Trustworthy and Explainable Causal Representation Learning in Transformers

Yang Liu[1]*, Yinghao Zhang[1]*, Lin Liu[2], Jiuyong Li[2], Debo Cheng[3]†, Zaiwen Feng[1]†
[1]Huazhong Agricultural University,
[2]Adelaide University, [3]Hainan University
zaiwen.feng@mail.hzau.edu.cn, chengd@hainanu.edu.cn

Abstract

A prevalent approach to interpretable representation learning involves creating a mask that weights the significance of each input feature, followed by deriving a masked representation by applying this mask to the input representation. However, the identifiability of these learned masked representations is often uncertain, making the origin of these representations ambiguous or unreliable. Furthermore, the approaches to interpreting Transformer based on attention weights have been criticized for their faithfulness. To address these limitations, we propose a causal framework that directly learns identifiable and explainable representations from attention weights, rather than relying on importance masks. Our framework leverages identifiability theory and causal representation learning to extract explainable representations within a subspace of input representations, effectively transforming frozen representation learning methods into self-explaining systems. Experimental results on real-world datasets demonstrate that, compared to well-established state-of-the-art methods, our approach provides identifiable and more trustworthy explanations while guaranteeing faithfulness.

1 Introduction

Explainable AI (XAI) has garnered significant attention as a way to explain model outputs and enhance transparency, especially due to the recent advancements in the Transformer model across various tasks (Vaswani et al., 2017; Yang et al., 2025). In natural language processing (NLP), most existing XAI methods apart from those focusing on natural language explanations, typically involve learning a mask that quantifies the importance of each input feature (Zhao et al., 2024). This mask is then applied to the original input representation to derive an explainable representation, referred to in this paper as explainable **masked representations**. However, there is no theoretical guarantee that these explainable masked representations from data are identifiable, which is crucial for providing trustworthy explanations.

Identifiability means that the variables of interest are uniquely determined and expressed through the true observed distribution (Wu and Fukumizu, 2021), that is, the learned representations are trustworthy and derived from data, not meaningless or erroneous (Lewbel, 2019). Two fundamental criteria for evaluating XAI methods are faithfulness and trustworthiness, with identifiability being essential for achieving both. Faithfulness measures how well an interpretation reflects the decision-making process actually used by the model, whereas trustworthiness concerns whether the explanation is grounded in a reliable source and whether it remains consistent with human-annotated evidence. These notions are related but distinct: faithfulness focuses on correspondence to model behavior, while trustworthiness emphasizes whether the learned explanation itself has a well-founded origin (Xu and Yang, 2025; Zhao et al., 2024).

As a necessary condition for trustworthiness, identifiability has become increasingly important in the field of XAI, especially as many methods use one neural network to explain another (Zaigrajew et al., 2025; Wu et al., 2023; Zhang et al., 2023; Møller et al., 2024). However, most XAI methods do not ensure that explainable masked representations from data are identifiable (Xu and Yang, 2025; Lundberg and Lee, 2017), implying that these methods may violate both principles. (Zhang et al., 2023) addresses the identifiability of learned masks, but does not further consider masked representations. This raises the question: can masked representations be both identifiable and explainable? Current researches on mask-based interpretability offer no clear answer.

*These authors contributed equally to this work.
†Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 27482–27501

XAI methods based on attention weights utilize the intermediate layer parameters of a neural network to compute feature importance (Xu and Yang, 2025). However, the faithfulness of explanations for model predictions based on attention weights is often debated. Studies have shown that self-attention can exhibit bias towards specific positions or irrelevant tokens, hindering its ability to accurately reflect model predictions based on the actual input features (Bai et al., 2021; Serrano and Smith, 2019). Conversely, other studies have suggested that attention weights can provide faithful explanations in certain contexts (Wang et al., 2025; Jacovi and Goldberg, 2020). Given that Transformer models contain multiple layers of multihead attention and feed-forward network (FFN) layers, directly weighting and summing these layers to obtain feature importance inevitably leads to unfaithful explanations.

A fundamental challenge in XAI is ensuring both trustworthiness and faithfulness in model explanations. A promising approach to address this issue is to construct interpretable models within a subspace of input representations, while simultaneously emulating the reasoning process of the black-box model to achieve more faithful approximations (Jacovi and Goldberg, 2020).

Motivated by these insights and the principle of identifiability, we propose a causal surrogate model, illustrated in Fig.1. This model captures the reasoning process of the Transformer within a subspace of input representations, leveraging recent advancements in causal inference and structural causal models (SCMs) (Pearl, 2009) to enhance interpretability in large language models (LLMs) (Wu et al., 2023; Zhou et al., 2023). By incorporating nonlinear independent component analysis (ICA) (Hyvarinen et al., 2019) and identifiable variational autoencoder (VAE) (Khemakhem et al., 2020a), we connect causal representation learning with Transformer explanation through an identifiable surrogate framework.

To provide identifiable explanations for Transformers, we introduce an XAI framework for Transformers, termed the Causal Identifiable Explanation Model (CIEM)[1]. Departing from traditional mask-based explanation methods, CIEM leverages causal representation learning to extract identifiable and interpretable representations directly from our proposed causal model. A distinguishing feature of CIEM is its ability to selectively explain different layers of Transformer models, effectively transforming them into self-explaining systems with broad applicability to other representation learning models. CIEM demonstrates the feasibility of learning identifiable and trustworthy representations within a subspace of input representations, which contributes to narrowing the deficiency in model transparency. Our main contributions are summarized as follows:

- We first examine identifiability in the evaluation of trustworthiness in explainable representations and propose an XAI framework for Transformers that learns identifiable representations instead of importance masks.
- We propose a novel causal model (CIEM) to learn identifiable and explainable representations, providing theoretically-grounded explanations for Transformers. CIEM integrates causal representation learning and identifiability theory, ensuring that learned explanations are grounded in the true data distribution rather than being arbitrary or misleading.
- Extensive experiments demonstrate that, compared to state-of-the-art approaches, CIEM provides both identifiable and more trustworthy explanations while ensuring faithfulness.

2 Related Work

Recent research has explored neural network-based XAI methods. One prominent direction is rationalization extraction, which employs a selector and a predictor network to learn coherent and continuous importance masks (Liu et al., 2025; Yuan et al., 2025; Zhang et al., 2023; Liu et al., 2022). However, these approaches are often limited to rationalization extraction methods and do not generalize beyond them. Other studies have aimed to directly learn masks of important features by designing frameworks that enhance faithfulness (Sun et al., 2025; Wu et al., 2023; Nie et al., 2025). (Zhang et al., 2023) discusses the identifiability of importance masks but does not address the identifiability of masked representations. Beyond mask-learning approaches, prior Transformer explanation methods also include attention-based attribution, as well as perturbation and gradient-based attribution methods (Luo et al., 2024). These lines of work mainly target attribution quality or faithfulness, whereas the trustworthiness and identifiability of learned

[1] https://github.com/LORD-ROY/CIEM

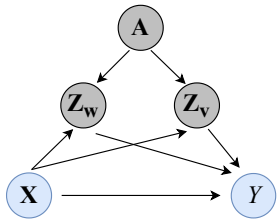

Figure 1: The causal DAG representing the data generation process for CIEM. $\mathbf{X}$ is the observational data. $\mathbf{A}$ is the output of a specific layer $f_{\phi_k}(\cdot)$ of the multilayer representation learning method. $\mathbf{Z_w}$ and $\mathbf{Z_v}$ represent the weight and value matrices, respectively. Y is jointly generated by $\mathbf{Z_w}$, $\mathbf{Z_v}$ and $\mathbf{X}$ through $f_h(\cdot)$. Arrows represent the causal relationships between variables.

representations are usually left implicit. As a result, trustworthiness remains a significant gap in the field. Additional discussions on XAI methods can be found in the Appendix C.

Causal inference has recently been integrated into XAI to establish causal relationships between input features and model outputs (Carloni et al., 2025; Rohekar et al., 2024; Zhang et al., 2023; Wu et al., 2023; Xu et al., 2024). (Wu et al., 2023) proposes a causal framework to improve faithfulness, but it does not analyze trustworthiness or identifiability. Developing a causal approach that ensures both faithfulness and trustworthiness remains an open challenge.

Identifiable Representation Learning. The application of independent component analysis (ICA) to representation learning has led to nonlinear ICA, which has been increasingly applied in causal inference. For instance, (Khemakhem et al., 2020a) provides a theoretical framework for learning identifiable representations within variational autoencoders (iVAEs). Additional details on the implementation of identifiable VAE (iVAE) (Khemakhem et al., 2020a) can be found in the Appendix B.2.

3 Causal Model in Representation Space of Attention

3.1 Preliminaries

Notations. We use $f_{\phi_k}(\cdot)$ to denote the first k layers in a Transformer to be interpreted and $f_\phi(\mathbf{X})$ represents the final output of the interpreted Transformer, where the input $\mathbf{X} \in \mathbb{R}^{n \times d_m}$ has n tokens, each with an embedding dimension of d_m. Let $f(\cdot)$ be an LLM and $f_h(\cdot)$ denote its head. The *head layer* of the LLM $f_h(\cdot)$ maps $f_\phi(\mathbf{X})$ to probability distributions over the vocabulary or class labels, such that $f(\mathbf{X}) = f_h(f_\phi(\mathbf{X}))$. Taking a post-hoc interpretation, the final output $f(\mathbf{X})$ is denoted as Y. Throughout this paper, bold uppercase letters denote matrix variables (e.g., $\mathbf{X}$), while bold lowercase letters represent vectors (e.g., $\mathbf{b}$).

Self-Attention mechanism. Each head's attention weights are computed identically as follows:

$$\mathbf{Attention} = softmax(\frac{\mathbf{X}\mathbf{W_Q}\mathbf{W_K^T}\mathbf{X^T}}{\sqrt{d_k}})\mathbf{X}\mathbf{W_V}, \quad (1)$$

where $\mathbf{W_Q}, \mathbf{W_K} \in \mathbb{R}^{d_m \times d_k}$ and $\mathbf{W_V} \in \mathbb{R}^{d_m \times d_v}$ are neural network parameters.

Structural causal model. A structural causal model (SCM) is defined as a directed acyclic graph (DAG) consisting of nodes and edges (Pearl, 2009). Each node represents either an exogenous or endogenous variable, while each edge signifies a causal relationship between two variables. A formal definition of a structural causal model is provided in the Appendix B.1. The causal diagram for our model is shown in Fig. 1.

3.2 The Proposed Causal Model for Attention Representation Learning

Relying solely on the attention weights in Eq 1 may not offer a faithful explanation of the model's predictions, since $f_{\phi_k}(\mathbf{X})$ incorporates the outputs of FFN layers—particularly in advanced Mixture of Experts (MoE) models, which contain numerous such layers (Liu et al., 2024). To better understand the reasoning process of the self-attention mechanism in Eq. 1, we define the weight matrix produced by the $softmax(\cdot)$ function as $\mathbf{Z_w} \in \mathbb{R}^{n \times n}$, and denote the value projection matrix $\mathbf{W_V}$ as $\mathbf{Z_v}$. Here, $\mathbf{Z_w}(i, j)$ denotes the importance of the j-th token to the i-th token. Given the original input $\mathbf{X}$, the matrix product $\mathbf{Z_w}\mathbf{X}$ integrates the relative importance of different tokens, and the subsequent linear transformation by $\mathbf{Z_v}$ yields the final attention-based representation.

Based on the above analysis, we propose an interpretable module to learn representations $\mathbf{Z_w}$ and $\mathbf{Z_v}$ through $f_{\phi_k}(\mathbf{X})$ for each Transformer layer, rather than relying solely on the self-attention outputs. This distinction is important because directly extracting or reweighting the original attention components may inherit biases from raw attention and does not place the explanation in an identifiable surrogate framework. CIEM instead learns $\mathbf{Z_w}$ and $\mathbf{Z_v}$ as latent variables aligned with the behavior of the interpreted Transformer, so that the resulting explanation can be analyzed for both faithfulness

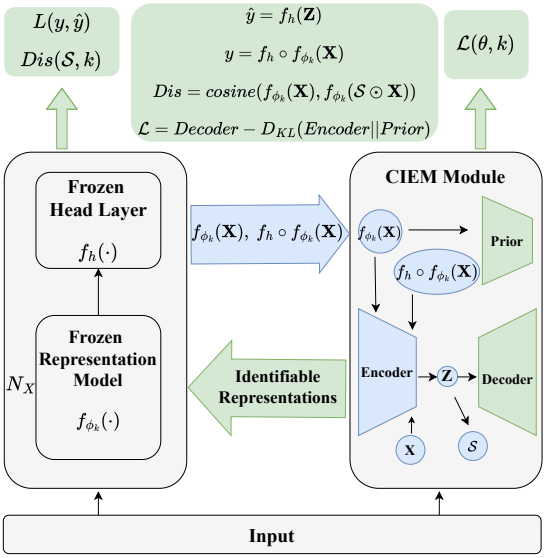

Figure 2: Overview of the CIEM Framework. A frozen representation learning model is presented on the left, while a pluggable CIEM module is shown on the right, together forming a self-explaining system. All arrows indicate the flow of information within this system. Blue components represent the structures and parameters used during inference, whereas green components denote the additional structures and parameters utilized during training. N_X refers to a specific layer in the Transformer. The Decoder, Prior, and Encoder in the CIEM module correspond to Eq. 5, 6, 7, respectively.

and identifiability. For instance, one may augment only $f_\phi(\cdot)$ with such an interpretable module, or apply distinct interpretable modules to each $f_{\phi_k}(\cdot)$. We further provide a theoretical guarantee that the representations $\mathbf{Z_w}$ and $\mathbf{Z_v}$ learned by our CIEM model are identifiable in Sec. 4.

We now discuss how to ensure that the interpretation module to have a reasoning process similar to that of the interpreted Transformer. Fig. 1 illustrates the reasoning process of our proposed interpretable module. The graph explicitly separates attention-related and value-related factors while retaining their joint influence on the final prediction. This separation is useful because it makes it possible to analyze how token interaction and value-related information contribute to the output within a tractable framework, without claiming that the original Transformer itself exposes explicit causal variables. $\mathbf{A}_k$ denotes the attention representation $f_{\phi_k}(\mathbf{X})$ given $\mathbf{X}$, while the weight matrix $\mathbf{Z_w}$ and the value matrix $\mathbf{Z_v}$ are potential factors learned from the input representation $\mathbf{X}$ with the auxiliary of $\mathbf{A}_k$. The final output Y is determined by the combination of $\mathbf{Z_w}$, $\mathbf{Z_v}$, and $\mathbf{X}$. The underlying causal mechanisms are formalized as follows:

$$\begin{aligned} \mathbf{A}_k &= f_{\phi_k}(\mathbf{X}), \quad \mathbf{Z_w} = q_1(\mathbf{X}, \mathbf{A_k}), \\ \mathbf{Z_v} &= q_2(\mathbf{X}, \mathbf{A_k}),\ Y = f_h(\mathbf{Z_w}, \mathbf{X}, \mathbf{Z_v}), \end{aligned} \tag{2}$$

where $q_1(\cdot)$ and $q_2(\cdot)$ denote the underlying causal mechanisms, and $\mathbf{A}_k$ is derived from the interpreted $f_{\phi_k}(\cdot)$. Although $\mathbf{X}$ in Fig. 1 is not a direct cause of $\mathbf{A}_k$, $\mathbf{A}_k$ can be derived from $f_{\phi_k}(\cdot)$ without requiring any additional processing. Therefore, $\mathbf{A}_k$ is directly utilized in our model. Eq. 2 illustrates the source of $\mathbf{A}_k$. Y stems from $f_h(\cdot)$ and will be used as a condition in the data generation process in Sec. 4. Additionally, we transform the potential exogenous variables into additive noise in Eq. 5. Overall, $f(\cdot)$ is transformed into a self-explaining system by integrating an optional interpretable module, while keeping all parameters of the original model fixed.

The importance score of the j-th token across all tokens can be obtained by summing the values in the j-th column of the weight matrix $\mathbf{Z_w}$. Formally, the importance score for token j is given by:

$$S(x_j) = Norm(\sum_{i=1}^{n} \mathbf{Z_w}(i, j)\ ),\ j = 1, \cdots, n \tag{3}$$

where the index pair (i, j) denotes the role of the $j\text{-}th$ token with respect to the $i\text{-}th$ token, and $Norm$ refers to $Max\text{-}Min$ normalization. Let $\mathcal{S} = \{S(x_j)\}_{1 \le j \le n}$ denote the set of importance scores for all tokens, which corresponds to the mask produced by the relevant interpretability method. $\mathbf{Z_v}$ is treated in CIEM as a value-related latent representation rather than as a directly visualized token-level explanation. Accordingly, we do not evaluate $\mathbf{Z_v}$ in isolation. Instead, its explanatory role is assessed indirectly through the composite representation $\mathbf{Z} = \mathbf{Z_w}(\mathbf{X}\mathbf{Z_v})$ and its ability to preserve the prediction behavior of the original model.

The overview of CIEM model is presented in Fig. 2, where the parameters of the interpreted Transformer are fixed. The decoder, prior, and posterior, represented by trapezoids, are all neural networks designed to learn the distinct distributions of $\mathbf{Z_w}$ and $\mathbf{Z_v}$, corresponding to Eq. 5, 6, 7 in Sec 4.1, respectively. Appendix B.2, D.2 details the implementation of distinct distributions for the learned representations. The green components are used for training, while only the blue components are required for inference. In Sec.4, we will explain the right half of Fig.2 and the data generation

process shown in Fig.1, and provide a theoretical justification for the identifiability of our generative model.

4 Identifiability and Faithfulness

4.1 Generative Model and Identifiability

CIEM introduces an auxiliary generative model not to replace the explained Transformer, but to construct a surrogate representation space in which identifiability can be analyzed. In particular, the auxiliary variable $\mathbf{A} = f_{\phi_k}(\mathbf{X})$ is needed to define a conditional prior over latent variables, since unconditional latent-variable models are generally unidentifiable (Khemakhem et al., 2020a). The conditional variable $Y = f(\mathbf{X})$ further ties the learned representation to the prediction being explained, so that the surrogate representation remains aligned with the behavior of the original model. Our objective is to learn a generative model on an augmented dataset $\mathcal{D} = \{(\mathbf{X}_i, Y_i, \mathbf{A}_i^k)\}_{i=1}^N$ and to jointly generate representations $\mathbf{Z_w}$ and $\mathbf{Z_v}$ simultaneously, following the causal schema illustrated in Fig.1. Here, $\mathbf{A}_i^k$ refers to $f_{\phi_k}(\mathbf{X}_i)$ given a pre-trained Transformer, and Y_i is the output of the overall model $f(\mathbf{X}_i)$. The training dataset $\mathcal{D}$ is thus determined by the input $\mathbf{X}$ and the interpreted Transformer $f_{\phi_k}(\cdot)$. Directly learning $\mathbf{Z_w}$ and $\mathbf{Z_v}$ is nontrivial due to their structural complexity. To address this, we adopt a factorized learning strategy. Specifically, rather than learning $\mathbf{Z_v}$ directly, we learn the product $\mathbf{XZ_v} \in \mathbb{R}^{n \times d_m}$, which provides a more interpretable and accessible representation compared to the raw network parameters in $\mathbf{Z_v}$. For $\mathbf{Z_w}$, we introduce two low-rank matrices $\mathbf{Z_{wa}}$ and $\mathbf{Z_{wb}} \in \mathbb{R}^{n \times r}$, and define $\mathbf{Z_w}$ as their matrix product: $\mathbf{Z_{wb}}\mathbf{Z_{wa}}^T$. This formulation enables the simultaneous learning of $\mathbf{Z_w}$ and $\mathbf{Z_v}$ through the composite representation $\mathbf{Z_w}(\mathbf{XZ_v})$. To simplify the presentation, we denote the composite representation $\mathbf{Z_w}(\mathbf{XZ_v})$ as $\mathbf{Z}$, i.e., $\mathbf{Z} = \mathbf{Z_w}(\mathbf{XZ_v})$, where the dimension of $\mathbf{Z}$ is $d_z := \dim(\mathbf{Z}) = 2r + d_m$.

we specify our generative model $p_\theta(\mathbf{X}, \mathbf{Z}|\ \mathbf{A}, Y)$ in Eq. 4 and show its identifiability. For simplicity, $\mathbf{A}$ will be used in place of $\mathbf{A_k}$ in the subsequent equations.

$$p_\theta(\mathbf{X}, \mathbf{Z}|\mathbf{A}, Y) = p_g(\mathbf{X}|\mathbf{Z}, Y)p_\lambda(\mathbf{Z}|\mathbf{A}, Y) \quad (4)$$

$$p_g(\mathbf{X}|\ \mathbf{Z}, Y) = p_\epsilon(\mathbf{X} - g(\mathbf{Z}, Y)) \quad (5)$$

$$p_\lambda(\mathbf{Z}|\ \mathbf{A}, Y) \sim \mathcal{N}(\mathbf{Z}; \mu(\mathbf{A}, Y), \sigma^2(\mathbf{A}, Y)) \quad (6)$$

where $\theta := (g, \mu, \sigma)$ represents the set of model parameters, and $\mathbf{A}$ and Y are treated as auxiliary and conditional variables, respectively. The conditional prior distribution over the latent variable $\mathbf{Z}$ is denoted as $p_\lambda(\mathbf{Z} \mid \mathbf{A}, Y)$ and is defined as a factorized Gaussian belonging to the exponential family. Its natural parameters depend on $\mathbf{A}$ and Y, given by $\lambda(\mathbf{A}, Y) := \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$. The decoder $p_g(\mathbf{X}|\ \mathbf{Z}, Y)$ implies that $\mathbf{X}$ can be decomposed as $\mathbf{X} = g(\mathbf{Z}, Y) + \epsilon$, where ϵ denotes a noise term that is independent of $\mathbf{Z}$, Y, and $g(\cdot)$, and follows a distribution $p_\epsilon(\epsilon)$. The latent representation $\mathbf{Z}$ used in $g(\mathbf{Z}, Y)$ is sampled from the posterior distribution, ensuring that it is dependent on the observed input $\mathbf{X}$, thereby improving the efficiency and relevance of the generative process. As defined in Eq.7, the posterior distribution over $\mathbf{Z}$ (i.e., the encoder) is modeled as a factorized Gaussian. This posterior integrates the generative mechanisms $q_1(\cdot)$ and $q_2(\cdot)$ as described in Eq.2.

$$q(\mathbf{Z}|\ \mathbf{X}, \mathbf{A}, Y) \sim \mathcal{N}(\mathbf{Z}; \mu'(\mathbf{X}, \mathbf{A}, Y), \sigma'^2(\mathbf{X}, \mathbf{A}, Y)) \quad (7)$$

To optimize our CIEM model, we derive a modified evidence lower bound (ELBO) based on the standard VAE formulation:

$$\begin{aligned}\log p_\theta(\mathbf{X}|\mathbf{A}, Y) \geq \mathcal{L}(\theta, k) := \mathbb{E}_{z\sim q} \log p_g(\mathbf{X}|\ \mathbf{Z}, Y) \\ - D_{KL}(q(\mathbf{Z}|\mathbf{X}, \mathbf{A}, Y)||p_\lambda(\mathbf{Z}|\mathbf{A}, Y))\end{aligned} \quad (8)$$

where D_{KL} represents the Kullback-Leibler (KL) divergence. Since both the conditional prior $p_\lambda(\mathbf{Z}|\mathbf{A}, Y)$ and the conditional posterior $q(\mathbf{Z}|\mathbf{X}, \mathbf{A}, Y)$ are factorized Gaussian distributions, the decoder $p_g(\mathbf{X}|\mathbf{Z}, Y)$, which represents the likelihood distribution given the posterior of $\mathbf{Z}$, is also defined as a factorized Gaussian distribution. For the explicit formulation of $p_g(\mathbf{X} \mid \mathbf{Z}, Y)$ and the simplified expression of the ELBO in Eq.8, please refer to AppendixD.2.

Note that $\mathbf{A}$ is a required auxiliary variable, while Y is an optional conditional variable. Specifically, our method can be applied without conditioning on Y if the explained model does not include a *head layer*. However, in practice, the explained model typically contains a *head layer*, so we condition on Y to ensure that the learned representation $\mathbf{Z}$ is more closely tied to the model's output. In our framework, Y is generated by the representation $\mathbf{Z}$ through the *head layer* $f_h(\cdot)$ of the explained model, without the need for additional training parameters. This approach enhances the faithfulness of the learned representations. For details on the implementation of the unconditional schema and

further information on auxiliary and conditional variables, refer to Appendix D.2. The identifiability result in this section applies to the CIEM surrogate generative model in Eq. 4, rather than to the full Transformer itself. Our goal is to characterize when the learned latent representation $\mathbf{Z}$ is identifiable within this surrogate framework while staying aligned with the explained model. Building on the theory of iVAE (Khemakhem et al., 2020b), the identifiability of our model is formally defined as follows.

Definition 4.1. *Let θ and $\tilde{\theta}$ be two parameters in the model parameter space Θ, and $\sim$ be an equivalence relation on Θ. We say that the generative model in Eq. 4 is identifiable if θ and $\tilde{\theta}$ satisfy the following condition:*

$$\forall\, \theta, \tilde{\theta} \in \Theta : p_\theta(\mathbf{X}|\mathbf{A}, Y) = p_{\tilde{\theta}}(\mathbf{X}|\mathbf{A}, Y) \Rightarrow \theta \sim \tilde{\theta} \quad (9)$$

Suppose θ and $\tilde{\theta}$ are the true and learned parameters, respectively. If the conditional marginal distributions of θ and $\tilde{\theta}$ are identical, then their conditional joint distributions must also be the same, i.e., $p_\theta(\mathbf{X}, \mathbf{Z}|\mathbf{A}, Y) = p_{\tilde{\theta}}(\mathbf{X}, \mathbf{Z}|\mathbf{A}, Y)$, which implies that the learned representation $\mathbf{Z}$ is trustworthy. The equivalence relation $\sim$ on Θ is defined by our model parameters.

Definition 4.2. *The equivalence relation $\sim$ on Θ is defined as*

$$\begin{aligned} &\exists\, \mathbf{Q},\ \mathbf{b}, \\ &g^{-1}(\mathbf{X}, Y) = \mathbf{Q}\, \tilde{g}^{-1}(\mathbf{X}, Y) + \mathbf{b},\ \forall\, (\mathbf{X}, Y) \in \mathcal{D}, \end{aligned} \quad (10)$$

where $\mathbf{Q}$ is an invertible $d_z \times d_z$ diagonal matrix and $\mathbf{b}$ is a d_z-dimensional vector.

We ensure that the conditional marginal distributions of θ and $\tilde{\theta}$ are approximately the same through Eq. 8. If θ and $\tilde{\theta}$ satisfy the condition in Eq. 10, they will satisfy Eq. 9. This, in turn, implies that the generative model defined in Eq.4 is identifiable, thereby establishing the trustworthiness of the learned representation $\mathbf{Z}$. The following theorem, inherited from iVAE (Khemakhem et al., 2020b), states the conditions under which our model satisfies Eq. 10 and thus ensures identifiability.

Theorem 4.1. *If the generative model defined in Eq.4 satisfies the following conditions:*

(i) The function g in Eq.5 is injective.

(ii) The function g in Eq.5 is differentiable.

(iii) $\lambda(\mathbf{A}, Y)$ is non-degenerate, i.e., there exist $2d_z$+1 points $(\mathbf{A}_0, Y_0), (\mathbf{A}_1, Y_1), \ldots, (\mathbf{A}_{2d_z+1}, Y_{2d_z+1}) \in \mathcal{D}$ such that the $2d_z \times 2d_z$ matrix $\mathbf{L} := [\gamma_1, \cdots, \gamma_{2d_z}]$ is invertible, where $\gamma_i := \lambda(\mathbf{A}_i, Y_i) - \lambda(\mathbf{A}_0, Y_0)$.

Then the parameters θ and $\tilde{\theta}$ satisfy Eq.10.

The identifiability of our model ensures that the true parameter θ can be identified (learned) by $\tilde{\theta}$ up to an affine transformation. For a detailed proof, see Definition 4.2 and Theorem 4.1 in the Appendix A. Appendix D.2 provides a more in-depth discussion of how identifiability plays a critical role in interpretable methods. Practically, condition *(i)* is compatible with Transformer models built from continuous operators and smooth activations such as GELU or SwiGLU; condition *(ii)* is satisfied by any network capable of backpropagation; and condition *(iii)* requires sufficiently rich auxiliary and conditional variation together with enough data to identify the latent structure. These are sufficient conditions for the identifiability of the CIEM surrogate generative model, rather than claims that every regular LLM automatically satisfies them exactly. In practice, some modern Transformer models may approximate these conditions well enough for CIEM to be useful, while violations can still arise for models with strongly non-injective activations, insufficient training, or inadequate data.

4.2 Faithfulness Module

The importance mask $\mathcal{S}$, generated by Eq. 3, is used to enhance the faithfulness of $\mathbf{Z_w}$ and to promote a clear distinction between $\mathbf{Z_w}$ and $\mathbf{Z_v}$. The masked representation produced by $\mathcal{S}$ should yield attention outputs that closely resemble those produced by the original representation (Møller et al., 2024). In other words, when passed through the attention mechanism $f_{\phi_k}(\cdot)$, both the masked and original representations should result in similar outputs. To quantify this similarity, we define the following distance measure:

$$Dis(\mathcal{S}, k) = cosine(f_{\phi_k}(\mathcal{S} \odot \mathbf{X}), f_{\phi_k}(\mathbf{X})) \quad (11)$$

where $\odot$ denotes element-wise product, $\mathcal{S} \odot \mathbf{X}$ is the masked representations for mask $\mathcal{S}$, and $cosine$ is cosine similarity. Incorporating this faithfulness module into the generative modeling process, we define the final loss function of our CIEM framework as:

$$\min_\theta \mathbb{E}_{X\sim\mathcal{D}}\Big[L(y, \hat{y}) - \alpha\mathcal{L}(\theta, k) - \tau Dis(\mathcal{S}, k)\Big] \quad (12)$$

where $L(y, \hat{y})$ is an optional cross-entropy loss with $y = f(\mathbf{X})$ and $\hat{y} = f_h(\mathbf{Z})$, and $\alpha, \tau > 0$ are

Methods	Explanation 1				Explanation 2				Explanation 3			
	Acc(↑)	P(↑)	R(↑)	F1(↑)	Acc(↑)	P(↑)	R(↑)	F1(↑)	Acc(↑)	P(↑)	R(↑)	F1(↑)
LIME	60.4	29.78	44.69	35.74	60.12	27.07	43.68	33.43	60.48	26.82	44.32	33.42
KernelSHAP	57.84	26.32	39.49	31.59	58.3	24.61	39.71	30.39	58.99	24.81	40.99	30.91
IG	59.47	28.53	42.80	34.24	59.86	26.72	43.11	32.99	60.22	26.47	43.73	32.98
RawAttention	56.09	23.96	35.95	28.75	56.18	21.75	35.09	26.85	57.02	22.15	36.6	27.6
Occlusion	61.11	30.74	46.12	36.89	59.94	26.82	43.27	33.12	60.4	26.72	44.15	33.29
CIMI	65.23	36.31	54.48	43.58	66.21	35.31	56.97	43.6	65.82	34.05	56.27	42.43
CIEM(our)	**68.24**	**40.38**	**60.59**	**48.46**	**69.7**	**39.98**	**64.51**	**49.36**	**68.83**	**38.12**	**62.99**	**47.5**

Table 1: Trustworthiness results (%) on manually labeled datasets with the top 40% of the importance mask used for evaluation.

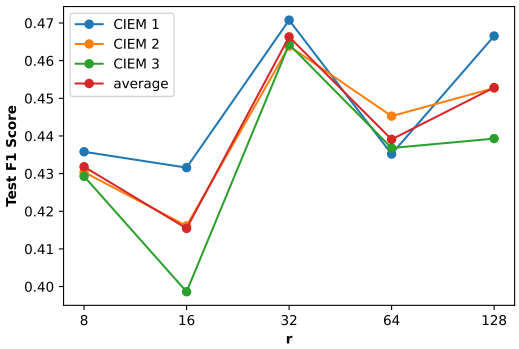

Figure 3: F1 scores of CIEM on test sets with different values of r, while keeping α and τ at their initial values of 1.0.

weighting hyperparameters that balance the ELBO and faithfulness terms.

5 Experiments

5.1 Experimental setup

Datasets. **e-SNLI** (Camburu et al., 2018) is utilized to assess trustworthiness, combining its *contradiction* and *entailment* annotations for evaluation. Faithfulness experiments are conducted on **SST-2** (Socher et al., 2013), **IMDB** (Maas et al., 2011), and two categories of **AGNews** (Zhang et al., 2015). Data processing details are in Appendix E.1.

Metrics. For the trustworthiness experiments, we evaluate the generated explanations based on Accuracy, Precision, Recall, and F1 score, assessing how closely the explanations align with manually labeled important features. For the faithfulness experiments, we adopt Comprehensiveness (COMP) (DeYoung et al., 2020) and Sufficiency (SUFF) (DeYoung et al., 2020) as the primary metrics for faithfulness (Chan et al., 2022). COMP and SUFF assign zero to the rows corresponding to the N most and N least important features in the original representation, respectively, and compare the changes in output probabilities before and after removal. COMP evaluates the importance of the key features, while SUFF assesses the insignificance of the unimportant features. However, we observed that using a fixed selection of N features across different datasets led to suboptimal results for longer datasets. To address this, we optimized the experimental setup by selecting N% of the most and least important features, ensuring that N% is less than 50% to avoid overlap in feature selection. This adjustment improves fairness across datasets.

Fig. 4 illustrates the trend of F1 scores as the percentage of selected tokens increases. The results show that the F1 scores of CIEM and CIMI consistently improve when the selected percentage is below 50%. This suggests that the percentage of highlighted tokens does not significantly affect the relative performance between methods. Additional trends and an explanation for selecting the Top-40% of important features can be found in Appendix E.2.

Table 3 shows example explanations generated by the different methods, and more examples can be found in Appendix E.6. These qualitative examples provide illustrative support for the quantitative results, but they do not replace a dedicated human evaluation protocol.

5.2 Sensitivity Analysis and Ablation Studies

Ablation studies are conducted sequentially on the parameters r, τ, and α. First, sensitivity experiments are performed on r, with α and τ fixed at an initial value of 1.0. Fig. 3 shows the experimental results for different values of r on the test sets. The results indicate that both excessively small and large values of r yield suboptimal outcomes. Based on these observations, we select $r = 32$ as the optimal value.

Next, we conduct an optimization process over a broad range of τ values to investigate how the scaling relationship between τ and α influences the results. Finally, we refine our search by experimenting with a narrower range of α values to identify the optimal α. Appendix E.3 presents the results

Methods	e-SNLI		SST-2		IMDB		AGNews	
	COMP(↑)	SUFF(↓)	COMP(↑)	SUFF(↓)	COMP(↑)	SUFF(↓)	COMP(↑)	SUFF(↓)
LIME	0.445	0.347	0.396	0.118	0.205	0.173	0.13	0.058
KernelSHAP	0.382	0.381	0.285	0.189	0.197	0.207	0.106	0.067
IG	0.579	0.224	0.464	0.097	0.431	0.047	0.325	0.004
RawAttention	0.172	0.506	0.248	0.252	0.231	0.227	0.153	0.091
Occlusion	0.534	0.266	0.321	0.103	0.194	0.087	0.103	**0.004**
CIMI	**0.688**	0.244	**0.648**	**0.037**	**0.684**	0.042	0.401	0.004
CIEM(our)	0.618	**0.174**	0.646	0.047	0.64	**0.041**	**0.414**	0.012

Table 2: Faithfulness results on real-world datasets, with the best overall method highlighted in bold.

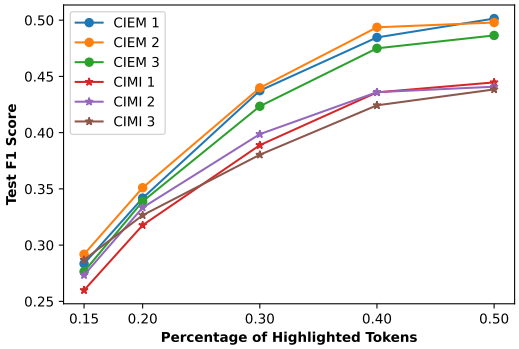

Figure 4: Test F1 scores of CIEM and CIMI with different percentages of highlighted tokens when $r = 32, \alpha = 0.9, \tau = 1.0$.

of these ablation studies, showing that the optimal values for α and τ are 0.9 and 1.0, respectively.

Additionally, Appendix E.3 includes experiments with $\alpha = 0$ and $\tau = 0$. When $\tau = 0$, our model becomes a purely VAE-based causal representation learning model, and the experimental results still outperform the baseline methods in Table 1. However, when $\alpha = 0$, the results decline significantly, indicating that identifiability theory and causal inference play a crucial role in learning trustworthy explanations.

Baselines. We compare CIEM against several well-known and recent methods. CIMI is a recent causality-inspired model for learning importance masks (Wu et al., 2023). Raw attention uses attention weights to quantify feature importance (Vaswani et al., 2017). LIME (Ribeiro et al., 2016), Occlusion (Zeiler and Fergus, 2014) and KernelSHAP (Lundberg and Lee, 2017) are widely recognized interpretable methods. Integrated Gradients (IG) calculates feature importance based on gradients (Sundararajan et al., 2017). A brief introduction to the baselines is presented in Appendix B.3.

Implementations. For a fair comparison, we use the open-source version of BERT-base-uncased (Devlin, 2018) with a *head layer* (MLPs) as the target model for explanation, where we explain $f_\phi(\cdot)$. This choice keeps the comparison with baselines controlled in the main experiments, and the reported results in this paper therefore focus on explanations of the final layer rather than a systematic comparison across multiple layers or larger Transformer backbones. For calculating the COMP and SUFF metrics, we use N% values from [10, 20, 30, 40, 45]. The parameters α, τ, and r are set to 0.9, 1, and 32 after optimization, respectively. Additional implementation details can be found in Appendix D.1.

5.3 Trustworthiness and Faithfulness Experiments

We perform trustworthiness experiments using three manually labeled explanation sets derived from the e-SNLI dataset. The details of these sets are provided in Appendix E.1, Table 5. Since e-SNLI provides manually labeled evidence at the token level, the resulting Accuracy, Precision, Recall, and F1 scores should be interpreted as measuring agreement with annotated tokens rather than as a full human-centered evaluation of explanation quality. For the generated importance mask $\mathcal{S}$, we select the Top-40% of features to calculate the metrics, as shown in Table 1.

For the trustworthiness experiments, as shown in Table 1, our method consistently outperforms all other methods across all metrics for all three manually labeled datasets. By leveraging attention information in a more rational and efficient manner, our approach generates identifiable representations that can be interpreted as clear and relevant explanations for the original input. Because these metrics evaluate agreement with manually annotated evidence, the results suggest that CIEM identifies tokens that are more consistent with standardized human rationales on e-SNLI. At the same time, they should be interpreted as relative evidence under a token-level proxy rather than as a complete

CIEM	LIME	IG	RawAttention
Label: entailment **Pred**: entailment Premise: A woman is petting a dog outside. Hypothesis: A person and an animal are interacting out of doors.	**Label**: entailment **Pred**: entailment Premise: A woman is petting a dog outside. Hypothesis: A person and an animal are interacting out of doors.	**Label**: entailment **Pred**: entailment Premise: A woman is petting a dog outside. Hypothesis: A person and an animal are interacting out of doors.	**Label**: entailment **Pred**: entailment Premise: A woman is petting a dog outside. Hypothesis: A person and an animal are interacting out of doors.

Table 3: Examples of explanations generated by four different methods. Each example evaluates the relationship between the Premise and Hypothesis (contradiction or entailment). The underlined portions highlight the important features that have been manually annotated, while the features in different colors represent explanations generated by the four methods.

assessment of human explanation quality.

In the faithfulness experiments, as observed in Table 2, our method continues to outperform most other methods across the majority of metrics on four real-world datasets. COMP and SUFF metrics reflect the ability of an interpretable method to identify significant and insignificant features within the explained model. These results highlight the competitive edge of our proposed method compared to state-of-the-art interpretable techniques that prioritize faithfulness. A key strength of our approach lies in the identifiability and trustworthiness of the generated explanations. By focusing on both faithfulness and identifiability, our method ensures that the identified representations not only significantly influence model predictions but also have a clear and reliable origin. Importantly, CIEM is not designed to optimize only a single attribution score. Instead, the empirical pattern is consistent with our overall objective: competitive faithfulness is maintained while trustworthiness and identifiability are explicitly strengthened. This also helps explain why some scores should be read as part of a trade-off rather than as an attempt to dominate every baseline on every metric.

5.4 Evaluation and Computational Efficiency

The trustworthiness and faithfulness experiments have demonstrated the ability of $\mathbf{Z_w}$ to reliably and accurately capture the crucial features used by the interpreted Transformer for prediction. Because $\mathbf{Z_v}$ is a value-related latent representation rather than a standalone token-level attribution, we evaluate it indirectly through the composite representation $\mathbf{Z} = \mathbf{Z_w}(\mathbf{X}\mathbf{Z_v})$. Table 4 shows that the accuracies of $f(\mathbf{X})$ and $f_h(\mathbf{Z})$ are very close, indicating that the learned representation preserves the predictive behavior of the original model and that $\mathbf{Z_v}$ contributes useful value-related information to the explanation.

Experiments on synthetic datasets in the Appendix E.5 demonstrate the identifiability of our model and explain why our identifiability generally does not extend to other baseline methods.

Furthermore, results in Appendix E.4 indicate that our model is computationally efficient, making it a practical solution for real-world applications. Considering both computational efficiency and the long datasets in Table 4, CIEM demonstrates good scalability.

6 Conclusion

In this paper, we propose a causal framework, the Causal Identifiable Explanation Model (CIEM), designed to transform multi-layer representation learning models, such as Transformers, into self-explaining systems. Unlike traditional methods that rely on masked representations, our approach leverages identifiability theory and causal representation learning to derive identifiable representations, ensuring that the learned explanations are both interpretable and trustworthy. Extensive experiments on both real and synthetic datasets demonstrate that, compared to state-of-the-art methods, CIEM consistently produces more trustworthy and reliable explanations while guaranteeing faithfulness. Our findings highlight the importance of identifiability in enhancing model interpretability and trustworthiness.

Limitations

The identifiability of CIEM's explanation requires the model to be interpreted to satisfy the three assumptions outlined in Theorem 4.1. These assumptions are sufficient conditions for the CIEM surrogate generative model rather than guarantees for every explained model, and violations may reduce explanation effectiveness. In addition, CIEM introduces extra training cost through its auxiliary generative model. Empirically, our study is limited to BERT-base-uncased, does not yet cover some stronger attribution-style baselines or systematic layer-wise analysis, and evaluates trustworthiness mainly through e-SNLI token-level agreement rather than full human evaluation.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable? In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pages 25–34.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. Advances in Neural Information Processing Systems, 31.

Gianluca Carloni, Andrea Berti, and Sara Colantonio. 2025. The role of causality in explainable artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(2):e70015.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. arXiv preprint arXiv:2105.02657.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4443–4458, Online. Association for Computational Linguistics.

Mengnan Du, Ninghao Liu, Fan Yang, Shuiwang Ji, and Xia Hu. 2019. On attribution of recurrent neural network predictions via additive decomposition. In The world wide web conference, pages 383–393.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 12963–12971.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. 2019. Nonlinear ica using auxiliary variables and generalized contrastive learning. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 859–868. PMLR.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online. Association for Computational Linguistics.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020a. Variational autoencoders and nonlinear ica: A unifying framework. In International Conference on Artificial Intelligence and Statistics, pages 2207–2217. PMLR.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. 2020b. Variational autoencoders and nonlinear ica: A unifying framework. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), volume 108 of Proceedings of Machine Learning Research, pages 2207–2217. PMLR.

Diederik P Kingma. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Arthur Lewbel. 2019. The identification zoo: Meanings of identification in econometrics. Journal of Economic Literature, 57:835–903.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Wei Liu, Zhiying Deng, Zhongyu Niu, Jun Wang, Haozhao Wang, and Ruixuan Li. 2025. Exploring practical gaps in using cross entropy to implement maximum mutual information criterion for rationalization. Transactions of the Association for Computational Linguistics, 13:577–594.

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. 2022. Fr: Folded rationalization with a unified encoder. Advances in Neural Information Processing Systems, 35:6954–6966.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. ACM Computing Surveys, 56(9):1–36.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Qianren Mao, Jianxin Li, Senzhang Wang, Yuanning Zhang, Hao Peng, Min He, and Lihong Wang. 2019. Aspect-based sentiment classification with attentive neural turing machines. In IJCAI, pages 5139–5145.

Bjørn Leth Møller, Christian Igel, Kristoffer Knutsen Wickstrøm, Jon Sporring, Robert Jenssen, and Bulat Ibragimov. 2024. Finding NEM-u: Explaining unsupervised representation learning through neural network generated explanation masks. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 36048–36071. PMLR.

Zhenyu Nie, Zheng Xiao, Tao Wang, Anthony Theodore Chronopoulos, Răzvan Andonie, and Amir Mosavi. 2025. Tips: A text interaction evaluation metric for learning model interpretation. Expert Systems with Applications, 287:128184.

Judea Pearl. 2009. Causality: Models, Reasoning and Inference, 2nd edition. Cambridge University Press, USA.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. 2024. Causal interpretation of self-attention in pre-trained transformers. Advances in Neural Information Processing Systems, 36.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2931–2951.

Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, page 901–904, New York, NY, USA. Association for Computing Machinery.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. 2020. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). arXiv preprint arXiv:2001.04872.

Ying Sun, Hengshu Zhu, and Hui Xiong. 2025. Toward faithful neural network intrinsic interpretation with shapley additive self-attribution. IEEE Transactions on Neural Networks and Learning Systems, 36(9):16294–16308.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR.

Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual explanations for neural recommenders. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, page 1627–1631, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Jie Wang, Haidong Shao, Jing He, Le Liu, Jingqiang Ma, and Bin Liu. 2025. A novel interpretable fault diagnosis method using multi-image feature extraction and attention fusion. Pattern Recognition Letters, 189:38–47.

Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018. Aspect sentiment classification with both word-level and clause-level attention networks. In IJCAI, volume 2018, pages 4439–4445.

Chenwang Wu, Xiting Wang, Defu Lian, Xing Xie, and Enhong Chen. 2023. A causality inspired framework for model interpretation. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 2731–2741.

Pengzhou Wu and Kenji Fukumizu. 2021. β-intact-vae: Identifying and estimating causal effects under limited overlap. arXiv preprint arXiv:2110.05225.

Biao Xu and Guanci Yang. 2025. Interpretability research of deep learning: A literature survey. Information Fusion, 115:102721.

Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. 2024. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. In The Twelfth International Conference on Learning Representations, pages 1 – 22.

Shuting Yang, Zehui Liu, Wolfgang Mayer, Ningpei Ding, Ying Wang, Yu Huang, Pengfei Wu, Wanli Li, Lin Li, Hong-Yu Zhang, and Zaiwen Feng. 2025. Shizishangpt: An agricultural large language model integrating tools and resources. In Web Information Systems Engineering – WISE 2024, pages 284–298, Singapore. Springer Nature Singapore.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4094–4103.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. Advances in Neural Information Processing Systems, 34:12822–12835.

Libing Yuan, Shuaibo Hu, Kui Yu, and Le Wu. 2025. Boosting explainability through selective rationalization in pre-trained language models. arXiv preprint arXiv:2501.03182.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. 2025. Interpreting clip with hierarchical sparse autoencoders. arXiv preprint arXiv:2502.20578.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 818–833. Springer.

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. 2023. Towards trustworthy explanation: On causal rationalization. In International Conference on Machine Learning, pages 41715–41736. PMLR.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. Advances in neural information processing systems, 28.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2):1–38.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4227–4241.

Appendix: *Trustworthy and Explainable Causal Representation Learning in Transformers*

This appendix accompanies the paper *Towards Trustworthy and Explainable Representation Learning in Transformers*. It provides additional details on the proof of the theorems, backgrounds, related work, model implementations and experiments.

A Proof

A.1 Proof of Equivalence Relation

g^{-1} and g^{-1} satisfy Eq. 10, so $\sim$ is reflexive.

If g^{-1} and $\tilde{g}^{-1}$ satisfy Eq. 10, $\tilde{g}^{-1}$ and g^{-1} satisfy Eq. 10 since $\mathbf{Q}$ is invertible. $\sim$ is symmetric.

Let g^{-1}, $\hat{g}^{-1}$, $\tilde{g}^{-1} \in \Theta$, s.t. $g^{-1} \sim \hat{g}^{-1}$ and $\hat{g}^{-1} \sim \tilde{g}^{-1}$. Then $\exists\ \mathbf{Q_1},\ \mathbf{Q_2}$ and $\mathbf{b_1}, \mathbf{b_2}$ s.t.

$$\begin{aligned} g^{-1}(\mathbf{X}, Y) &= \mathbf{Q_1}\ \hat{g}^{-1}(\mathbf{X}, Y) + \mathbf{b_1} \quad and \\ \tilde{g}^{-1}(\mathbf{X}, Y) &= \mathbf{Q_2}\ \hat{g}^{-1}(\mathbf{X}, Y) + \mathbf{b_2} \\ &= \mathbf{Q_2}\ \mathbf{Q_1^{-1}}(g^{-1}(\mathbf{X}, Y) - \mathbf{b_1}) + \mathbf{b_2} \\ &= \mathbf{Q_2}\ \mathbf{Q_1^{-1}} g^{-1}(\mathbf{X}, Y) \\ &\quad - \mathbf{Q_2}\ \mathbf{Q_1^{-1}}\ \mathbf{b_1} + \mathbf{b_2} \\ &= \mathbf{Q_3}\ g^{-1}(\mathbf{X}, Y) + \mathbf{b_3} \end{aligned} \tag{13}$$

Thus $g^{-1} \sim \tilde{g}^{-1}$. $\sim$ is transitivity. Eq. 10 is an equivalence relation.

A.2 Proof of Theorem 4.1

The proof of Theorem 4.1 is roughly equivalent to Theorem 1 in (Khemakhem et al., 2020a), with the difference that we learn the latent representation conditional on Y. Maximizing the conditional marginal distribution of $\mathbf{X}$ in Eq. 8 yields the existence of two sets of parameters (g, λ) and $(\tilde{g}, \tilde{\lambda})$ such that $p_{g,\lambda}(\mathbf{X}|\mathbf{A}, Y) = p_{\tilde{g},\tilde{\lambda}}(\mathbf{X}|\mathbf{A}, Y)$ for all points $(\mathbf{X}, \mathbf{A}, Y)$. Then, the integral form of the conditional marginal distribution:

$$\begin{aligned} &\int_{\mathcal{Z}} p_\lambda(\mathbf{Z}|\mathbf{A}, Y) p_g(\mathbf{X}|\mathbf{Z}, Y) dz \\ &= \int_{\mathcal{Z}} p_{\tilde{\lambda}}(\mathbf{Z}|\mathbf{A}, Y) p_{\tilde{g}}(\mathbf{X}|\mathbf{Z}, Y) dz \end{aligned} \tag{14}$$

p_g and $p_{\tilde{g}}$ can be transformed into additive noise forms in Eq. 5. Using (i) and (ii) from Theorem 1, variable substitution is performed on $\mathbf{Z}$:

$$\begin{aligned} &\int_{\mathcal{Z}} p_\lambda(\mathbf{Z}|\mathbf{A}, Y) p_\epsilon(\mathbf{X} - g(\mathbf{Z}, Y)) dz \\ &= \int_{\mathcal{Z}} p_{\tilde{\lambda}}(\mathbf{Z}|\mathbf{A}, Y) p_\epsilon(\mathbf{X} - \tilde{g}(\mathbf{Z}, Y)) dz \end{aligned} \tag{15}$$

$$\begin{aligned} &\int_{\mathcal{X}} p_\lambda(g^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)\ vol(J_{g^{-1}}(\overline{\mathbf{x}}, Y)) p_\epsilon(\mathbf{X} - \overline{\mathbf{x}}) d\overline{\mathbf{x}} \\ &= \int_{\mathcal{X}} p_{\tilde{\lambda}}(\tilde{g}^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)\ vol(J_{\tilde{g}^{-1}}(\overline{\mathbf{x}}, Y)) p_\epsilon(\mathbf{X} - \overline{\mathbf{x}}) d\overline{\mathbf{x}} \end{aligned} \tag{16}$$

In Eq. 16, J represents the Jacobian matrix and $vol(C) = \sqrt{CC^T}$ We introduce

$$\begin{aligned} &P_{g,\lambda}(\overline{\mathbf{x}}) \\ &= p_\lambda(g^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)\ vol(J_{g^{-1}}(\overline{\mathbf{x}}, Y)) \mathbb{I}_{\mathcal{X}}(\overline{\mathbf{x}}) \\ &P_{\tilde{g},\tilde{\lambda}}(\overline{\mathbf{x}}) \\ &= p_{\tilde{\lambda}}(\tilde{g}^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)\ vol(J_{\tilde{g}^{-1}}(\overline{\mathbf{x}}, Y)) \mathbb{I}_{\mathcal{X}}(\overline{\mathbf{x}}) \end{aligned} \tag{17}$$

on g and $\tilde{g}$. Then

$$\begin{aligned} &\int_{\mathbb{R}^{n \times d_m}} P_{g,\lambda}(\overline{\mathbf{x}}) p_\epsilon(\mathbf{X} - \overline{\mathbf{x}}) d\overline{\mathbf{x}} \\ &= \int_{\mathbb{R}^{n \times d_m}} P_{\tilde{g},\tilde{\lambda}}(\overline{\mathbf{x}}) p_\epsilon(\mathbf{X} - \overline{\mathbf{x}}) d\overline{\mathbf{x}} \end{aligned} \tag{18}$$

By applying a convolution operation to both sides of Eq. 18, we obtain $P_{g,\lambda}(\mathbf{x}) = P_{\tilde{g},\tilde{\lambda}}(\mathbf{x})$ where x is a general variable. The detailed convolution operation can be easily obtained by referring to Sec. B.2.2 in (Khemakhem et al., 2020a). $P_{g,\lambda}(\mathbf{x})$ or $P_{\tilde{g},\tilde{\lambda}}(\mathbf{x})$ is a noise-free equation, i.e., we transform the conditional marginal distribution with noise (Eq. 14) into a noise-free distribution.

Meanwhile, $p_\lambda(g^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)$ and $p_{\tilde{\lambda}}(\tilde{g}^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)$ are defined in Eq. 6 as a factorized Gaussian distribution with natural parameters $\lambda(\mathbf{A}, Y)$ and $\tilde{\lambda}(\mathbf{A}, Y)$, respectively. $T(\mathbf{Z}) = (\mathbf{Z}, \mathbf{Z^2})$ is a sufficient statistic for $p_\lambda(g^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)$ and $p_{\tilde{\lambda}}(\tilde{g}^{-1}(\overline{\mathbf{x}}, Y)|\mathbf{A}, Y)$. The condition (iii) of Theorem 1 assumes that there exists $2d_z$+1 data points $(\mathbf{A}_0, Y_0), (\mathbf{A}_1, Y_1), \ldots, (\mathbf{A}_{2d_z+1}, Y_{2d_z+1}) \in \mathcal{D}$ that satisfy Eq. 17. We insert each of these $2d_z$+1 data points into the equation $P_{g,\lambda}(\mathbf{x}) = P_{\tilde{g},\tilde{\lambda}}(\mathbf{x})$. Subsequently, the logarithm of both sides of each equation is taken. We subtract the logarithmic equation for the first data point $(\mathbf{A}_0, Y_0)$ with each of the last $2d_z$ logarithmic equations to obtain $2d_z$ equations. For the $i = 1, \cdots, 2d_z$ equation with $\gamma_i(\mathbf{A}_i, Y_i) := \lambda(\mathbf{A}_i, Y_i) - \lambda(\mathbf{A}_0, Y_0)$,

$$\begin{aligned} &\left\langle T(g^{-1}(\mathbf{x}, Y_i)),\ \gamma_i(\mathbf{A_i}, Y_i) \right\rangle \\ &= \left\langle T(\tilde{g}^{-1}(\mathbf{x}, Y_i)),\ \tilde{\gamma}_i(\mathbf{A_i}, Y_i) \right\rangle \\ &+ \beta(\mathbf{A}_i, Y_i) - \beta(\mathbf{A}_0, Y_0) \end{aligned} \tag{19}$$

where $\beta(\mathbf{A_i}, Y_i)$ is the portion of each logarithmic equation that does not contain g^{-1} or $\tilde{g}^{-1}$.

The invertible matrix $\mathbf{L} := [\gamma_1, \cdots, \gamma_{2d_z}]$ is defined in condition (iii) of Theorem 1. Represent the $2d_z$ equations in Eq. 19 as matrix form:

$$\begin{aligned} \mathbf{L}^T\, T(g^{-1}(\mathbf{x}, Y_i)) &= \tilde{\mathbf{L}}^T\, T(\tilde{g}^{-1}(\mathbf{x}, Y_i)) + \mathbf{c} \\ T(g^{-1}(\mathbf{x}, Y_i)) &= \mathbf{L}^{-T}\, \tilde{\mathbf{L}}^T\, T(\tilde{g}^{-1}(\mathbf{x}, Y_i)) \\ &\quad + \mathbf{L}^{-T}\, \mathbf{m} \\ T(g^{-1}(\mathbf{x}, Y_i)) &= \mathbf{G}\, T(\tilde{g}^{-1}(\mathbf{x}, Y_i)) + \mathbf{n} \end{aligned} \tag{20}$$

where $\mathbf{G} = \mathbf{L}^{-T}\, \tilde{\mathbf{L}}^T, \mathbf{n} = \mathbf{L}^{-T}\, \mathbf{m}$ and $\mathbf{m} = [\beta(\mathbf{A}_1, Y_1) - \beta(\mathbf{A}_0, Y_0), \cdots, \beta(\mathbf{A}_{2d_z}, Y_{2d_z}) - \beta(\mathbf{A}_0, Y_0)]^T$. Following a similar decomposition process as Appendix B in (Sorrenson et al., 2020), the invertible matrix $\mathbf{G}$ can be blocked as:

$$\mathbf{G} = \begin{pmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{U} & \mathbf{Q^2} \end{pmatrix} \tag{21}$$

where $\mathbf{Q}$ is an invertible $d_z \times d_z$diagonal matrix. Then, rewriting Eq. 20

$$\begin{pmatrix} g^{-1}(\mathbf{x}, Y_i) \\ (g^{-1}(\mathbf{x}, Y_i))^2 \end{pmatrix} = \begin{pmatrix} \mathbf{Q} & \mathbf{O} \\ \mathbf{U} & \mathbf{Q^2} \end{pmatrix} \begin{pmatrix} \tilde{g}^{-1}(\mathbf{x}, Y_i) \\ (\tilde{g}^{-1}(\mathbf{x}, Y_i))^2 \end{pmatrix} + \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} \tag{22}$$

Finally, it can be obtained that

$$g^{-1}(\mathbf{X}, Y) = \mathbf{Q}\, \tilde{g}^{-1}(\mathbf{X}, Y) + \mathbf{b} \tag{23}$$

where $\mathbf{b}$ is a d_z-dimensional vector.

B More Background

B.1 Structural Causal Model

A structural causal model (SCM)(Pearl, 2009) provides a formal representation of causal mechanisms. Formally, an SCM is defined as a tuple $\{\mathcal{U}, \mathcal{X}, \mathcal{F}, P(\mathcal{U})\}$, where $\mathcal{U}$ denotes a set of exogenous variables, $\mathcal{X}$ represents a set of endogenous variables, $\mathcal{F}$ comprises a set of deterministic functions. Each function $f \in \mathcal{F}$ describes the value of an endogenous variable $X \in \mathcal{X}$ as a function of its direct causes. $P(\mathcal{U})$ denotes the probability distribution over an exogenous variables $U \in \mathcal{U}$. Let $\mathbf{Pa}_i$ and U_i be all direct causes and exogenous variables of the endogenous variable X_i, respectively, and $f_i \in \mathcal{F}$ be its causal function. X_i can be obtained from $f_i(\mathbf{Pa}_i, U_i)$. All endogenous and exogenous variables, along with the directed edges representing causal relationships, constitute a directed acyclic graph (DAG).

B.2 VAE, CVAE and iVAE

A conventional variational autoencoder (VAE) (Kingma, 2013) is a generative model that learns a latent representation of observed data $\mathbf{X}$. A plethora of VAE variants have been proposed to extend the capabilities of the basic VAE model (Khemakhem et al., 2020a; Sohn et al., 2015). Viewing model optimization through a statistical lens, VAE can be conceptualized as an autoencoder augmented with a regularization term. Specifically, the VAE employs variational Bayes to reconstruct the evidence lower bound (ELBO) of a standard autoencoder. The ELBO of log-likelihood is obtained as:

$$\begin{aligned} &\log p(x) \geq \log p(x) - D_{KL}(q(z|x)||p(z|x)) \\ &= \mathbb{E}_{z\sim q} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p(z)). \end{aligned} \tag{24}$$

The encoder $q_\phi(z|x)$ is designed to approximate the true posterior distribution $p(z|x)$ of the latent variable z given the observed data x. The decoder $p_\theta(x|z)$ is employed to generate a new data point $\tilde{x}$ by decoding a latent variable z that is sampled from the approximate posterior distribution $q_\phi(z|x)$. D_{KL} represents the KL divergence. Both the decoder and encoder are typically parameterized by neural networks. The first term of Eq. 24 can be simplified to the mean squared error between the newly generated data point $\tilde{x}$ and the original data point x. Since both the variational posterior $q_\phi(z|x)$ and prior $p(z)$ are defined as factorized Gaussian distributions, the second term of Eq. 24 is also straightforward to compute. Therefore, Eq. 24 can be simplified into an efficient and easily computable form (Kingma, 2013).

Conditional VAE (Sohn et al., 2015) enhances the correlation between the learned representation and a conditional variable C compared to the standard VAE, enabling the latent representation to generate data corresponding to the given variable C. The variational ELBO of CVAE is derived as:

$$\begin{aligned} \log p(x|c) \geq\ &\mathbb{E}_{z\sim q} \log p(x|z, c) \\ &- D_{KL}(q(z|x, c)||p(z|c)). \end{aligned} \tag{25}$$

Since the encoder has captured the dependency between the condition C and the latent representation, CVAE assumes a standard normal prior $z \sim \mathcal{N}(0, I)$, as in the standard VAE.

iVAE (Khemakhem et al., 2020a) introduces a novel framework for VAE identifiability by leveraging nonlinear ICA and an auxiliary variable u,

resulting in the first identifiability results for VAE. The variational ELBO of iVAE is derived as:

$$\log p(x|u) \geq \mathbb{E}_{z\sim q} \log p_f(x|z) - D_{KL}(q(z|x,u)||p_\lambda(z|u)), \quad (26)$$

where, as mentioned, $X = f(Z) + \epsilon$, where Z follows an exponential family distribution with sufficient statistic T and natural parameter $\lambda(U)$, and ϵ represents additive noise. Unlike the CVAE, the decoder of iVAE is conditioned solely on the latent variable Z, reflecting an independence assumption between the input X and the generative process (Khemakhem et al., 2020a). The model is identifiable if its functional parameters (f, T, λ) can be uniquely recovered from the data, up to some trivial transformations, ensuring a meaningful interpretation of the learned model.

B.3 A Brief Description of the Baselines

The main ideas behind the baselines are as follows.

- LIME (Ribeiro et al., 2016) samples points around a specified input example and uses model evaluations at these points to train a simpler interpretable model, such as a linear model.
- KernelSHAP (Lundberg and Lee, 2017) computes Shapley Values by weighting the kernel and regularisation terms within the LIME framework.
- Integrated Gradients(IG) (Sundararajan et al., 2017) approximates the integral of the gradient of the model output to assign an importance score to each input feature.
- Occlusion (Zeiler and Fergus, 2014) replaces one token at a time with a reference value and measures the impact on model performance to assess the importance of each token.
- RawAttention (Vaswani et al., 2017) calculates feature importance by weighting and summing the attention weights in Transformer.
- CIMI (Wu et al., 2023) is a causality-inspired framework designed with three faithfulness modules for enhancing the faithfulness of explanations.

C More Related Work

XAI Methods. A common and effective approach to generating feature importance in XAI involves using attention weights. Several studies have assessed feature importance through weighted summation scores of raw attention and applied this method to tasks such as sentiment classification and question answering (Mao et al., 2019; Shen et al., 2018; Wang et al., 2018). However, raw attention can contain redundant information (Bai et al., 2021), prompting the development of improved attention attribution methods to explain self-attention mechanisms more effectively (Hao et al., 2021; Tran et al., 2021). (Chrysostomou and Aletras, 2021) introduces a task-scaling mechanism to enhance the faithfulness of attention-based explanations, while (Rohekar et al., 2024) models self-attention as a structural equation model specific to each input sequence.

Interpretable methods for NLP tasks. Interpretable methods in NLP based on feature importance can be classified into four categories depending on the way the feature importance is calculated (Luo et al., 2024). The input perturbation methods perturb the input representations via a binary mask, and then utilize the perturbations and the corresponding outputs to train a transparent surrogate model (Ribeiro et al., 2016; Lundberg and Lee, 2017). Since these perturbations consider each feature as independent, the faithfulness of the input perturbations is questioned. The rationale extraction methods train a selector for selecting coherent and consecutive rationales, and then train a predictor to predict the correct outcome (Lei et al., 2016; Yu et al., 2021). The rationale selector is usually an end-to-end model and employs different optimization methods, e.g. reinforcement learning (Lei et al., 2016), adversarial network (Yu et al., 2019), to improve the faithfulness of the extracted rationale. The attribution methods use model gradients or other hidden states to identify the contribution of each feature, which are usually sensitive to changes in the input (Bach et al., 2015; Du et al., 2019; Sundararajan et al., 2017).

D CIEM Details and Discussion

D.1 More Implementations

Implementations. For LIME and KernelSHAP, the results of each experiment may have some error (within 10%). The results of KernelSHAP for

LIME in Tables 1 and 2 are the average of 10 independent experiments. For IG, independent experiments on each dataset yield the same results. For CIMI, we set up the same configuration as its open source version. Each independent experiment outputs the same results after training is completed. In the case of our proposed method, CIEM, the optimal reparameterization parameter (ϵ) is determined post-training. This post-training selection of ϵ is crucial to ensure consistent output for identical input instances.

For baseline selection, we do not choose the rationalization extraction methods. Primarily, these methods require two neural network models to serve as a selector and a predictor respectively in their experimental setup (Lei et al., 2016; Zhang et al., 2023), which is unfair to other categories of methods. Secondly, the datasets employed in experiments of rationalization extraction methods are specifically designed to explain continuous and coherent rationalization. Thus, comparisons with human-annotated datasets may not be entirely justified.

For CIEM training on all datasets, we employ the Adam optimizer with a learning rate of 2e-4 and a batch size of 64, and train 10 epochs. All random seeds in our code, except for those used in reparameterization, are fixed to 0. The random seeds for reparameterization are set to the optimal value determined for each experiment. In trustworthiness experiments, we treat the Top-40% of explanations generated by each method as ground truth and calculate Accuracy, Precision, Recall, and F1-score against human-annotated explanations. In faithfulness experiments, the other three datasets are set to r = 32 in order to be consistent with the settings of the e-SNLI dataset. All experiments are carried out with PyTorch on A100 GPU.

D.2 More Details on CIEM

Unconditional schema of CIEM. CIEM provides a generative model of unconditional schema, which is an unsupervised representation learning. The unconditional schema of CIEM can be obtained from Eq. 4-6.

$$\begin{aligned} p_\theta(\mathbf{X},\ \mathbf{Z}|\ \mathbf{A}) &= p_g(\mathbf{X}|\ \mathbf{Z})p_\lambda(\mathbf{Z}|\ \mathbf{A}) \\ p_g(\mathbf{X}|\ \mathbf{Z}) &= p_\epsilon(\mathbf{X} - g(\mathbf{Z})) \\ p_\lambda(\mathbf{Z}|\ \mathbf{A}) &\sim \mathcal{N}(\mathbf{Z}; \mu(\mathbf{A}), \sigma^2(\mathbf{A})) \\ q(\mathbf{Z}|\ \mathbf{X}, \mathbf{A}) &\sim \mathcal{N}(\mathbf{Z}; \mu'(\mathbf{X}, \mathbf{A}), \sigma'^2(\mathbf{X}, \mathbf{A})) \end{aligned} \tag{27}$$

where $\theta := (g, \mu, \sigma)$ represents the model parameters, and $\mathbf{A}$ is auxiliary variable. The conditional prior $p_\lambda(\mathbf{Z}|\ \mathbf{A})$, conditional posterior $q(\mathbf{Z}|\ \mathbf{X}, \mathbf{A})$ and likelihood distribution $p_g(\mathbf{X}|\ \mathbf{Z})$ are all defined as factorized Gaussian distributions in the same way as the conditional schema. $\lambda(\mathbf{A}) := (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ is the natural parameters of $p_\lambda(\mathbf{Z}|\ \mathbf{A})$ dependent on $\mathbf{A}$. The identifiable theory still holds in the unconditional schema.

The importance of identifiability to interpretability. An important sense of identifiability is clarity of source, which is crucial in interpretability. An example is provided in (Khemakhem et al., 2020a) to illustrate that a well-fitted representation may not actually be the true distribution. Especially with the establishment of numerous neural network-based XAI methods, identifiability becomes even more important. The primary objective of interpretability is to elucidate the reasons behind a model's predictions. When a model is unidentifiable, it becomes challenging to obtain accurate insights into its internal mechanisms, potentially resulting in explanations that are either inaccurate or misleading. Identifiability plays a crucial role in numerous real-world contexts. For instance, in domains like healthcare, it is essential to comprehend the rationale behind a model's decision-making process for purposes such as risk assessment, liability tracing, and other critical applications.

E More Experiments

E.1 More Details of Data Processing

e-SNLI. e-SNLI is a dataset designed by (Camburu et al., 2018) to examine generated natural language explanations in an automated manner, which is a subset of SNLI (Bowman et al., 2015) containing manually labeled explanations. The original training set of e-SNLI is divided into two files due to its excessive size, and we only take the data from the first file. Each example of the dataset contains two sentences representing the premise and hypothesis and a label with a value of contradiction or entailment. We use prompts to construct the dataset in the form *Instruction: Classify the relationship between Premise and Hypothesis into two categories: contradiction and entailment. Premise: Sentence1. Hypothesis: Sentence2*, and label entailment as positive and contradiction as negative. Table 4 shows all the details of our split dataset. For more flexible computation, we create a short version of the e-SNLI dataset by filtering texts with token lengths

Dataset	instances			Len			Token Len			Accuracy	Accuracy
	Train	Val	Test	Train	Val	Test	Train	Val	Test	$f(\mathbf{X})$	$f_h(\mathbf{Z})$
e-SNLI	8000	2000	2000	35.85	37.01	36.84	47.70	49.01	48.83	90.9%	90.85%
SST-2	6920	872	1821	19.32	19.56	19.24	25.27	25.52	25.14	90.17%	90.28%
IMDB	10421	2605	2606	98.67	98.73	98.93	110.42	110.34	110.77	92.98%	93.02%
AGNews	20000	4000	3782	30.79	31.1	30.61	42.25	42.66	41.99	96.64%	96.67%

Table 4: Dataset division and length. Token length and classification accuracy based on bert-base-uncased.

	Instances	Word Len	Token Len	Tagged Word	Tagged Token	Ratio
Explanation 1	197	22.31	28.33	5.31	6.74	23.78%
Explanation 2	197	22.31	28.33	4.98	6.26	22.11%
Explanation 3	197	22.31	28.33	4.97	6.12	21.59%

Table 5: 200 randomly selected instances (three dropped) from e-SNLI test set, each containing three different manually labeled explanations. The table shows the length of the explanations and their percentage of the original sentence.

less than 150. Table 5 shows the details of the test sets used in the trustworthiness experiments.

SST-2. The Stanford Sentiment Treebank (SST) (Socher et al., 2013) is a standard sentiment classification dataset, and we use its binary version SST-2. Every entry in its original dataset passes our filtering, so Table 4 shows the details of its original version.

IMDB. IMDB (Maas et al., 2011) is a large sentiment analysis dataset containing 50K movie reviews. it contains a lot of unfavorable information for training, such as HTML tags. We remove HTML tags from all data and filter to retain data with token length less than 150. Among all the filtered data, we separate the training/validation/testing sets in the ratio of 4:1:1 and show them in Table 4.

AGNews. AGNews (Zhang et al., 2015) contains 4 categories, each with 30,000 training and 1,900 testing samples, for a total of 120,000 training samples and 7,600 testing samples. We extract data from AGNews with categories *World* and *Sci/Tech* as positive and negative, respectively. Subsequently, we filter the data in the training set with token length less than 150, sampling 20,000 as the training set and another 4,000 as the validation set, and filter the test set with token length less than 150 as the test set. Table 4 shows all the details.

E.2 How to choose the importance percentage?

For the metrics calculation in all trustworthiness experiments, we did not take into account the prompts added during the dataset construction, i.e., the percentage of highlighted tokens applied only to the premises and hypotheses of the original dataset.

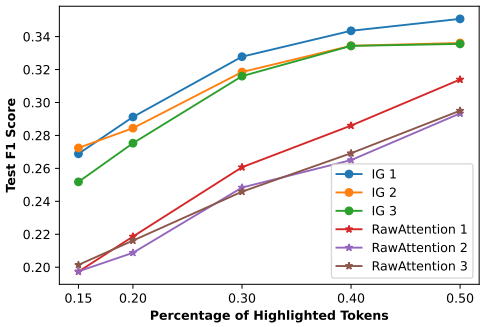

Figure 5: Test F1 scores of IG and RawAttention with different percentages of highlighted tokens.

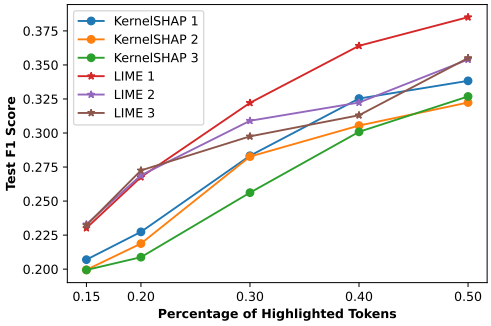

Figure 6: Test F1 scores of LIME and KernelSHAP with different percentages of highlighted tokens.

Fig. 4, 5, and 6 show how the F1 scores are affected by the importance ratio for different methods on three manually labeled datasets. All methods exhibit increasing F1 scores as the percentage decreased below 50%. The relative performance of different methods remain consistent regardless of the percentage of highlighted tokens, which suggests that the percentage is not a determining factor

in the relative performance among different methods. Furthermore, the percentage should be maintained between 10% and 50% to avoid skewing the F1 score due to excessively high precision or recall.

Table. 5 shows that the percentage of highlighted tokens should be higher than 20%. Fig. 4, 5, and 6 also illustrate that when the percentage of highlighted tokens increases from 40% to 50%, the performance of different methods shows little improvement. Considering the above analyses, we finally choose 40% as the percentage of highlighted tokens in the trustworthiness experiment.

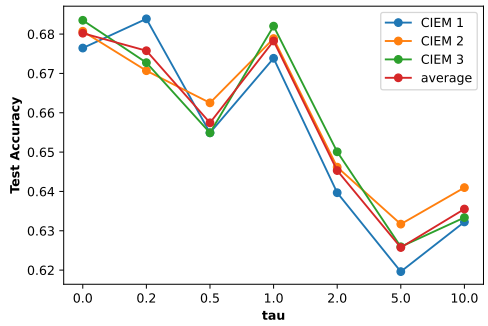

Figure 7: Test Accuracy of CIEM with different τ when $r = 32, \alpha = 1.0$.

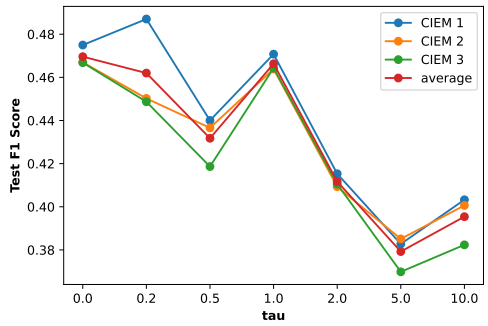

Figure 8: Test F1 scores of CIEM with different τ when $r = 32, \alpha = 1.0$.

E.3 More Ablation Studies

Fig. 7-10 illustrate the optimization process of τ and α. Observations reveal that similar magnitudes of τ and α yield better results. The optimization ultimately converged to $\tau = 1.0$ and $\alpha = 0.9$.

Fig. 12 and 13 present the ablation studies conducted to evaluate the impact of our proposed module on faithfulness. By comparing the results when setting $\alpha = 1.0, \tau = 0$ and $\alpha = 0, \tau = 1.0$ against the optimal experimental setting, we demonstrate that our module consistently enhances the faithfulness of the generated explanations. In the specific case where $\alpha = 1.0$ and $\tau = 0$, our VAE-based model continues to exhibit robust performance.

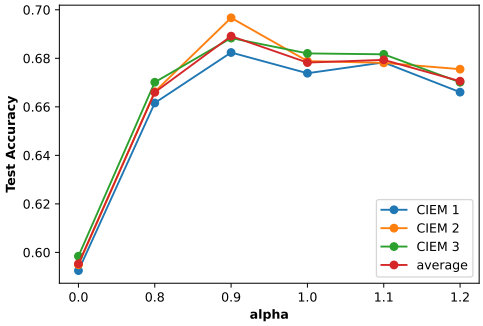

Figure 9: Test Accuracy of CIEM with different α when $r = 32, \tau = 1.0$.

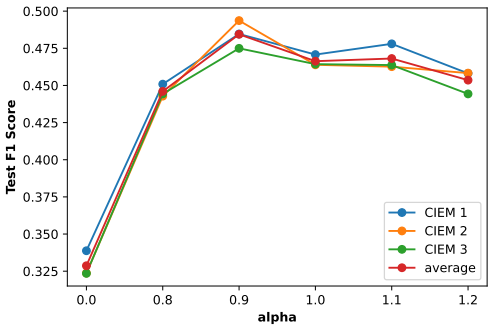

Figure 10: Test F1 scores of CIEM with different α when $r = 32, \tau = 1.0$.

This finding underscores the utility of causal inference in developing interpretable models.

E.4 Computational Cost

Table 6 presents the mean computational cost per instance for trustworthiness and faithfulness experiments conducted on each individual dataset. Our proposed CIEM method demonstrates significantly lower latency and superior responsiveness compared to perturbation-based methods, such as LIME and KernelSHAP. The high computational complexity inherent in these perturbation-based methods often results in prolonged explanation generation times. In contrast, our approach effectively mitigates this limitation, offering a computationally efficient solution.

E.5 The Experiments on Synthetic Datasets

It is necessary to discuss whether identifiability is suitable for the compared baselines. The fundamental assumption of identifiability is the existence of a latent variable or representation that can be identified in the method. These compared baselines typically operate on a learning paradigm of importance masks, meaning there is no latent representation with dimensionality matching the input and output representations of the original model.

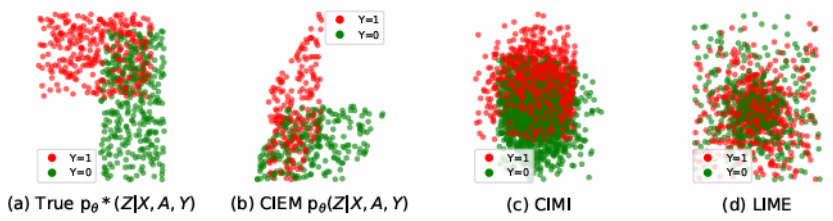

Figure 11: The visualization focuses on the densest region of the data distribution. (a) The true distribution of latent variables. (b) The distribution of latent variables learned by CIEM. (c) The distribution of latent variables learned by CIMI. (d) The distribution of latent variables learned by LIME.

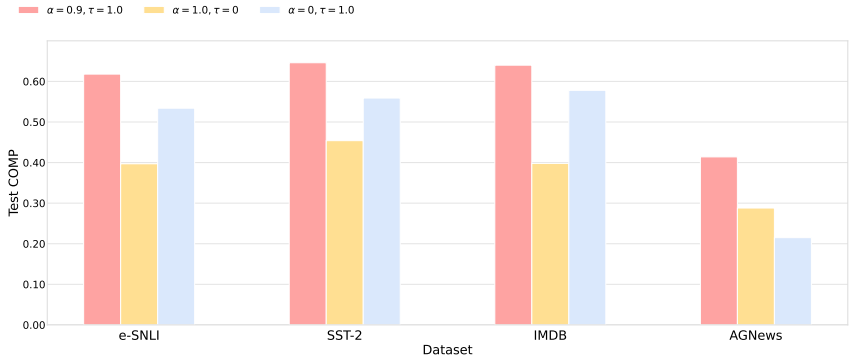

Figure 12: COMP scores of CIEM on test sets with different α and τ.

Consequently, our identifiability results generally do not extend to the baselines.

To address the identifiability results of the compared baselines, it is necessary to construct a plausible latent representation for them. As mentioned in the third paragraph of the introduction, this plausible latent representation (defined in the paper as masked representation) is derived from the mask applied to the original input representation. We conduct identifiability experiments for the baselines based on this masked representation using the same synthetic data.

We follow the following data generation process for simulation experiments. Our primary methodology involves generating the latent variable $\mathbf{Z}$ using its known data generation process. Subsequently, $\mathbf{X}$, $\mathbf{A}$, and Y are generated from $\mathbf{Z}$. With this generated dataset ($\mathbf{X}$, $\mathbf{A}$, Y), a CIEM is trained to learn the latent variable $\mathbf{Z}$. The identifiability of our model is then demonstrated by comparing the learned latent variable with the true latent variable.

$$\begin{aligned} &\mathbf{Z} \sim \mathcal{N}(\mathbf{Z}; \mu', \sigma'^2), \\ &\mathbf{X}|\mathbf{Z} \sim \mathcal{N}(\mathbf{X}; u(\mathbf{Z}), v(\mathbf{Z})), \\ &\mathbf{A} = s(\mathbf{X}), \\ &Y|\mathbf{X}, \mathbf{Z} \sim Bern(Logi(l(\mathbf{X}, \mathbf{Z}))) \end{aligned} \tag{28}$$

Our estimated latent variable $\mathbf{Z}$ is generated from a specific conditional posterior, which follows a factorized Gaussian distribution with parameters $\mu' = 0$, $\sigma'^2 = 1$, and a dimensionality of $dim(\mathbf{Z}) = 30$. The functions u, v, s, and l are linear. We generate synthetic data $\mathbf{X}, \mathbf{A}, Y$ following Eq. 28 and then learn the latent variable $\mathbf{Z}$ from the generated data. As illustrated in Fig. 11, a linear transformation effectively bridges the gap between the learned latent variable distribution and the true latent variable distribution.

E.6 More Examples

Table 7 presents additional examples of CIEM explanations. We select two instances, each with three distinct human annotations, enabling a comparison of subtle differences in explanations for the same instance.

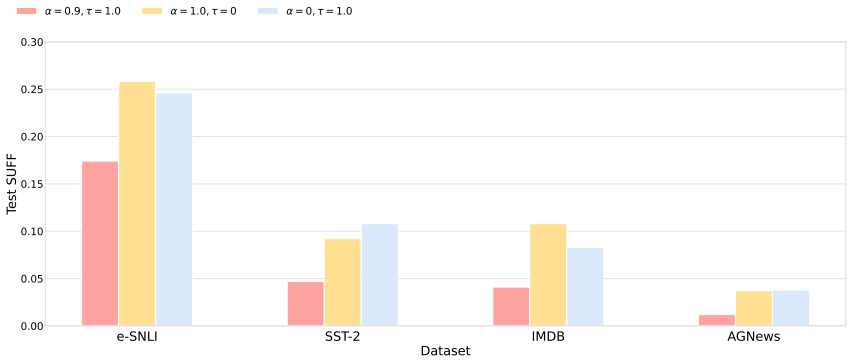

Figure 13: SUFF scores of CIEM on test sets with different α and τ.

Method	Time in seconds (e-SNLI)	
	Trustworthiness	Faithfulness
LIME	0.105	0.191
KernelSHAP	0.104	0.192
IG	0.021	0.057
RawAttention	0.006	0.038
Occlusion	0.138	0.255
CIMI	0.015	0.041
CIEM(Our)	**0.006**	**0.034**

Table 6: Runtime for each instance of the trustworthiness and faithfulness experiments.

CIEM	**CIMI**	**IG**	**RawAttention**
Label: entailment **Pred**: entailment Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	**Label**: entailment **Pred**: entailment Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	**Label**: entailment **Pred**: entailment Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	**Label**: entailment **Pred**: entailment Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.
Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.
Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.	Premise: A man wearing a red vest is walking past a black and green fence. Hypothesis: A man is wearing a vest.

Table 7: Examples of explanations generated by four different methods. Each example is judging the relationship between Premise and Hypothesis (contradiction or entailment). The underlined portions represent important features that are manually annotated, while the features in different colors represent explanations generated by different methods. Our method matches the manual annotations best.