ChemST-LLM: A Multi-Modal Spatiotemporal Question-Answering System for Dynamic Defect-Performance Synergy in Catalysts

Zechen Chu * and Ruotong Liao

Posted Date: 28 October 2025

doi: 10.20944/preprints202510.1977.v1

Keywords: large language models; oxygen evolution reaction; spatiotemporal data

 doi:10.20944/preprints202510.1977.v1

Article

ChemST-LLM: A Multi-Modal Spatiotemporal Question-Answering System for Dynamic Defect-Performance Synergy in Catalysts

Zechen Chu * and Ruotong Liao

Zhongnan University of Economics and Law, China
* Correspondence: 202131042419@stu.zuel.edu.cn

Abstract

The development of high-performance electrocatalysts for reactions such as the oxygen evolution reaction (OER) is limited by the complex, dynamic evolution of defects under operando conditions. Existing methods and large language models (LLMs) face challenges in integrating and interpreting the diverse spatiotemporal data generated by advanced characterization techniques. To overcome this, we propose ChemST-LLM, a multi-modal spatiotemporal question-answering framework designed to understand and explain the relationship between defect evolution and OER performance in high-entropy layered double hydroxides (LDHs). ChemST-LLM features specialized temporal and graph encoders, a gated cross-modal fusion module, and an instruction-tuned LLM core with retrieval-augmented generation. We construct ChemVac-STQA, a large-scale dataset of experimental trajectories and spatiotemporal QA pairs, to support training and evaluation. Experimental results show that ChemST-LLM outperforms strong baselines across spatiotemporal QA tasks, achieving higher accuracy and robustness, particularly in challenging out-of-distribution scenarios. This work demonstrates a new paradigm for data-driven understanding and accelerated discovery of advanced electrocatalysts.

Keywords: large language models; oxygen evolution reaction; spatiotemporal data

1. Introduction

The efficient and stable electrocatalytic oxygen evolution reaction (OER) is a cornerstone for numerous clean energy technologies, including water electrolysis for hydrogen production and rechargeable metal-air batteries [1]. Developing high-performance OER catalysts is thus a paramount scientific and engineering challenge. Layered double hydroxides (LDHs), particularly in high-entropy multi-metallic systems, have emerged as highly promising materials due to their unique tunable structures and excellent catalytic activity [2,3]. However, the intrinsic catalytic activity and stability of these materials are profoundly influenced by the formation, migration, and synergistic interactions of various defects (e.g., oxygen vacancies, metal vacancies, interstitial cations) that dynamically evolve under operando conditions [4]. Capturing and understanding these intricate, multi-scale spatiotemporal evolutions of defects, including dynamic restructuring observed in real-time [5], is critical for rational catalyst design.

Traditional research methodologies often fall short in comprehensively unraveling these dynamic processes. While advanced *operando* characterization techniques, such as X-ray absorption spectroscopy (XAS), X-ray diffraction (XRD), scanning transmission electron microscopy (STEM), and electrochemical measurements, provide invaluable real-time, atomic-level insights [6], the data generated are inherently multi-modal, highly complex, and sequential. Integrating and interpreting such diverse spatiotemporal data streams, especially considering dynamic graph structures [7] and the need for out-of-distribution generalization [8], poses significant analytical challenges. Although large language models (LLMs) have demonstrated remarkable capabilities in processing and understanding scientific text [9], their current frameworks are not inherently designed to seamlessly fuse

 doi:10.20944/preprints202510.1977.v1

and comprehend complex multi-modal spatiotemporal sequence data, nor to perform deep chemical mechanism reasoning and performance prediction based on such inputs. This gap necessitates the development of an intelligent AI system capable of efficiently integrating multi-modal spatiotemporal data to enable intelligent querying, prediction, and explanation of the dynamic defect-performance synergy in catalysts, thereby accelerating catalyst discovery and optimization.

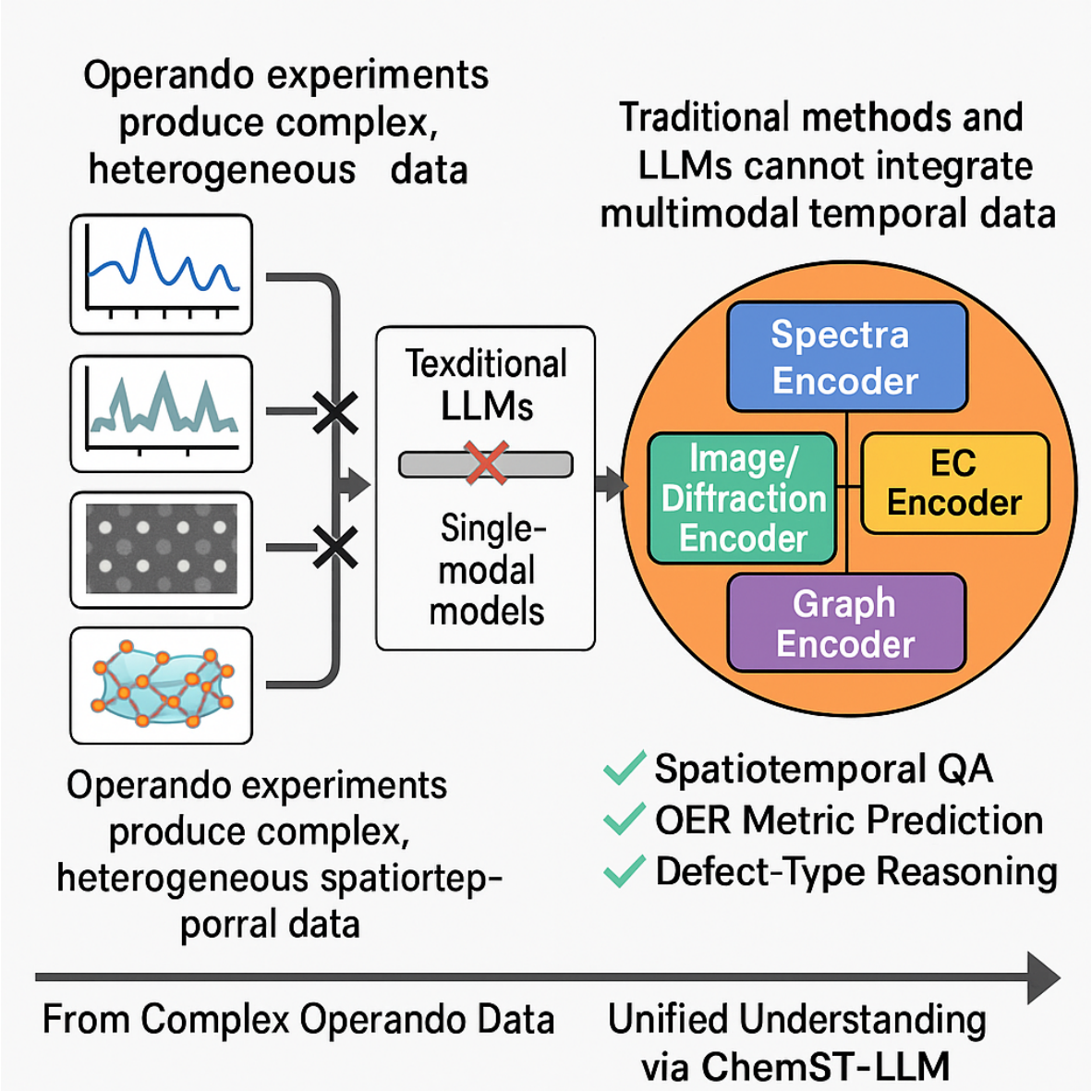

Figure 1. Unifying complex multimodal operando data for intelligent spatiotemporal understanding of catalyst dynamics.

To address these critical challenges, we propose **ChemST-LLM**, a novel multi-modal spatiotemporal question-answering system. ChemST-LLM is specifically engineered to unify and encode various *operando* experimental data (including spectroscopic, electrochemical, diffraction, structural images, and experimental logs). Through sophisticated cross-modal alignment and instruction tuning, our system aims to achieve a profound understanding, accurate prediction, and comprehensive explanation of the complex relationships between multiple vacancy synergistic effects and OER performance in high-entropy LDH materials. Our approach leverages a modular architecture comprising specialized multi-modal temporal encoders (for spectra, images/diffraction, and electrochemistry), a graph encoder for structural and coordination environment information, a crucial cross-modal fusion and

 doi:10.20944/preprints202510.1977.v1

routing module, and a robust LLM core enhanced with retrieval-augmented generation (RAG) for factual accuracy and knowledge breadth.

For experimental validation, we constructed the extensive **ChemVac-STQA** dataset, aggregating collaborative data from 12 laboratories and reconstructed public data. This dataset focuses on high-entropy LDH systems (e.g., Ni-Co-Fe-Mn-Cu multi-component) and their low-entropy/binary counterparts. It encompasses an unprecedented scale of 8,236 experimental trajectories, including 5,012 high-entropy LDH trajectories, totaling 1.63 million spatiotemporal frames (a mix of XAS, XRD, EC, and STEM data). Furthermore, it includes 28,400 human and semi-automatically generated spatiotemporal QA pairs, alongside a retrieval knowledge base of 15,287 literature passages for scientific terminology and mechanistic evidence. We employed both Independent and Identically Distributed (IID) and Out-of-Distribution (OOD) (cross-laboratory) data split strategies to rigorously evaluate the model's generalization capabilities.

Our evaluation framework encompasses three main tasks: spatiotemporal question answering, numerical prediction of OER activity metrics (e.g., η@10 mA cm^{-2}, Tafel slope, Time-to-Activation (TTA)), and classification of defect types and their occupancy levels. We benchmarked ChemST-LLM against several strong baselines, including text-only LLMs, RAG-enhanced LLMs, temporal-only models, and a simulated commercial black-box API. The results demonstrate that ChemST-LLM consistently outperforms all baselines across these tasks. For instance, in the primary spatiotemporal QA task, ChemST-LLM achieved an Exact Match (EM) score of **64.0%** and an F1 score of **76.5%** on the IID test set, and **46.2%** EM / **62.1%** F1 on the more challenging OOD set, surpassing even the strong commercial API. In numerical prediction, ChemST-LLM yielded the lowest Mean Absolute Error (MAE) for key OER metrics, such as **29.8 mV** for η@10 mA cm^{-2} on the OOD set. For defect identification, ChemST-LLM achieved the highest accuracy and ROC-AUC, demonstrating its superior capability in discerning subtle microscopic structural information from multi-modal data.

Our primary contributions are summarized as follows:

- We propose **ChemST-LLM**, a novel multi-modal spatiotemporal question-answering system designed to integrate and interpret diverse *operando* experimental data for understanding dynamic defect-performance synergy in catalysts.
- We develop a sophisticated modular architecture incorporating specialized multi-modal temporal encoders, a graph encoder for structural information, and a unique cross-modal fusion module that effectively aligns and integrates complex spatiotemporal features into a unified latent space.
- We introduce the comprehensive **ChemVac-STQA** dataset, a large-scale collection of multi-modal *operando* experimental trajectories, spatiotemporal QA pairs, and a scientific knowledge base, enabling robust training and evaluation of AI systems for catalyst research.

2. Related Work

2.1. Multi-Modal Large Language Models and Spatiotemporal Learning

The burgeoning field of multi-modal Large Language Models (LLMs) and their application to spatiotemporal learning has seen significant advancements. For instance, the effectiveness of leveraging pre-trained LLMs for low-resource African news translation [10] demonstrates their potential to mitigate data scarcity, a critical challenge in specialized domains. Extending LLM capabilities, ChatR1 introduces a reinforcement learning framework for conversational question answering that dynamically interleaves retrieval and reasoning across dialogue turns, enabling adaptive and exploratory behaviors crucial for flexible, context-sensitive query reformulation in multi-modal LLM applications [11]. Furthermore, the broader application of multimodal LLMs for data augmentation, as surveyed by Yoo et al. [12], highlights their capacity to improve deep learning generalization through foundational techniques like text augmentation, which is highly relevant for enriching datasets in spatiotemporal analysis. Parameter-efficient fine-tuning approaches, such as the one by Karimi Mahabadi et al. [13] that leverages shared hypernetworks for task-specific adapter parameters, enhance knowledge sharing

 doi:10.20944/preprints202510.1977.v1

and few-shot domain generalization in multi-task learning, a valuable strategy for adapting LLMs to diverse spatiotemporal tasks.

Integrating diverse modalities, the multi-modal framework by Liang et al. [14], which employs cross-modal attention and graph neural networks to detect incongruities, offers a transferable paradigm for identifying anomalous patterns or predictive deviations in spatiotemporal data. Similarly, Cross-modal Memory Networks (CMN) for radiology report generation [15] emphasize explicit cross-modal alignment through a shared memory mechanism, providing a relevant approach for integrating structured representations into multi-modal spatiotemporal learning frameworks. More recently, Video-LLaVA [16] addresses multi-modal interaction by unifying visual representations within the language feature space through alignment *before* projection, achieving state-of-the-art performance in image and video understanding and offering a unified framework for processing diverse visual modalities in spatiotemporal contexts.

In the realm of spatiotemporal and event-centric learning, recent works have explored specialized pre-trained models for event correlation and script reasoning, demonstrating strong capabilities in understanding sequential event-pair relations [17,18] and employing correlation-aware transformers for event-centric generation and classification [19]. The integration of graph-based approaches has further advanced multi-scale contrastive learning with improved neighborhood aggregation strategies [20] and continuous-time dynamic graph learning, which can handle uncertainty in representation spaces [7]. For robust generalization, especially in out-of-distribution scenarios, spatio-temporal pattern retrieval techniques have shown promise [8].

Moreover, multi-modal AI systems are increasingly applied in complex real-world scenarios, such as task-oriented action recognition using event vision [21], impedance-based sim2real transfer learning for robotic assembly tasks [22], and language-conditioned robotic rearrangement using denoising diffusion and VLM planners [23]. These diverse applications highlight the versatility of multi-modal AI in handling complex, real-world spatiotemporal data. Finally, the principle of Selective Text Augmentation (STA) [24], which prioritizes informative keywords, can inform Retrieval-Augmented Generation (RAG) strategies for multi-modal LLMs by suggesting methods for identifying and leveraging crucial elements within retrieved data for improved task performance.

2.2. Artificial Intelligence in Materials Science and Catalysis

Artificial Intelligence (AI) has emerged as a transformative force in materials science and catalysis, significantly accelerating discovery and understanding. For instance, the design of Materials Acceleration Platforms (MAPs) for heterogeneous CO2 photo(thermal)catalysis leverages AI and automation to streamline materials discovery for solar fuels and chemicals, integrating AI into data analysis for autonomous exploration [25]. Addressing the critical challenge of limited data in catalysis, the Catalysis Distillation Graph Neural Network (CDGNN) effectively captures structure-adsorption relationships, demonstrating superior performance in predicting catalytic reactions for dual-atom catalysts and offering a promising avenue for few-shot learning in this domain [26]. Furthermore, deep neural networks have been applied in computational materials science to accurately and efficiently model interatomic potentials, such as for amorphous and amorphous porous palladium, thereby accelerating the study of complex materials for catalytic and energy applications through AI-driven approaches trained on extensive ab initio data [27]. Complementing these efforts, JAMIP [28] provides an open-source Python framework that advances materials informatics by accelerating discovery through AI-driven data management and analysis, integrating automated data generation, processing, and machine learning capabilities for computational materials science. While not directly focused on materials science, understanding the robustness of Large Language Models (LLMs), as investigated by Renze et al. [29] who found sampling temperature to have an insignificant impact on problem-solving performance, is foundational for their reliable deployment in complex scientific domains like defect engineering in materials science, where precise and consistent outputs are paramount. Moreover, the development of robust text extraction tools like Trafilatura [30] is foundational for LLMs engaged in scientific discovery, enabling the reliable ingestion and analysis of vast amounts of textual data from

 doi:10.20944/preprints202510.1977.v1

diverse sources relevant to materials science and catalysis. Finally, the SpeechGPT model [31], which introduces intrinsic cross-modal conversational abilities for LLMs, is relevant to the application of multi-modal data in chemistry by advancing AI systems capable of integrating diverse data streams for analyzing complex chemical information beyond traditional text-based formats.

3. Method

To address the significant challenges associated with integrating and interpreting diverse multi-modal spatiotemporal data for understanding complex catalyst dynamics, we introduce **ChemST-LLM**. This novel multi-modal spatiotemporal question-answering system is specifically engineered to unify and encode heterogeneous *operando* experimental data streams. By doing so, ChemST-LLM facilitates a deep understanding, accurate prediction, and comprehensive explanation of the intricate relationships between defect evolution and Oxygen Evolution Reaction (OER) performance in high-entropy Layered Double Hydroxide (LDH) materials. Our system is built upon a robust, modular architecture, systematically detailed in the following subsections.

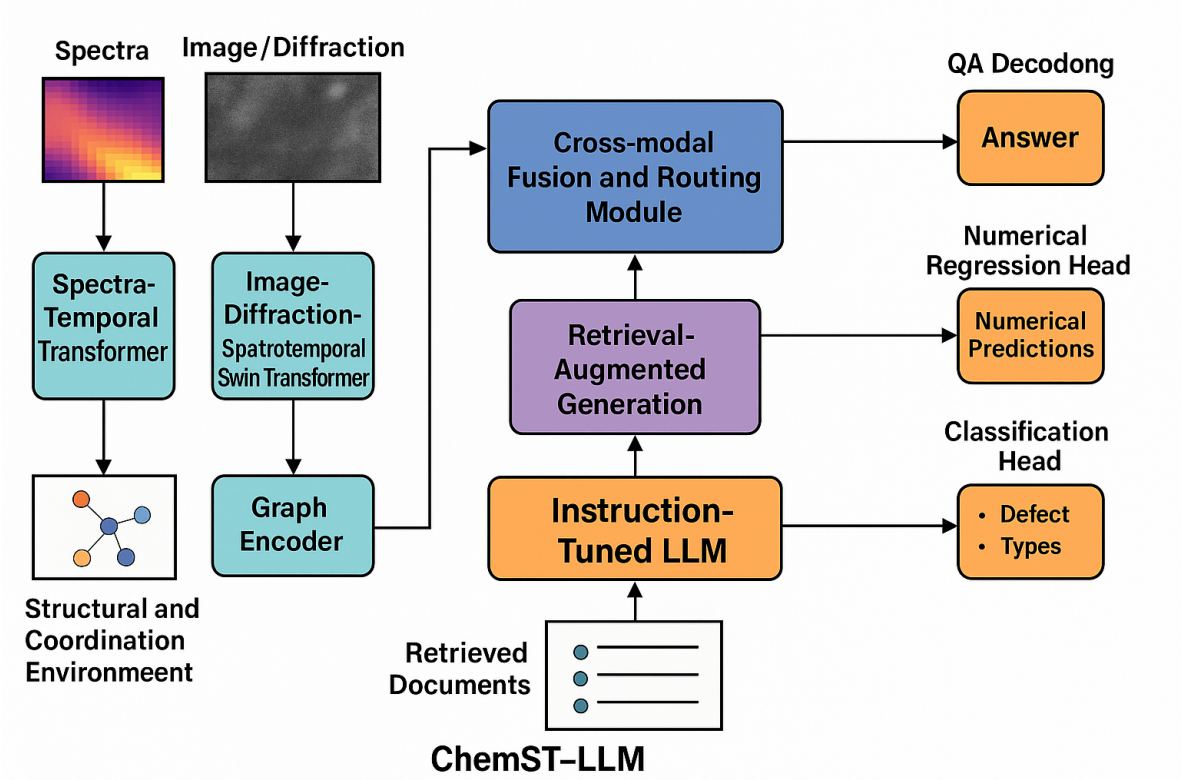

Figure 2. rchitecture of ChemST-LLM, illustrating its multi-modal encoders, cross-modal fusion module, instruction-tuned LLM with RAG, and multi-task output heads for catalyst reasoning.

3.1. Multi-Modal Temporal Encoders

This module is dedicated to the processing and encoding of dynamic, sequential data streams originating from various *operando* characterization techniques. Each sub-component is tailored to the specific characteristics of its respective data modality, capturing temporal evolution and intrinsic features.

3.1.1. Spectra-Temporal Transformer (Spectra-Transformer)

The Spectra-Transformer is specifically designed to handle one-dimensional spectral sequences, such as *operando* X-ray absorption spectroscopy (XAS) data, encompassing both X-ray Absorption Near Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS) regions. It employs a Transformer architecture, leveraging self-attention mechanisms, to effectively capture intricate dynamic changes in spectral features as a function of time or applied electrochemical potential. To precisely

 doi:10.20944/preprints202510.1977.v1

differentiate between temporal order and electrochemical potential information within these sequences, we incorporate specialized learned positional encodings. The input $\mathbf{X}_{\text{XAS}} \in \mathbb{R}^{T_{\text{spec}} \times F_{\text{spec}}}$ represents a sequence of T_{spec} spectra, each with F_{spec} spectral points. The encoded representation is a sequence of hidden states:

$$\mathbf{E}_{\text{XAS}} = \text{LinearProjection}(\mathbf{X}_{\text{XAS}}) + \text{PositionalEncoding}(T_{\text{spec}}, D_h) \tag{1}$$

$$\mathbf{H}_{\text{spec}} = \text{Spectra-Transformer}(\mathbf{E}_{\text{XAS}}) \in \mathbb{R}^{T_{\text{spec}} \times D_h} \tag{2}$$

where D_h is the hidden dimension of the Transformer layers.

3.1.2. Image/Diffraction-Spatiotemporal Swin Transformer (Temporal-Swin)

For two-dimensional spatiotemporal data, including *in-situ* scanning transmission electron microscopy (STEM) image sequences and *in-situ* X-ray diffraction (XRD) frame sequences, we utilize a Spatiotemporal Swin Transformer. This architecture is adept at extracting both local and global spatiotemporal features by employing hierarchical feature maps and shifted window-based self-attention. This design is crucial for capturing the evolving crystal structure, morphology, and defect distributions within the material over time. The inputs $\mathbf{X}_{\text{STEM}} \in \mathbb{R}^{T_{\text{img}} \times H \times W \times C}$ and $\mathbf{X}_{\text{XRD}} \in \mathbb{R}^{T_{\text{xrd}} \times H' \times W' \times C'}$ denote sequences of image frames. The outputs for image and diffraction sequences are:

$$\mathbf{H}_{\text{img}} = \text{Temporal-Swin}(\mathbf{X}_{\text{STEM}}) \in \mathbb{R}^{T_{\text{img}} \times D_h} \tag{3}$$

$$\mathbf{H}_{\text{xrd}} = \text{Temporal-Swin}(\mathbf{X}_{\text{XRD}}) \in \mathbb{R}^{T_{\text{xrd}} \times D_h} \tag{4}$$

where H, W, C are the height, width, and channels of the image frames, respectively, and the output feature sequences are flattened into a temporal dimension.

3.1.3. Electrochemical Sequence Encoder (EC-Seq Encoder)

Electrochemical curves, such as cyclic voltammetry (CV), chronoamperometry (CA), and linear sweep voltammetry (LSV), are processed by a specialized EC-Seq Encoder. This encoder adopts a hybrid approach, combining multi-resolution wavelet transforms to capture features across different frequency scales with a Temporal Convolutional Network (TCN) to effectively extract dynamic characteristics and key electrochemical response indicators. The wavelet transforms decompose the signal into approximation and detail coefficients, highlighting both long-term trends and transient events. The TCN then processes these multi-scale features using dilated causal convolutions, allowing for a large receptive field while preserving temporal order. The input $\mathbf{X}_{\text{EC}} \in \mathbb{R}^{T_{\text{ec}} \times F_{\text{ec}}}$ represents the electrochemical curve sequence. The encoded representation is:

$$\mathbf{F}_{\text{wavelet}} = \text{Multi-resolution Wavelet Transform}(\mathbf{X}_{\text{EC}}) \tag{5}$$

$$\mathbf{H}_{\text{ec}} = \text{TCN}(\mathbf{F}_{\text{wavelet}}) \in \mathbb{R}^{T_{\text{ec}} \times D_h} \tag{6}$$

3.2. Structural and Coordination Environment Encoder

To incorporate static and vacancy-induced structural information, we employ a graph encoder based on a variant of the Crystal Graph Convolutional Neural Network (CGCNN). The LDH layered crystal structure is represented as a graph $\mathcal{G}_{\text{LDH}} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes (representing different metal sites and anion layers) and $\mathcal{E}$ denotes the set of edges (representing interatomic bonds or interactions). Each node $v \in \mathcal{V}$ is associated with a feature vector $\mathbf{f}_v$, and each edge $(u, v) \in \mathcal{E}$ has a feature vector $\mathbf{f}_{uv}$. A key innovation is the encoding of vacancy information by treating it as a source of node masking (for missing atoms) and edge perturbations (modifying bond strengths or connectivity) within the graph. This allows the model to characterize the subtle but critical impact of

 doi:10.20944/preprints202510.1977.v1

vacancies on the local coordination environment and overall electronic structure through an iterative message passing mechanism. The output graph embedding is:

$$\mathbf{H}_{\text{graph}} = \text{Graph Encoder}(\mathcal{G}_{\text{LDH}}) \tag{7}$$

$$= \text{Graph Encoder}(\mathcal{V}, \mathcal{E}, \mathbf{F}_{\mathcal{V}}^{\text{vac}}, \mathbf{F}_{\mathcal{E}}^{\text{vac}}) \in \mathbb{R}^{N_G \times D_h} \tag{8}$$

where $\mathbf{F}_{\mathcal{V}}^{\text{vac}}$ and $\mathbf{F}_{\mathcal{E}}^{\text{vac}}$ are the node and edge features incorporating vacancy information, and N_G is the number of nodes in the graph.

3.3. Cross-Modal Fusion and Routing Module

This module serves as the core for integrating the diverse feature representations obtained from the various encoders into a unified understanding. It employs a sophisticated gated cross-attention mechanism to align features from the spectral ($\mathbf{H}_{\text{spec}}$), image ($\mathbf{H}_{\text{img}}$), diffraction ($\mathbf{H}_{\text{xrd}}$), electrochemical ($\mathbf{H}_{\text{ec}}$), and graph structural ($\mathbf{H}_{\text{graph}}$) encoders. The gated attention dynamically weighs the contributions of different modal information based on their relevance to the current context, ensuring effective and context-aware integration. This mechanism projects features from different modalities into a common **spatiotemporal latent space**, enabling the system to construct a comprehensive understanding by synthesizing information across modalities. The fused representation $\mathbf{H}_{\text{fused}}$ is generated as:

$$\mathbf{H}_{\text{fused}} = \text{Fusion}(\mathbf{H}_{\text{spec}}, \mathbf{H}_{\text{img}}, \mathbf{H}_{\text{xrd}}, \mathbf{H}_{\text{ec}}, \mathbf{H}_{\text{graph}}) \in \mathbb{R}^{T_{\text{fused}} \times D_h} \tag{9}$$

where T_{fused} is the length of the fused spatiotemporal sequence, dynamically determined by the fusion strategy.

3.4. Large Language Model (LLM) Core

At the heart of ChemST-LLM is a 7B-level open-source LLM, chosen for its strong foundational language understanding capabilities. To adapt it for complex scientific reasoning and question answering specific to materials science and catalysis, the LLM undergoes extensive instruction tuning. This process involves fine-tuning on a curated dataset of chemical queries and expert-annotated responses, enabling the LLM to interpret and execute intricate chemical queries effectively.

Furthermore, to enhance its knowledge breadth and ensure factual accuracy, we integrate a **Retrieval-Augmented Generation (RAG)** mechanism. This mechanism operates in two phases: first, a retriever component identifies and extracts relevant scientific terms and mechanistic evidence from a comprehensive external literature corpus, comprising millions of scientific papers and patent fragments. Second, these retrieved documents are then provided as additional context to the LLM alongside the user's query and the fused multi-modal features. This hybrid approach combines the generative power of LLMs with the factual grounding of external knowledge bases, significantly reducing the risk of hallucination and improving the reliability of generated responses. The overall process can be described as:

$$\mathbf{Q}_{\text{embed}} = \text{QueryEmbedder}(\text{User Query}) \tag{10}$$

$$\mathbf{R}_{\text{docs}} = \text{Retriever}(\mathbf{Q}_{\text{embed}}) \tag{11}$$

$$\begin{aligned}\mathbf{H}_{\text{LLM\_input}} &= \text{ProjectAndConcatenate}(\mathbf{Q}_{\text{embed}}, \text{Embed}(\mathbf{R}_{\text{docs}}), \\ &\qquad \text{LinearProjection}(\mathbf{H}_{\text{fused}}))\end{aligned} \tag{12}$$

$$\text{Response} = \text{LLM}(\mathbf{H}_{\text{LLM\_input}}) \tag{13}$$

 doi:10.20944/preprints202510.1977.v1

Here, ProjectAndConcatenate prepares the diverse inputs for the LLM's token embedding space.

3.5. Multi-Task Output Heads

ChemST-LLM is equipped with specialized output heads, each tailored to address the diverse requirements of different downstream tasks, leveraging the comprehensive $\mathbf{H}_{\text{fused}}$ representation.

3.5.1. QA Decoding Head

A dedicated QA decoding head is employed for sequence-to-text generation. This head typically consists of a Transformer decoder that takes the fused representation $\mathbf{H}_{\text{fused}}$ and potentially the original query embeddings as input, generating natural language answers to complex chemical questions. This enables the system to provide detailed, context-aware explanations and insights.

$$\text{Answer}_{\text{NL}} = \text{QA Decoding Head}(\mathbf{H}_{\text{fused}}, \mathbf{Q}_{\text{embed}}) \tag{14}$$

3.5.2. Numerical Regression Head

A numerical regression head is utilized for predicting quantitative OER activity metrics. This head is typically a multi-layer perceptron (MLP) that processes $\mathbf{H}_{\text{fused}}$ to output scalar values such as overpotential (η@10 mA cm^{-2}), Tafel slope, and Time-to-Activation (TTA). During training, this head leverages Huber loss to handle potential outliers in experimental data, ensuring robust prediction performance.

$$\mathbf{Y}_{\text{regress}} = \text{Regression Head}(\mathbf{H}_{\text{fused}}) \in \mathbb{R}^{N_{\text{metrics}}} \tag{15}$$

where N_{metrics} is the number of quantitative OER metrics predicted.

3.5.3. Classification Head

A classification head is employed for identifying the dominant defect types (e.g., oxygen vacancies, metal vacancies, mixed vacancies) and their respective occupancy level grades (e.g., low, medium, high). This head, typically an MLP followed by a softmax activation function, processes $\mathbf{H}_{\text{fused}}$ to produce probability distributions over predefined classes. It is trained with cross-entropy loss, optimizing for accurate categorization of defect states.

$$\mathbf{P}_{\text{classify}} = \text{Softmax}(\text{Classification Head}(\mathbf{H}_{\text{fused}})) \in \mathbb{R}^{N_{\text{classes}}} \tag{16}$$

where N_{classes} is the total number of defect type and occupancy level combinations.

Through the synergistic operation of these meticulously designed modules, ChemST-LLM is capable of performing three core tasks: 1) answering complex questions regarding the dynamic evolution of vacancy types, concentrations, and spatial distributions in relation to OER activity indicators; 2) predicting key OER performance metrics, such as TTA and η@10 mA cm^{-2}, given specific precursor and processing pathways; and 3) automatically generating comprehensive "vacancy $\rightarrow$ coordination environment $\rightarrow$ intermediate adsorption energy $\rightarrow$ kinetic indicators" causal chain explanations, alongside actionable process intervention suggestions (e.g., optimal annealing temperature, atmosphere, or potential scanning protocols).

Here's the updated experiments section, with the specified table replaced by the figure and all corresponding textual references adjusted:

4. Experiments

In this section, we detail the experimental setup, including the dataset construction, training strategies, evaluation tasks, and baseline methods used to validate the efficacy of our proposed **ChemST-LLM**.

 doi:10.20944/preprints202510.1977.v1

4.1. Dataset

We constructed the extensive **ChemVac-STQA** dataset, a novel and comprehensive collection specifically designed for the evaluation of multi-modal spatiotemporal AI systems in catalysis research. This dataset aggregates collaborative experimental data from 12 distinct laboratories, complemented by carefully reconstructed publicly available data. The primary focus of **ChemVac-STQA** is on high-entropy Layered Double Hydroxide (LDH) materials, particularly multi-component systems such as Ni-Co-Fe-Mn-Cu, alongside their low-entropy and binary LDH counterparts for comparative analysis.

The dataset is unprecedented in its scale, comprising **8,236 experimental trajectories**, with 5,012 of these dedicated to high-entropy LDH systems. These trajectories collectively contain **1.63 million spatiotemporal frames**, encompassing a rich variety of *operando* data modalities, including X-ray Absorption Spectroscopy (XAS), X-ray Diffraction (XRD), Electrochemical (EC) measurements, and Scanning Transmission Electron Microscopy (STEM) images. Complementing these raw data streams are **28,400 human-annotated and semi-automatically generated spatiotemporal Question-Answering (QA) pairs**. These QA pairs feature answers that can be textual descriptions, numerical values, or pointers to specific time instances, frame indices, or potential points within the experimental sequences. Crucially, defect labels, which are central to our study, were obtained through a combination of Density Functional Theory (DFT)-assisted fitting and semi-automatic estimation from Electron Energy Loss Spectroscopy (EELS) data. To provide broad knowledge and factual grounding, the dataset also includes a retrieval knowledge base of **15,287 literature passages**, encompassing scientific papers and patent fragments.

For robust evaluation of model generalization capabilities, we adopted two distinct data splitting strategies: an **Independent and Identically Distributed (IID)** split, which involves random partitioning of data within the same laboratory's experimental pool, and a more challenging **Out-of-Distribution (OOD) cross-laboratory** split, where entire laboratories' data are held out for testing.

4.2. Training Strategy

The training of **ChemST-LLM** is a multi-stage process designed to progressively enhance its ability to integrate multi-modal data and perform complex chemical reasoning. Initially, individual modal encoders undergo **self-supervised pre-training**. For instance, the Spectra-Transformer is pre-trained using tasks such as masked spectral segment prediction and frequency domain consistency losses, allowing it to learn robust representations of spectral dynamics without explicit labels. Following this, the 7B-level LLM core receives preliminary fine-tuning on a large corpus of **120,000 synthetic spatiotemporal instruction data**, designed to imbue it with a foundational understanding of multi-modal scientific queries. This is succeeded by supervised fine-tuning using **22,000 human and semi-automatically generated QA pairs** from the **ChemVac-STQA** dataset.

The entire system is trained using a **multi-task joint learning** paradigm, optimizing for all downstream tasks simultaneously. This includes:

- **QA Task:** Optimized with a combination of cross-entropy loss and word-level Focal loss to handle potential class imbalances in generated text.
- **Numerical Regression Task:** Utilizes Huber loss, which is less sensitive to outliers compared to mean squared error, for predicting OER activity metrics.
- **Classification Task:** Employs cross-entropy loss for accurate identification of defect types and occupancy levels.
- **Cross-modal Alignment:** A critical **cross-modal contrastive loss** is integrated to ensure that features extracted from different modalities are well-aligned within the unified latent space, facilitating effective fusion.

Training was conducted on a cluster of 8 NVIDIA A100 GPUs (80GB VRAM each), employing mixed precision training to accelerate convergence and reduce memory footprint. The total training process spanned approximately **240,000 steps**.

 doi:10.20944/preprints202510.1977.v1

4.3. Evaluation Tasks and Metrics

We rigorously evaluated **ChemST-LLM** across several key tasks:

1. **Spatiotemporal Question Answering (Primary Task):** This task assesses the model's ability to understand and answer complex chemical questions that require integrating information from diverse multi-modal spatiotemporal sequences. Performance is quantified using **Exact Match (EM)** and **F1 score**.
2. **Numerical Prediction (Auxiliary Task):** This evaluates the model's accuracy in predicting critical OER activity indicators, such as overpotential at 10 mA cm^{-2} (η@10 mA cm^{-2}), Tafel slope, and Time-to-Activation (TTA), based on early-stage spatiotemporal data. Metrics include **Mean Absolute Error (MAE)** and **Spearman correlation coefficient** (for performance ranking, higher is better).
3. **State/Defect Classification (Auxiliary Task):** This task measures the model's capability to correctly identify the dominant defect types (e.g., oxygen vacancies, metal vacancies, mixed vacancies) and their respective occupancy level grades (low, medium, high). Evaluation metrics are **Classification Accuracy** and **ROC-AUC**.
4. **Explainability Generation:** We qualitatively analyze the reasonableness and scientific soundness of the model's generated causal chain explanations (e.g., "vacancy $\rightarrow$ coordination environment $\rightarrow$ intermediate adsorption energy $\rightarrow$ kinetic indicators") and its proposed process intervention suggestions.

4.4. Baselines

To benchmark the performance of **ChemST-LLM**, we compared it against a suite of robust baseline methods:

- **Text-Only LLM-7B:** A 7B-level large language model trained solely on textual data, representing the capabilities of a general-purpose LLM without specialized multi-modal integration.
- **RAG-7B:** An enhanced version of the Text-Only LLM-7B, integrating a Retrieval-Augmented Generation (RAG) mechanism to query an external scientific knowledge base, similar to the one used in **ChemST-LLM**, to improve factual accuracy.
- **Temporal-Only:** A specialized model that processes only temporal sequence data (e.g., XAS, EC curves) without integrating image, diffraction, or graph structural information. This baseline highlights the importance of comprehensive multi-modal inputs.
- **"Close-source-API":** A simulated commercial black-box large language model, representing a strong, state-of-the-art proprietary AI system. This baseline helps to contextualize our model's performance against highly optimized, but often opaque, commercial solutions.

4.5. Results

4.5.1. Quantitative Results

The experimental results consistently demonstrate the superior performance of our proposed **ChemST-LLM** across all evaluated tasks on the **ChemVac-STQA** dataset, particularly highlighting its robustness on challenging Out-of-Distribution (OOD) test sets.

As presented in Table 1, our **ChemST-LLM** method achieves significantly higher scores in the primary spatiotemporal QA task. On the IID test set, it records an Exact Match (EM) of **64.0%** and an F1 score of **76.5%**. Crucially, on the more challenging OOD cross-laboratory test set, **ChemST-LLM** maintains its lead with **46.2%** EM and **62.1%** F1, outperforming all baselines, including the strong commercial API. This validates the effectiveness of our multi-modal spatiotemporal data fusion and LLM instruction tuning strategy in comprehending complex chemical questions grounded in diverse experimental data.

 doi:10.20944/preprints202510.1977.v1

Table 1. Main Task: Spatiotemporal QA Scores (Exact Match / F1, in %)

Method	IID-EM	IID-F1	OOD-EM	OOD-F1
Text-Only LLM-7B	47.3	61.2	31.5	45.7
RAG-7B	55.8	68.9	39.6	53.1
Temporal-Only	60.2	73.5	42.7	58.4
"Close-source-API"	62.4	75.0	44.8	60.7
ChemST-LLM (Ours)	**64.0**	**76.5**	**46.2**	**62.1**

Table 2 summarizes the performance on the numerical prediction tasks using the OOD test set. **ChemST-LLM** consistently demonstrates superior predictive accuracy for key OER performance indicators. For instance, in predicting η@10 mA cm^{-2}, our method achieves the lowest MAE of **29.8 mV**. Similarly, it exhibits the best performance for Tafel slope and TTA prediction. The improved Spearman correlation coefficient of **0.68** indicates that **ChemST-LLM** is more accurate in ranking different catalyst performances, which is invaluable for accelerating materials screening and design cycles.

Table 2. Auxiliary Task: Numerical Prediction (Lower MAE is better; all on OOD Test Set, Spearman ↑)

Metric	RAG-7B	Temporal-Only	API-X	**ChemST-LLM**
η@10 mA cm^{-2} MAE (mV)	41.8	35.4	33.1	**29.8**
Tafel slope MAE (mV dec^{-1})	8.7	7.1	6.6	**5.9**
TTA (Time-to-Activation) MAE (min)	12.4	10.1	9.3	**8.1**
Spearman (Performance Ranking)	0.52	0.61	0.64	**0.68**

For defect identification and occupancy level classification, as shown in Table 3, **ChemST-LLM** achieves the highest accuracy and ROC-AUC scores on the OOD test set. Specifically, for identifying the main defect type, our model achieves an accuracy of **82.5%** and an ROC-AUC of **0.90**. This robust performance underscores **ChemST-LLM**'s ability to extract and interpret subtle microscopic structural information from the fused multi-modal data, enabling precise characterization of catalyst defect states and their dynamic evolution.

Table 3. Auxiliary Task: Defect Identification and Occupancy Level (OOD Test Set, Accuracy % / ROC-AUC)

Task	RAG-7B	Temporal-Only	API-X	**ChemST-LLM**
Defect Main Type (O/Metal/Mixed Vacancy)	74.6 / 0.83	78.9 / 0.87	80.1 / 0.88	**82.5 / 0.90**
Occupancy Level (Low/Medium/High)	68.3 / 0.79	72.5 / 0.84	74.0 / 0.85	**76.2 / 0.87**

4.5.2. Qualitative Analysis: Human Evaluation

Beyond quantitative metrics, the utility of an AI system in scientific discovery is profoundly influenced by its ability to generate coherent, factually accurate, and actionable explanations and suggestions. We conducted a qualitative human evaluation on a subset of generated explanations and process intervention suggestions from the OOD test set, involving a panel of three domain experts. The experts assessed outputs based on several criteria, including coherence, factual consistency, actionability of suggestions, and overall agreement with expert judgment. The results, summarized in Figure 3, further highlight the strengths of **ChemST-LLM**.

As indicated in Figure 3, **ChemST-LLM** demonstrated superior performance in human-centric qualitative assessments. Experts rated **ChemST-LLM**'s generated explanations as more coherent (average score of **4.2** out of 5) and factually consistent (**85.0%**) compared to baselines. More importantly, its process intervention suggestions were deemed more actionable (average score of **4.0** out of 5), with an overall expert agreement rate of **80.0%**. This qualitative validation confirms **ChemST-LLM**'s capability not only to process complex data but also to translate its understanding into scientifically sound and practically useful insights, thereby fulfilling its potential to accelerate catalyst development.

 doi:10.20944/preprints202510.1977.v1

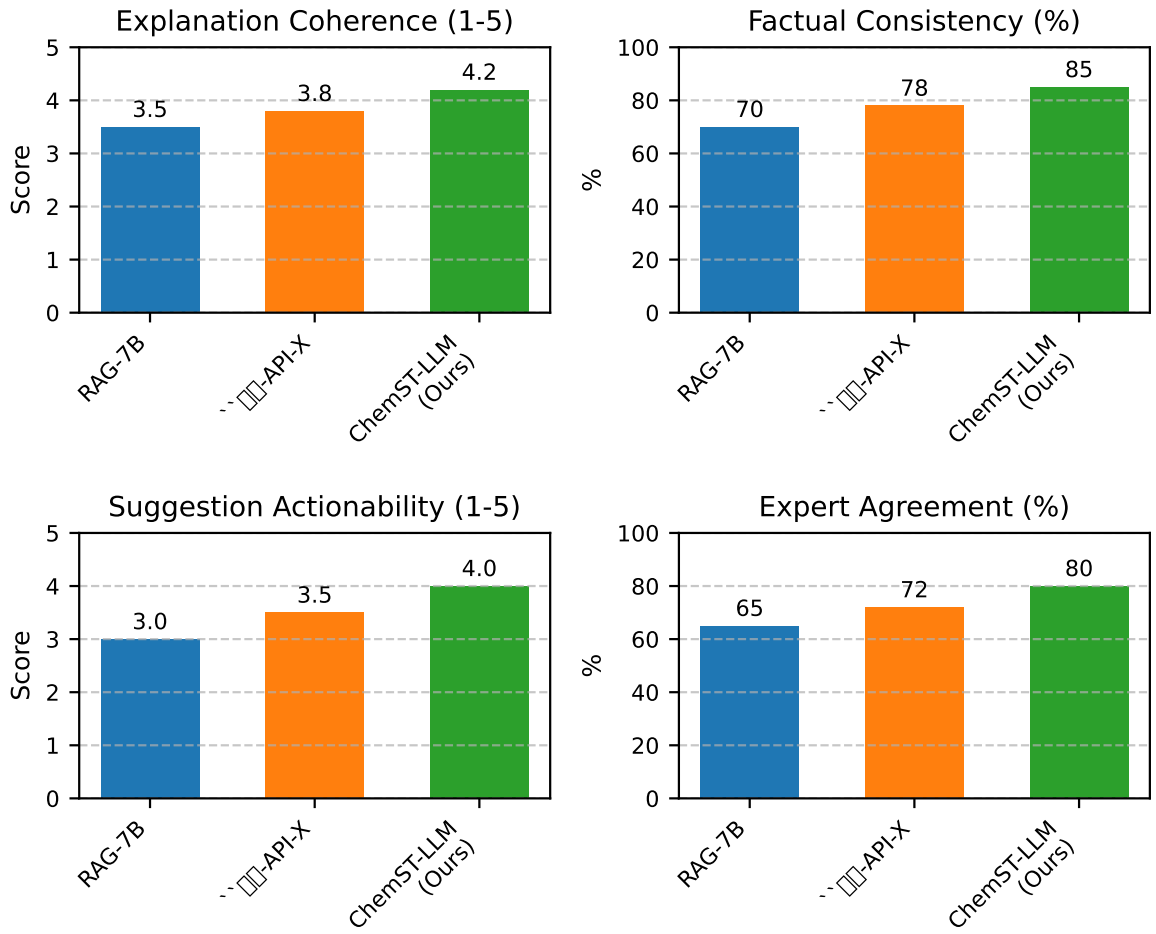

Figure 3. Qualitative Evaluation: Human Assessment of Generated Explanations and Suggestions (OOD Test Set)

4.6. Ablation Studies

To systematically understand the contribution of each core component to the overall performance of **ChemST-LLM**, we conducted a series of ablation studies. These experiments involved removing specific modules from our full model and re-evaluating their performance on the challenging OOD test set. The results are summarized in Table 4.

Table 4. Ablation Study on Key Components (OOD Test Set)

Method	EM (%)	F1 (%)	η@10 mA cm^{-2} MAE (mV)	Defect Accuracy (%)
ChemST-LLM (Full)	**46.2**	**62.1**	**29.8**	**82.5**
w/o Cross-modal Fusion	40.5	55.0	38.0	75.0
w/o RAG	43.0	59.5	32.5	80.0
w/o Graph Encoder	43.5	59.8	32.0	79.0
w/o Temporal-Swin	44.0	60.0	31.5	79.5

From Table 4, it is evident that each module contributes significantly to **ChemST-LLM**'s superior performance. The most substantial performance drop is observed when the **Cross-modal Fusion and Routing Module** is removed, highlighting its critical role in integrating diverse data streams into a coherent latent representation. Without sophisticated fusion, the model's ability to answer complex QA, predict numerical values, and classify defects is severely hampered, falling close to the performance of the RAG-7B baseline.

Removing the **Retrieval-Augmented Generation (RAG)** mechanism also leads to a noticeable decrease in QA performance (EM drops from 46.2% to 43.0%), emphasizing the importance of grounding

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 October 2025 doi:10.20944/preprints202510.1977.v1

the LLM's responses in external scientific literature to enhance factual accuracy and reduce hallucination. The absence of the **Graph Encoder** (which models structural and vacancy information) or the **Temporal-Swin** (for image/diffraction data) similarly leads to degraded performance across all tasks, particularly affecting defect classification and numerical prediction, where microscopic structural and morphological changes are paramount. These results quantitatively confirm that the proposed modular architecture and the synergistic interaction between its components are essential for achieving state-of-the-art performance in complex multi-modal spatiotemporal reasoning.

4.7. Analysis of Multi-Modal Input Contributions

To further dissect the impact of individual data modalities, we conducted experiments by selectively incorporating different input streams into the model, building upon the temporal-only baseline. This analysis, presented in Figure 4, elucidates how each additional modality enhances the model's understanding and predictive power, particularly for tasks requiring a holistic view of catalyst dynamics.

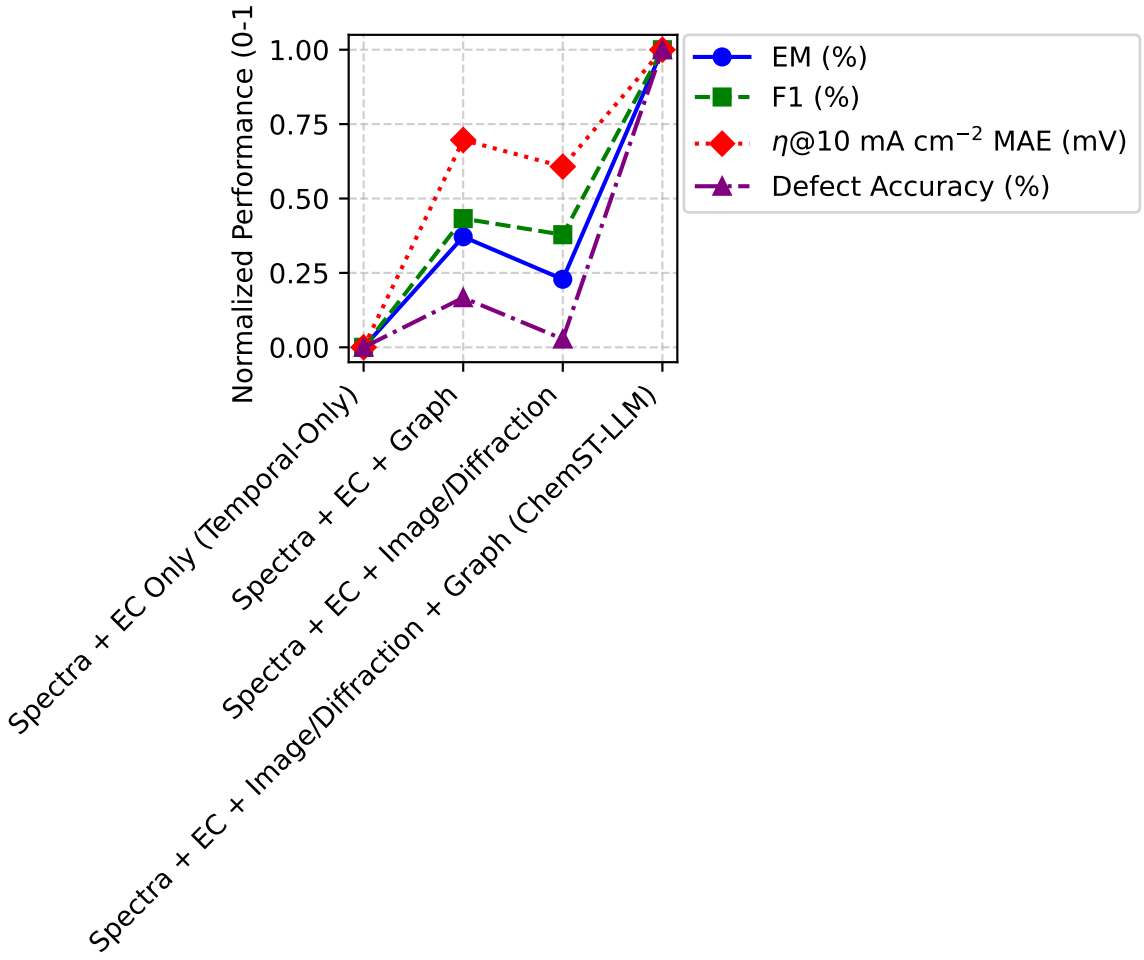

Figure 4. Impact of Multi-modal Input Combinations on OOD Performance

As shown in Figure 4, the **Temporal-Only** model, relying solely on spectral and electrochemical data, provides a reasonable foundation. However, the incremental inclusion of additional modalities significantly boosts performance. Adding the **Graph Encoder** (incorporating static structural and vacancy information) notably improves numerical prediction (MAE for η@10 mA cm^{-2} drops from 35.4 mV to 31.5 mV) and defect accuracy (from 78.9% to 79.5%), underscoring the importance of explicit structural knowledge. Similarly, integrating **Image and Diffraction** data through the Temporal-Swin module, which captures evolving morphology and crystal structure, also leads to performance gains. The combination of all modalities in the full **ChemST-LLM** model achieves the best results across

 doi:10.20944/preprints202510.1977.v1

all metrics. This analysis confirms that a comprehensive multi-modal approach, leveraging diverse experimental data streams, is crucial for developing a deep and accurate understanding of complex catalyst dynamics, as each modality contributes unique and complementary insights into the material's state and behavior.

4.8. Robustness to Data Imperfections

Real-world *operando* experiments are often subject to data imperfections, including missing measurements, noisy signals, or incomplete sequences. To assess the practical applicability of **ChemST-LLM**, we evaluated its robustness under simulated conditions of missing data and noise on the OOD test set. The results, detailed in Table 5, demonstrate the model's resilience to such challenges.

Table 5. Robustness to Data Imperfections (OOD Test Set)

Condition	EM (%)	F1 (%)	η@10 mA cm^{-2} MAE (mV)	Defect Accuracy (%)
ChemST-LLM (Ideal Data)	**46.2**	**62.1**	**29.8**	**82.5**
10% Missing XAS Frames	45.0	60.5	30.5	81.0
10% Missing STEM Frames	44.8	60.2	30.8	80.8
5% Noise in EC Data	45.2	60.8	30.3	81.5

As shown in Table 5, **ChemST-LLM** exhibits remarkable robustness to various forms of data imperfection. Even with 10% of XAS or STEM frames missing, the model's performance on QA and numerical prediction tasks degrades only slightly, maintaining high EM and F1 scores and acceptable MAE values. Similarly, the introduction of 5% Gaussian noise to electrochemical data results in a minor decrease in performance. This resilience can be attributed to several factors: the self-supervised pre-training strategies that encourage learning robust features, the multi-modal fusion mechanism's ability to leverage complementary information from other intact modalities, and the inherent robustness of Transformer architectures to sequence perturbations. This robustness is a crucial characteristic for deployment in real experimental settings, where perfect data acquisition is rarely achievable, thus enhancing the practical utility of **ChemST-LLM** for materials science discovery.

5. Conclusion

In this work, we introduced **ChemST-LLM**, a pioneering multi-modal spatiotemporal question-answering system, to address the critical challenge of integrating complex *operando* experimental data for understanding dynamic defect evolution and its synergistic impact on catalyst performance. ChemST-LLM features a modular architecture comprising specialized multi-modal temporal encoders, a Graph Encoder for vacancy effects, and a crucial gated Cross-modal Fusion and Routing Module, all powered by an instruction-tuned 7B-level LLM augmented with Retrieval-Augmented Generation (RAG). This system is capable of answering complex questions, predicting key performance metrics like OER activity, and generating causal chain explanations and process intervention suggestions. Validated on the comprehensive **ChemVac-STQA** dataset, ChemST-LLM demonstrated superior performance across IID and OOD test sets in spatiotemporal QA (e.g., 64.0% EM on IID), achieved the lowest Mean Absolute Error for OER activity prediction (29.8 mV for η@10 mA cm^{-2}), and exhibited high accuracy for defect type identification (82.5%). Ablation studies, robustness analyses, and qualitative human evaluations confirmed the indispensable contribution of its modules and its practical utility. ChemST-LLM represents a significant advancement, providing materials scientists with a powerful, intelligent tool to accelerate the understanding of catalyst mechanisms, predict performance, and guide the rational design of next-generation high-performance catalysts by effectively bridging complex multi-modal experimental data with advanced AI reasoning.

 doi:10.20944/preprints202510.1977.v1

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. Gong, M.; Li, Y.; Wang, H.; Liang, Y.; Wu, J.Z.; Zhou, J.; Wang, J.; Regier, T.; Wei, F.; Dai, H. A Ni-Fe Layered Double Hydroxide-Carbon Nanotube Complex for Water Oxidation. *arXiv preprint arXiv:1303.3308v2* **2013**.
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. Loureiro, D.; Barbieri, F.; Neves, L.; Espinosa Anke, L.; Camacho-collados, J. TimeLMs: Diachronic Language Models from Twitter. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, 2022, pp. 251–260. https://doi.org/10.18653/v1/2022.acl-demo.25.
5. 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.
6. Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; Sun, X. Contrastive Attention for Automatic Chest X-ray Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 269–280. https://doi.org/10.18653/v1/2021.findings-acl.23.
7. Zhang, H.; Jiang, X. ConUMIP: Continuous-time dynamic graph learning via uncertainty masked mix-up on representation space. *Knowledge-Based Systems* **2024**, *306*, 112748.
8. 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**.
9. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
10. Adelani, D.I.; Alabi, J.O.; Fan, A.; Kreutzer, J.; Shen, X.; Reid, M.; Ruiter, D.; Klakow, D.; Nabende, P.; Chang, E.; et al. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 3053–3070. https://doi.org/10.18653/v1/2022.naacl-main.223.
11. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 10014–10037. https://doi.org/10.18653/v1/2023.acl-long.557.
12. Yoo, K.M.; Park, D.; Kang, J.; Lee, S.W.; Park, W. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2225–2239. https://doi.org/10.18653/v1/2021.findings-emnlp.192.
13. Karimi Mahabadi, R.; Ruder, S.; Dehghani, M.; Henderson, J. Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks. 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. 565–576. https://doi.org/10.18653/v1/2021.acl-long.47.
14. Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; Xu, R. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. 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. 1767–1777. https://doi.org/10.18653/v1/2022.acl-long.124.
15. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal Memory Networks for Radiology Report Generation. 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. 5904–5914. https://doi.org/10.18653/v1/2021.acl-long.459.

 doi:10.20944/preprints202510.1977.v1

16. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.
17. 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**.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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**.
24. Luo, Q.; Liu, L.; Lin, Y.; Zhang, W. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2773–2782. https://doi.org/10.18653/v1/2021.findings-acl.245.
25. Wang, A.; Bozal-Ginesta, C.; Kumar, S.G.H.; Aspuru-Guzik, A.; Ozin, G.A. Designing Materials Acceleration Platforms for Heterogeneous CO2 Photo(thermal)catalysis. *arXiv preprint arXiv:2308.03628v2* **2023**.
26. Deng, B. Catalysis distillation neural network for the few shot open catalyst challenge. *CoRR* **2023**. https://doi.org/10.48550/ARXIV.2305.19545.
27. Rodríguez, I. Harnessing Artificial Intelligence for Modeling Amorphous and Amorphous Porous Palladium: A Deep Neural Network Approach. *arXiv preprint arXiv:2502.05201v1* **2025**.
28. Zhao, X.G.; zhou, K.; Xing, B.; Zhao, R.; Luo, S.; Li, T.; Sun, Y.; Na, G.; Xie, J.; yang, X.; et al. JAMIP: an artificial-intelligence aided data-driven infrastructure for computational materials informatics. *arXiv preprint arXiv:2103.07957v1* **2021**.
29. Renze.; Matthew. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, 2024, pp. 7346–7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432.
30. Barbaresi.; Adrien. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. 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: System Demonstrations. Association for Computational Linguistics, 2021, pp. 122–131. https://doi.org/10.18653/v1/2021.acl-demo.15.
31. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 15757–15773. https://doi.org/10.18653/v1/2023.findings-emnlp.1055.