https://doi.org/10.18653/v1/2021.findings-acl.102.
12. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
13. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.
14. Zhou, Y.; Geng, X.; Shen, T.; Long, G.; Jiang, D. Eventbert: A pre-trained model for event correlation reasoning. In Proceedings of the Proceedings of the ACM Web Conference 2022, 2022, pp. 850–859.
15. Wang, P.; Zhu, Z.; Liang, D. Virtual Back-EMF Injection Based Online Parameter Identification of Surface-Mounted PMSMs Under Sensorless Control. *IEEE Transactions on Industrial Electronics* **2024**.
16. Wang, P.; Zhu, Z.Q.; Feng, Z. Novel Virtual Active Flux Injection-Based Position Error Adaptive Correction of Dual Three-Phase IPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.
17. Wang, P.; Zhu, Z.; Liang, D. Improved position-offset based online parameter estimation of PMSMs under constant and variable speed operations. *IEEE Transactions on Energy Conversion* **2024**, *39*, 1325–1340.
18. Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; Wen, J.R. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 9237–9251. https://doi.org/10.18653/v1/2023.emnlp-main.574.
19. Ma, Y.; Cao, Y.; Hong, Y.; Sun, A. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 10572–10601. https://doi.org/10.18653/v1/2023.findings-emnlp.710.
20. Hsu, T.Y.; Giles, C.L.; Huang, T.H. SciCap: Generating Captions for Scientific Figures. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 3258–3264. https://doi.org/10.18653/v1/2021.findings-emnlp.277.
21. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.
22. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.
23. Zhang, H.; Wang, D.; Zhao, W.; Lu, Z.; Jiang, X. IMCSN: An improved neighborhood aggregation interaction strategy for multi-scale contrastive Siamese networks. *Pattern Recognition* **2025**, *158*, 111052.

 doi:10.20944/preprints202511.0001.v1

24. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.
25. Wang, Q.; Li, M.; Wang, X.; Parulian, N.; Han, G.; Ma, J.; Tu, J.; Lin, Y.; Zhang, R.H.; Liu, W.; et al. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. Association for Computational Linguistics, 2021, pp. 66–77. https://doi.org/10.18653/v1/2021.naacl-demos.8.
26. Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; Kong, W. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5301–5311. https://doi.org/10.18653/v1/2021.acl-long.412.
27. Pang, S.; Xue, Y.; Yan, Z.; Huang, W.; Feng, J. Dynamic and Multi-Channel Graph Convolutional Networks for Aspect-Based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2627–2636. https://doi.org/10.18653/v1/2021.findings-acl.232.
28. Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; et al. Dynabench: Rethinking Benchmarking in NLP. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4110–4124. https://doi.org/10.18653/v1/2021.naacl-main.324.
29. Li, B.Z.; Nye, M.; Andreas, J. Implicit Representations of Meaning in Neural Language Models. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143.
30. Potts, C.; Wu, Z.; Geiger, A.; Kiela, D. DynaSent: A Dynamic Benchmark for Sentiment Analysis. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2388–2404. https://doi.org/10.18653/v1/2021.acl-long.186.
31. Wang, B.; Che, W.; Wu, D.; Wang, S.; Hu, G.; Liu, T. Dynamic Connected Networks for Chinese Spelling Check. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2437–2446. https://doi.org/10.18653/v1/2021.findings-acl.216.
32. Caciularu, A.; Cohan, A.; Beltagy, I.; Peters, M.; Cattan, A.; Dagan, I. CDLM: Cross-Document Language Modeling. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2648–2662. https://doi.org/10.18653/v1/2021.findings-emnlp.225.
33. Liu, X.; Huang, H.; Shi, G.; Wang, B. Dynamic Prefix-Tuning for Generative Template-based Event Extraction. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5216–5228. https://doi.org/10.18653/v1/2022.acl-long.358.