UERLens: Understanding Event Relations in Large Language Models

Yong Guan[1], **Zhiyuan Li**[1], **Shaoru Guo**[2*]
[1]School of Control and Computer Engineering, North China Electric Power University, Beijing, China
[2]School of Computer and Information Technology, Shanxi University, China
yongguan@ncepu.edu.cn, zhiyuan.li@ncepu.edu.cn, gsr@sxu.edu.cn

Abstract

Events exhibit rich semantic relations that are essential for understanding the unfolding of real-world processes. Although large language models (LLMs) have achieved strong performance on event relation extraction, how event relations are internally represented and utilized remains unclear. In this paper, we present UERLens, an interpretability framework for understanding event relations in LLMs. Specifically, we first construct UERBench, a counterfactual dataset for event relation analysis that covers causal, temporal, and sub-event relations. Based on counterfactual pairs, we identify relation-sensitive internal features by comparing model activations. We then examine the functional role of these features through model manipulation, including model intervention and model training. Experimental results show that event relations are encoded through structured and layer-specific internal features. Disabling relation-sensitive features leads to performance drops of over 22%, while enhancing them yields improvements of up to 7%. Furthermore, leveraging these interpretable features to train a lightweight classifier significantly improves event relation extraction, achieving F1 gains of up to 24% for causal relations.[1]

1 Introduction

Event relations capture structured semantic dependencies between events, such as causal, temporal, and sub-event relations (Cheng et al., 2025). In natural language, such relations are often expressed implicitly across sentences. For example, when a narrative mentions a `power outage` and a `train delay`, a meaningful relation can be inferred from context without explicit connectives, as shown in Figure 1. In the era of large language models (LLMs), understanding event relations goes beyond recognizing isolated events and

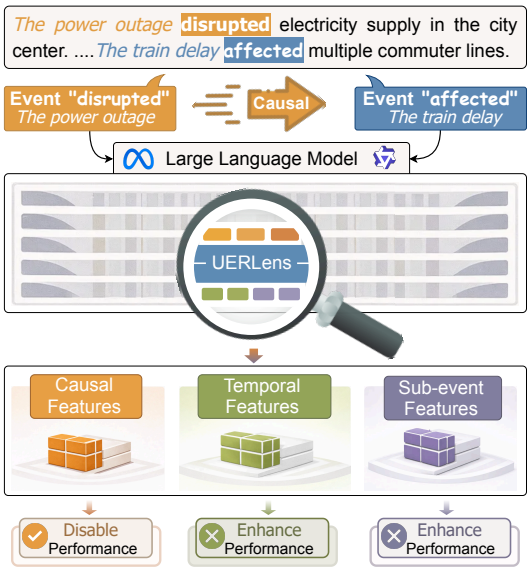


Figure 1: Interpretable event relations in LLMs, where relations are captured by relation-sensitive features.

is crucial for higher-level reasoning and coherent, temporally and causally grounded generation (Li et al., 2025; Mumuni and Mumuni, 2025; Guo et al., 2023).

Existing studies on event relation extraction (ERE), which aims to extract relations between events, have primarily focused on improving model performance. These efforts range from fine-tuning LLMs (Zhu et al., 2023), to instruction-based alignment (Qi et al., 2024), as well as single-agent (Hu et al., 2025) and multi-agent (Guan et al., 2025). In terms of interpretability, prior work has examined the explainability of event detection (Guan et al., 2023) or leveraged event knowledge to generate explanatory analyses (Du et al., 2021; Yao et al., 2024). While these studies provide useful insights into how event information supports interpretability, comparatively little attention has been paid to interpreting event relations themselves, particularly at the level of

*Corresponding authors
[1]https://github.com/ncepu-eai/UERLens

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 463–471

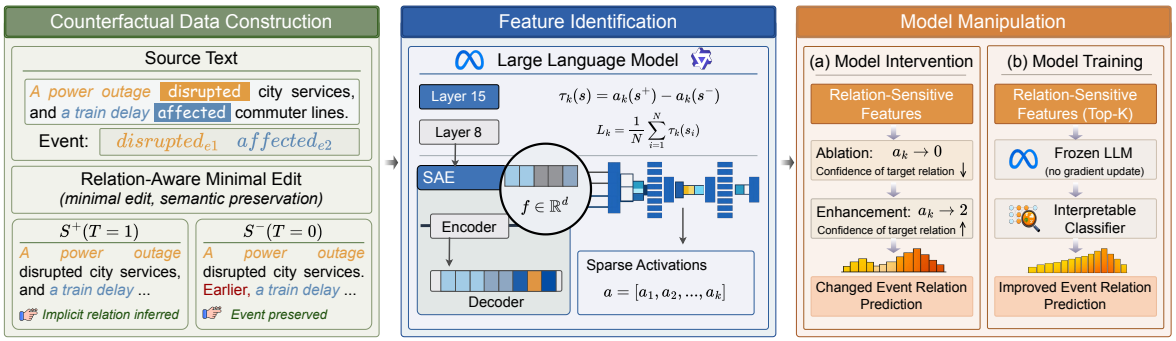


Figure 2: Overview of the UERLens framework. Starting from source event descriptions, UERLens constructs counterfactual pairs to specific event relations, identifies relation-sensitive internal features via activation comparison, and analyzes their functional roles through model intervention and training.

internal model representations. As a result, two key questions remain unexplored: *(i) do LLMs encode event relations in identifiable and relation-sensitive internal features?* and *(ii) do such features play a functional role in supporting event relation understanding and model performance?*

To address this gap, we propose UERLens, an interpretability framework for analyzing the internal representation and utilization of event relations in LLMs. UERLens consists of three main components. First, we construct a large-scale counterfactual dataset covering three fundamental event relation types, namely causal, temporal, and sub-event relations, to enable controlled analysis of relation-specific effects. Second, we adopt sparse auto-encoders (SAEs) to decompose intermediate model activations and identify internal features that are strongly associated with specific event relations. Third, we investigate the functional role of relation-sensitive features through model manipulation, including model intervention and model training. Together, these components enable an in-depth analysis of understanding event relations in LLMs by linking internal representations to observable changes in model behavior and performance. Our main contributions are summarized as follows:

- We propose UERLens, an interpretability framework for analyzing event relations in LLMs, moving beyond token- or entity-level analysis to focus on structured relations such as causal, temporal, and sub-event relations.
- We construct UERBench, a counterfactual dataset for event relation analysis, enabling controlled and fine-grained investigation of relation-specific internal representations.
- We demonstrate that relation-sensitive internal features play a functional role in understanding event relation and can be leveraged to improve event relation extraction using interpretable classifiers.

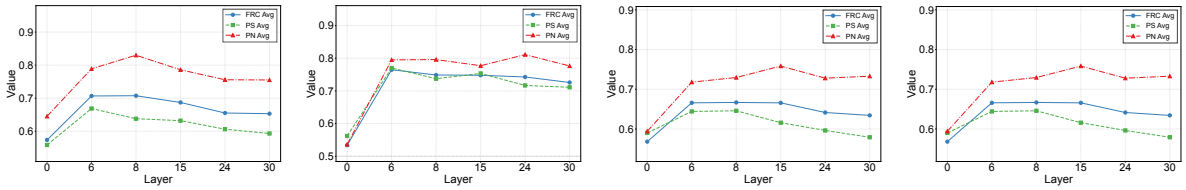
2 Methodology

The framework of UERLens consists of three components as shown in Figure 2. We first construct a large-scale counterfactual dataset to control for the presence of specific event relations. We then identify internal features associated with different event relations by comparing model activations on counterfactual sentence pairs. Finally, we examine the influence of these features on model behavior through targeted intervention and model training.

2.1 Counterfactual Data Construction

To enable controlled analysis of event relation representations, we construct UERBench, a counterfactual dataset grounded in the event schema of MAVEN-ERE (Wang et al., 2022). The core idea is to isolate the presence of a specific event relation while preserving the underlying events and their semantic content, so that differences in model activations can be attributed to the relation itself.

Following established counterfactual analysis practices (Jing et al., 2025), each instance in UERBench consists of a paired sample (s^+, s^-). In s^+, a target event relation (e.g., causal, temporal, or sub-event) holds between two events, whereas in s^- the same events are retained but the target relation is removed through minimal linguistic modification. This pairing enables direct comparison under minimal semantic variation. Detailed construction principles are provided in Appendix A.1.



(a) Causal relation. (b) Temporal relation. (c) Subevent relation. (d) Average of each relation.

Figure 3: Layer-wise distribution of relation-sensitive features for different event relations.

Relations	Sub-relation	#Docs	#Rels
Causal	Cause	1,529	6,797
	Precondition	2,770	9,519
Sub-event	Subevent	1,488	9,193
Temporal	Begins-on	86	253
	Ends-on	52	105
	Contains	2,082	5,923
	Before	1,804	2,913
	Overlap	862	2,887
	Simultaneous	1,348	4,170

Table 1: Data statistics of UERBench. #Docs denotes the number of documents and #Rels denotes the number of relations.

2.2 Feature Identification

To identify internal model features that are sensitive to specific event relations, we analyze model activations of LLMs on counterfactual pairs (s^+, s^-) using SAEs (Muhamed et al., 2025; Tolooshams et al., 2025).

Let $\mathbf{f} \in \mathbb{R}^d$ denote the hidden-state activation at a given LLM layer. An SAE maps $\mathbf{f}$ to a sparse activation vector $\mathbf{a}$, where each dimension a_k corresponds to a latent base vector. For a counterfactual pair (s^+, s^-), we measure the sensitivity of base vector k to the target relation via its latent effect:

$$\tau_k(s) = a_k(s^+) - a_k(s^-) \quad (1)$$

We aggregate this effect over all N samples to obtain the expected average latent effect.

$$L_k = \frac{1}{N} \sum_{i=1}^{N} \tau_k(s_i) \quad (2)$$

At last, we select the top-ranked vectors as relation-sensitive features for subsequent model manipulation and training.

2.3 Model Manipulation

After identifying relation features, we examine their functional role through model manipulation, including model intervention and training.

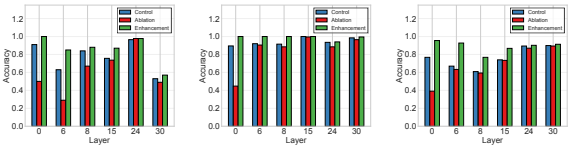
Model Intervention. We perform model interventions by directly modifying the activations of selected relation-sensitive features during forward propagation. Given an input sentence, hidden states at a target layer are encoded into sparse activations via the SAE. For a base vector k, we either set its activation a_k to zero (ablation) or to a fixed high value (enhancement) to suppress or amplify the corresponding feature. The modified activations are decoded and propagated through the remaining layers to obtain the final prediction.

Model Training. We further examine whether relation-sensitive features alone provide informative and discriminative representations for event relation extraction. Rather than aiming to optimize performance, this experiment serves as a complementary validity check to model intervention. We select the top-K features for each relation type. The activations of these features are concatenated into feature vectors and used to train a lightweight and interpretable classifier, namely a support vector machine (SVM). While SVM serves as a concrete instantiation, the framework is not restricted to a specific classifier.

3 Experiments

3.1 Experiment Setup

Dataset Analysis. All experiments are conducted on UERBench, a counterfactual dataset constructed from MAVEN-ERE documents for event relation analysis (Table 1). The dataset covers three relation types, including causal, temporal, and sub-event relations, with a total of eight sub-relations. It exhibits a diverse distribution across relation types, with temporal relations being the most abundant, followed by causal and sub-event



(a) Causal. (b) Temporal. (c) Subevent.

Figure 4: Effects of feature intervention on ERE task.

relations. Except for the Begins-on and Ends-on sub-relations, which contain around 200 instances, all other sub-relations have more than 2,800 instances.

Evaluation Metrics. Following prior work (Jing et al., 2025), we evaluate how well internal features capture event relations using a counterfactual evaluation protocol. Feature sensitivity is quantified with three metrics: Probability of Sufficiency (PS), Probability of Necessity (PN), and Feature Representation Confidence (FRC). Intuitively, PS measures how likely a feature is activated when the target relation is introduced, while PN measures how likely the feature is deactivated when the relation is removed. FRC combines PS and PN via a harmonic mean, penalizing features that are only sufficient or only necessary.

For event relation extraction, we report standard classification metrics, including precision (P), recall (R), and F1 score (F1).

Backbone Models. We conduct experiments on LLaMA-3.1-8B (Grattafiori et al., 2024). For SAEs, we adopt OpenSAE (THU-KEG, 2025) and use the released checkpoints corresponding to all 32 transformer layers [2].

3.2 Feature Identification Analysis

Do LLMs encode event relations in identifiable and relation-sensitive internal features?

Yes. LLMs encode event relations through identifiable and relation-sensitive internal features, which exhibit clear and consistent layer-wise organization across different relation types.

Figure 3 presents the layer-wise distribution of relation-sensitive features. Specifically, relation-sensitive features tend to concentrate in the middle to upper layers of the model, with substantially higher FRC scores than those observed at the embedding layer. This indicates that event relation knowledge is not encoded at the surface lexical level, but emerges at intermediate representational stages. Moreover, while all three relations exhibit similar layer-wise trends, their overall representation strength differs: temporal relations show the strongest feature sensitivity, followed by causal relations, with sub-event relations exhibiting slightly lower but still stable scores. Across relations, the Probability of Necessity is generally higher than the Probability of Sufficiency, suggesting that these features are often required for correct relation prediction, even if they are not solely sufficient on their own.

[2] https://github.com/THU-KEG/OpenSAE

Relations	Method	P	R	F1
Causal	ICL	71.54	63.71	67.40
	RFB	49.06	49.17	47.65
	RSFT	73.26	72.33	72.05
Subevent	ICL	65.82	52.83	58.61
	RFB	52.97	52.17	48.70
	RSFT	70.24	68.17	67.33
Temporal	ICL	64.07	61.27	62.64
	RFB	50.17	50.17	49.66
	RSFT	63.84	63.17	62.72

Table 2: Model performance (%) using randomly selected features (RFB) and relation-sensitive features identified via model intervention (RSFT).

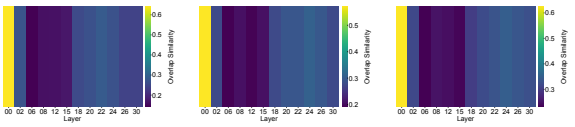
3.3 Model Manipulation

Do relation-sensitive features play a functional role in event relation understanding and model performance?

Yes. Intervening on relation-sensitive features leads to systematic and consistent changes in ERE, indicating that these features play a functional role in event relation understanding.

Results of Model Intervention. Figure 4 shows that intervening on relation-sensitive features consistently alters ERE across causal, temporal, and sub-event relations. Enhancing these features strengthens the corresponding relation predictions, while ablating them leads to clear performance degradation, indicating a functional contribution to event relation understanding. Temporal and causal relations exhibit stronger sensitivity, with sub-event relations showing slightly weaker but stable effects.

For comparison, we conduct control experiments by intervening on non-target relation features. As reported in Appendix C, such interven-



(a) C–S (b) C–T (c) S–T

Figure 5: Pairwise overlap of relation-sensitive features across relations: causal vs. subevent (C–S), causal vs. temporal (C–T), and subevent vs. temporal (S–T).

Layer	FRC	PS	PN
0	0.5740	0.5840	0.5844
6	0.6612	0.6392	0.7164
12	0.6092	0.5936	0.6592
15	0.6372	0.5996	0.7192
24	0.4907	0.4944	0.5452

Table 3: Layer-wise results on Qwen model.

tions lead to substantially weaker and less consistent effects, indicating that the observed behavioral changes are specific to relation-sensitive features rather than resulting from generic activation perturbations.

Results of Model Training. Table 2 reports the performance of classifiers trained on RFB and RSFT. Consistent with the intervention results, classifiers trained on RSFT achieve substantially higher results across all relation types, despite using a lightweight and interpretable model. In particular, for causal relations, RSFT achieves F1 score of 72.05 compared to 47.65 for RFB. Subevent and temporal relations show similar patterns, with F1 improvements of 18.63 and 13.06, respectively. These results indicate that the identified sensitive features form informative and discriminative representations that can be directly exploited by simple models, providing complementary evidence for their functional relevance.

3.4 Cross-Relation Analysis

Figure 5 shows the pairwise overlap of relation-sensitive features across causal, temporal, and subevent relations. We observe a clear layer-wise pattern. At the embedding layer, feature overlap is relatively high across all relation pairs, indicating shared lexical or surface-level representations.

As representations progress into intermediate layers, the overlap decreases substantially. This suggests that the model increasingly separates relation types and forms more relation-specific features. This stage likely corresponds to the emergence of relation-level semantic distinctions.

In higher layers, partial overlap re-emerges, particularly for relation pairs involving temporal relations. Among the three pairs, subevent-temporal relations show the highest overall similarity, followed by causal-temporal and causal-subevent. This pattern suggests that temporal reasoning shares internal features with both causal progression and event decomposition, while causal and sub-event relations remain more distinct. Overall, these results indicate that event relations are encoded in a structured manner. They are initially shared at the lexical level, differentiated in intermediate layers, and partially integrated again at higher levels of abstraction.

3.5 Cross-Model Generalization

To evaluate whether relation-sensitive features generalize beyond a specific model, we conduct additional experiments on Qwen (1.7B), which differs from LLaMA-3.1-8B in both architecture and parameter scale. We follow the same experimental protocol for feature identification and compute FRC, PS, and PN across layers.

Table 3 presents the layer-wise results on Qwen. We observe a consistent trend with the LLaMA-based analysis. Relation-sensitive features exhibit relatively low strength at the embedding layer, become more prominent in intermediate layers (Layers 6–15), and decrease in deeper layers. This consistency across models suggests that event relation information is not encoded at the surface lexical level, but instead emerges at intermediate representational stages.

4 Conclusion

In this paper, we propose UERLens, an interpretability framework for understanding event relation in large language models. We identify relation-sensitive internal features with clear layer-wise structure and demonstrate their functional impact on model predictions. Our analysis reveals that event relations are encoded in a structured manner across layers, emerging at intermediate representational stages rather than being tied to surface-level cues. We further show that these features provide informative representations that can be leveraged by simple models, and that the observed patterns generalize across different datasets and model architectures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62406162, 62506218), and Fundamental Research Program of Shanxi Province (202403021211092).

Limitations

Despite the promising results, this work has several limitations that suggest directions for future research.

First, this work focuses on event relation understanding and does not cover other event related tasks such as event detection, event argument extraction, or cross-document event linking. While our framework is general and potentially applicable to these settings, extending UERLens to a broader range of event understanding tasks remains future work.

Second, our analysis is limited to text-based large language models. Event relations in real-world scenarios are often grounded in multimodal signals such as images or videos. Investigating how event relations are represented and manipulated in multimodal large models is an important direction for future research.

Ethical Considerations

Data Privacy. All data used in this work are derived from publicly available datasets. The dataset does not contain personal, sensitive, or private information.

AI Assistants in Research or Writing. We make limited use of AI assistants during this research. Specifically, GPT-5 is used for language polishing and code debugging during the development process.

References

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *Preprint*, arXiv:2411.10371.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. ExCAR: Event graph knowledge enhanced explainable causal reasoning. In *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)*, pages 2354–2363.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yong Guan, Jiaoyan Chen, Freddy Lecue, Jeff Pan, Juanzi Li, and Ru Li. 2023. Trigger-argument based explanation for event detection. In *Findings of the Association for Computational Linguistics*, pages 5046–5058.

Yong Guan, Hao Peng, Lei Hou, and Juanzi Li. 2025. MMD-ERE: Multi-agent multi-sided debate for event relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6889–6896.

Shaoru Guo, Chenhao Wang, Yubo Chen, Kang Liu, Ru Li, and Jun Zhao. 2023. EventOA: An event ontology alignment benchmark based on FrameNet and Wikidata. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10038–10052.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2025. LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28220–28239.

Bobo Li, Xudong Han, Jiang Liu, Yuzhe Ding, Liqiang Jing, Zhaoqi Zhang, Jinheng Li, Xinya Du, Fei Li, Meishan Zhang, Min Zhang, Aixin Sun, Philip S. Yu, and Hao Fei. 2025. Event extraction in large language model. *Preprint*, arXiv:2512.19537.

Aashiq Muhamed, Mona T. Diab, and Virginia Smith. 2025. Decoding dark matter: Specialized sparse autoencoders for interpreting rare concepts in foundation models. In *Findings of the Association for Computational Linguistics*, pages 1604–1635.

Alhassan Mumuni and Fuseini Mumuni. 2025. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *Preprint*, arXiv:2501.03151.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. ADELIE: Aligning large language models on information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

THU-KEG. 2025. Opensae: Open-sourced sparse autoencoder towards interpreting large language models.

Bahareh Tolooshams, Ailsa Shen, and Anima Anandkumar. 2025. Sparse autoencoder neural operators: Model recovery in function spaces. *Preprint*, arXiv:2509.03738.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941.

Zhenhe Yao, Changhua Pei, Wenxiao Chen, Hanzhang Wang, Liangfei Su, Huai Jiang, Zhe Xie, Xiaohui Nie, and Dan Pei. 2024. Chain-of-event: Interpretable root cause analysis for microservices through automatically learning weighted event causal graph. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, FSE 2024, page 5061.

Fangqi Zhu, Jun Gao, Changlong Yu, Wei Wang, Chen Xu, Xin Mu, Min Yang, and Ruifeng Xu. 2023. A generative approach for script event prediction via contrastive fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Data Construction

A.1 Counterfactual Construction Principles

The construction of counterfactual sentences follows three principles.

- **Minimal edit.** Only the smallest linguistic unit that realizes the target relation is modified, such as replacing or removing a relational trigger, while leaving the rest of the sentence unchanged.
- **Semantic preservation.** The propositional content, argument structure, and discourse context are retained, ensuring that s^+ and s^- remain truth-conditionally equivalent except for the target relation.
- **Logical plausibility.** All counterfactual sentences are required to be grammatically correct and consistent with commonsense reasoning.

Our dataset is grounded in the event relation taxonomy of MAVEN-ERE, which categorizes relations into three major types: causal, temporal, and sub-event relations. For each relation type, we construct counterfactual pairs by selectively removing or altering relation-specific triggers while keeping the involved events unchanged.

B Implementation Details

Counterfactual Data Construction. Negative (counterfactual) instances in UERBench are generated using GPT-5 by minimally removing or altering relation-specific expressions while preserving the underlying events and their semantic content.

Feature Identification Setup. Neuron positioning experiments are conducted on six representative layers of the backbone LLM, namely Layers 0, 6, 8, 15, 24, and 30. For each layer, we compute FRC scores for all SAE base vectors and retain the top-50 vectors per layer, resulting in 300 candidate vectors in total. To further reduce noise and ensure semantic relevance, we employ an LLM-based filtering step to select the most representative 20 vectors from each layer, yielding 120 relation-sensitive base vectors used in subsequent experiments.

Model Intervention Setup. For model intervention experiments, we select the top-20 relation-sensitive base vectors from each of the six layers, resulting in 120 vectors in total. Each base vector is evaluated on 1,200 instances, including 600 positive and 600 negative instances for the target relation. To reduce stochastic variation, each instance is evaluated over 10 repeated forward passes. For enhancement interventions, the activation value of the target vector is set to 2, while for ablation interventions it is set to 0. In the control setting, vector activations are left unchanged.

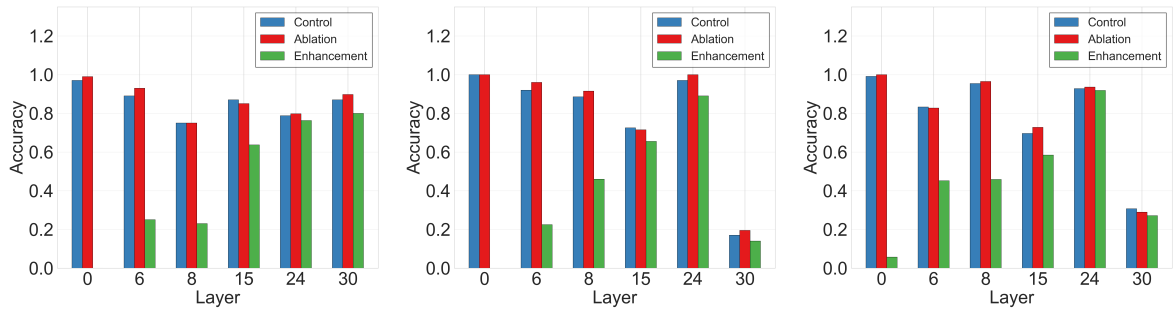
Cross-Relation Overlap Setup. To analyze feature overlap across relation types, we conduct pairwise comparisons among causal, temporal, and sub-event relations. For each relation type, we select the top-500 base vectors per layer according to FRC scores. Overlap analysis is performed across 12 layers (Layers 0, 2, 6, 8, 12, 15, 18, 20, 22, 24, 26, and 30), resulting in 6,000 vectors per relation type. Feature overlap is measured layer-wise to examine how different event relations share or diverge in their internal representations.

Model Training Setup. To evaluate whether relation-sensitive features provide advantages beyond feature dimensionality alone, we compare classifiers trained on two feature sets with identical size and layer distribution. Specifically, the RFB uses 120 features randomly sampled across six layers (20 per layer). Both the feature selection and classifier training are repeated five times with different random seeds, and the reported results are averaged across runs. In contrast, RSFT selects the same number of features per layer based on intervention-derived relation sensitivity.

C Intervention on Non-Target Relation Features

This section reports additional control experiments that examine the effects of intervening on non-target relation features. The results are shown in Figure 6. Specifically, for each event relation type, we apply ablation and enhancement interventions to features that are not associated with the target relation and evaluate the resulting changes in model predictions.

Across causal, temporal, and sub-event relations, we observe that intervening on non-target features leads to substantially weaker and less consistent effects compared to interventions on relation-sensitive features reported in the main text. In many layers, both ablation and enhancement of non-target features result in minimal



(a) Non-Causal relation. (b) Non-Temporal relation. (c) Non-Subevent relation.

Figure 6: Effects of intervening on non-target relation features for different event relation types.

changes or unstable fluctuations in prediction performance, and in some cases even introduce contradictory effects across layers.

Notably, while some layers exhibit non-trivial changes under non-target intervention, these effects do not follow a coherent layer-wise pattern and vary significantly across relation types. This contrasts sharply with the systematic and consistent trends observed when intervening on relation-sensitive features.